From Particle Filters to Malliavin Filtering with Application to Target Tracking

Sergio Pequito [*,**]

[*] *Instituto Superior Tecnico*
Institute for System and Robotics
Department of Electronic and Computer Engineering
[**] *Carnegie Mellon University*
Department of Electronic and Computer Engineering

Abstract
In this paper we address the nonlinear and non-Gaussian filtering problem. It is typically crucial to process data on-line as it arrives, both from the point of view of storage costs as well as for rapid adaptation to changing signal characteristics. Hereafter, we review both optimal and suboptimal Bayesian algorithms for nonlinear/non-Gaussian tracking problems, with a focus on particle filters. Particle filters are sequential Monte Carlo methods based on point mass (or "particle") representations of probability densities, which can be applied to any state-space model and which generalize the traditional Kalman filtering methods. Several variants of the particle filter such as SIR are introduced within a generic framework of the sequential importance sampling (SIS) algorithm. For the sake of completeness we present some mathematical preliminaries and different implementations of particle filters. Moreover, related theoretical and practical issues are addressed in detail, and we end this paper we some new results using Malliavin calculus.

Keywords: Stochastic filtering, Bayesian filtering, Sequential Monte Carlo methods, Particle Filters, Malliavin Calculus

Contents

⋆ Support for this research was provided by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program under Grant SFRH / BD / 33779 / 2009

1. INTRODUCTION

This paper is the final report to the course of Detection, Estimation and Filtering.

1.1 Initial proposal

New tools and Techniques in Particle Filters

≪Optimal estimation problems for non-linear non-Gaussian state-space models do not typically admit analytic solutions. Since their introduction in 1993, particle filtering methods have become a very popular class of algorithms to solve these estimation problems numerically in an online manner, i.e recursively as observations become available, and are now routinely used in fields as diverse as computer vision, econometrics, robotics and navigation. The objective of this tutorial is to provide a complete, up-to-date survey of this field as of 2008. Basic and advanced particle methods for filtering as well as smoothing are presented.≫

In order to better understand the motivation behind the choice of the topic, here is some of the motivation introductory text

1.2 Monte Carlo Methods and Monte Carlo Filtering

In recent decades, Monte Carlo techniques have been rediscovered independently in statistics, physics, and engineering. Roughly speaking, Monte Carlo technique is a kind of stochastic sampling approach aiming to tackle the complex systems which are analytically intractable. The power of Monte Carlo methods is that they can attack the difficult numerical integration problems. One of the attractive merits of sequential Monte Carlo approaches lies in the fact that they allow on-line estimation by combining the powerful Monte Carlo sampling methods with Bayesian inference, at an expense of reasonable computational cost. In particular, the sequential Monte Carlo approach has been used in parameter estimation and state estimation, for the latter of which it is sometimes called particle filter. The basic idea of particle filter is to use a number of independent random variables called particles, sampled directly from the state space, to represent the posterior probability, and update the posterior by involving the new observations; the "particle system" is properly located, weighted, and propagated recursively according to the Bayesian rule. In retrospect, the earliest idea of Monte Carlo method used in statistical inference but the formal establishment of particle filter seems fair to be due to Gordon, Salmond and Smith [Gordon et al. (1993)], who introduced certain novel resampling technique to the formulation. Almost in the meantime, a number of statisticians also independently rediscovered and developed the sampling-importance-resampling (SIR). The rediscovery and renaissance of particle filters in the mid-1990s after a long dominant period, partially thanks to the ever increasing computing power. Recently, a lot of work has been done to improve the performance of particle filters. Some potential future directions, will be considering combining these methods with Monte Carlo sampling techniques, as we will discuss later in the paper. The attention of this paper, however, is still on the Monte Carlo methods and particularly sequential Monte Carlo estimation.

1.3 Outline of Paper

In the section 2 we introduce some elementary mathematical background do cope with future concepts. In section 3 we tackle some of the basic concepts and we do the problem formulation in section 4 as well as the theoretical solution. In order to be able implement a solution we need to introduce some appropriate numerical scheme, for that we address the different possibilities in the section 5 where we give special focus to Monte Carlo methods and how to reduce the variance. Gathering everything referred previously we end up with the main topic of this paper, the particle filters, explored in section 6. For the sake of completeness we also present some of the theorectical results and issues and how to cope with some in section 7. Finally, we show some brief introduction in section 8 to extend the results using Malliavin calculus, that will originate a paper to be submitted to American Control Conference 2011. At last, we present some of the preliminary results in section 9 and conclude with conclusions and further research directions in section 10.

2. MATHEMATICAL PRELIMINARIES AND PROBLEM FORMULATION

2.1 Preliminaries

Definition 1. Let S be a set and $\mathcal{F}$ be a family of subsets of S. $\mathcal{F}$ is a $\sigma-$algebra if

(1) $\in \mathcal{F}$;
(2) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$;
(3) $A_1, A_2, ... \in \mathcal{F}$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

This means that a $\sigma-$algebra is closed under the complement and union of countably infinitely many sets. □

Definition 2. A probability space is defined by the elements $\{\Omega, \mathcal{F}, P\}$ where $\mathcal{F}$ is a $\sigma-$algebra of Ω and P is a complete, $\sigma-$additive probability measure of all $\mathcal{F}$. In other words, P is a set function whose arguments are random events (element of $\mathcal{F}$) such that axioms of probability hold. □

Definition 3. Let $p(x) = \frac{dP(x)}{d\mu}$ denote Radon-Nikodym density of probability distribution $P(x)$ w.r.t. a measure μ. When $x \in X$ is discrete and μ is a counting measure μ, $p(x)$ is a probability mass function (pmf); when x is continuous and μ is a Lebesgue measure, $p(x)$ is a probability density function (p.d.f.). □

Intuitively, the true distribution $P(x)$ can be replaced by empirical distribution given the simulated samples

$$\hat{P}(x) = \frac{1}{N_p} \sum_{i=1}^{N_p} \delta\left(x - x^{(i)}\right) \tag{1}$$

where $\delta(.)$ is a Radon-Nikodym density w.r.t. μ of the point-mass distribution concentrated at the point x. When $x \in X$ is discrete, $\delta(x - x^{(i)}$ is 1 for $x = x^{(i)}$ and 0 elsewhere. When $x \in X$ is continuous, $\delta(x - x^{(i)}$ is a Dirac-delta function, $\delta(x - x^{(i)} = 0$ for all $x \neq x^{(i)}$, and $\int_X d\hat{P}(x) = \int_X \hat{p}(x)dx = 1$

2.2 Notations

- p.d.f. - probability density function

- pmf - probability mass function
- E[.] - expected value
- Var[.] - variance
- Cov[.] - covariance

...

2.3 Stochastic Filtering Problem

Before we run into the mathematical formulation of stochastic filtering problem, it is necessary to clarify some basic concepts[Jazwinski (1970)]:

- *Filtering* is an operation that involves the extraction of information about a quantity of interest at time t by using data measured up to and including t.

Now, let us consider the following generic stochastic filtering problem in a dynamic state-space form

$$\dot{x}_t = f\left(t, x_t, u_t, w_t\right), \tag{2}$$
$$y_t = g\left(t, x_t, u_t, v_t\right), \tag{3}$$

where equations (2) and (3) are called state equation and measurement equation, respectively; x_t represents the state ector, y_t is the measurement vector, u_t represents the system input vector (as driving force) in a controlled environment: $f : \mathbb{R}^{N_x} \mapsto \mathbb{R}^{N_x}$ and $g : \mathbb{R}^{N_x} \mapsto \mathbb{R}^{N_y}$ are two vector-valued functions, which are potentially time-varying; w_t and v_t represent the process (dynamical) noise and measurement noise respectively, with appropriate dimensions. The above formulation is discussed in the continuous-time domain, in practice however, we are more concerned about the discrete-time filtering. In this context, the following practical filtering problem is concerned:

$$x_{n+1} = f(x_n, w_n), \tag{4}$$
$$y_n = g(x_n, v_n), \tag{5}$$

where w_n and v_n can be viewed as white noise random sequences with unknown statistics in the discrete-time domain. The state equation (4) characterizes the state transition probability $p(x_{n+1}|x_n)$, whereas the measurement equation (5) describes the probability $p(y_n|x_n)$ which is further related to the measurement noise model.

The equations (4) and (5) reduce to the following special case where a linear Gaussian dynamic system is considered:

$$\begin{aligned} x_{n+1} &= F_{n+1,n}x_n + w_n, \\ y_n &= G_n x_n + v_n, \end{aligned} \tag{6}$$

for which the analytic filtering solution is given by the Kalman filter, in which the sufficient statistics of mean and state-error correlation matrix are calculated and propagated. In equations (6), $F_{n+1,n}$, G_n are called transition matrix and measurement matrix, respectively.

Given the initial density $p(x_0)$, transition probability $p(x_{n+1}|x_n)$, and likelihood $p(y_n|x_n)$, the objective of the filtering is to estimate the optimal current state at time n given the observations up to time n, which is in essence the amount to estimating the posterior density $p(x_n|y_{0:n})$ or $p(x_{0:n}|y_{0:n})$. Although the posterior density provides a complete solution of the stochastic filtering problem still remains intractable since the density is a function rather than a finite-dimensional point estimate. We should also keep in mind that most of physical system are not finite dimensional, thus the infinite-dimensional system can only be modeled approximately by a finite-dimensional filter, in other words, the filter can only be suboptimal in this sense. Nevertheless, in the context of nonlinear filtering, it is still possible to formulate the exact finite-dimensional filtering solution[Arulampalam et al. (2002)].

2.4 Stochastic Differential Equations and Filtering

In the following, we will formulate the continuous-time stochastic filtering problem by Stochastic Differential Equation (SDE) theory. Suppose $\{x_t\}$ is a Markov process with an infinitesimal generator, rewriting state-space equations (2)-(3) in the following form:

$$\begin{aligned} dx_t &= f(t, x_t)dt + \sigma(t, x_t)dw_t, \\ dy_t &= g(t, x_t)dt + dv_t \end{aligned} \tag{7}$$

where $f(t, x_t)$ is often called nonlinear drift and $\sigma(t, x_t)$ called volatility or diffusion coefficient. Again, the noise processes $\{w_t,\ v_t,\ t \geq 0\}$ are two Wiener processes. $x_t \in \mathbb{R}^{N_x}$, $y_t \in \mathbb{R}^{N_x}$. First, let's look at the state equation (aka diffusion equation). For all $t \geq 0$, we define a backward diffusion operator L_t as

$$L_t = \sum_{i=1}^{N_x} f_t^i \frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{N_x} a_t^{ij} \frac{\partial^2}{\partial x_i \partial x_j}$$

where $a_t^{ij} = \sigma^i\left(t, x_t\right)\sigma^j\left(t, x_t\right)$. Operator L corresponds to an infinitesimal generator of the diffusion process $\{x_t,\ t \geq 0\}$. The goal now is to deduce conditions under which one can find a recursive and finite-dimensional (close form) scheme to compute the conditional probability distribution $p(x_t|\mathcal{Y}_t)$, given the filtration $\mathcal{Y}_t$ produced by the observation process. Let's define an innovations process

$$e_t = y_t - \int_0^t E\left[g(s, x_s)|y_{0:n}\right] ds$$

where $E\left[g(s, x_s)|\mathcal{Y}_t\right]$ is described as

$$\begin{aligned} \hat{g}\left(x_t\right) &= E\left[g(t, x_t)|\mathcal{Y}_t\right] = \\ &= \int_{-\infty}^{\infty} g\left(x_t\right) p\left(x_t|\mathcal{Y}_s\right) dx. \end{aligned}$$

For any test function $\phi \in \mathbb{R}^{N_x}$, the forward diffusion operator $\tilde{L}$ is defined as

$$\tilde{L}_t\phi = -\sum_{i=1}^{N_x} f_t^i \frac{\partial \phi}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{N_x} a_t^{ij} \frac{\partial^2 \phi}{\partial x_i \partial x_j}$$

which essentially is the Fokker-Planck operator. Given initial condition $p(x_0)$ at $t = 0$ as boundary condition it turns out that the p.d.f. of diffusion process satisfies the Fokker-Planck-Kolmogorov equation (FPK; aka Komolgorov forward equation)

$$\frac{\partial p\left(x_t\right)}{\partial t}=\tilde{L}_t p\left(x_t\right).$$

By involving the innovation process and assuming $E\left[v_t\right]=\Sigma_{v,t}$, we have the following Kushner's equation

$$dp\left(x_t|\mathcal{Y}_t\right)=\tilde{L}_t p\left(x_t|\mathcal{Y}_t\right)dt+p\left(x_t|\mathcal{Y}_t\right)e_t\Sigma_{v,t}^{-1}dt,(t\geq 0) \quad (8)$$

which reduces to the FPK equation when there are no observations or filtration $\mathcal{Y}_t$. Integrating (8), we have

$$p\left(x_t|\mathcal{Y}_t\right)=p\left(x_0\right)+\int_0^t p\left(x_s|\mathcal{Y}_s\right)ds+\int_0^t \tilde{L}_s p\left(x_s|\mathcal{Y}_s\right)e_s\Sigma_{v,s}^{-1}ds. \quad (9)$$

Given the conditional p.d.f. (9), suppose we want to calculate $\hat{\phi}\left(x_t\right)=E\left[\phi(x_t)|\mathcal{Y}_t\right]$ for any nonlinear function $\phi\in\mathbb{R}^{N_x}$. By interchanging the order of integrations, we have

$$\begin{aligned}
&\hat{\phi}\left(x_t\right)=\int_{-\infty}^{\infty}\phi\left(x\right)p\left(x_t|Y_t\right)dx\\
&=\int_{-\infty}^{\infty}\phi\left(x\right)p\left(x_0\right)dx\\
&+\int_0^t\int_{-\infty}^{\infty}\phi\left(x\right)\tilde{L}_s p\left(x_s|Y_s\right)dxds\\
&+\int_0^t\int_{-\infty}^{\infty}\phi\left(x\right)p\left(x_s|Y_s\right)e_s\Sigma_{v,s}^{-1}dxds=\\
&=E\left[\phi\left(x_0\right)\right]+\int_0^t\int_{-\infty}^{\infty}p\left(x_s|Y_s\right)L_s\phi\left(x\right)dxds\\
&+\int_0^t\left[\int_{-\infty}^{\infty}\phi\left(x\right)g\left(s,x\right)p\left(x_s|Y_s\right)dx\right.\\
&\left.-\hat{g}\left(x_s\right)\int_{-\infty}^{\infty}\phi\left(x\right)p\left(x_s|Y_s\right)dx\right]\Sigma_{v,s}^{-1}ds.
\end{aligned}$$

The Kushner equation lends itself a recursive form of filtering solution, but the conditional mean requests all of higher-order conditional moments and thus leads to an infinite-dimensional system. On the other hand, under some mild conditions, the unnormalized conditional density of x_t given $\mathcal{Y}_s$, denoted as $\pi(x_t|\mathcal{Y}_t)$, is the unique solution of the following stochastic partial differential equation (PDE) the so-called Zakai equation

$$d\pi\left(x_t|Y_t\right)=\tilde{L}\pi\left(x_t|Y_t\right)dt+g\left(t,x_t\right)\pi\left(x_t|Y_t\right)dy_t$$

with the same $\tilde{L}$ defined before. Zakai equation and Kushner equation have a *one-to-one* correspondence, but Zakai equation is much simpler, hence we are usually turned to solve the Zakai equation instead of Kushner equation.In the early history of nonlinear filtering, the common way is to discretize the Zakai equation to seek the numerical solution.

3. BAYESIAN STATISTICS AND BAYESIAN ESTIMATION

3.1 Bayesian Statistics

Bayesian theory is a branch of mathematical probability theory that allows people to model the uncertainty about the world and the outcomes of interest by incorporating prior knowledge and observational evidence. Bayesian analysis, interpreting the probability as a conditional measure of uncertainty, is one of the popular methods to solve the inverse problems. Before running into Bayesian inference and Bayesian estimation, we first introduce some fundamental Bayesian statistics.

Definition 4. (Bayesian Sufficient Statistics). Let $p(x,\mathcal{Y})$ denote the probability density of x conditioned on measurements $\mathcal{Y}$. A statistics, $\Psi\left(x\right)$, is said to be sufficient if the distribution of x conditionally on Ψ does not depend on $\mathcal{Y}$. In other words, $p\left(x,\mathcal{Y}\right)=p\left(x,\mathcal{Y}'\right)$ for any two sets $\mathcal{Y}$ and $\mathcal{Y}'$ s.t. $\Psi\left(\mathcal{Y}\right)=\Psi\left(\mathcal{Y}'\right)$. □

The sufficient statistics $\Psi(x)$ contains all of information brought by x about $\mathcal{Y}$. Sufficiency principle and likelihood principle are two axiomatic principles in the Bayesian inference. Among different intractable problems in Bayesian inference, we are concerned with the following one:

- *Expectation*: Given the conditional p.d.f., some averaged statistics of interest can be calculated

$$E_{p(x|y)}\left[f\left(x\right)\right]=\int_X f\left(x\right)p\left(x|y\right)dx$$

In Bayesian inference, all of uncertainties (including states, parameters which are either time-varying or fixed but unknown priors) are treated as random variables. The inference is performed within the Bayesian framework given all of available information. And the objective of Bayesian inference is to use priors and causal knowledge, quantitatively and qualitatively, to infer the conditional probability, given finite observations.

3.2 Recursive Bayesian Estimation

In the following, we present a detailed derivation of recursive Bayesian estimation, which underlies the principle of sequential Bayesian filtering. Two assumptions are used to derive the recursive Bayesian filter:

(1) The states follow a first-order Markov process

$$p(x_n|x_{0:n-1})=p(x_n|x_{n-1}).$$

(2) the observations are independent of the given states.

For notation simplicity, we denote Y_n as a set of observations $y_{0:n}:=\{y_0,...,y_n\}$; let $p(x_n|Y_n)$ denote the conditional p.d.f. of x_n. From Bayes rule we have

$$
\begin{aligned}
p\left(x_n|\mathcal{Y}_n\right) &= \frac{p\left(\mathcal{Y}_n|x_n\right)p\left(x_n\right)}{p\left(\mathcal{Y}_n\right)} \\
&= \frac{p\left(y_n,\mathcal{Y}_{n-1}|x_n\right)p\left(x_n\right)}{p\left(y_n,\mathcal{Y}_{n-1}\right)} \\
&= \frac{p\left(y_n|\mathcal{Y}_{n-1},x_n\right)p\left(\mathcal{Y}_{n-1}|x_n\right)p\left(x_n\right)}{p\left(y_n|\mathcal{Y}_{n-1}\right)p\left(\mathcal{Y}_{n-1}\right)} \\
&= \frac{p\left(y_n|\mathcal{Y}_{n-1},x_n\right)p\left(x_n|\mathcal{Y}_{n-1}\right)p\left(\mathcal{Y}_{n-1}\right)p\left(x_n\right)}{p\left(y_n|\mathcal{Y}_{n-1}\right)p\left(\mathcal{Y}_{n-1}\right)p\left(x_n\right)} \\
&= \frac{p\left(y_n|Y_{n-1},x_n\right)p\left(x_n|\mathcal{Y}_{n-1}\right)}{p\left(y_n|\mathcal{Y}_{n-1}\right)}
\end{aligned}
\tag{10}
$$

As shown in (10), the posterior density $p(x_n|\mathcal{Y}_n)$ is described by three terms:

- **Prior:** The prior $p(x_n|\mathcal{Y}_{n-1})$ defines the knowledge of the model
$$p\left(x_n|\mathcal{Y}_{n-1}\right) = \int p\left(x_n|x_{n-1}\right)p\left(x_{n-1}|\mathcal{Y}_{n-1}\right)dx_{n-1}$$
where $p(x_n|x_{n-1})$ is the transition density of the state.
- **Likelihood:** the likelihood $p(y_n|x_n)$ essentially determines the measurement noise model in the measurement equation.
- **Evidence:** The denominator involves an integral
$$p\left(y_n|\mathcal{Y}_{n-1}\right) = \int p\left(y_n|x_n\right)p\left(x_n|\mathcal{Y}_{n-1}\right)dx_n$$

Calculation or approximation of these three terms are the essences of the Bayesian filtering.

4. BAYESIAN OPTIMAL FILTERING

Bayesian filtering is aimed to apply the Bayesian statistics and Bayes rule to probabilistic inference problems, and specifically the stochastic filtering problem. In the past few decades, numerous authors have investigated the Bayesian filtering in a dynamic state space framework [Arnaud et al. (2008)].

4.1 Optimal Filtering

An optimal filter is said "optimal" only in some specific sense, in other other words, one should define a criterion which measures the optimality. In the sequel, the *criterium* that we are interested for measuring the optimality is:

- **Minimum mean-squared error (MMSE):** It can be defined in terms of prediction or filtering error (or equivalently the trace of state-error covariance)
$$E\left[\left\|x_n-\hat{x}_n\right\|^2|y_{0:n}\right] = \int \left\|x_n-\hat{x}_n\right\|^2 p\left(x_n|y_{0:n}\right)dx_n$$
which is aimed to find the conditional mean $\hat{x}_n = E\left[x_n|y_{0:n}\right] = \int x_n p\left(x_n|y_{0:n}\right)dx_n$.

As a remark MMSE method require the estimation of the posterior distribution (density), i.e. full knowledge of the prior, likelihood and evidence. The criterion of optimality used for Bayesian filtering is the Bayes risk of MMSE. Bayesian filtering is optimal in a sense that it seeks the posterior distribution which integrates and uses all of available information expressed by probabilities (assuming they are quantitatively correct). However, as time proceeds, one needs infinite computing power and unlimited memory to calculate the "optimal" solution, except in some special cases (e.g. linear Gaussian or conjugate family case). Hence in general, we can only seek a suboptimal or locally optimal solution.

4.2 Kalman Filter

Kalman filter, or Kalman-Bucy filter , consists of an iterative prediction-correction process. In the prediction step, the time update is taken where the one-step ahead prediction of observation is calculated; in the correction step, the measurement update is taken where the correction to the estimate of current state is calculated. In a stationary situation, the matrices A_n, B_n, C_n, D_n in (4) and (5) are constant, Kalman filter is precisely the Wiener filter for stationary least-squares smoothing.

Kalman filter is also optimal in the sense that it is unbiased $E[\hat{x}_n] = E[x_n]$ and is a minimum variance estimate. Kalman filter has a very nice Bayesian interpretation [Chen (2003)].

4.3 Optimum Nonlinear Filtering

In practice, the use of Kalman filter is limited by the ubiquitous nonlinearity and non-Gaussianity of physical world. Hence since the publication of Kalman filter, numerous efforts have been devoted to the generic filtering problem, mostly in the Kalman filtering framework.

In general, the nonlinear filtering problem *per si* consists in finding the conditional probability distribution (or density) of the state given the observations up to current time. Strictly speaking, the number of variables replacing the density function is infinite, but not all of them are of equal importance. Thus it is advisable to select the important ones and reject the remainder. The solutions of nonlinear filtering problem have two categories: global method and local method. In the global approach, one attempts to solve a PDE instead of an ODE in linear case, e.g. Zakai equation, Kushner-Stratonovich equation, which are mostly analytically intractable. Hence the numerical approximation techniques are needed to solve the equation. In special scenarios (e.g. exponential family) with some assumptions, the nonlinear filtering can admit the tractable solutions. In the local approach, finite sum approximation (e.g. Gaussian sum filter) or linearization techniques (i.e. EKF) are usually used. In the EKF, by defining

$$\hat{F}_{n+1,n} = \left.\frac{df(x)}{dx}\right|_{x=\hat{x}_n}, \quad \hat{G}_n = \left.\frac{dg(x)}{dx}\right|_{x=\hat{x}_{n|n-1}},$$

the equations (2)-(3) can be linearized into (4)-(5), and the conventional Kalman filtering technique is further employed. Because EKF always approximates the posterior $p(x_n|y_{0:n})$ as a Gaussian, it works well for some types of nonlinear problems, but it may provide a poor performance in some cases when the true posterior is non-Gaussian (e.g. heavily skewed or multimodal). Gelb [Gelb (1974)] provided an early overview of the uses of EKF. It is noted that the estimate given by EKF is usually biased since in general $E[f(x)] \neq f(E[x])$.

In summary, a number of methods have been developed for nonlinear filtering problems:

- Linearization methods: first-order Taylor series expansion (i.e. EKF), and higher-order filter.
- Numerical approximation methods, as to be discussed hereafter.

Among many other that can easily found in the literature[Chen (2003)].

5. NUMERICAL APPROXIMATION METHODS

5.1 Multigrid Method and Point-Mass Approximation

If the state is discrete and finite (or it can be discretized and approximated as finite), grid-based methods can provide a good solution and optimal way to update the filtered density $p(z_n|y_{0:n})$ (To discriminate from the continuous valued state x, we denote the discrete-valued state as z from now on). Suppose the discrete state $z \in \mathbb{N}$ consists of a finite number of distinct discrete states $\{1, 2, ..., N_z\}$. For the state space z_{n-1}, let $w^i_{n-1|n-1}$ denote the conditional probability of each z^i_{n-1} given measurement up to $n-1$, i.e. $p(z_{n-1} = z^i|y_{0:n-1}) = w^i_{n-1|n-1}$. Then the posterior p.d.f. at $n-1$ can be represented as

$$p\left(z_{n-1}|y_{0:n-1}\right) = \sum_{i=1}^{N_z} w^i_{n|n-1}\delta\left(z_n - z^i_n\right),$$

$$p\left(z_n|y_{0:n}\right) = \sum_{i=1}^{N_z} w^i_{n|n}\delta\left(z_n - z^i_n\right),$$

where

$$w^i_{n|n-1} = \sum_{j=1}^{N_z} w^j_{n-1|n-1} p\left(z^i_n|z^j_{n-1}\right)$$

$$w^i_{n|n} = \frac{w^i_{n|n-1} p\left(y_n|z^i_n\right)}{\sum_{j=1}^{N_z} w^j_{n|n-1} p\left(y_n|z^j_n\right)}$$

If the state space is continuous, the approximate-grid based method can be similarly derived. Namely, we can always discretized the state space into N_z discrete cell states, then a grid-based method can be further used to approximate the posterior density. The grid must be sufficient dense to obtain a good approximation, especially when the dimensionality of N_x is high, however the increase of N_z will increase the computational burden dramatically. If the state space is not finite, then the accuracy of grid-based method is not guaranteed.The disadvantage of grid-based method is that it requires the state space cannot be partitioned unevenly to give a great resolution to the state with high density. For further details the reader is referred to [Arulampalam et al. (2002)].

5.2 Monte Carlo Sampling Approximation

Monte Carlo methods use statistical sampling and estimation techniques to evaluate the solutions to mathematical problems. Only Monte Carlo sampling methods are discussed hereafter. A detailed background of Monte Carlo methods can refer to the book [Arnaud Doucet and Gordon (2001)] and survey papers [Arnaud et al. (2008)and Arulampalam et al. (2002)]. The underlying mathematical concept of Monte Carlo approximation is simple. Consider a statistical problem estimating a Lebesque-Stieltjes integral:

$$\int_X f\left(x\right) dP\left(x\right),$$

where $f(x)$ is an integrable function in a measurable space. As a brute force technique, Monte Carlo sampling uses a number of (independent) random variables in a probability space $(\Omega, \mathcal{F}, P)$ to approximate the true integral. Provided one draws a sequence of N_p i.i.d. random samples $\{x^{(1)}, ..., x^{(Np)}\}$ from probability distribution $P(x)$, then the Monte Carlo estimate of $f(x)$ is given by

$$\hat{f}_{N_p} = \frac{1}{N_p}\sum_{i=1}^{N_p} f\left(x^{(i)}\right)$$

for which $E\left[\hat{f}_{N_p}\right] = E\left[f\right]$ and $Var\left[\hat{f}_{N_p}\right] = \frac{1}{N_p} Var\left[f\right] = \frac{\sigma^2}{N_p}$. By the Kolmogorov Strong Law of Large Numbers (under some mild regularity conditions), $\hat{f}_{N_p}(x)$ converges to $E[f(x)]$ almost surely (a.s.) and its convergence rate is assessed by the Central Limit Theorem

$$\sqrt{N_p}\left(\hat{f}_{N_p} - E\left[f\right]\right) \sim \mathcal{N}\left(0, \sigma^2\right)$$

where σ^2 is the variance of $f(x)$. Namely, the error rate is of order $O(N_p^{-1/2})$, which is slower than the order $O(N_p^{-1}$ for deterministic quadrature in one-dimensional case. One crucial property of Monte Carlo approximation is the estimation accuracy is independent of the dimensionality of the state space, as opposed to most deterministic numerical methods. The variance of estimate is inversely proportional to the number of samples. There are two fundamental problems arising in Monte Carlo sampling methods:

(1) How to draw random samples $\{x^{(i)}\}$ from a probability distribution $P(x)$?;
(2) How to estimate the expectation of a function w.r.t. the distribution or density, i.e. $E[f(x)] = f(x)dP(x)$?

The first problem is a design problem, and the second one is an inference problem invoking integration. Besides, several central issues are concerned in the Monte Carlo sampling:

- **Consistency:** An estimator is consistent if the estimator converges to the true value almost surely as the number of observations approaches infinity.
- **Unbiasedness:** An estimator is unbiased if its expected value is equal to the true value.
- **Efficiency:** An estimator is efficient if it produces the smallest error covariance matrix among all unbiased estimators, it is also regarded optimally using the information in the measurements. A well-known efficiency criterion is the Cramér-Rao bound.
- **Robustness:** An estimator is robust if it is insensitive to the gross measurement errors and the uncertainties of the model.
- **Minimal variance:** Variance reduction is the central issue of various Monte Carlo approximation methods,

most improvement techniques are variance-reduction oriented.

In the rest of this subsection, we will provide a brief introduction of many popular Monte Carlo method relevant to our paper. No attempt is made here to present a complete and rigorous theory. For more theoretical details or applications, reader is referred to the book [Arnaud Doucet and Gordon (2001)].

Importance Sampling The objective of importance sampling is aimed to sample the distribution in the region of "importance" in order to achieve computational efficiency. This is important especially for the high-dimensional space where the data are usually sparse, and the region of interest where the target lies in is relatively small in the whole data space. The idea of importance sampling is to choose a proposal distribution $q(x)$ in place of the true probability distribution $p(x)$, which is hard-to-sample. The support of $q(x)$ is assumed to cover that of $p(x)$. Rewriting the integration problem as

$$\int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx. \tag{11}$$

Monte Carlo importance sampling is to use a number of (say N_p) independent samples drawn from $q(x)$ to obtain a weighted sum to approximate (11):

$$\hat{f} = \frac{1}{N_p}\sum_{i=1}^{N_p} W\left(x^{(i)}\right) f\left(x^{(i)}\right)$$

where $W\left(x^{(i)}\right) = \frac{p\left(x^{(i)}\right)}{q\left(x^{(i)}\right)}$ are called the importance weights (or importance ratios). If the normalizing factor of $p(x)$ is not known, the importance weights can be only evaluated up to a normalizing constant, hence $W\left(x^{(i)}\right) \propto \frac{p\left(x^{(i)}\right)}{q\left(x^{(i)}\right)}$. To ensure that $\sum_{i=1}^{N_p} W\left(x^{(i)}\right) = 1$, we normalize the importance weights to obtain

$$\hat{f} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p} W\left(x^{(i)}\right) f\left(x^{(i)}\right)}{\frac{1}{N_p}\sum_{i=1}^{N_p} W\left(x^{(i)}\right)} = \frac{1}{N_p}\sum_{i=1}^{N_p} \tilde{W}\left(x^{(i)}\right) f\left(x^{(i)}\right)$$

where $\tilde{W}\left(x^{(i)}\right) = \frac{W\left(x^{(i)}\right)}{\sum_{j=1}^{N_p} W\left(x^{(j)}\right)}$ are called the normalized importance weights. The variance of the importance sampler estimate (11) is given by

$$\begin{aligned}
&Var_q\left[\hat{f}\right] = \frac{1}{N_p}Var_q\left[f(x)W(x)\right] \\
&= \frac{1}{N_p}Var_q\left[f(x)p(x)/q(x)\right] \\
&= \frac{1}{N_p}\int\left(\frac{f(x)p(x)}{q(x)} - E_p\left[f\left(x\right)\right]\right)^2 q(x)dx \\
&= \frac{1}{N_p}\int\left(\left(\frac{f(x)p(x)}{q(x)}\right)^2 - 2p(x)f\left(x\right)E_p\left[f\left(x\right)\right]\right)dx \\
&+ \frac{\left(E_p\left[f\left(x\right)\right]\right)^2}{N_p} \\
&= \frac{1}{N_p}\int\left(\frac{f(x)p(x)}{q(x)}\right)^2 dx + \frac{\left(E_p\left[f\left(x\right)\right]\right)^2}{N_p}
\end{aligned}$$

The variance can be reduced when an appropriate $q(x)$ is chosen to

- match the shape of $p(x)$ so as to approximate the true variance;
- match the shape of $|f(x)|p(x)$ so as to further reduce the true variance.

Importance sampling estimate given by $\hat{f}$ is biased (thus a.k.a. biased sampling) but consistent, namely the bias vanishes rapidly at a rate $O(N_p)$. Provided q is appropriately chosen, as $N_p \to \infty$, from the Weak Law of Large Numbers, we know

$$\hat{f} \to \frac{E_q\left[W\left(x\right)f\left(x\right)\right]}{E_q\left[W\left(x\right)\right]}$$

It was also shown that if $E\left[\tilde{W}\left(x\right)\right] < \infty$ and

$$E\left[\tilde{W}\left(x\right)f^2\left(x\right)\right] < \infty,$$

then $\hat{f} \to E_p\left[f\right]$ a.s. and the Lindeberg-Levy Central Limit Theorem still holds

$$\sqrt{N_p}\left(\hat{f} - E_p\left[f\right]\right) \sim \mathcal{N}\left(0, \Sigma_f\right)$$

where

$$\Sigma_f = Var_q\left[\tilde{W}\left(x\right)\left(f\left(x\right) - \left(\hat{f} - E_p\left[f\left(x\right)\right]\right)\right)\right].$$

Remarks:

- Importance sampling is useful in two ways: (i) it provides an elegant way to reduce the variance of the estimator (possibly even less than the true variance); and (ii) it can be used when encountering the difficulty to sample from the true probability distribution directly.
- As shown in many empirical experiments, importance sampler (proposal distribution) should have a heavy tail so as to be insensitive to the outliers. The super-Gaussian distributions usually have long tails, with kurtosis bigger than 3.
- Although theoretically the bias of importance sampler vanishes at a rate $O(N_p)$, the accuracy of estimate is not guaranteed even with a large N_p. If $q(\cdot)$ is not close to $p(\cdot)$, it can be imagined that the weights are very uneven, thus many samples are almost useless because of their negligible contributions. In a highdimensional space, the importance sampling

estimate is likely dominated by a few samples with large importance weights.

For a nice/elementary introduction to this subject the reader is referred to [Anderson (1999)].

Control Variates Suppose we want to approximate $E[f]$ using a simple Monte Carlo average $\bar{f}$. If there is another payoff g for which we know $E[g]$, can use $\bar{g}-E[g]$ to reduce error in $\bar{f}-E[f(X)]$. This is achieved by defining a new estimator

$$\hat{f}=\bar{f}-\lambda\left(\bar{g}-E[g]\right),$$

that is unbiased since

$$E\left[\hat{f}\right]=E\left[\bar{f}\right]=E[f].$$

On the other hand, for a single sample we have the following variance

$$Var[f-\lambda(\bar{g}-E[g])]=Var[f]-2\lambda Cov[f,g]+\lambda^2 Var[g],$$

and for a average of N samples we end up with

$$\begin{aligned} &Var[f-\lambda(\bar{g}-E[g])]= \\ &\quad N^{-1}\left(Var[f]-2\lambda Cov[f,g]+\lambda^2 Var[g]\right). \end{aligned} \tag{12}$$

To minimize this, the optimum value for λ is

$$\lambda=\frac{Cov[f,g]}{Var[g]}. \tag{13}$$

And the resulting variance is when we replace (13) in (12) we have

$$N^{-1}Var[f]\left(1-\frac{(Cov[f,g])^2}{Var[f]\,Var[g]}\right)=N^{-1}Var[f]\left(1-\rho^2\right)$$

Where ρ is the correlation between f and g. The challenge is the choose a good g which is well correlated with f - the covariance, and hence the optimal ρ, can be estimated from the data.

Sequential Importance Sampling A good proposal distribution is essential to the efficiency of importance sampling, hence how to choose an appropriate proposal distribution $q(x)$ is the key to apply a successful importance sampling. However, it is usually difficult to find a good proposal distribution especially in a high-dimensional space. A natural way to alleviate this problem is to construct the proposal distribution sequentially, which is the basic idea of sequential importance sampling (SIS) Arnaud Doucet and Gordon (2001). In particular, if the proposal distribution is chosen in a factorized form

$$q(x_{0:n}|y_{0:n})=q(x_0)\prod_{i=1}^{n}q(x_t|x_{0:t-1},y_{0:t}),$$

then the importance sampling can be performed recursively. At this moment, we consider a simplified (unconditional p.d.f.) case for the ease of understanding. According to the "telescope" law of probability, we have the following:

$$\begin{aligned} p(x_{0:n})&=p(x_0)\,p(x_1|x_0)\ldots p(x_n|x_0,\ldots,x_{n-1}),\\ q(x_{0:n})&=q(x_0)\,q(x_1|x_0)\ldots q(x_n|x_0,\ldots,x_{n-1}). \end{aligned}$$

Hence the importance weight $W(x_{0:n})$ can be written as

$$W(x_{0:n})=\frac{p(x_0)\,p(x_1|x_0)\ldots p(x_n|x_0,\ldots,x_{n-1})}{q(x_0)\,q(x_1|x_0)\ldots q(x_n|x_0,\ldots,x_{n-1})}$$

which can be recursively calculated as

$$W_n(x_{0:n})=W_{n-1}(x_{0:n-1})\frac{p(x_n|x_{0:n-1})}{q_n(x_n|x_{0:n-1})}$$

Remarks:

- The advantage of SIS is that it doesn't rely on the underlying Markov chain. Instead, many i.i.d. replicates are run to create an importance sampler, which consequently improves the efficiency. The disadvantage of SIS is that the importance weights may have large variances, resulting in inaccurate estimate
- SIS method can be also used in a non-Bayesian computation, such as evaluation of the likelihood function in the missing-data problem
- It was shown in that the unconditional variance of the importance weights increases over time, which is the so-called weight degeneracy problem: Namely, after a few iterations of algorithm, only few or one of $W(x^{(i)})$ will be nonzero. This is disadvantageous since a lot of computing effort is wasted to update those trivial weight coefficients. In order to cope with this situation, resampling step is suggested to be used after weight normalization.

Sampling-Importance Resampling The Sampling Importance Resampling (SIR) is motivated from the Bootstrap and jackknife techniques. Bootstrap technique is referred to a collection of computationally intensive methods that are based on resampling from the observed data. The intuition of bootstrapping is to evaluate the properties of an estimator through the empirical cumulative distribution function (c.d.f.) of the samples instead of the true c.d.f.. In the statistics literature, Rubin first applied SIR technique to Monte Carlo inference, in which the resampling is inserted between two importance sampling steps. The resampling step is aimed to eliminate the samples with small importance weights and duplicate the samples with big weights. The generic principle of SIR proceeds as follows:

- Draw N_p random samples $\{x^{(i)}\}_{N_p}^{i=1}$ from proposal distribution $q(x)$;
- Calculate importance weights $W^{(i)}\propto p(x)/q(x)$ for each sample $x^{(i)}$;
- Normalize the importance weights to obtain $\tilde{W}^{(i)}$
- Resample with replacement N times from the discrete set $\{x^{(i)}\}_{N_p}^{i=1}$, where the probability of resampling from each $x^{(i)}$ is proportional to $\tilde{W}^{(i)}$

Remarks (on features):

- Resampling usually (but not necessarily) occurs between two importance sampling steps. In resampling step, the particles and associated importance weights $\{x^{(i)},\tilde{W}^{(i)}\}$ are replaced by the new samples with equal importance weights (i.e. $\tilde{W}^{(i)}=1/N_p$). Resampling can be taken at every step or only taken if regarded necessary.
- Resampling step plays an critical role in importance sampling since (i) if importance weights are uneven distributed, propagating the "trivial" weights through the dynamic system is a waste of comput-

ing power; (ii) when the importance weights are skewed, resampling can provide chances for selecting important samples and rejuvenate the sampler for the future use, though resampling doesn't necessarily improve the current state estimate because it also introduces extra Monte Carlo variation.
- Resampling schedule can be deterministic or dynamic. In deterministic framework, resampling is taken at every k time step (usually $k = 1$). In a dynamic schedule, a sequence of thresholds (that can be constant or time-varying) are set up and the variance of the importance weights are monitored; resampling is taken only when the variance is over the threshold.

The validity of inserting a resampling step in SIS algorithm has been justified, since resampling step also brings extra variation, some special schemes are needed. There are many types of resampling methods available in the literature:

(1) Multinomial resampling: the procedure reads as follows
- Produce a uniform distribution $u \sim \mathcal{U}(0,1)$, construct a c.d.f. for importance weights, calculate $s_i = \sum_{j=1}^{i} \tilde{W}^{(j)}$
- Find s_i s.t. $s_{i-1} \leq u < s_i$, the particle with index i is chosen;
- Given $\{x^{(i)}, \tilde{W}^{(i)}\}$, for $j = 1, ..., N_p$, generate new samples $x^{(j)}$ by duplicationg $x^{(i)}$ according to the associated $\tilde{W}^{(i)}$;
- Reset $W^{(i)} = 1/N_p$.

Multinomial resampling uniformly generates N_p new independent particles from the old particle set. Each particle is replicated N_i times (N_i can be zero), namely each $x^{(i)}$ produces N_i children. Note that here $\sum_{i=1}^{N_p} N_i = N_p$, $E[N_i] = N_p \tilde{W}^{(i)}$, $Var[N_i] = N_p \tilde{W}^{(i)}(1 - \tilde{W}^{(i)})$.

(2) Residual resampling: suggests a partially deterministic resampling method. The two-step selection procedure is as follows:
- For each $i = 1, ..., N_p$ retain $k_i = \left\lfloor N_p \tilde{W}^{(i)} \right\rfloor$ copies of $x_n^{(i)}$;
- Let $N_r = N_p - k_1 - ... - K_{N_p}$, obtain N_r i.i.d. draws from $\{x_n^{(i)}\}$ with probabilities proportional to $N_p \tilde{W}^{(i)} - k_i$ $(i = 1, ..., N_p)$
- Reset $W^{(i)} = 1/N_p$.

Residual resampling procedure is computationally cheaper than the conventional SIR and achieves a lower sampler variance, and it doesn't introduce additional bias. Every particle in residual resampling is replicated.

(3) Systematic resampling (or Minimum variance sampling)the procedure proceeds as follows:
$u \sim U(0,1)/N_p; j = 1; l = 0; i = 0;$
do while $u < 1$
if $l > u$ then
$u = u + 1/N_p$; output $x^{(i)}$
else
pick k in $\{j, ..., N_p\}$ $i = x^{(k)}, l = l + W^{(i)}$
switch $(x^{(k)}, W^{(k)})$ with $(x^{(j)}, W^{(j)})$
$j = j + 1$
end if
end do

The systematic resampling treats the weights as continuous random variables in the interval $(0, 1)$, which are randomly ordered. The number of grid points $\{u + k/Np\}$ in each interval is counted. Every particle is replicated and the new particle set is chosen to minimize $Var[N_i] = E[(N_i - E[N_i])^2]$. The complexity of systematic resampling is $O(N_p)$.

Remarks (on weakness):

- SIR only achieves approximate draws from the posterior as $N_p \to \infty$.
- Although resampling can alleviate the weight degeneracy problem, it unfortunately introduces other problems: after one resampling step, the simulated trajectories are not statistically independent any more, thus the convergence result due to the original central limit theorem is invalid; resampling causes the samples that have high importance weights to be statistically selected many times, thus the algorithm suffers from the loss of diversity.
- Resampling step also limits the opportunity to parallelize since all of the particles need to be combined for selection.

6. SEQUENTIAL MONTE CARLO ESTIMATION: PARTICLE FILTERS

With the background knowledge of stochastic filtering, Bayesian statistics, and Monte Carlo techniques, we are now in a good position to discuss the theory and paradigms of particle filters. In this section, we focus the attention on the sequential Monte Carlo approach for sequential state estimation. Sequential Monte Carlo technique is a kind of recursive Bayesian filter based on Monte Carlo simulation, it is also called bootstrap filter. The working mechanism of particle filters is following: The state space is partitioned as many parts, in which the particles are filled according to some probability measure. The higher probability, the denser the particles are concentrated. The particle system evolves along the time according to the state equation, with evolving p.d.f. determined by the FPK equation. Since the p.d.f. can be approximated by the point-mass histogram, by random sampling of the state space, we get a number of particles representing the evolving p.d.f.. However, since the posterior density model is unknown or hard to sample, we would rather choose another distribution for the sake of efficient sampling. To avoid intractable integration in the Bayesian statistics the posterior distribution or density is empirically represented by a weighted sum of N_p samples drawn from the posterior distribution

$$p(x_n|\mathcal{Y}_n) \approx \frac{1}{N_p} \sum_{n=1}^{N_p} \delta\left(x_n - x_n^{(i)}\right) \equiv \hat{p}(x_n|\mathcal{Y}_n)$$

where $x^{(i)}$ are assumed to be i.i.d. drawn from $p(x_n|\mathcal{Y}_n)$. When N_p is sufficiently large, $\hat{p}(x_n|\mathcal{Y}_n)$ approximates the

true posterior $p(x_n|\mathcal{Y}_n)$. By this approximation, we can estimate the mean of a nonlinear function

$$\begin{aligned} E[f(x_n)] &\approx \int f(x_n)\hat{p}(x_n|\mathcal{Y}_n)\,dx_n \\ &= \frac{1}{N_p}\sum_{n=1}^{N_p}\int f(x_n)\,\delta\left(x_n - x_n^{(i)}\right)dx_n \\ &= \frac{1}{N_p}\sum_{n=1}^{N_p} f\left(x_n^{(i)}\right) \equiv \hat{f}_{N_p}(x). \end{aligned}$$

Since it is usually impossible to sample from the true posterior, it is common to sample from an easy-to-implement distribution, the so called proposal distribution denoted by $q(x_n|\mathcal{Y}_n)$, hence

$$\begin{aligned} E[f(x_n)] &= \int f(x_n)\frac{p(x_n|\mathcal{Y}_n)}{q(x_n|\mathcal{Y}_n)}q(x_n|\mathcal{Y}_n)\,dx_n \\ &= \int f(x_n)\frac{W(x_n)}{p(\mathcal{Y}_n)}q(x_n|\mathcal{Y}_n)\,dx_n \\ &= \frac{1}{p(\mathcal{Y}_n)}\int f(x_n)W_n(x_n)\,q(x_n|\mathcal{Y}_n)\,dx_n \end{aligned} \tag{14}$$

where

$$W_n(x_n) = \frac{p(Y_n|x_n)\,p(x_n)}{q(x_n|Y_n)}.$$

Finally (14) can be rewritten as

$$\begin{aligned} E[f(x_n)] &= \frac{\int f(x_n)W_n(x_n)\,q(x_n|Y_n)\,dx_n}{\int W_n(x_n)\,q(x_n|Y_n)\,dx_n} \\ &= \frac{E_{q(x_n|Y_n)}[W_n(x_n)\,f(x_n)]}{E_{q(x_n|Y_n)}[W_n(x_n)]}. \end{aligned} \tag{15}$$

By drawing the i.i.d. samples $\{x_n^{(i)}\}$ from $q(x_n|Y_n)$, we can approximate (15) by

$$\begin{aligned} E[f(x_n)] &\approx \frac{\frac{1}{N_p}\sum_{i=1}^{N_p} W_n\left(x_n^{(i)}\right) f\left(x_n^{(i)}\right)}{\frac{1}{N_p}\sum_{i=1}^{N_p} W_n\left(x_n^{(i)}\right)} \\ &= \sum_{i=1}^{N_p}\tilde{W}_n\left(x_n^{(i)}\right) f\left(x_n^{(i)}\right) = \hat{f}(x), \end{aligned}$$

where

$$\tilde{W}_n\left(x_n^{(i)}\right) = \frac{W_n\left(x_n^{(i)}\right)}{\sum_{j=1}^{N_p} W_n\left(x_n^{(j)}\right)}.$$

Suppose the proposal distribution has the following factorized form

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

The posterior $p(x_{0:n}|y_{0:n})$ can be factorized as

$$p(x_{0:n}|y_{0:n}) = p(x_{0:n-1}|y_{0:n-1})\,\frac{p(y_n|x_n)\,p(x_n|x_{n-1})}{p(y_n|y_{0:n-1})}.$$

Thus the importance weights $W_n^{(i)}$ can be updated recursively

$$\begin{aligned} W_n^{(i)} &= \frac{p\left(x_{0:n}^{(i)}|y_{0:n}\right)}{q\left(x_{0:n}^{(i)}|y_{0:n}\right)} \\ &\propto \frac{p\left(y_n|x_n^{(i)}\right)p\left(x_n^{(i)}|x_{n-1}^{(i)}\right)p\left(x_{0:n-1}^{(i)}|y_{0:n-1}\right)}{q\left(x_n^{(i)}|x_{0:n-1}^{(i)},y_{0:n}\right)q\left(x_{0:n-1}^{(i)}|,y_{0:n-1}\right)} \\ &= W_{n-1}^{(i)}\frac{p\left(y_n|x_n^{(i)}\right)p\left(x_n^{(i)}|x_{n-1}^{(i)}\right)}{q\left(x_n^{(i)}|x_{0:n-1}^{(i)},y_{0:n}\right)}. \end{aligned} \tag{16}$$

6.1 Sequential Importance Sampling (SIS) Filter

In practice, we are more interested in the current filtered estimate $p(x_n|y_{0:n})$ instead of $p(x_{0:n}|y_{0:n})$. Provided $q(x_n^{(i)}|x_{0:n-1}^{(i)},y_{0:n})$ is assumed to be equivalent to $q(x_n^{(i)}|x_{0:n-1}^{(i)},y_n)$, (16) can be simplified as

$$W_n^{(i)} = W_{n-1}^{(i)}\frac{p\left(y_n|x_n^{(i)}\right)p\left(x_n^{(i)}|x_{n-1}^{(i)}\right)}{q\left(x_n^{(i)}|x_{0:n-1}^{(i)},y_n\right)}.$$

As discussed earlier, the problem of the SIS filter is that the distribution of the importance weights become more and more skewed as time increases. Hence, after some iterations, only very few particles have non-zero importance weights. This phenomenon is often called weight degeneracy or sample impoverishment. An intuitive solution is to multiply the particles with high normalized importance weights, and discard the particles with low normalized importance weights, which can be be done in the resampling step. To monitor how bad is the weight degeneration, we need a measure. A suggested measure for degeneracy, the so-called effective sample size, N_{eff} , was introduced as

$$\begin{aligned} N_{eff} &= \frac{N_p}{1 + Var_{q(.|y_{0:n})}\left[\tilde{W}(x_{0:n-1})\right]} \\ &= \frac{N_p}{E_{q(.|y_{0:n})}\left[\left(\tilde{W}(x_{0:n-1})\right)^2\right]} \le N_p \end{aligned}$$

The second equality above follows from the facts that $Var[\xi] = E[\xi^2] - E[\xi]^2$ and $E_q[\tilde{W}] = 1$. In practice, the true N_{eff} is not available, thus its estimate, $\hat{N}_{eff}$, is alternatively given:

$$\hat{N}_{eff} = \frac{1}{\sum_{i=1}^{N_p}\left(\tilde{W}_n^{(i)}\right)^2}$$

When $\hat{N}_{eff}$ is below a predefined threshold N_T (say $N_p/2$ or $N_p/3$), the resampling procedure is performed. The idea is following: when the $\hat{N}_{eff} < N_T$ (where N_T can be either a predefined value or the median of the weights), then each sample is accepted with probability $\min\{1, W_n^{(i)}/N_T\}$; all the accepted samples are given a new weight $W_n^{(j)} = \max\{N_T, W_n^{(i)}\}$, and the rejected samples are restarted

and rechecked at the all previously violated thresholds. It is obvious that this procedure is computational expensive as n increases.

6.2 Bootstrap/SIR filter

The Bayesian bootstrap filter due to Gordon, Salmond and Smith [Gordon et al. (1993)], is very close in spirit to the sampling importance resampling (SIR) filter developed independently in statistics by different researchers, with a slight difference on the resampling scheme. Here we treat them as the same class for discussion. The key idea of SIR filter is to introduce the resampling step as we have discussed. The resampling step is flexible and varies from problems as well as the selection scheme and schedule. It should be noted that resampling does not really prevent the weight degeneracy problem, it just saves further calculation time by discarding the particles associated with insignificant weights. What it really does is artificially concealing the impoverishment by replacing the high important weights with many replicates of particles, thereby introducing high correlation between particles. A generic algorithm of Bayesian bootstrap/SIR filter using transition prior density as proposal distribution, where the resampling step is performed at each iteration using any available resampling method discussed earlier.

Remarks

- Both SIS and SIR filters use importance sampling scheme. The difference between them is that in SIR filter, the resampling is always performed (usually between two importance sampling steps); whereas in SIS filter, importance weights are calculated sequentially, resampling is only taken whenever needed, thus SIS filter is less computationally expensive.
- The choice of proposal distributions in SIS and SIR filters plays an crucial role in their final performance
- Resampling step is suggested to be done after the filtering, because resampling brings extra random variation to the current samples. Normally (eps.in off-line processing), the posterior estimate (and its relevant statistics) should be calculated before resampling.
- As suggested by some authors, in the resampling stage, the new importance weights of the surviving particles are not necessarily reset to $1/N_p$, but rather abide certain procedures
- To alleviate the sample degeneracy in SIS filter, we can change the weights update formula as

$$W_n^{(i)} = \left(W_{n-1}^{(i)}\right)^{\alpha} \frac{p\left(y_n|x_n^{(i)}\right) p\left(x_n^{(i)}|x_{n-1}^{(i)}\right)}{q\left(x_n^{(i)}|x_{0:n-1}^{(i)}, y_n\right)}$$

  where the scalar $0 < \alpha < 1$ plays a role as annealing factor that controls the impact of previous importance weights.

7. THEORETICAL AND PRACTICAL ISSUES

7.1 Choices of Proposal Distribution

The potential criteria of choosing a good proposal distribution should include:

- The support of proposal distribution should cover that of posterior distribution, in other words, the proposal should have a broader distribution
- The proposal distribution has a long-tailed behavior to account for outliers.
- Ease of sampling implementation, preferably with linear complexity
- Taking into account of transition prior and likelihood, as well as most recent observation data.
- Achieving minimum variance.
- Being close (in shape) to the true posterior.

However, achieving either of these goals is not easy and we don't know what the posterior suppose to look like. Theoretically, it was shown that the choice of proposal distribution $q\left(x_n|x_{0:n-1}^{(i)}, y_{0:n}\right) = p\left(x_n|x_{n-1}^{(i)}, y_n\right)$ minimizes the variance of importance weights $W_n^{(i)}$ conditional upon $x_{0:n-1}^{(i)}$ and $y_{0:n}$. By this, the importance weights can be recursively calculated as $W_n^{(i)} = W_{n-1}^{(i)} p\left(y_n|x_n^{(i)}\right)$. However, this optimal proposal distribution suffers from certain drawbacks: It requires sampling from $p\left(x_n|x_{n-1}^{(i)}, y_n\right)$ and evaluating the integral $p\left(y_n|x_{n-1}^{(i)}\right) = \int p\left(y_n|x_n\right) p\left(x_n|x_{n-1}^{(i)}\right) dx_n$. On the other hand, it should be also pointed out that there is no universal choice for proposal distribution, which is usually problem dependent. Choosing an appropriate proposal distribution requires a good understanding of the underlying problem. In the following, we present some rules of thumb available in the literature and discuss their features.

Prior Distribution Prior distribution was first used for proposal distribution because of its intuitive simplicity. If $q\left(x_n|x_{0:n-1}, y_{0:n}\right) = p\left(x_n|x_{n-1}\right)$, the importance weights are updated by

$$W_n^{(i)} = W_{n-1}^{(i)} p\left(y_n|x_n^{(i)}\right) \tag{17}$$

which essentially neglects the effect of the most recent observation y_n. This kind of proposal distribution is easy to implement, but usually results in a high variance because the most recent observation y_n is neglected in $p(x_n|x_{n-1})$. The problem becomes more serious when the likelihood is peaked and the predicted state is near the likelihood's tail, in other words, the measurement noise model is sensitive to the outliers. From (17), we know that importance weights are proportional to the likelihood model. It is obvious that $W(x)$ will be very uneven if the likelihood model is not flat. In the Gaussian measurement noise situation, the flatness will be determined by the variance. If Σ_v is small, the distribution of the measurement noise is peaked, hence $W(x)$ will be peaked as well, which makes the the sample impoverishment problem more severe. Hence we can see that, choosing transition prior as proposal is really a brute force approach whose result can be arbitrarily bad, though it was widely used in the literature and sometimes produced reasonably good results (really depending on the noise statistics).

7.2 Convergence and Asymptotic Results

As discussed earlier, although the convergence of Monte Carlo approximation is quite clear, the convergence be-

havior of sequential Monte Carlo method or particle filter is different and deserves special attention. Many authors have explored this issue from different perspectives, but most results are available in the probability literature. A review of convergence results on particle filtering methods has been recently given by Crisan and Doucet from practical point of view [Chen (2003)]. We summarize the main results from their survey paper.

Almost Sure Convergence: If the the transition kernel $K(x_t|x_{t-1})$ is *Feller*, importance weights are upper bounded, and the likelihood function is continuous, bounded, and strictly positive, then with $N_p \to \infty$ the filtered density given by particle filter converges asymptotically to the true posterior.

Mean Square Convergence: If likelihood function is bounded, for any bounded function $\phi \in \mathbb{R}^{N_x}$, then for $t \geq 0$, there exists a $C_{t|t}$ independent of N_p s.t.

$$E\left[\left(\left(\hat{P}_{t|t},\phi\right)-\left(P_{t|t},\phi\right)\right)^2\right] \leq C_{t|t}\frac{\|\phi\|^2}{N_p}, \qquad (18)$$

where $\left(\hat{P}_{t|t},\phi\right) = \int \phi\left(x_{0:n}\right)P\left(dx_{0:t}|y_{0:t}\right)$, $\|\phi\| = \sup_{x_{0:t}} |\phi\left(x_{0:t}\right)|$.
It should be cautioned that, it seems at the first sight that particle filtering method beats the curse of dimensionality, as the rate of convergence, $1/N_p$, is independent on the state dimension N_x. This is nevertheless not true because in order to assure (18) holds, the number of particles N_p needs to increase over the time since it depends on $C_{t|t}$, a term that further relies on N_x. As discussed in, in order to assure the uniform convergence, both $C_{t|t}$ and the approximation error accumulates over the time. In a high-dimensional space (order of tens or higher), particle filters still suffer the problem of curse of dimensionality. Empirically, we can estimate the requirement of the number of particles, although this bound in practice is loose and usually data/problem dependent. Suppose the minimum number is determined by the effective volume (variance) of the search space (proposal) against the target space (posterior). If the proposal and posterior are uniform in two N_x-dimensional hyperspheres with radii r and R $(R > r)$ respectively, the effective particle number N_{eff} is approximately measured by the the volume ratio in the proposal space against posterior space, namely

$$N_{eff} \approx N_p \times (r/R)^{N_x}$$

when the ration is low $r << R$, the effective number decreases exponentially as N_x increases; on the other hand, if we want to keep the effective number as a constant, we need to increase N_p exponentially as N_x increases. An important asymptotic result is the error bound of the filter. According to the Cramér-Rao theorem, the expected square error of an estimate is generally given by

$$\begin{aligned}\mathcal{E}\left(x\right) &= E\left[\left(x-\hat{x}\right)^2\right] \\ &\geq \frac{\left[1+\frac{dE[x-\hat{x}]}{dx}\right]}{J\left(x\right)} + \left(E\left[x-\hat{x}\right]\right)^2\end{aligned}$$

where $J(x)$ is the Fisher information matrix defined by

$$J\left(x\right) = E\left[\left(\frac{\partial}{\partial x}\log p\left(x,y\right)\right)\left(\frac{\partial}{\partial x}\log p\left(x,y\right)\right)^T\right].$$

If the estimate is unbiased (namely $E[\hat{x}-x]=0$), then (x) is equal to the variance, and we have

$$\mathcal{E}\left(x\right) = J^{-1}\left(x\right) \qquad (19)$$

and the estimate satisfying (19) is called Fisher efficient. Kalman filter is Fisher-efficient under LQG circumstance in which the state-error covariance matrix plays a similar role as the inverse Fisher information matrix. Naturally, the issue is also interesting within the particle filtering framework. Recently, it has been established that under some regularity conditions, the particle filters also satisfy the Cramér-Rao bound

$$\begin{aligned}E\left[\tilde{x}_n\tilde{x}_n^T\right] &\geq P_n \\ E\left[\left\|\tilde{x}_n\right\|^2\right] &\geq tr\left(P_n\right)\end{aligned}$$

where $\tilde{x}_n = x_n - \hat{x}_{n|n}$ is the one-step ahead prediction error, and

$$\begin{aligned}P_{n+1} &= F_n\left(P_n^{-1}+R_n^{-1}\right)^{-1}F_n^T + G_nQ_nG_n^{-1}, \\ P_0^{-1} &= E\left[\frac{\partial}{\partial x_0 x_0}\log p\left(x_0\right)\right], \\ F_n &= E\left[\frac{\partial}{\partial x_n}f\left(x_n,w_n\right)\right], \\ R_n^{-1} &= E\left[\frac{\partial}{\partial x_n x_n}\log p\left(y_n|x_n\right)\right], \\ G_n^T &= E\left[\frac{\partial}{\partial w_n}f\left(x_n,w_n\right)\right], \\ Q_n^{-1} &= E\left[\frac{\partial}{\partial w_n w_n}\log p\left(w_n\right)\right].\end{aligned}$$

The upper bound is time-varying and can be recursively updated by replacing the expectation with Monte Carlo average. For derivation details and discussions, see [Arnaud Doucet and Gordon (2001)] for more general unified treatment (filtering, prediction, smoothing) and extended situations.

7.3 Bias-Variance

Let's first consider the exact Monte Carlo sampling. The true and Monte Carlo state-error covariance matrices are defined by

$$\begin{aligned}\Sigma &= E_p\left[\left(x-\mu\right)\left(x-\mu\right)^T\right], \\ \Sigma_{\hat{\mu}} &= E_p\left[\left(x-\hat{\mu}\right)\left(x-\hat{\mu}\right)^T\right],\end{aligned}$$

where $\mu = E_p\left[x\right]$, $\hat{\mu} = \frac{1}{N_p}\sum_{i=1}^{N_p} x^{(i)}$, where $\left\{x^{(i)}\right\}$ are i.i.d. samples drawn from true p.d.f. $p(x)$. It can be proved that

$$\begin{aligned}\Sigma_{\hat{\mu}} &= \left(1+\frac{1}{N_p}\right)\Sigma \\ &= \Sigma + Var_p\left[\hat{\mu}\right].\end{aligned}$$

Hence, the uncertainty from the exact Monte Carlo sampling part is the order of N_p^{-1}. In practice,we usually calculate the sample variance in place of true variance, for Monte Carlo simulation, we have

$$\Sigma_{\hat{\mu}} = \frac{1}{N_p - 1} \sum_{i=1}^{N_p} \left(x^{(i)} - \hat{\mu}\right) \left(x^{(i)} - \hat{\mu}\right)^T.$$

It should be cautioned that $\hat{\Sigma}_{\hat{\mu}}$ is an unbiased estimate of Σ instead of $\Sigma_{\hat{\mu}}$ is given by $(1 + N_p^{-1})\Sigma_{\hat{\mu}}$. Second, we particularly consider the importance sampling where the i.i.d. samples are drawn from the proposal distribution. Recalling some notations defined earlier, it must be cautioned again that although $\hat{f}_{N_p}$ is unbiased (i.e. $E_p[f(x)] = E_p[\hat{f}_{N_p}(x)]$), however, $\hat{f}$ is biased (i.e. $E_p[f(x)] \neq E_p[\hat{f}_{N_p}(x)]$. In practice, with moderate sample size, it was shown that the bias is not negligible. The bias accounts for the following sources: limited simulated samples, limited computing power and limited memory (calculation of posterior $p(x_{0:n}|y_{0:n})$ needs storing the data up to n), not to mention the sampling inaccuracy as well as the existence of noise. In the Monte Carlo filtering context, suppose $\hat{x}_n$ is an estimate given by the particle filter, by writing

$$x_n - \hat{x}_n = (x_n - E_q[\hat{x}_n|y_{0:n}]) + (E_q[\hat{x}_n|y_{0:n}] - \hat{x}_n),$$

we may calculate the expected gross error

$$\begin{aligned}
\mathcal{E} &= E_q\left[tr\left((x_n - \hat{x}_n)(x_n - \hat{x}_n)^T\right)|y_{0:n}\right] \\
&= tr\left(E_q\left[(x_n - \hat{x}_n)(x_n - \hat{x}_n)^T|y_{0:n}\right]\right) \\
&= tr\left(\underbrace{E_q\left[(\hat{x}_n - E_q[\hat{x}_n|y_{0:n}])(\hat{x}_n - E_q[\hat{x}_n|y_{0:n}])^T|y_{0:n}\right]}_{Covariance}\right. \\
&\left. + \underbrace{(E_q[\hat{x}_n|y_{0:n}] - x_n)(E_q[\hat{x}_n|y_{0:n}] - x_n)^T}_{Bias^2}\right)
\end{aligned} \tag{20}$$

where

$$E_q[x_n|y_{0:n}] = \int x_n W(x_n) q(x_n|y_{0:n})\, dx_n,$$

and

$$W(x_n) = \frac{p(x_n|y_{0:n})}{q(x_n|y_{0:n})}.$$

If $p = q$, the bias vanishes to zero at a rate $O(N_p)$, then $\mathcal{E}$ only accounts for variance, and the state-error covariance is the true covariance. If $p \neq q$, $\mathcal{E}$ generally consists of both bias and variance where the bias is a nonzero constant. Hence, equation (20) represents the bias-(co)variance dilemma. When the loss $\mathcal{E}$ is fixed, the bias and variance is a trade-off. Generally, we can define the bias and variance of importance sampling or MCMC estimate as:

$$\begin{aligned}
Bias &= E_q\left[\hat{f}(x)\right] - E_p[f(x)], \\
Var &= E_q\left[\left(\hat{f}(x) - E_p\left[\hat{f}(x)\right]\right)^2\right]
\end{aligned}$$

where $\hat{f}(x)$ is given by the weighted importance sampling. The quality of approximation is measured by a loss function $\mathcal{E}$, as decomposed by

$$\begin{aligned}
\mathcal{E} &= E_q\left[\left(\hat{f}(x) - E_p\left[\hat{f}(x)\right]\right)^2\right] \\
&= Bias^2 + Var.
\end{aligned}$$

7.4 Robustness

Robustness (both algorithmic robustness and numerical robustness) issue is important for the discrete-time filtering. In many practical scenarios, the filter might encounter the possibility of divergence where the algorithmic assumption is violated or the numerical problem is encountered (e.g., ill-conditioned matrix factorization). We focus our attention on the particle filters. There are two fundamental problems concerning the robustness in particle filters. First, when there is an outlier, the importance weights will be very unevenly distributed and it usually requires a large number of N_p to assure the accuracy of empirical density approximation. Hence the measurement density $p(y_n|x_n)$ is supposed to insensitive to the x_n. Second, the empirical distribution from the samples often approximates poorly for the long-tailed distribution, either for proposal distribution or for posterior. This is imaginable because the probability sampling from the tail part of distribution is very low, and resampling somehow makes this problem more severe. Many results have shown that even the mixture distribution can not well describe the tail behavior of the target distribution. Hence, outliers will possibly cause the divergence of filter or produce a very bad performance. Recently, it has been shown that the sample size estimate given by (89) is not robust, the approximated expression might be infinitely wrong for certain $f(x)$, $p(x)$ and $q(x)$. It can be derived that

$$\begin{aligned}
Var_q\left[\hat{f}\right] &= \frac{1}{N_p} Var_q[f(x)W(x)] \\
&= \frac{1}{N_p} E_q\left[(f(x) - E_p[f(x)])^2 W^2(x)\right] + O\left(N_p^{-2}\right)
\end{aligned}$$

where $W(x) = p(x)/q(x)$. For a large N_p, the true effective sample size is given as

$$\begin{aligned}
N_{eff}^{'} &= \frac{Var_p[f]}{Var_q\left[\hat{f}\right]} \\
&\approx \frac{N_p E_p\left[(f(x) - E_p[f(x)])^2\right]}{E_q\left[(f(x) - E_p[f(x)])^2 W^2(x)\right]}
\end{aligned}$$

The expression of N_{eff} (derived by using first two moments of $W(x)$ and $f(x)$) can be very poor (for two simple cases, one leads to $\frac{N_{eff}^{'}}{N_{eff}} \to 0$ and the other $\frac{N_{eff}^{'}}{N_{eff}} \to \infty$. A more robust effective sample size estimate has been proposed

$$N_{eff} = \frac{N_p \sum_{i=1}^{N_p} \left(f(x^{(i)}) - E_p[f(x)]\right)^2 W(x^{(i)})}{\sum_{i=1}^{N_p} \left(f(x^{(i)}) - E_p[f(x)]\right)^2 W^2(x^{(i)})}.$$

Another critical issue is the estimate of the important weights within the IS, SIS, SIR framework. Note that $W(x) = p(x)/q(x)$ is a function instead of a point estimatate. Being a function usually implies certain prior knowledge, e.g. smoothness, non-negativeness, finite support. However, when we use a finite number of random (uneven) samples to represent this function, the inaccuracy (both bias and variance) is significant. This problem becomes more severe if the outliers come in.

7.5 Evaluation and Implementation

We should keep in mind that designing particular particle filter is problem dependent. In other words, there is no general rule or universal good particle filter. For instance, there always a tradeoff between the fact that we prefer to keep the spread of particles wide (to prevent missing hypothesis), and the case like target tracking, where we instead prefer to keep the support of particles bounded (to improve the accuracy). To give another example, in many cases we want the particle filter robust to the outliers, thereby an insensitive likelihood model is preferred, however in some case where the cost is unaffordable even the likelihood is low, a risk-sensitive model is needed. On the other hand, one particle filter Algorithm A works well (better than another particle filter Algorithm B) doesn't necessarily mean that it has the gain over Algorithm B on the other problems. Hence it is not fair to conclude that Algorithm A is superior to Algorithm B for only one particular problem being tested. Justification of the superiority of certain algorithm over the others even on a specific problem is also unfair without Monte Carlo simulations. One of the merits about particle filter is the implementation complexity is $O(N_p)$, independent of the state dimension N_x. As to the evaluation criteria of Monte Carlo or particle filters, a straightforward indicator of performance of different algorithms can be seen from the MSE between the estimate and true value. Due to the Monte Carlo nature, variance is an important criterion, e.g. (co)variance of estimate and variance of importance weights, both of which are calculated based on Monte Carlo averaging results (say 100 or 1000 independent runs). This requirement is deemed necessary when comparing different particle filters performance, otherwise it is unfair to say one is better than the others or the opposite. Other evaluation issues include sampling and resampling efficiency, tradeoff between performance and computational complexity, parallel architecture, ease of implementation, etc.

8. OTHER FORMS OF BAYESIAN FILTERING: MALLIAVIN ESTIMATOR

8.1 Extended Abstract

In Malliavin Calculus [Nualart (2006)], suppose that ones goal is to evaluate

$$\begin{aligned} &E\left[f'\left(X\right)\right]=\int f'\left(x\right)p\left(x\right)dx \\ &=-\int f\left(x\right)\frac{p'\left(x\right)}{p\left(x\right)}p\left(x\right)dx=E\left[f\left(X\right)\frac{p'\left(x\right)}{p\left(x\right)}\right] \\ &=E\left[f\left(X\right)\frac{p'\left(x\right)}{p\left(x\right)}\right]=E\left[f\left(X\right)H\left(X,1\right)\right], \end{aligned}$$

where

$$H\left(X,1\right)=\frac{p'\left(x\right)}{p\left(x\right)}.$$

One of the main purposes is to evaluate this last quantity in a efficient way. The main advantage of Malliavin calculus is that: The same simulated paths give good estimates for densities at any point. That is, one can compute the density over the whole real line with the same number of paths. And allowing one to get efficient computational methods. For more details, one is refered to [Mrad et al. (2006); E. Fournié and N.Touzi (1999); E. Fournié (2001)]

Suppose that we have a diffusion process like in (7), the solution is given by

$$X_t=X_0+\int_0^t b\left(X_s\right)ds+\int_0^t \sigma\left(X_s\right)dW_s,\ t\in[0,T].$$

If Hörmander hypothesis is satisfied then X_T density exists and is smooth. Using Malliavin calculus to develop an expression for $H(X,1)$ that can be simulated. So we get a Monte Carlo for (with variance reduction - using control variate).

$$E\left[f'\left(X\right)\right]=\frac{1}{N}\sum f\left(\tilde{X}^i\right)H\left(\tilde{X}^i,1\right),$$

where $\tilde{X}^i$ are independent Euler approximations of X.

For our purpose, we want to calculate the posterior, leading to the choice of $f'\left(X\right)=\delta\left(X\right)$ then $E\left[\delta\left(X\right)\right]=p\left(x\right)$ (p.d.f. of X).

Form this we can have the following theorem

Theorem 5. Let $\varphi,\frac{d\varphi}{dx}\in L^2\left(\mathbb{R}\right),\varphi\left(0\right)=1$ a localization function and r a parameter and $c\in L^2\left(\mathbb{R}\right)$ a "control variate". The density function

$$p\left(x\right)=E\left[\xi_{c,r}\left(x\right)\right],$$

where

$$\xi_{c,r}\left(x\right)=\left(1_{\{X\geq x\}}-c\left(x\right)\right)H\left(X,\varphi\left(\frac{X-x}{r}\right)\right),$$

moreover,

$$H\left(X,\varphi\left(\frac{X-x}{r}\right)\right)=\varphi\left(\frac{X-x}{r}\right)H\left(X,1\right)-\frac{1}{r}\varphi'\left(\frac{X-x}{r}\right),$$

and

$$H\left(X,1\right)=\frac{\int_0^T dW_t}{\int_0^T D_sXds}+\frac{\int_0^T\int_0^T D_tD_sXdsdt}{\left(\int_0^T D_sXds\right)^2}.$$

The variance of $\xi_{c,r}\left(x\right)$ is minimized for

$$c\left(x\right)=\frac{E\left[1_{\{X\geq x\}}H\left(X,\varphi\left(\frac{X-x}{r}\right)\right)^2\right]}{E\left[H\left(X,\varphi\left(\frac{X-x}{r}\right)\right)^2\right]},$$

and

$$r=\sqrt{\frac{\int_0^\infty\varphi'\left(z\right)^2dz}{E\left[H\left(X,1\right)^2\right]\int_0^\infty\varphi\left(z\right)^2dz}},$$

with $\varphi\left(x\right)=e^{-\lambda|x|},\lambda>0$. □

In the previous stated we still have to have in account the need to discretize the following quantities [Bouchard et al. (2002)]

$$D_sX_t=\begin{cases}\sigma e^{\int_s^t\bar{b}'(X_v)dv+\int_s^t\sigma'(X_v)dW_v} & ,\ s\leq t\\ 0 & s>t\end{cases}$$

where $\bar{b}'(X_v) = b'(X_v) - \frac{1}{2}\sigma'(X_v)^2$, and

$$D_s D_t X_T = D_s(X_t)\,\sigma'(X_t)\, e^{\int_t^T \bar{b}'(X_v)dv + \int_t^T \sigma'(X_v)dW_v}$$
$$+ \left[\sigma'(X_t)\, 1_{\{t \le s\}} + \int_t^T \bar{b}''(X_v)\, D_s X_v dv \right.$$
$$\left. + \int_t^T \sigma''(X_v)\, D_s X_v dW_v \right] D_t X_T.$$

9. SIMULATION RESULTS

From the previous stated in section (8), we have now a tool to compute the posterior density function $p\left(x_k | x_{k-1}^i\right)$. In order to have a sanity check suppose that we have the following diffusion process

$$dX_t = \sigma dW_t$$

and with initial condition given by $\delta(x_0 = 0)$. The solution exists and is given by a closed formula as a Gaussian with suitable parameters. Hereafter follows a graphic (Fig. 1) that shows for 1000 paths generation the approximate result.

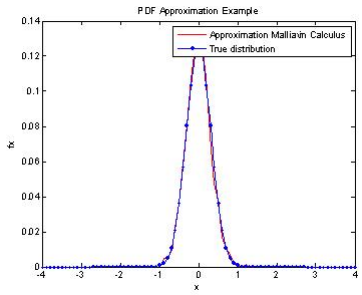

Figure 1. Sanity check example

Finally we explore the efficiency of the particle filter using Malliavin calculus against the classical SIR and KF. For this propose assume the following system

$$x_k = 1 + \sin\left(4 * 10^{-2} * \pi * k\right) + 0.5 * x_{k-1} + \sqrt{5} * w$$

$$y_k = \begin{cases} \dfrac{x_k}{5} + v & k \le 30 \\ -2 + \dfrac{x_k}{2} + v & k > 30 \end{cases}$$

where w, v are assumed to be WGN.

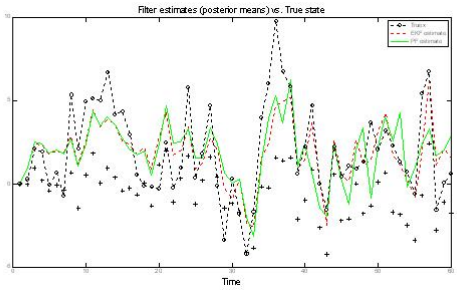

Figure 2. PF with 1000 particles

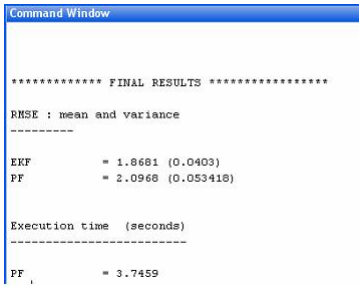

Figure 3. Time and Variance of PF with 1000 particles

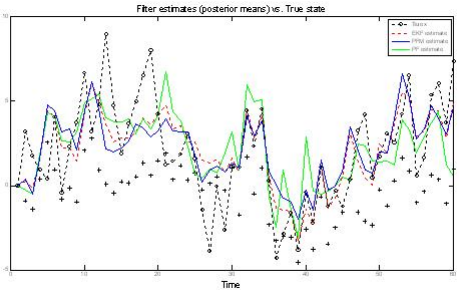

Figure 4. PFM with 20 particles and 50 Brownian paths

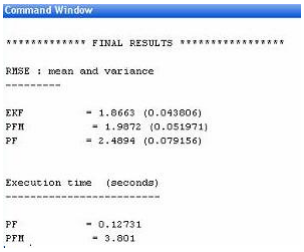

Figure 5. Time and variance of PFM with 20 particles and 50 Brownian paths

9.1 Discussion of results

Keeping in mind that we have a linear system we should chose as a optimality criteria the variance and the time consumed. For both, particle filter (PF) and particle filter using Malliavin calculus (PMF) we did residual resampling and the presented results in Fig. 3 and Fig. 5 says respect to an average of 1000 runs. In these figures we can infere that for the same computational time, PMF outperforms the PF since the results of the variance are close to the optimum given by the Kalman Filter (EKF). For visual understanding of the performance, one is pointed to Fig. 2 and Fig. 4.

10. CONCLUSIONS AND FURTHER RESEARCH

In the present paper we try to survey the basics of Sequential Monte Carlo methods, aka, Particle filters. For that we introduce the problem statement and the different Monte Carlo techniques and how to use those in a sequential manner. We briefly explore the mathematical issues of the method and briefly presented some different approach for Monte Carlo methods, to be developed and full presented in a paper to be submitted to American Control Conference 2011.

REFERENCES

Anderson, E.C. (1999). *Monte Carlo Methods and Importance Sampling - Lectures*. Berkeley University, CA, USA.

Arnaud, Doucet, and Johansen (2008). A tutorial on particle filtering and smoothing: Fifteen years later. Technical report.

Arnaud Doucet, N.d.F. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Arulampalam, M.S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174–188. doi: 10.1109/78.978374.

Bouchard, B., Vi, U.P., and Touzi, N. (2002). Discrete time approximation and monte-carlo simulation of backward stochastic differential equations.

Chen, Z. (2003). Bayesian filtering: From kalman filters to particle filters, and beyond. Technical report, McMaster University.

E. Fournié, J.M.Larsy, J.L.P.L. (2001). Applications of malliavin calculus to monte carlo methods in finance ii.

E. Fournié, J.M.Larsy, J.L.P.L. and N.Touzi (1999). Applications of malliavin calculus to monte carlo methods in finance i.

Gelb, A. (1974). *Applied optimal estimation*. MIT Press.

Gordon, N.J., Salmond, D.J., and Smith, A.F.M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2), 107–113.

Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Mrad, M., Touzi, N., and Zeghal, A. (2006). Monte carlo estimation of a joint density using malliavin calculus, and application to american options. *Comput. Econ.*, 27(4), 497–531. doi:http://dx.doi.org/10.1007/s10614-005-9005-3.

Nualart, D. (2006). *The Malliavin Calculus and Related Topics*. Springer-Verlag.