

STOCHASTIC NONLINEAR FILTERING IN CONTINUOUS TIME

by

Elia Liitiäinen

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science

(Automation and Systems Technology)

at the

HELSINKI UNIVERSITY OF TECHNOLOGY

12th July 2005

DISCARD THIS PAGE

LIST OF FIGURES

Figure	Page
8.1 RMS of the filters calculated over 50 simulations.	72
8.2 A realization of the filtering processes.	73
8.3 Results of experiment 1, RMS of the filters over 1000 simulations.	76
8.4 Experiment 2, an example of a situation where the linear-regression filter fails. . .	77

DISCARD THIS PAGE

LIST OF TABLES

Table	Page
8.1 Results of the phase tracking experiment with constant frequency.	63
8.2 Results of the phase tracking experiment with unknown frequency.	66

DISCARD THIS PAGE

TABLE OF CONTENTS

	Page
LIST OF FIGURES	1
LIST OF TABLES	2
1 Introduction	5
2 Martingale Theory	7
2.1 Basic Definitions	7
2.2 Discrete-time Martingales	8
2.3 Martingale Theory for Continuous-parameter Processes	11
2.4 Potentials	12
2.5 The Dual Predictable Projection	17
2.6 Quadratic Variations	18
2.7 Stochastic Integration with Respect to Brownian motion	19
2.8 A Representation Theorem for Martingales	20
3 Stochastic Differential Equations	22
3.1 Strong Solutions	22
3.2 Infinitesimal Generators	23
3.3 Numerical Methods for Solving SDEs	24
4 Girsanov's Theorem	26
5 Nonlinear Filtering - A General Case	29
5.1 The Innovation Process	29
5.2 The Kushner-Stratonovitch Equation	33
6 Filtering of Stochastic Differential Equation Models	39
6.1 A Modification of the Kushner-Stratonovitch Equation	39

	Page
6.2 The Kallianpur-Striebel Formula	42
6.3 The Zakai Equation	44
7 Numerical Methods for Continuous-time Nonlinear Filtering	47
7.1 Time Discretization	47
7.2 Interacting Particle Systems Approximations	51
7.3 Galerkin's Method	53
7.4 Gaussian Filters	55
8 Applications	61
8.1 Simulations with Galerkin's Method	61
8.1.1 Phase Tracking with a Known Frequency	62
8.1.2 Phase Tracking with an Unknown Frequency	65
8.1.3 Conclusions	67
8.2 Simulations with Gaussian Filters	68
8.2.1 Vertically Falling Body	69
8.2.2 Frequency Tracking	74
8.2.3 Conclusions	78
LIST OF REFERENCES	80

Chapter 1

Introduction

In many engineering problems being able to estimate hidden states from noisy measurements is important. In the simple case where everything is linear there exists an optimal closed-form solution called the Kalman filter which was introduced in [17]. However, most real-world problems are nonlinear making numerical approximations necessary. Unfortunately finding efficient methods which work also in high dimensions is not easy and there has been lots of research on this subject.

In this thesis the continuous-time nonlinear filtering problem is investigated. A literature survey on the basics of the theory is given. In the second part of the thesis the theory is applied to practice in three different problems.

In the theoretical part three different methods for nonlinear filtering are derived. The particle filter introduced for the first time in [10] is a standard tool that has gained lots of attention during the last 10 years. Galerkin's method for continuous-time filtering on the other hand is much more rare and has rarely been used in engineering applications. The third method, the Gaussian linear-regression filter, is a generalization of the famous Extended Kalman Filter.

The theoretical approach of this thesis is mainly standard. However, to derive the Gaussian filters a linear regression approach was adopted. This approach, presented for example in [23], unifies different approximation methods and is intuitively clear. A heuristic approach for estimating the quality of linear approximations is proposed basing on the theory.

In the experimental part of the work the theory is tested on four simulations which model real-world phenomena. Proposals for practical implementation of Galerkin's method are given based on the experiments. It is also pointed out that by replacing the EKF by a slightly more sophisticated linear-regression filtering algorithm, improvements are possible often without additional computational cost. This idea is applied to the problem of frequency and phase tracking, where traditionally EKF-based techniques have been used.

The structure of the thesis is as follows. In chapters 2, 3 and 4 an introduction to stochastic analysis and the theory of martingales is given. Main emphasis is on fixing notation and concepts.

The nonlinear filtering problem is investigated in a very general setting in chapter 5. In this chapter the famous Kushner-Stratonovitch equation is derived by using the theory of martingales. In chapter 6 a more specific model is introduced. This model allows the Kushner-Stratonovich equation to be written in a more concrete form. In addition two other important tools are introduced: the Kallianpur-Striebel formula and the Zakai equation.

In chapter 7 three numerical methods for solving the nonlinear filtering problem are introduced. These are based on either approximating straightforwardly the Kushner-Stratonovitch equation or on using time-discretization and the Kallianpur-Striebel formula. The approach in this chapter is more practical than in previous chapters and no analysis of convergence is given.

In chapter 8 results of numerical simulations are given and analysed. For clarity the experiments are divided into two parts. In the first one main emphasis is on testing Galerkin's method. In the second one the linear-regression filtering technique and the particle filter are investigated through two different experiments.

Chapter 2

Martingale Theory

In this chapter we will go through basics of martingale theory. Important tools for nonlinear filtering are introduced.

2.1 Basic Definitions

We denote by (Ω, \mathcal{F}, P) a probability space. We will assume the space to be complete. A continuous-time stochastic process $X = (X_t)$ is a mapping $\mathbf{R}^+ \times \Omega \mapsto \mathbf{R}$ so that each X_t is a random variable. Correspondingly discrete-time processes are mappings $X = (X_n) : \mathbf{Z}^+ \times \Omega \mapsto \mathbf{R}$ with each X_n a random variable.

Definition 2.1 A stochastic process $X = (X_t)$ is called measurable, if it is measurable as a mapping $X : [0, \infty] \times \Omega \mapsto \mathbf{R}$ with respect to the σ -algebra $\mathcal{B}(\mathbf{R}^+) \times \mathcal{F}$. X is called integrable if $X_t \in L^1(\Omega, \mathcal{F}, P)$ for all $t > 0$.

We denote by (\mathcal{F}_t) a filtration on (Ω, \mathcal{F}) . $\mathcal{F}_t^{X,0}$ is defined as $\sigma(X_s, s \leq t)$, the σ -algebra generated by the process (X_t) . \mathcal{F}_t^X denotes the σ -algebra generated by $\mathcal{F}_t^{X,0} \cup \{A \in \mathcal{F} : P(A) = 0\}$. $\tilde{\mathcal{F}}_t^{X,0}$ is the usual augmentation of $\mathcal{F}_t^{X,0}$ and thus right continuous and complete.

Definition 2.2 Let $X = (X_t)$ be a stochastic process and (\mathcal{F}_t) a filtration. If each X_t is \mathcal{F}_t measurable, the process is called adapted to (\mathcal{F}_t) .

By (X_t, \mathcal{F}_t) we mean a measurable stochastic process adapted to (\mathcal{F}_t)

Definition 2.3 (X_t, \mathcal{F}_t) is called progressively measurable, if for each $t > 0$ it is $\mathcal{B}([0, t]) \times \mathcal{F}_t$ measurable as a mapping $X : [0, t] \times \Omega \mapsto \mathbf{R}$.

Definition 2.4 $X = (X_t)$ is called continuous if the paths $X_t(\omega) : \mathbf{R}^+ \mapsto \mathbf{R}$ are continuous a.s. Similar definitions hold for right and left continuity and increasing/decreasing processes.

A right continuous process with left limits is called cadlag.

Definition 2.5 Processes $X = (X_t)$ and $Y = (Y_t)$ are called indistinguishable if $P(\{\omega : X_t(\omega) = Y_t(\omega) \forall t \geq 0\}) = 1$. X is said to be a version of Y if $P(X_t = Y_t) = 1$ for all $t \geq 0$.

The following proposition is quite obvious and shows that measurability poses no problems.

Proposition 1 Let (X_t, \mathcal{F}_t) be right-continuous at every point on Ω . Then it is progressively measurable.

Next we introduce the concept of stopping times which is very important.

Definition 2.6 A random variable $\tau : \Omega \rightarrow \mathbf{R}^+$ is called a (finite) stopping time with respect to (\mathcal{F}_t) , if $\{\tau \leq t\} \in \mathcal{F}_t$ for all $t \geq 0$.

$\mathcal{F}_\tau = \{A \in \mathcal{F} | A \cap \{\tau \leq t\} \in \mathcal{F}_t \forall t \geq 0\}$ is the σ -algebra generated by the stopping time τ . It is easy to prove that X_τ is \mathcal{F}_τ measurable if $X = (X_t)$ is progressively measurable and \mathcal{F} -measurable, if X is measurable. Often it is useful to allow the value ∞ for a stopping time.

Definition 2.7 A stochastic process (X_t, \mathcal{F}_t) is called L^p -bounded if $\sup_t E[|X_t|^p] < \infty$.

Generalizations to higher dimensions are obvious. It should be clear from the context when we use vector valued processes. The definitions in this section have obvious discrete-time versions.

To fix some notation, the i :th component of a vector x will be denoted by x^i . We will use the notation x^* for transpose. We adopt the convention $\inf\{\emptyset\} = \infty$.

2.2 Discrete-time Martingales

The theory of discrete-time martingales is the basis for understanding continuous-time processes. Many of the results for continuous-time martingales can be derived using corresponding discrete-time results.

Definition 2.8 The process $(X_n, \mathcal{F}_n)_{n \geq 0}$ is a supermartingale, if $X_n \in L^1(\Omega, \mathcal{F}, P)$ for $n \geq 0$ and $E[X_n | \mathcal{F}_m] \leq X_m$ for $n \geq m \geq 0$. Correspondingly, it is a submartingale if $-X_n$ is a supermartingale and a martingale if it has both properties.

The following theorem is useful as it allows us to decompose a stochastic process into noise (the martingale part) and a regular component.

Theorem 2.9 (Doob decomposition). Let (X_n, \mathcal{F}_n) be an integrable stochastic process. Then it has the following unique decomposition:

$$X_n = X_0 + M_n + A_n,$$

where (M_n, \mathcal{F}_n) is a martingale and $E[A_n | \mathcal{F}_{n-1}] = A_n$. Moreover, $M_0 = 0$ and $A_0 = 0$.

Proof Define

$$M_n = X_n - X_0 - \sum_{i=1}^n E[X_i - X_{i-1} | \mathcal{F}_{i-1}].$$

Clearly M_n has the required properties. We define $A_n = X_n - M_n$.

Suppose $X_n = X_0 + M_n + A_n = X_0 + \tilde{M}_n + \tilde{A}_n$ are two different decompositions of X . Then $M_n - \tilde{M}_n$ is a martingale and \mathcal{F}_{n-1} measurable. Because $M_0 = \tilde{M}_0$, it is clear that $M_n = \tilde{M}_n$ a.s.

Corollary 2.10 Let (X_n, \mathcal{F}_n) be a supermartingale. Then X_n has the Doob decomposition

$$X_n = X_0 + M_n + A_n,$$

where (A_n) is a.s. decreasing.

The following lemma can be proved using Fubini's theorem (see [29] for the proof):

Lemma 2.11 Let $X, Y \geq 0$ be random variables for which the following inequality holds for all $c \geq 0$:

$$cP(X \geq c) \leq \int_{\{\omega \in \Omega: X(\omega) \geq c\}} Y dP.$$

Then $\|X\|_p \leq q\|Y\|_p$ where $1/p + 1/q = 1$ and $p > 1$

Theorem 2.12 (Doob's maximal inequality). Let (X_n, \mathcal{F}_n) be a nonnegative submartingale.

Then

$$cP\left(\sup_{n \in [0, N]} X_n \geq c\right) \leq \int_{\{\omega \in \Omega: \sup_{n \in [0, N]} X_n(\omega) \geq c\}} X_N dP.$$

Proof This theorem is a fundamental one and can be found in most introductory books on stochastic calculus.

The next strong theorem is important. It follows straightforwardly from lemma 2.11 and theorem 2.12 taking into account that if (X_n) is a martingale, then $|X_n|$ is a submartingale.

Theorem 2.13 (Doob's L^p -inequality). Let (X_n, \mathcal{F}_n) be a martingale. For all $p > 1$ we have the inequality

$$\| \sup_{n \in [0, N]} |X_n| \|_p \leq q \|X_N\|_p.$$

The supermartingale convergence theorem is another important theorem.

Theorem 2.14 (Doob's Supermartingale Convergence Theorem). Let (X_n, \mathcal{F}_n) be a supermartingale and $\sup_{n \geq 0} E[|X_n|] < \infty$. Then there exists a random variable $X_\infty \in L^1(\Omega, \mathcal{F}, P)$ so that (P-a.s.)

$$X_\infty = \lim_{n \rightarrow \infty} X_n.$$

There exist also other convergence results. The one for square integrable martingales is especially useful from our point of view.

Theorem 2.15 Let (X_n, \mathcal{F}_n) be a square-integrable martingale for which $\sup_{n \geq 0} X_n^2 < \infty$. Then there exists a random variable $X_\infty \in L^2(\Omega, \mathcal{F}, P)$ so that $\lim_{n \rightarrow \infty} \|X_\infty - X_n\|_2 = 0$ and $\lim_{n \rightarrow \infty} X_n = X_\infty$ a.s.

The proofs can be found in [29].

2.3 Martingale Theory for Continuous-parameter Processes

In this section we will recall some basic theorems and definitions for continuous-parameter martingales. The main point is to present a couple of theorems which are skipped in many books on stochastic calculus but are still important.

The definition of continuous-parameter submartingales, supermartingales and martingales is similar to that of discrete-parameter processes.

Theorem 2.16 Assume that (X_t, \mathcal{F}_t) is a supermartingale and the filtration (\mathcal{F}_t) satisfies the usual conditions. If the function $f(t) = E[X_t]$ is right-continuous, then (X_t) has a cadlag version.

Proof See [29].

Next we examine stopped supermartingales.

Theorem 2.17 Let (X_t, \mathcal{F}_t) be a stochastic process and τ a stopping time. Then the stopped process $(X_{t \wedge \tau})$ is adapted to (\mathcal{F}_t) if and only if it is adapted to $(\mathcal{F}_{t \wedge \tau})$. In addition, $(X_{t \wedge \tau})$ is a (\mathcal{F}_t) -martingale if and only if it is a $(\mathcal{F}_{t \wedge \tau})$ -martingale.

Proof The proof is quite simple, see [16].

Theorem 2.18 Let (X_t, \mathcal{F}_t) be a cadlag martingale and τ a stopping time. If (\mathcal{F}_t) satisfied the usual conditions, then $(X_{t \wedge \tau})$ is a (\mathcal{F}_t) -martingale.

The next theorem will also be used regularly.

Theorem 2.19 Let (\mathcal{F}_t) satisfy the usual conditions. Assume that (X_t, \mathcal{F}_t) is a cadlag martingale and $S \geq T \geq 0$ bounded stopping times. Then

$$E[X_T | \mathcal{F}_S] = X_S.$$

Proof See [29]. This is the famous optional stopping theorem.

The inequalities and convergence theorems for discrete time martingales may easily be generalized to the continuous-parameter case.

Definition 2.20 A measurable process (X_t, \mathcal{F}_t) is called a local martingale, if there exists a sequence of (finite) stopping times (τ_n) so that $\tau_n \rightarrow \infty$ a.s. and $(X_{t \wedge \tau_n})$ is a martingale for all $n > 0$.

Definition 2.21 Suppose that $X = (X_t, \mathcal{F}_t)$ is a continuous process. X is a semimartingale, if

$$X_t = X_0 + M_t + A_t,$$

where $M_0 = A_0 = 0$, (M_t) is a continuous local (\mathcal{F}_t) -martingale and (A_t) is P-a.s. of finite variation.

Definition 2.22 The space of continuous L^2 -bounded martingales (with respect to some filtration) is denoted by M_c^2 .

Theorem 2.23 Given a filtration satisfying the usual conditions, M_c^2 is a hilbert space with the inner product

$$(M_t, N_t) = E[M_\infty N_\infty].$$

2.4 Potentials

Definition 2.24 A positive supermartingale (X_n, \mathcal{F}_n) is a discrete-parameter potential, if $E[X_n] \rightarrow 0$ when $n \rightarrow \infty$.

Definition 2.25 (X_t, \mathcal{F}_t) is a continuous-parameter potential if it is cadlag, positive and $E[X_t] \rightarrow 0$.

Theorem 2.26 Let (X_n, \mathcal{F}_n) be a potential. Then there exists an L^1 -bounded, increasing (\mathcal{F}_n) -adapted process A_n so that

$$X_n = E[A_\infty | \mathcal{F}_n] - A_n.$$

Here $A_n \rightarrow A_\infty$ a.s. and in the L^1 -norm. In addition, $A_0 = 0$, A_n is \mathcal{F}_{n-1} measurable and the decomposition is (a.s.) unique.

Proof Let $X_n = X_0 - A_n + M_n$ be the Doob decomposition for X_n . Clearly (A_n) is increasing and positive. $A_n \rightarrow A_\infty$ P-a.s., where A_∞ is \mathcal{F}_∞ measurable. Using the monotone convergence theorem we see that $E[A_\infty] < \infty$. By Doob's Supermartingale Convergence Theorem $M_n \rightarrow M_\infty$, where $M_\infty \in L^1(\Omega, \mathcal{F}_\infty, P)$. Noting that $X_n \rightarrow 0$ P-a.s., it is obvious that $X_0 + M_\infty = A_\infty$ P-a.s. from which the claim follows.

We are more interested in the continuous-parameter case. In the rest of the chapter we suppose that the filtration (\mathcal{F}_t) satisfies the usual conditions unless otherwise said.

Definition 2.27 Let (A_t, \mathcal{F}_t) be an increasing, L^1 -bounded cadlag process and $A_0 = 0$. It is called natural if

$$E\left[\int_0^\infty M_t - M_t^- dA_t\right] = 0$$

for all bounded cadlag-martingales (M_t) .

Definition 2.28 A cadlag supermartingale (X_t, \mathcal{F}_t) is of class D, if for all $\epsilon > 0$ there exists $\delta > 0$ so that

$$P(A) < \delta \Rightarrow \int_A |X_\tau| dP < \epsilon$$

for all finite stopping times τ .

Now we are ready to formulate and prove the continuous-parameter version of theorem 2.26.

Theorem 2.29 Let (X_t, \mathcal{F}_t) be a potential of class D. Then

$$X_t = E[A_\infty | \mathcal{F}_t] - A_t,$$

where (A_t) has the following properties:

1. (A_t) is adapted to (\mathcal{F}_t) , cadlag and increasing
2. $\sup_{t \geq 0} E[|A_t|] < \infty$
3. $E[A_\infty - A_t] \rightarrow 0$ as $t \rightarrow \infty$
4. $A_t \rightarrow A_\infty$ P-a.s.
5. $A_0 = 0$.

Proof We define discrete-parameter potentials by

$$X_n^j = X_{n2^{-j}} \quad (n \geq 0, j > 0).$$

Let (A_n^j) be the corresponding increasing L^1 -bounded processes. By taking the limit for each j , we get a sequence $(A_\infty^j) \in L^1(\Omega, \mathcal{F}_\infty, P)$. Next we show that the collection (A_∞^j) is uniformly integrable.

Define the stopping times

$$\tau_N^j = \inf_{n \geq 0} \{A_{n+1}^j > N\}.$$

Choose $0 < \epsilon < 1$ and N so that

$$P(A_\infty^j > N) \leq \frac{1}{N} E[A_\infty^j] = \frac{1}{N} E[X_0] < \epsilon$$

Next we see that

$$E[A_\infty^j - A_{\tau_N^j}^j] = E[X_{\tau_N^j}^j] = \int_{\{\tau_N^j < \infty\}} X_{\tau_N^j}^j dP = \lim_{M \rightarrow \infty} \int_{\{A_\infty^j > N\}} X_{\tau_N^j \wedge M}^j dP < \delta,$$

where $\delta > 0$ may be chosen arbitrarily small by choosing a small enough ϵ . The last equality follows from the fact that the sequence $(X_{\tau_N^j \wedge M}^j)_{M \geq 0}$ is uniformly integrable. Using again the Tsebychev-inequality it is seen that

$$\begin{aligned} P(A_\infty^j > N + M) &= P(A_\infty^j - A_{\tau_N^j}^j + A_{\tau_N^j}^j > N + M) \\ &\leq P(A_\infty^j - A_{\tau_N^j}^j > M) \leq \frac{1}{M} E[A_\infty^j - A_{\tau_N^j}^j] < \frac{\delta}{M}. \end{aligned}$$

The uniform integrability of (A_∞^j) is seen from the following inequality which holds for all $j > 0$:

$$\int_{\{A_\infty^j > N+M\}} A_\infty^j dP \leq \delta + \int_{\{A_\infty^j > N+M\}} A_{\tau_N^j}^j dP \leq \delta + \frac{N\delta}{M}.$$

Using the uniform integrability it may be proved that there exists a subsequence so that $(A_\infty^{j_k}) \rightarrow A_\infty$ in the weak topology of $L^1(\Omega, \mathcal{F}_\infty, P)$.

Let $r > s$ be rational numbers of the form $l2^{-r}$, where $l \geq 0$ and $r > 0$. Then (P-a.s.)

$$\begin{aligned} E[A_\infty | \mathcal{F}_s] - X_s &= \lim_{k \rightarrow \infty}^{\text{weak}} E[A_\infty^{j_k} | \mathcal{F}_s] - X_s \\ &\leq \lim_{k \rightarrow \infty}^{\text{weak}} E[A_\infty^{j_k} | \mathcal{F}_r] - X_r = E[A_\infty | \mathcal{F}_r] - X_r. \end{aligned}$$

We choose a cadlag version of $E[A_\infty|\mathcal{F}_t]$ and define $A_t = E[A_\infty|\mathcal{F}_t] - X_t$ completing the proof. The concept of natural processes may seem unnecessary at this point but is important in the next section where we define the dual predictable projection.

Theorem 2.30 (A_t) in the previous theorem is natural.

Proof Let (M_t) be any bounded cadlag martingale. By equation (7.1) in [7]

$$E\left[\int_0^\infty M_t dA_t\right] = E[M_\infty A_\infty].$$

The previous equation is a quite straightforward consequence of the Lebesgue Dominated Convergence Theorem. We approximate M_{t-} by

$$M_t^j = M_{n2^{-j}} \quad \text{if } t \in [n2^{-j}, (n+1)2^{-j}) \quad (n \geq 0, j > 0).$$

Then

$$\begin{aligned} \int_0^\infty M_t^j dA_t &= E\left[\sum_{n=0}^\infty M_{n2^{-j}}(A_{2^{-j}(n+1)} - A_{2^{-j}n})\right] = \sum_{n=0}^\infty E[M_{n2^{-j}}(X_{2^{-j}(n+1)} - X_{2^{-j}n})] \\ &= \sum_{n=0}^\infty E[M_{n2^{-j}}(A_{n+1}^j - A_n^j)] = \sum_{n=0}^\infty E[M_{(n+1)2^{-j}}A_{n+1}^j - M_{n2^{-j}}A_n^j] = E[A_\infty^j M_\infty], \end{aligned}$$

where (A_n^j) is defined as in the previous theorem. Note that in the fourth equality, the fact that A_{n+1}^j is $\mathcal{F}_{n2^{-j}}$ measurable was used. By choosing an appropriate weakly convergent subsequence (j_k) and using the Lebesgue Dominated Convergence Theorem it follows that

$$E\left[\int_0^\infty M_{t-} dA_t\right] = E[M_\infty A_\infty].$$

Theorem 2.31 (A_t) in theorem 2.29 is unique in the sense that if (A_t) and (\tilde{A}_t) satisfy the properties given in theorems 2.29 and 2.30, then they are indistinguishable.

Proof

$$X_t = E[A_\infty|\mathcal{F}_t] - A_t = E[\tilde{A}_\infty|\mathcal{F}_t] - \tilde{A}_t$$

Let M_t be any bounded cadlag martingale. We again approximate M_{t-} by

$$M_t^j = M_{n2^{-j}} \quad \text{if } t \in [n2^{-j}, (n+1)2^{-j}) \quad (n \geq 0, j > 0).$$

Because A_t is natural,

$$\begin{aligned} E[M_\infty A_\infty] &= \lim_{j \rightarrow \infty} E\left[\sum_{n=0}^{\infty} M_n^j (A_{(n+1)2^{-j}} - A_{n2^{-j}})\right] \\ &= \lim_{j \rightarrow \infty} E\left[\sum_{n=0}^{\infty} M_n^j (\tilde{A}_{(n+1)2^{-j}} - \tilde{A}_{n2^{-j}})\right] = E[\tilde{A}_\infty M_\infty]. \end{aligned}$$

The claim follows because $M_\infty \in L^1(\Omega, \mathcal{F}_\infty, P)$ may be chosen freely.

Theorem 2.32

$$E[A_\infty^2] = E\left[\int_0^\infty X_t + X_{t^-} dA_t\right]$$

Proof Define the cadlag martingales

$$M_t = E[A_\infty | \mathcal{F}_t]$$

and

$$M_t^n = E[A_\infty I_{\{\omega: A_\infty(\omega) < n\}} | \mathcal{F}_t].$$

Both martingales are a.s. nonnegative. Also $M_t^n \leq M_t^{n+1} \leq M_t$ and $M_t^n \leq n$ P-a.s. By Doob's maximal inequality

$$P\left(\sup_{t \in [0, T]} |M_t - M_t^n| > \epsilon\right) \leq \frac{1}{\epsilon} E[M_T - M_T^n] \leq \frac{1}{\epsilon} E[A_\infty I_{\{\omega: A_\infty(\omega) > n\}}].$$

for all $T > 0$. Thus $\sup_{t \in \mathbf{R}^+} |M_t - M_t^n| \rightarrow 0$ in probability. We may choose a subsequence (n_k) so that $\sup_{t \in \mathbf{R}^+} |M_t - M_t^{n_k}| \rightarrow 0$ P-a.s. as $k \rightarrow \infty$.

$$\int_0^\infty X_t + X_{t^-} dA_t = \int_0^\infty M_t + M_{t^-} - A_t - A_{t^-} dA_t = \int_0^\infty M_t + M_{t^-} dA_t - A_\infty^2.$$

By the Monotone Convergence Theorem and theorem 7 in section 2.4.7. in [7]

$$\begin{aligned} E\left[\int_0^\infty M_t + M_{t^-} dA_t\right] &= E\left[\lim_{k \rightarrow \infty} \int_0^\infty M_t^{n_k} + M_{t^-}^{n_k} dA_t\right] \\ &= \lim_{k \rightarrow \infty} E\left[\int_0^\infty M_t^{n_k} + M_{t^-}^{n_k} dA_t\right] = \lim_{k \rightarrow \infty} 2E[M_\infty^{n_k} A_\infty] = 2E[A_\infty^2]. \end{aligned}$$

2.5 The Dual Predictable Projection

The dual predictable projection is useful in filtering theory. It is for example used to decompose an observation process into an innovation process and a noise component.

Definition 2.33 Let (X_t, \mathcal{F}_t) be a stochastic process. It is called predictable, if it is measurable w.r.t. the σ -algebra generated by sets of the form $A \times (s, t]$ and $B \times \{0\}$, where $s > 0$, $A \in \mathcal{F}_s$ and $B \in \mathcal{F}_0$.

We need the following theorem:

Theorem 2.34 Let (A_t, \mathcal{F}_t) be an increasing nonnegative L^1 -bounded cadlag process. It is predictable if and only if it is natural.

Proof See [7]. The proof is not given here because to prove the result, a fair amount of stopping time theory is needed. In our case the direction \Leftarrow is more relevant.

Theorem 2.35 Let (V_t) be an increasing nonnegative L^1 -bounded cadlag process. (\mathcal{F}_t) is a filtration containing all the null sets in \mathcal{F} but is not necessarily right-continuous. (V_t) need not be adapted to (\mathcal{F}_t) . Then there exists a unique natural and predictable (w.r.t. (\mathcal{F}_{t^+})) process (U_t, \mathcal{F}_t) so that

$$E[V_t - V_s | \mathcal{F}_s] = E[U_t - U_s | \mathcal{F}_s] \quad (t \geq s \geq 0).$$

Proof Clearly $E[V_\infty - V_t | \mathcal{F}_{t^+}]$ is a potential of class D. Choose (U_t) as the process with the properties given in theorem 2.29. It may be shown using the monotone class theorem that because (U_t) is predictable, adapted to (\mathcal{F}_{t^+}) and $U_0 = 0$, it is also adapted to (\mathcal{F}_t) . Also

$$E[V_t - V_s | \mathcal{F}_{s^+}] = E[U_t - U_s | \mathcal{F}_{s^+}]$$

(U_t) is called the dual predictable projection of (V_t) .

Theorem 2.36 Let (V_t) be a process with the properties given in theorem 2.35 and (U_t) its dual predictable projection. If $\sup_{t \in \mathbf{R}^+} E|V_t|^2 < \infty$, then (U_t) is L^2 -bounded.

Proof Define $M_t = E[V_\infty | \mathcal{F}_t]$. Recall that

$$X_t = E[V_\infty - V_t | \mathcal{F}_t] = E[U_\infty | \mathcal{F}_t] - U_t.$$

Now by theorem 2.32 and its proof

$$E[U_\infty^2] \leq E\left[\int_0^\infty M_t + M_{t-} dU_t\right] = E\left[2 \int_0^\infty M_{t-} dU_t\right]$$

The inequality follows from the fact that $X_t \leq M_t$. By Fatou's lemma

$$\begin{aligned} E\left[\int_0^\infty M_{t-} dU_t\right] &= E\left[\lim_{\delta \rightarrow 0} \sum_{i=0}^{\infty} M_{i\delta} (U_{(i+1)\delta} - U_{i\delta})\right] \\ &\leq \liminf_{\delta \rightarrow \infty} E\left[\sum_{i=0}^{\infty} M_{i\delta} (V_{(i+1)\delta} - V_{i\delta})\right] \\ &\leq (E[\sup_{t \in \mathbf{R}^+} M_t^2])^{1/2} (E[V_\infty^2])^{1/2} < \infty. \end{aligned}$$

The last inequality follows from Doob's L^p -inequality.

2.6 Quadratic Variations

In this section we will recall the definition and some other things concerning quadratic variations. The purpose is mainly to fix concepts and notation.

Theorem 2.37 Let $M = (M_t, \mathcal{F}_t)$ be a M_c^2 -martingale. Then there exists an (\mathcal{F}_t) -adapted increasing continuous process $\langle M \rangle$ so that

$$M_t^2 - \langle M \rangle_t$$

is a uniformly integrable martingale and $\langle M \rangle_0 = 0$. Moreover, $\langle M \rangle$ is unique.

Proof See for example [16].

Using localisation, a similar result for local martingales may be proved.

Theorem 2.38 If $M = (M_t, \mathcal{F}_t)$ is a continuous local martingale, there exists an (\mathcal{F}_t) adapted increasing continuous process $\langle M \rangle$ so that

$$M_t^2 - \langle M \rangle_t$$

is a local martingale and $\langle M \rangle_0 = 0$. $\langle M \rangle$ is unique.

$\langle M \rangle$ is the quadratic variation of (M_t) .

Definition 2.39 Let $M = (M_t)$ and $N = (N_t)$ be local martingales. The covariation between M and N is defined as

$$\langle M, N \rangle = \frac{1}{4}(\langle M + N \rangle - \langle M - N \rangle).$$

2.7 Stochastic Integration with Respect to Brownian motion

For an introduction to stochastic integration with respect to local martingales, see [16]. In this section we present an extended definition for integrals with respect to Brownian motion. This extension removes the predictability requirement.

Definition 2.40 A continuous process (W_t) is called a Brownian motion with respect to the filtration (\mathcal{F}_t) , if

1. $W_0 = 0$
2. $W_t - W_s$ is independent of \mathcal{F}_s for all $t \geq s \geq 0$.
3. $W_t - W_s$ is normally distributed with mean 0 and variance $t - s$.

Correspondingly, (W_t) is an \mathbf{R}^n -Brownian motion, if the components (W_t^i) are independent Brownian motions w.r.t. (\mathcal{F}_t) . Unless otherwise said, we choose (\mathcal{F}_t^W) as the filtration (\mathcal{F}_t) in the definition 2.40.

When integrating with respect to a local martingale, the property of local predictability of the integrand is essential. However, in the case of Brownian motion the definition can be extended to the following class of processes:

Definition 2.41 Let (\mathcal{F}_t) be a filtration satisfying the usual conditions. L_W^2 is defined as the class of measurable (\mathcal{F}_t) -adapted processes $X : [0, \infty] \times \Omega \rightarrow \mathbf{R}$ with the property

$$\int_{\mathbf{R}^+ \times \Omega} |X|^2 \lambda(dt) \times P(d\omega) < \infty.$$

L_W^2 is a Hilbert space of equivalence classes.

Definition 2.42 For $i \geq 0$ let H_i be \mathcal{F}_{t_i} -measurable uniformly bounded functions. Functions of the form

$$s = \sum_{i=0}^{\infty} H_i I_{(t_i, t_{i+1}]},$$

where $s \in L^2_W$ and $t_{i+1} > t_i$, are called simple functions.

We define the subspace

$$\mathcal{L} = \{s \in L^2_W : s \text{ is simple}\}.$$

The following theorem shows that for every equivalence class in L^2_W , there corresponds a predictable process.

Theorem 2.43 \mathcal{L} is dense in L^2_W .

Proof The proof is given in [16].

Thus stochastic integration may be defined in the space L^2_W and is a M^2_c -martingale for those functions.

2.8 A Representation Theorem for Martingales

Theorem 2.44 Let $W = (W_t)$ be an R^n -valued Brownian motion. Then every cadlag L^2 -bounded martingale (M_t, \mathcal{F}_t^W) has a representation

$$M_t = M_0 + \int_0^t f_s^* dW_s,$$

where $*$ denotes transpose and $f_t \in L^2_W$ (see previous section, definition of L^2_W for vector valued processes is similar).

Proof Let \mathcal{S} be the subspace of functions in $L^2(\Omega, \mathcal{F}_\infty^W, P)$ which are of the form

$$a + \int_0^\infty f_t^* dW_t,$$

where f_t has the required properties and $a \in R$. \mathcal{S} is closed because stochastic integration preserves norm. Set $s_t = [s_t^1, \dots, s_t^n]^* \in L^2((R^+)^n)$, where s_t^i 's are bounded step functions. The

idea of the proof is to show that if $Y \in L^2(\Omega, \mathcal{F}_\infty^W, P)$, $E[Y] = 0$ and $Y \perp \mathcal{S}$, then $Y=0$. But this is easy once we have recalled that by Ito's formula

$$Z_\infty = \exp\left(\int_0^\infty s_t dW_t - \int_0^\infty s_t^2 dt\right) = 1 + \sum_{j=0}^n \int_0^\infty Z_t dW_t^j.$$

In addition, $E[Z_\infty^2] < \infty$. Thus $Z_\infty \in \mathcal{S}$. Choose a bounded random variable $Y \in \mathcal{S}$. By choosing appropriate step functions, for any $t_j > t_{j-1} \dots > t_0$ we see that

$$E\left[\exp\left(\sum_{j=1}^n a_j^* W_{t_j}\right) Y\right] = 0$$

for all $a_j \in \mathbf{R}^n$. But Y defines a finite measure on $(\Omega, \mathcal{F}^{W_{t_1}, \dots, W_{t_j}})$ and it can be proved that $E[Y|W_{t_1} \dots W_{t_j}] = 0$ for example by analytic continuation. This can only be true if $Y = 0$.

Now we have proved that all functions in $L^2(\Omega, \mathcal{F}_\infty^W, P)$ with mean 0 belong to \mathcal{S} which proves the claim of the theorem.

Chapter 3

Stochastic Differential Equations

Stochastic differential equations are a standard modelling tool in continuous time. Even though they do not allow as general models as Markov chains in discrete-time, they are still quite general. In continuous-time filtering, both coming observation and hidden states are modelled as stochastic differential equations.

3.1 Strong Solutions

A stochastic differential equation is an equation of the form

$$X_t = \zeta + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s \quad [3.1]$$

with the initial condition $X_0 = \zeta \in L^2$. ζ is assumed to be independent of (F_t^W) . $b : [0, \infty) \times \mathbf{R}^n \rightarrow \mathbf{R}^n$ and $\sigma : [0, \infty) \times \mathbf{R}^n \rightarrow \mathbf{R}^{n \times r}$ are Borel measurable. (W_t) is an \mathbf{R}^n -dimensional Brownian motion. By (\mathcal{F}_t) we denote the augmented σ -algebra generated by (W_t) and X_0 .

Definition 3.1 If $A \in \mathbf{R}^n \times \mathbf{R}^r$, we define

$$\|A\|^2 = \sum_{i=1}^n \sum_{j=1}^r |A^{ij}|^2$$

Definition 3.2 A stochastic process (X_t, \mathcal{F}_t) is a strong solution to the stochastic differential equation 5.3, if

1. (X_t) is continuous.
2. $\int_0^t \sum_{i=1}^n \|b^i(s, X_s)\| ds + \int_0^t \|\sigma(s, X_s)\|^2 ds < \infty$ P-a.s. for all $t \geq 0$.
3. $X_t = \zeta + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s$

Theorem 3.3 Suppose that there exists $K > 0$ so that for all $t \geq 0$

$$\|b(t, x) - b(t, y)\| + \|\sigma(t, x) - \sigma(t, y)\| \leq K\|x - y\|$$

and

$$\|b(t, x)\| + \|\sigma(t, x)\| \leq K(1 + \|x\|).$$

Then the stochastic differential equation 3.1 has a unique strong solution. In addition,

$$\sup_{t \in [0, T]} E[X_t^2] < \infty \quad (T > 0).$$

The proof of this important result can be found in any book on stochastic calculus. It is also possible to define weak solutions but from our point of view the assumptions in theorem 3.3 are not too restrictive.

3.2 Infinitesimal Generators

Assume that (X_t) is a strong solution to equation 3.1, where $b(s, X_s)$ and $\sigma(s, X_s)$ satisfy the Lipschitz and growth conditions. Let $f \in C_c^2(\mathbf{R}^n)$, where $C_c^2(\mathbf{R}^n)$ is the vector space of continuous twice differentiable functions with compact support. By Ito's formula

$$\begin{aligned} f(X_t) &= f(X_0) + \sum_{i=1}^n \int_0^t b^i(s, X_s) \partial_{x_i} f(X_s) ds + \sum_{i=1}^n \sum_{j=1}^n \int_0^t \partial_{x_i} f(X_s) \sigma^{ij}(s, X_s) dW_s^j \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \int_0^t \partial_{x_i x_j} f(X_s) d\langle X^i, X^j \rangle_s. \end{aligned}$$

On the other hand,

$$\langle X^i, X^j \rangle_t = \int_0^t (\sigma(s, X_s) \sigma^*(s, X_s))^{ij} ds.$$

Define the linear operator

$$Af = \sum_{i=1}^n b^i(s, x) \partial_{x_i} f(x) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\sigma(s, X_s) \sigma^*(s, X_s))^{ij} \partial_{x_i x_j} f(x).$$

Then $f(X_t) - f(X_0) - \int_0^t Af(X_s) ds$ is a martingale.

Definition 3.4 The operator A is called the (infinitesimal) generator for (X_t) .

The operator A is of course related to the theory of analytic semigroups. Discussion on this point can for example be found in [16] or [29]. The infinitesimal generator is certainly very useful, because it gives a nice decomposition of (X_t) into martingale and finite variation parts.

3.3 Numerical Methods for Solving SDEs

In this section we derive briefly two numerical methods for approximating solutions of stochastic differential equations. The derivations are meant to be as simple as possible, as we do not want to go too deep into this complicated subject.

Consider the equation

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s. \quad [3.2]$$

In most cases there are no closed form solutions and numerical approximations of the process (X_t) are necessary. This means that (X_t) is replaced by a simpler process which is close to the original one and is such that realizations of it can be simulated. The first method which comes to mind is a generalization of the deterministic Euler method to the stochastic case. This is defined by choosing a step size h and defining the Markov chain

$$X_n^h = X_{n-1}^h + hb((n-1)h, X_{(n-1)h}) + \sigma((n-1)h, X_{(n-1)h})(W_{nh} - W_{(n-1)h}).$$

Next suppose that the initial condition is known perfectly. Then we have the following general error bound which also shows that the Euler method is not necessarily as good as its deterministic counterpart and thus not entirely satisfying.

Theorem 3.5 If the conditions

$$E[\|X_0\|^2] < \infty$$

$$\|b(t, x) - b(t, y)\| + \|\sigma(t, x) - \sigma(t, y)\| \leq K_1 \|x - y\|$$

$$\|b(t, x)\| + \|\sigma(t, x)\| \leq K_2 \|x\|$$

$$\|a(t, x) - a(s, x)\| + \|b(s, x) - b(t, x)\| \leq K_3(1 + \|x\|)|s - t|^{\frac{1}{2}}$$

hold for all $t, s \geq 0$ and $x \in \mathbb{R}^n$, then

$$E[\|X_{nh} - X_n^h\|^2] = O(h)$$

Proof The proof is given in [18].

Time-discretization of an SDE is not the most error-sensitive phase of a nonlinear filter. However, it may often be a good idea to use an order 1 method instead of the simple Euler-scheme. If we take one more term from the Ito-Taylor expansion of (X_t) , we get the method

$$\begin{aligned} X_n^i &= X_{n-1}^i + hb^i((n-1)h, X_{n-1}) \\ &+ \sum_{r=1}^n \sum_{l=1}^n \sum_{k=1}^n \sigma^{kl}((n-1)h, X_{n-1}) \partial_{x_k} \sigma^{ir}((n-1)h, X_{n-1}) \int_{(n-1)h}^{nh} \int_{(n-1)h}^t 1 dW_s^l dW_t^r. \end{aligned}$$

It can be proved under quite general assumptions that $E[\|X_{nh} - X_n^h\|^2] = O(h^2)$. In case b and σ are smooth enough the Milstein scheme is preferable to the Euler method. However, evaluation of multiple Ito-integrals is not totally simple. See [18] for more information on this subject. In numerical experiments we will use the Euler method because of its simplicity (except in one deterministic case). Also our models are relatively simple and there was no need for higher order approximations. In fact, if b is a constant matrix, then the error bound for the Euler method is of order 1 instead of 0.5.

Higher order approximations can be obtained by taking more terms from the Ito-Taylor expansion. There exist also numerous implicit and explicit algorithms for solving SDEs. An extensive introduction to these matters can be found in [18].

Chapter 4

Girsanov's Theorem

In this chapter we present Girsanov's Theorem. It is necessary when deriving the filtering equations and the Kallianpur-Striebel formula. We do not present the most general case, for a more general version of the theorem see [16].

Stochastic processes are defined with respect to a filtration (\mathcal{F}_t) that satisfies the usual conditions.

Theorem 4.1 Let $M = (M_t)$ be a continuous local martingale. Then

$$Z_t = \exp\left(M_t - \frac{1}{2}\langle M \rangle_t\right)$$

is a local martingale. Moreover, if $E[Z_t] = 1$ for all $t \geq 0$, then (Z_t) is a martingale.

Proof By Ito's formula,

$$Z_t - 1 = \int_0^t Z_s dM_s - \frac{1}{2} \int_0^t Z_s d\langle M \rangle_s + \frac{1}{2} \int_0^t Z_s d\langle M - \frac{1}{2}\langle M \rangle \rangle_s = \int_0^t Z_s dM_s$$

from which the first claim follows. Choose a sequence of stopping times $(\tau_n) \rightarrow \infty$ so that $(Z_{\tau_n \wedge t})$ is a martingale. By Fatou's lemma,

$$E[Z_t | \mathcal{F}_s] \leq \liminf_{n \rightarrow \infty} E[Z_{\tau_n \wedge t} | \mathcal{F}_s] = Z_s$$

when $s \leq t$. Thus Z_t is a supermartingale and the last claim follows for example from Doob's decomposition. Choose any numerable subset of $[0, \infty)$ and denote the corresponding discrete-time supermartingale by (Z_n) . Then $Z_n = Z_0 + M_n - A_n$, where (A_n) is monotone and negative. Because $E[Z_n] = Z_0 - E[A_n] = 1$, (A_n) must be a constant and (Z_n) a martingale.

Theorem 4.2 Let P and \tilde{P} be equivalent probability measures. Then the process defined by

$$L_t = \frac{d\tilde{P}|_{\mathcal{F}_t}}{dP|_{\mathcal{F}_t}}$$

is an (\mathcal{F}_t) -martingale (w.r.t. P).

Proof Choose any $A \in \mathcal{F}_s$ and $t \geq s \geq 0$. Then

$$\int_A L_t dP = \tilde{P}(A) = \int_A L_s dP$$

Theorem 4.3 Assume that P and \tilde{P} are equivalent probability measures on \mathcal{F}_∞ and the process L_t defined in the previous theorem is an uniformly integrable martingale. Then a continuous process (X_t, \mathcal{F}_t) is a local martingale w.r.t. \tilde{P} if and only if $(X_t L_t)$ is a P -local martingale.

Proof The proof is not hard and it is left to the reader. It can also be found in [16].

Theorem 4.4 Let (M_t, \mathcal{F}_t) be a continuous P -local martingale. Assume that the exponential local martingale

$$Z_t = \exp(M_t - \frac{1}{2}\langle M \rangle_t)$$

has a constant mean. Let P and \tilde{P} be probability measures on \mathcal{F}_∞ so that

$$\frac{d\tilde{P}}{dP} = Z_\infty$$

If $V = (V_t)$ is a continuous local martingale, then

$$X_t = V_t - \langle V, M \rangle_t$$

is a \tilde{P} -local martingale so that the quadratic variation of (X_t) is $\langle V \rangle$.

Proof That Z_∞ exists and is integrable follows from the supermartingale convergence theorem and Fatou's lemma.

$$\begin{aligned} Z_t X_t &= \int_0^t X_s dZ_s + \int_0^t Z_s dX_s + \langle Z, X \rangle_s \\ &= \int_0^t Z_s X_s dM_s + \int_0^t Z_s dV_s - \int_0^t Z_s d\langle V, M \rangle_s + \langle X, Z \rangle_s. \end{aligned}$$

Using the Kunita-Watanabe identity (see for example [28])

$$\langle X, Z \rangle_t = \int_0^t Z_s d\langle V, M \rangle_s.$$

Thus

$$X_t Z_t = \int_0^t Z_s X_s dM_s + \int_0^t Z_s dV_s.$$

From theorem 4.3 it follows that (X_t) is a \tilde{P} local martingale. Choose $t > 0$ and a partition $t = t_n > t_{n-1} \dots > t_0$. Then it is standard knowledge that

$$\sum_{i=1}^n (V_{t_i} - V_{t_{i-1}})^2 \rightarrow \langle V \rangle_t \tag{4.1}$$

in probability as $\max_i |t_i - t_{i-1}| \rightarrow 0$. This holds in both spaces (Ω, P) and (Ω, \tilde{P}) . By choosing an appropriate subsequence we may assume that the convergence is pointwise w.r.t. P . Now noting that P and \tilde{P} are equivalent probability measures, it is obvious that the limit 4.1 does not depend on the choice of measure. Thus $\langle X \rangle^{\tilde{P}} = \langle V \rangle^{\tilde{P}} = \langle V \rangle^P$.

This theorem is significant in filtering theory. It can be used for a change of probability measure. It is used to make a process of the form $W_t + S_t$, where W_t is a brownian motion, a Brownian motion with respect to a new probability measure.

Chapter 5

Nonlinear Filtering - A General Case

In this chapter we will derive the important Kushner-Stratonovitch equation. The theory is quite complicated, but the basic idea is clear. Consider a process (X_t) and a continuous measurement signal (y_t) . The measurements are given by

$$y_t = h(t, X_t) + v_t,$$

where v_t is white noise. Note that white noise exists only in an abstract sense and the notation is formal. v_t is thought as the derivative of a Brownian motion (in practice continuous-time noise is close to it). By integrating we get

$$Y_t = \int_0^t y_s ds = \int_0^t h(s, X_s) ds + V_t,$$

where V_t is a Brownian motion. Without losing any information, the filtering problem can be stated as finding the conditional distribution of (X_t) when the process (Y_t) is known.

5.1 The Innovation Process

We adopt the same quite general setting as in [16]. The observation process is defined as an \mathbf{R}^n -valued process

$$Y_t = S_t + W_t$$

The observation σ -algebra is defined as $\mathcal{Y}_t = \sigma(\sigma(Y_s, s \leq t) \cup \{A \in \mathcal{F} : P(A) = 0\})$. By (\mathcal{F}_t) we denote a filtration satisfying the usual conditions.

The following assumptions are assumed to hold in the rest of the thesis:

1.

$$S_t(\omega) = \int_0^t h_s(\omega) ds, \quad [5.1]$$

where (h_t) is adapted to (\mathcal{F}_t) , progressively measurable and

$$E\left[\int_0^\infty h_t^2 dt\right] < \infty.$$

2. $E[|h_t|] < \infty$ for all $t > 0$.

3. (W_t, \mathcal{F}_t) is an \mathbf{R}^n -Brownian motion.

The goal of nonlinear filtering is to compute $E[f(X_t)|\mathcal{Y}_t]$ where (X_t) is the state process and $f(x)$ a measurable function. (X_t) is assumed to be adapted to (\mathcal{F}_t) and take its values in a Polish space. The state-space is denoted by \mathcal{E} .

S_t may be written as

$$S_t = S_t^+ - S_t^- = \int_0^t h_s^+ ds - \int_0^t h_s^- ds.$$

\tilde{S}_t is defined as

$$\tilde{S}_t = U_t^1 - U_t^2,$$

where (U_t^1) is the dual predictable projection of (S_t^+) on (\mathcal{Y}_t) . Correspondingly, (U_t^2) is the dual predictable projection of S_t^- . It may be proved that

$$\tilde{S}_t = \int_0^t E[h_s|\mathcal{Y}_s] ds$$

where a progressively measurable version of the process $(E[h_t|\mathcal{Y}_t])$ is chosen. A proof that such a version exists is given in [16], Theorem 2.7.1.

Definition 5.1 The innovation process is defined as

$$v_t = Y_t - \tilde{S}_t.$$

Some basic properties of the innovation process could be proved without the restrictive assumption of equation 5.1. The next theorem is not so surprising once it is noted that (v_t) is a martingale.

Theorem 5.2 (v_t, \mathcal{Y}_t) is a Brownian motion.

Proof For simplicity only the proof of the scalar case is given here. That (v_t) is a continuous martingale is easy to prove by an elementary calculation. Choose $t > 0$ and points $t = t_n > t_{n-1} > \dots > t_0 = 0$. It is a well known fact that

$$\begin{aligned} \sum_{i=1}^n (v_i - v_{i-1})^2 &= \sum_{i=1}^n (S_{t_i} - \tilde{S}_{t_i} - S_{t_{i-1}} + \tilde{S}_{t_{i-1}})^2 \\ &\quad + \sum_{i=1}^n 2(S_{t_i} - \tilde{S}_{t_i} - S_{t_{i-1}} + \tilde{S}_{t_{i-1}})(W_{t_i} - W_{t_{i-1}}) \\ &\quad + \sum_{i=1}^n (W_{t_i} - W_{t_{i-1}})^2 \rightarrow \langle S \rangle_t \quad (\text{in probability}) \end{aligned}$$

as $n \rightarrow \infty$ and $\max_i |t_i - t_{i-1}| \rightarrow 0$. But the second term in the right side may be bounded by

$$\sum_{i=1}^n 2(S_{t_i} - \tilde{S}_{t_i} - S_{t_{i-1}} + \tilde{S}_{t_{i-1}})(W_{t_i} - W_{t_{i-1}}) \leq 2(\text{var}_t(S) + \text{var}_t(\tilde{S}))(\max_i |W_{t_i} - W_{t_{i-1}}|)$$

and thus converges to 0 in probability as $n \rightarrow \infty$ and $\max_i |t_i - t_{i-1}| \rightarrow 0$. The same holds for the first term. The last sum converges to $\langle W \rangle_t = t$. By Levy's Characterisation of Brownian Motion (see for example [28]) (v_t) is a Brownian Motion with respect to (\mathcal{Y}_t) .

The previous theorem shows why innovation processes are interesting. The proof of the next theorem follows closely that of theorem 8.3.1. in [16]. Again this result is not surprising taking into account the martingale representation theorem 2.44.

Theorem 5.3 Any square integrable cadlag martingale (M_t, \mathcal{Y}_t) ($M_0 = 0$) may be represented as

$$M_t = \int_0^t \psi_s^* dv_s, \tag{5.2}$$

where (ψ_t) satisfies the usual measurability and integrability conditions.

Proof First an exponential local martingale is defined as

$$Z_t = \exp\left(-\int_0^t E[h_s | \mathcal{Y}_s]^* dv_s - \frac{1}{2} \int_0^t \|E[h_s | \mathcal{Y}_s]\|^2 ds\right).$$

To make (Z_t) a martingale and to bound it, we define the stopping times

$$\tau_j = \inf_t \{t \geq 0 : Z_t > j \text{ or } Z_t < 1/j\}.$$

Choose any $j > 0$. We get the exponential martingale

$$Z_{t \wedge \tau_j} = \exp\left(- \int_0^{t \wedge \tau_j} E[h_s | \mathcal{Y}_s]^* dv_s - \frac{1}{2} \int_0^{t \wedge \tau_j} \|h_s\|^2 ds\right).$$

Now by Girsanov's theorem (theorem 4.4)

$$\tilde{Y}_t = v_t + \int_0^{t \wedge \tau_j} E[h_s | \mathcal{Y}_s] ds$$

is a Brownian motion with respect to the filtration (\mathcal{Y}_t) and the probability measure \tilde{P} . This may be seen by considering the covariations and quadratic variations of the components. By theorem 4.4 it is easy to see that they are the same as those of a Brownian Motion. It then follows from Levy's Characterisation of Brownian Motion that (\tilde{Y}_t) is a Brownian Motion.

Let $(\mathcal{F}_t^{\tilde{Y}})$ be the σ -algebra generated by (\tilde{Y}_t) augmented with 0-measurable sets. It can be proved that

$$\mathcal{F}_{t \wedge \tau_j}^{\tilde{Y}} = \tilde{\sigma}(\tilde{Y}_{t \wedge \tau_j}, 0 \leq s \leq t) = \tilde{\sigma}(Y_{t \wedge \tau_j}, 0 \leq s \leq t) = \mathcal{Y}_{t \wedge \tau_j},$$

where $\tilde{\sigma}$ denotes the completion of the generated σ -algebra. Next note that τ_j is an $(\mathcal{Y}_{t \wedge \tau_j})$ stopping time and thus an $(\mathcal{F}_t^{\tilde{Y}})$ stopping time.

Let (M_t, \mathcal{Y}_t) be a square integrable cadlag martingale (w.r.t. \tilde{P}) so that $M_0 = 0$. Then $(M_{t \wedge \tau_j})$ is a $(\mathcal{F}_t^{\tilde{Y}})$ -martingale by theorems 2.17 and 2.18.

Thus by theorem 2.44 (the integral is w.r.t. \tilde{P})

$$M_{t \wedge \tau_j} = \int_0^t \tilde{f}_s^* d\tilde{Y}_s,$$

where (\tilde{f}_t) is predictable and satisfies the integrability conditions. On the other hand,

$$M_{t \wedge \tau_j} = \int_0^{t \wedge \tau_j} \tilde{f}_s^* d\tilde{Y}_s = \int_0^{t \wedge \tau_j} f_s^* dY_s, \quad [5.3]$$

where (f_s) is adapted and measurable wrt. (\mathcal{Y}_s) . 5.3 follows straightforwardly from the basic formulas of stochastic integration.

We are interested in P -martingales (M_t, \mathcal{Y}_t) . Switching back to using P , by Ito's formula and representation 5.3 we get

$$Z_{t \wedge \tau_j} M_{t \wedge \tau_j} = Z_{t \wedge \tau_j} \int_0^{t \wedge \tau_j} f_s^* dv_s = \int_0^{t \wedge \tau_j} Z_{s \wedge \tau_j} f_s^* + Z_{s \wedge \tau_j} M_{s \wedge \tau_j}^* dv_s = \int_0^t (\Psi_s^j)^* dv_s.$$

because $M_{t \wedge \tau_j}$ is a martingale with respect to the measure \tilde{P} . This means that $Z_{t \wedge \tau_j} M_{t \wedge \tau_j}$ is a martingale in the space $(\Omega, P, \mathcal{Y}_t)$. On the other hand, a continuous martingale of finite variation is constant, which in this case means that all terms of finite variation could be omitted.

Clearly the sequence (Ψ_t^j) is unique in the sense that $(\Psi_{t \wedge \tau_j}^j) = (\Psi_{t \wedge \tau_j}^{j+1}) \lambda(dt) \times P(d\omega)$ -a.s. Thus by localisation we may define (Ψ_t) which satisfies all the conditions required.

5.2 The Kushner-Stratonovitch Equation

In this section we derive an equation for the conditional expectations $E[f(X_t)|\mathcal{Y}_t]$ where f is measurable and smooth enough. This is done by decomposing $f(X_t)$ into a martingale and a finite variation part. The finite variation part is easy to characterize but the martingale is more difficult.

Assume that $f(X_t)$ is cadlag and L^2 -bounded. We also assume that

$$f(X_t) = E[f(X_0)] + M_t(f) + A_t(f),$$

where $M_t(f) = f(X_t) - A_t(f) - E[f(X_0)]$ is a cadlag martingale. From $A(f) = (A_t(f))$ it is assumed that $A_0 = 0$ and

$$E[\text{var}_\infty(A(f))^2] < \infty$$

A typical choice for $(A_t(f))$ is the infinitesimal generator for a stochastic differential equation.

It can be quite easily proved that

$$E[f(X_t)|\mathcal{Y}_t] = E[f(X_0)] + \tilde{M}_t(f) + \tilde{A}_t(f),$$

where $(\tilde{A}_t(f))$ is obtained by taking the dual predictable projections of positive and negative parts of $(A_t(f))$. $(\tilde{M}_t(f), \mathcal{Y}_t)$ is an L^2 -bounded cadlag martingale. From theorem 2.36 it follows that

$$E[\text{var}_\infty(\tilde{A}(f))^2] < \infty$$

Using equation 5.2 we may write

$$\tilde{M}_t(f) = \int_0^t \psi_s(f)^* dv_s,$$

where $(\psi_t(f))$ is predictable, adapted to (\mathcal{Y}_t) and

$$E\left[\int_0^\infty \|\psi_s(f)\|^2 ds\right] < \infty.$$

Theorem 5.4 There exists a predictable process $(D_t(f), \mathcal{F}_t)$ for which

$$E\left[\int_0^\infty \|D_t(f)\|^2 ds\right] < \infty$$

and

$$M_t W_t - \int_0^t D_s(f) ds$$

is a martingale.

Proof This theorem is a straightforward consequence of theorem 6.8.1. and its proof in [16]. Note that Kallianpur defines the quadratic variation for all cadlag processes. We are mainly interested in continuous processes, though the proof of the Kushner-Stratonovitch equation is given assuming that $f(X_t)$ is cadlag.

Lemma 5.5 Assume that (f_t, \mathcal{Y}_t) is a predictable process so that

$$E\left[\int_0^\infty f_s^2 ds\right] < \infty.$$

Let (B_t, \mathcal{F}_t) and (Z_t, \mathcal{Y}_t) be cadlag processes for which

$$E[\text{var}_\infty(B)^2] < \infty$$

and

$$Z_t = \int_0^t f_s dv_s.$$

Let $T = t_k > t_{k-1} > \dots > t_0 = 0$ be a partition of the interval $[0, T]$. Then

$$E\left[\sum_{i=1}^k (B_{t_i} - B_{t_{i-1}}) \int_{t_{i-1}}^{t_i} f_s dv_s\right] \rightarrow 0$$

as $k \rightarrow \infty$ and $\max_i |t_i - t_{i-1}| \rightarrow 0$.

Proof We have the following obvious estimate

$$\left| \sum_{i=1}^k (B_{t_i} - B_{t_{i-1}}) \int_{t_{i-1}}^{t_i} f_s dv_s \right| \leq \max_i \left| \int_{t_{i-1}}^{t_i} f_s dv_s \right| \text{var}_T(B) \rightarrow 0 \quad \text{P-a.s. as } k \rightarrow \infty.$$

Moreover,

$$\max_i \left| \int_{t_{i-1}}^{t_i} f_s dv_s \right| \leq 2 \sup_{t \in [0, T]} \left| \int_0^t f_s dv_s \right|.$$

By Doob's inequality

$$2E\left[\left(\sup_{t \in [0, T]} \left| \int_0^t f_s dv_s \right|\right)^2\right] \leq 8E\left[\left(\int_0^T f_s dv_s\right)^2\right] = 8E\left[\int_0^T f_s^2 ds\right].$$

Using the Hölder inequality we see that

$$E[\text{var}_T(B) \sup_{t \in [0, T]} \left| \int_0^t f_s dv_s \right|] \leq 2(E[\text{var}_T(B)^2])^{\frac{1}{2}} (E[\int_0^T f_s^2 ds])^{\frac{1}{2}} < \infty.$$

Thus by the Dominated Convergence Theorem

$$E\left[\left|\sum_{i=1}^k (B_{t_i} - B_{t_{i-1}}) \int_{t_{i-1}}^{t_i} f_s dv_s\right|\right] \rightarrow 0$$

as $k \rightarrow \infty$.

Now we state the main result of this chapter.

Theorem 5.6 Choose any $f \in C_b(\mathcal{E})$. Then the following equality holds:

$$\begin{aligned} E[f(X_t)|\mathcal{Y}_t] &= E[f(X_0)] + \tilde{A}_t(f) + \int_0^t E[f(X_s)h_s|\mathcal{Y}_s]^* dv_s \\ &\quad - \int_0^t E[f(X_s)|\mathcal{Y}_s]E[h_s|\mathcal{Y}_s]^* dv_s + \int_0^t E[D_s(f)^*|\mathcal{Y}_s] dv_s \end{aligned} \quad [5.4]$$

Proof Choose any $S_\infty \in L^\infty(\Omega, \mathcal{Y}_\infty, P)$ and let (S_t) be the corresponding martingale. By theorem 5.3

$$S_t = \int_0^t \psi_s^* dv_s.$$

Using a density argument we may assume that ψ_s is bounded. To prove the formula 5.4 we only need to examine

$$E[\tilde{M}_t(f)S_t] = E[M_t(f)S_t] + E[(\tilde{M}_t(f) - M_t(f))S_t]. \quad [5.5]$$

The proof consists of two parts. First we will examine the term $E[M_t(f)S_t]$.

$$\begin{aligned} E[M_t(f)S_t] &= E[M_t(f)(\int_0^t \psi_s^* dY_s - \int_0^t \psi_s^* E[h_s|\mathcal{Y}_s] ds)] \\ &= E[M_t(f) \int_0^t \psi_s^* dW_s] + E[M_t(f) \int_0^t \psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds] \end{aligned}$$

The first term is easy

$$\begin{aligned} E[M_t(f) \int_0^t \psi_s^* dW_s] &= \sum_{i=0}^n E[\int_0^t \psi_s^i D_s^i(f) ds] \\ &= \sum_{i=0}^n E[\int_0^t \psi_s^i E[D_s^i(f)|\mathcal{Y}_s] ds] \\ &= E[S_t \int_0^t E[D_s(f)|\mathcal{Y}_s]^* dv_s]. \end{aligned}$$

Note that

$$E[\int_0^t (M_t(f) - M_s(f))\psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds] = 0.$$

Thus for the second term we have

$$\begin{aligned} &E[M_t(f) \int_0^t \psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds] \\ &= E[\int_0^t M_s(f)\psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds] \\ &= E[-\int_0^t A_s(f)\psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds] + E[\int_0^t f(X_s)\psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds]. \end{aligned}$$

On the other hand

$$E[\int_0^t f(X_s)\psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds] = E[S_t \int_0^t E[f(X_s)h_s|\mathcal{Y}_s]^* dv_s] - E[S_t \int_0^t E[f(X_s)|\mathcal{Y}_s]E[h_s|\mathcal{Y}_s]^* dv_s].$$

Next we examine the second term in 5.5. First note that

$$\begin{aligned} E[(\tilde{M}_t(f) - M_t(f))S_t] &= E[(E[f(X_t)|\mathcal{Y}_t] - f(X_t))S_t] + E[(A_t(f) - \tilde{A}_t(f))S_t] \\ &= E[(A_t(f) - \tilde{A}_t(f))S_t]. \end{aligned} \tag{5.6}$$

Choose points $t = t_k > t_{k-1} > \dots > t_0 = 0$. Equation 5.6 can be written as a sum

$$E[(A_t(f) - \tilde{A}_t(f))S_t] = E[\sum_{i=1}^k (\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f))S_{t_{i-1}} + \sum_{i=1}^k (\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f))(S_t - S_{t_{i-1}})],$$

where $\Delta A_{t_i}(f) = A_{t_i}(f) - A_{t_{i-1}}(f)$. Remembering that

$$E[\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f) | \mathcal{Y}_{t_i}] = 0$$

it is obvious that

$$E\left[\sum_{i=1}^k (\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f)) S_{t_{i-1}}\right] = 0.$$

Thus we have

$$\begin{aligned} E[(\tilde{M}_t(f) - M_t(f)) S_t] &= \sum_{i=1}^k E[(\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f)) \int_{t_{i-1}}^t \psi_s^* dW_s \\ &\quad + (\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f)) \int_{t_{i-1}}^t \psi_s^* (h_s - E[h_s | \mathcal{Y}_s]) ds]. \end{aligned}$$

On the other hand,

$$E[(\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f)) \int_{t_i}^t \psi_s^* dW_s] = E[(\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f)) E[\int_{t_i}^t \psi_s^* dW_s | \mathcal{F}_{t_i}]] = 0.$$

By lemma 5.5

$$\sum_{i=1}^k E[(\Delta A_{t_i}(f) - \Delta \tilde{A}_{t_i}(f)) \int_{t_{i-1}}^{t_i} \psi_s^* dv_s] \rightarrow 0$$

when $k \rightarrow \infty$ and $\max_{1 \leq i \leq k} |t_i - t_{i-1}| \rightarrow 0$. Using Fubini's theorem we can get rid of one more term:

$$E[\Delta \tilde{A}_{t_i}(f) \int_{t_i}^t \psi_s^* (h_s - E[h_s | \mathcal{Y}_s]) ds] = \int_{t_i}^t E[\Delta \tilde{A}_{t_i}(f) \psi_s^* (h_s - E[h_s | \mathcal{Y}_s])] ds = 0.$$

Thus what we have left is

$$E[(\tilde{A}_t(f) - A_t(f)) S_t] = \sum_{i=1}^k E[\Delta A_{t_i}(f) \int_{t_i}^t \psi_s^* (h_s - E[h_s | \mathcal{Y}_s]) ds].$$

By an elementary manipulation

$$\sum_{i=1}^k E[\Delta A_{t_i}(f) \int_{t_i}^t \psi_s^* (h_s - E[h_s | \mathcal{Y}_s]) ds] = \sum_{i=1}^{k-1} E[A_{t_i}(f) \int_{t_i}^{t_{i+1}} \psi_s^* (h_s - E[h_s | \mathcal{Y}_s]) ds].$$

Thus we have

$$\begin{aligned}
& |E[\sum_{i=1}^k \Delta A_{t_i}(f) \int_{t_i}^t \psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds - \int_0^t A_s(f) \psi_s^*(h_s - E[h_s|\mathcal{Y}_s]) ds]| \\
& \leq \sum_{i=0}^{k-1} E[\int_{t_i}^{t_{i+1}} |\psi_s^*(h_s - E[h_s|\mathcal{Y}_s])| (\text{var}_{t_{i+1}}(A(f)) - \text{var}_{t_i}(A(f))) ds] \\
& \leq E[\text{var}_t(A(f)) \max_{1 \leq i \leq k-1} \int_{t_i}^{t_{i+1}} |\psi_s^*(h_s - E[h_s|\mathcal{Y}_s])| ds] \rightarrow 0
\end{aligned}$$

by the Lebesgue Dominated Convergence Theorem.

Now by summing the terms from the first and the second parts of the proof we see the desired result.

At the first sight the KS-equation may seem a bit too abstract to be useful. But in the next section it will be seen that the terms have a concrete form once the model has a bit more structure. The KS-equation has lots of theoretical significance and knowing it is essential to anyone who is interested in the theory of continuous-time filters.

Chapter 6

Filtering of Stochastic Differential Equation Models

In this chapter we adopt the same general assumptions and notation as in the previous chapter. The theory is simpler than that of the previous section.

6.1 A Modification of the Kushner-Stratonovitch Equation

We will examine the model

$$X_t = X_0 + \int_0^t a(s, X_s, Y_s) ds + \int_0^t b(s, X_s, Y_s) dW_s \quad [6.1]$$

$$Y_t = \int_0^t c(s, X_s, Y_s) ds + V_t. \quad [6.2]$$

(W_t, \mathcal{F}_t) and (V_t, \mathcal{Y}_t) are independent Brownian motions. Assuming independence is not necessary but slightly simplifies the equations. We chose not to examine the correlated case because the uncorrelated case is general enough for most purposes. Also to derive the Kallianpur-Striebel formula the assumption that (X_t) and (W_t) are independent is necessary.

We also assume that X_0 is independent of (W_t) and (V_t) . To ensure the existence and uniqueness of solutions to equations 6.1 and 6.2, $a(s, x, y)$, $b(s, x, y)$ and $c(s, x, y)$ are assumed to satisfy the Lipschitz and growth conditions.

For $f \in C_b^2(\mathbf{R}^N)$, define $A_t f$ as

$$A_t f = \sum_{i=1}^n a^i(t, X_t, Y_t) \partial_{x_i} f + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (bb^*)^{ij}(t, X_t, Y_t) \partial_{x_i x_j} f.$$

We assume that $A_t f \in L_W^2$. This should not be a problem, see for example chapter 5 in [16] on the scalar case. By $A_t f(t, x, \omega)$ we mean the function of three variables when X_t is replaced by

x . The definition of $A_t f$ is of course motivated by the definition of the infinitesimal generator in section 3.2. If a and b are smooth enough, we may define

$$A_t^* f = - \sum_{i=1}^n \partial_{x_i} (a^i(t, x, Y_t(\omega)) f) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \partial_{x_i x_j} ((bb^*)^{ij}(t, x, Y_t(\omega)) f).$$

Assume that for any $t > 0$

$$P(X_t \in A | \mathcal{Y}_t) = \int_A \phi(t, x, \omega) dx,$$

where $\phi(t, x, \omega)$ is positive, measurable and fixing any (t, x) we get an \mathcal{Y}_t -measurable mapping.

Define

$$H(t, x, \omega) = \phi(x, \omega, t) c(t, x, Y_t(\omega)) + \phi(x, \omega, t) \int_{\mathbf{R}^N} c(t, x, Y_t(\omega)) \phi(x, \omega, t) dx.$$

$H(t, x, \omega)$ is well defined because

$$E \left[\int_0^T \int_{\mathbf{R}^N} \|c(t, x, Y_t(\omega))\| \phi(x, \omega, t) dx dt \right] \leq \int_0^T E[\|c(t, X_t(\omega), Y_t(\omega))\|] dt < \infty.$$

$H(t, \cdot, \omega)$ is continuous for almost all $t \geq 0$ and $\omega \in \Omega$.

Theorem 6.1 Assume that the assumptions stated earlier in this chapter hold. We also assume that ϕ, a, b and c are a.s. smooth enough for $t > 0$,

$$\int_0^t \int_{\mathbf{R}^N} |A^* \phi(s, x, \omega)| dx ds < \infty \quad (\text{a.s.})$$

and

$$\int_0^t \int_{\mathbf{R}^N} \int_{\Omega} \|H(s, x, \omega)\|^2 dP dx ds < \infty.$$

Then we have for each $t > 0$ the equation

$$\phi(t, x, \omega) = \phi(0, x, \omega) + \int_0^t A_s^* \phi(s, x, \omega) ds + \int_0^t H(s, x, \omega)^* dv_s \quad [6.3]$$

which holds almost surely for almost every x .

Proof Choose $f \in C_c^\infty$. By theorem 5.6

$$\begin{aligned}
E[f(X_t)|\mathcal{Y}_t] &= E[f(X_0)] + \int_0^t E[A_s f | \mathcal{Y}_s] ds + \int_0^t \int_{\mathbf{R}^N} f(x) c(s, x, Y_s(\omega)) \phi(s, x, \omega) dx dv_s \\
&\quad + \int_0^t \int_{\mathbf{R}^N} f(x) \phi(s, x, \omega) dx \int_{\mathbf{R}^N} c(s, x, Y_s(\omega)) \phi(s, x, \omega) dx dv_s \\
&= \int_{\mathbf{R}^N} \phi(0, x, \omega) f(x) dx + \int_0^t \int_{\mathbf{R}^N} A_s f(s, x, \omega) \phi(s, x, \omega) dx ds \\
&\quad + \int_0^t \int_{\mathbf{R}^N} f H(s, x, \omega) dx dv_s. \tag{6.4}
\end{aligned}$$

By the stochastic Fubini theorem (in [27] it is given in a very general setting)

$$\int_0^t \int_{\mathbf{R}^N} f H(s, x, \omega) dx dv_s = \int_{\mathbf{R}^N} \int_0^t f H(s, x, \omega) dv_s dx.$$

The same theorem ensures that both sides are well defined.

For the other term in the right hand side of the equation 6.4 we have

$$\int_0^t \int_{\mathbf{R}^N} A_s f(s, x, \omega) \phi(s, x, \omega) dx ds = \int_{\mathbf{R}^N} \int_0^t f(x) A_s^* \phi(s, x, \omega) ds dx \tag{6.5}$$

by Fubini's Theorem.

The right hand side of equation 6.3 belongs almost surely to $L^1(\mathbf{R}^N)$. Because $L^1(\mathbf{R}^N)$ is separable, it is easy to see using equations 6.4 and 6.5 by choosing a countable dense subset that equation 6.3 holds a.s. for almost every x .

Note that if there exists a (t, x) continuous versions of the terms in the equation, then the claim holds a.s. for all (t, x) . More discussion on this point can be found in [30]. In the simulation part these conditions hold even though no formal proof is given.

The existence and uniqueness of solutions to the KS-equation is a deep question and is not discussed here. In the literature on continuous-time nonlinear filtering the Kushner-Stratonovitch equation is often given in the modified form derived in this section. The modified equation is a stochastic partial differential equation and standard numerical techniques for solving PDEs can be used.

6.2 The Kallianpur-Striebel Formula

In the previous section we derived an equation for the conditional densities of $p(dx|\mathcal{Y}_t)$. An alternative and theoretically easier way to investigate is to use the Kallianpur-Striebel formula. We will examine the following slightly less general model than in section 6.1:

$$X_t = X_0 + \int_0^t a(s, X_s) ds + \int_0^t b(s, X_s) dW_s$$

$$Y_t = \int_0^t c(s, X_s) ds + V_t$$

The state-space is \mathbf{R}^N and Y_t takes its values in \mathbf{R}^M . This model is of course a special case of the previous one and thus we may adopt the same setting as before.

The following theorem tells us that the exponential local martingale is a martingale. This is important, because it will be used for a change of measure.

Theorem 6.2 Define

$$Z_t = \exp\left(-\int_0^t c(s, X_s)^* dV_s - \frac{1}{2} \int_0^t \|c(s, X_s)\|^2 ds\right).$$

Then Z_t is a martingale.

Proof By theorem 4.1 it suffices to prove that $E[Z_t] = 1$ for all $t > 0$. Denote by (\mathcal{G}_t) the σ -algebra

$$\mathcal{G}_t = \sigma(\mathcal{F}_t^V \cup (\cup_{t \in \mathbf{R}^+} \mathcal{F}_t^W) \cup \sigma(X_0)).$$

Clearly (V_t, \mathcal{G}_t) is a Brownian motion. Thus we may change the filtration to (\mathcal{G}_t) and examine the local martingale (Z_t, \mathcal{G}_t) . The set

$$A_N = \{\omega : \|X_s(\omega)\| \leq N \text{ for all } s \in [0, t]\}$$

belongs to \mathcal{G}_0 . It is easy to prove that

$$I_{A_N} \int_0^t c(s, X_s)^* dV_s = \int_0^t I_{A_N} c(s, X_s)^* dV_s.$$

By defining $Z_t^N = I_{A_N} Z_t$ we obtain

$$\begin{aligned} E[Z_t^N] &= \int_{\Omega} I_{A_N} Z_t dP = \int_{\Omega} \exp(-I_{A_N} \int_0^t c(s, X_s)^* dV_s - \frac{1}{2} I_{A_N} \int_0^t \|c(s, X_s)\|^2 ds) dP - P(A_N^c) \\ &= \int_{\Omega} \exp(-\int_0^t I_{A_N} c(s, X_s)^* dV_s - \frac{1}{2} \int_0^t I_{A_N} \|c(s, X_s)\|^2 ds) dP - P(A_N^c). \end{aligned}$$

It is clear that

$$P(A_N^c) \rightarrow 0$$

and

$$\int_{A_N} Z_t dP \rightarrow \int_{\Omega} Z_t dP$$

as $N \rightarrow \infty$. Because $I_{A_N} \|c(s, X_s)\| \leq K(1 + N)$ for some constant $K > 0$, it is not hard to prove that $E[Z_t^N] \rightarrow 1$ (the proof can be found in most books on stochastic calculus) which concludes the proof.

Choose $T > 0$ and define the probability measure \tilde{P} on \mathcal{F}_T by

$$\tilde{P}(A) = \int_A Z_T dP.$$

An interesting fact about this measure is that by theorem 4.4 (Y_t) is a Brownian motion in the space $(\Omega, \mathcal{F}_T, \tilde{P})$. Note also that if U is an integrable \mathcal{F}_s measurable random variable, then $E_{\tilde{P}}[U | \mathcal{Y}_s] = E_{\tilde{P}}[U | \mathcal{Y}_T]$. This follows from the facts that

$$\mathcal{Y}_T = \sigma(\mathcal{Y}_s \cup (\cup_{T \geq t > s} \sigma(Y_t - Y_s)))$$

and $Y_t - Y_s \perp \mathcal{F}_s$ (w.r.t. \tilde{P}). This formula is of course natural as (Y_t) is a Brownian motion and doesn't bring any new information about U after $t = s$.

Another thing which motivates the change of measure is the following:

Theorem 6.3 (X_t) and (Y_t) are independent under the measure \tilde{P} .

Proof Clearly $[W_t^* Y_t^*]^*$ is a Brownian motion with respect to the history (\mathcal{F}_t) and thus $F_T^W \perp F_T^Y$. It is also easy to see that $X_0 \perp F_T^{W,Y}$. The claim is clear after noting that (X_t) is adapted to the σ -algebra generated by X_0 and (W_t) .

By elementary manipulation

$$Z_t^{-1} = \exp\left(\int_0^t c(s, X_s) dV_s + \frac{1}{2} \int_0^t \|c(s, X_s)\|^2 ds\right) = \exp\left(\int_0^t c(s, X_s) dY_s - \frac{1}{2} \int_0^t \|c(s, X_s)\|^2 ds\right).$$

Now we will state a result which is important from both practical and theoretical point of views.

Theorem 6.4 (Kallianpur-Striebel)

For any bounded measurable function $\phi(x)$ we have

$$E_P[\phi(X_t)|\mathcal{Y}_t] = \frac{E_{\tilde{P}}[\phi(X_t)Z_t^{-1}|\mathcal{Y}_t]}{E_{\tilde{P}}[Z_t^{-1}|\mathcal{Y}_t]}$$

Proof The proof is straightforward. Choose any $A \in \mathcal{Y}_t$. Then

$$\begin{aligned} \int_A E_P[\phi(X_t)|\mathcal{Y}_t] E_{\tilde{P}}[Z_t^{-1}|\mathcal{Y}_t] d\tilde{P} &= \int_A \phi(X_t) dP \\ &= \int_A E_{\tilde{P}[\phi(X_t)Z_t^{-1}|\mathcal{Y}_t]} d\tilde{P} \end{aligned}$$

The Kallianpur-Striebel formula will be used to derive an algorithm for numerically solving the filtering problem.

6.3 The Zakai Equation

It is possible to use the Kushner-Stratonovitch equation for solving the filtering problem. Another approach is to use the Zakai equation which will be derived here for the model given in the previous section.

First we define an exponential local martingale (Z_t, \mathcal{Y}_t) :

$$\begin{aligned} Z_t &= \exp\left(\int_0^t E[c(s, X_s)|\mathcal{Y}_s]^* dY_s - \frac{1}{2} \int_0^t \|E[c(s, X_s)|\mathcal{Y}_s]\|^2 ds\right) \\ &= \exp\left(\int_0^t E[c(s, X_s)|\mathcal{Y}_s]^* dv_s + \frac{1}{2} \int_0^t \|E[c(s, X_s)|\mathcal{Y}_s]\|^2 ds\right). \end{aligned}$$

By Ito's formula

$$dZ_t = \sum_{i=1}^N Z_t E[c^i(t, X_t)|\mathcal{Y}_t] dv_t^i + Z_t \|E[c(t, X_t)|\mathcal{Y}_t]\|^2 dt. \quad [6.6]$$

Define

$$\pi_t(f) = E[f(X_t)|\mathcal{Y}_t]$$

and

$$\sigma_t(f) = Z_t E[f(X_t)|\mathcal{Y}_t],$$

where $f \in C_b(\mathbf{R}^N)$. Now again by Ito's formula we have

$$\sigma_t(f) = E[f(X_0)] + \int_0^t Z_s d\pi_s(f) + \int_0^t E[f(X_s)|\mathcal{Y}_s] dZ_s + \langle \pi(f), Z \rangle_t.$$

By equation 6.6

$$\int_0^t \pi_s(f) dZ_s = \int_0^t Z_s E[f(X_s)|\mathcal{Y}_s] E[c(s, X_s)|\mathcal{Y}_s]^* dv_s + \int_0^t \|E[c(s, X_s)|\mathcal{Y}_s]\|^2 Z_s E[f(X_s)|\mathcal{Y}_s] ds$$

and by the Kushner-Stratonovitch equation (equation 5.4)

$$\begin{aligned} \int_0^t Z_s d\pi_s(f) &= \int_0^t Z_s E[A_s f | \mathcal{Y}_s] ds - \int_0^t Z_s E[f(X_s)|\mathcal{Y}_s] E[c(s, X_s)|\mathcal{Y}_s]^* dv_s \\ &\quad + \int_0^t Z_s E[f(X_s)c(s, X_s)|\mathcal{Y}_s]^* dv_s. \end{aligned}$$

For the last term we get

$$\begin{aligned} \langle \pi(f), Z \rangle_t &= \sum_{i=1}^N \int_0^t Z_s E[c^i(s, X_s)|\mathcal{Y}_s] E[f(X_s)c^i(s, X_s)|\mathcal{Y}_s] ds \\ &\quad - \sum_{i=1}^N \int_0^t Z_s E[f(X_s)|\mathcal{Y}_s] \|E[c^i(s, X_s)|\mathcal{Y}_s]\|^2 ds. \end{aligned}$$

Recall that

$$v_t = Y_t - \int_0^t E[c(s, X_s)|\mathcal{Y}_s] ds.$$

What we have left after summing the terms is

$$\begin{aligned} \sigma_t(f) &= E[f(X_0)] + \int_0^t Z_s E[A_s f | \mathcal{Y}_s] ds + \int_0^t Z_s E[f(X_s)c^*(s, X_s)|\mathcal{Y}_s] dY_s \\ &= E[f(X_0)] + \int_0^t \sigma_s(A_s f) ds + \int_0^t \sigma_s(f(X_s)c^*(s, X_s)) dY_s. \end{aligned}$$

In practice, the solution of the Zakai equation is slightly easier to compute than that of equation 5.4. Linearity makes it theoretically easier than the Kushner-Stratonovitch equation. Again

uniqueness of solutions to this equation is a difficult question. A powerful tool for obtaining answers to it is the theory of stochastic differential equations in Hilbert spaces. A good reference on this interesting subject is [27].

However, in [13] it is stated that numerical instability may be a problem with the Zakai equation. On the other hand, according to the same paper the Kushner-Stratonovitch equation does not have similar problems. Because the additional work needed for solving the KS-equation is not significant we chose to use it for numerical calculations.

Chapter 7

Numerical Methods for Continuous-time Nonlinear Filtering

We will again examine the model (see chapter 6):

$$X_t = X_0 + \int_0^t a(s, X_s) ds + \int_0^t b(s, X_s) dW_s \quad [7.1]$$

$$Y_t = \int_0^t c(s, X_s) ds + V_t. \quad [7.2]$$

In addition to the Lipschitz and growth conditions we assume that a , b and c are smooth enough when necessary. Often in theory a boundeness assumption is made but in practice this may be too restrictive.

We will first derive a method for time discretization of the filtering problem. After that three numerical methods are presented. One of these is based on solving the Kushner-Stratonovitch equation and the other two are based on time discretization.

7.1 Time Discretization

At this point it is easier to examine the processes in a canonical probability space. For simplicity X_0 is assumed to be constant. Let (W_t) and (Y_t) be independent Brownian motions on some probability space. We define the spaces $(\Omega_W, \mathcal{F}_W, P_W)$ and $(\Omega_Y, \mathcal{F}_Y, P_Y)$ where $\Omega_W = \Omega_Y = C([0, T])$ and $\mathcal{F}_W = \mathcal{F}_Y$ is the Borel σ -algebra w.r.t. the sup-norm. P_W and P_Y are Wiener measures. Next we choose $\Omega = \Omega_W \times \Omega_Y$ and \mathcal{F} as the completion of $\mathcal{F}_T^W \times \mathcal{F}_T^Y$. By defining $\tilde{P} = P_W \times P_Y$ we get the probability space $(\Omega, \mathcal{F}, \tilde{P})$. Each $\omega \in \Omega$ is of the form (ω_W, ω_Y) . The filtration (\mathcal{F}_t) is that generated by (W_t) and (Y_t) augmented with null sets.

We would like to place emphasis on the fact that the observation process (Y_t) is a Brownian motion on this probability space. The basic properties of Brownian motions will be used without warning. For example the independence of increments will be used a lot.

(X_t) is defined as the solution of equation 7.1 on $(\Omega, \mathcal{F}, \tilde{P})$. Then from theorem 6.2 it follows that

$$Z_t = \exp\left(\int_0^t c(s, X_s)^* dY_s - \frac{1}{2} \int_0^t \|c(s, X_s)\|^2 ds\right)$$

is an (\mathcal{F}_t) -martingale and if we define the probability measure P by $dP = Z_T d\tilde{P}$ then

$$V_t = Y_t - \int_0^t c(s, X_s) ds$$

is a Brownian motion independent of (W_t) . Note that we use the same techniques as in section 6.2 but to another direction. For future use, we define

$$L_t^h = \exp\left(\int_t^{t+h} c(s, X_s)^* dY_s - \frac{1}{2} \int_t^{t+h} \|c(s, X_s)\|^2 ds\right).$$

The unnormalized conditional distribution is defined as the conditional measure

$$\pi_t(A) = E_{\tilde{P}}[I_A Z_t | \mathcal{Y}_t].$$

The relation between $E_P[I_A | \mathcal{Y}_t]$ and $\pi_t(A)$ is given by the Kallianpur-Striebel formula (equation 6.2). In practice we only have discrete observations. With this in mind we define

$$\mathcal{Y}_{nh}^h = \sigma(Y_{jh}, 0 \leq j \leq n)$$

and

$$\pi_{nh}^h(A) = E_{\tilde{P}}[I_A Z_{nh} | \mathcal{Y}_{nh}^h].$$

Correspondingly $\pi_{nh}^{h, y_0, \dots, y_{nh}}(A) = E_{\tilde{P}}[I_A Z_t | Y_0 = y_0, \dots, Y_{nh} = y_{nh}]$, where $y_{ih} \in \mathbf{R}^M$ for $i = 0 \dots n$. π_{nh}^h and $\pi_{nh}^{h, y_0, \dots, y_{nh}}$ define also measures on \mathbf{R}^N .

It is possible and in fact not very difficult to prove the recursive formula

$$\begin{aligned} & \int_{\mathbf{R}^N} f(x) \pi_{nh}^{h, y_0, \dots, y_{nh}}(dx) \\ &= \int_{\mathbf{R}^N} E_{\tilde{P}}[f(X_{nh}) L_{(n-1)h}^h | Y_0 = y_0, \dots, Y_{nh} = y_{nh}, X_{(n-1)h} = x] \pi_{(n-1)h}^{h, y_0, \dots, y_{(n-1)h}}(dx). \quad [7.3] \end{aligned}$$

The following theorem says a bit more by specifying the accurate form of the terms.

Theorem 7.1 For any $f \in C_b(\mathbf{R}^N)$,

$$\begin{aligned} \int_{\mathbf{R}^N} f(x) \pi_{nh}^{h, y_0, \dots, y_{nh}}(dx) &= \int_{\mathbf{R}^N} E_{\bar{P}}[f(X_{nh}) \exp(\frac{1}{h}(y_{nh} - y_{(n-1)h})^* \int_{(n-1)h}^{nh} c(s, X_s) ds \\ &\quad - \frac{1}{2h} \|\int_{(n-1)h}^{nh} c(s, X_s) ds\|^2) | X_{(n-1)h} = x] \pi_{(n-1)h}^{h, y_0, \dots, y_{(n-1)h}}(dx). \end{aligned}$$

Proof The proof given here is similar to that in [21]. Choose $\omega_W \in \Omega_W$. Then by a simple manipulation we get

$$\begin{aligned} &E_{\bar{P}}[Z_{nh}(\omega_W, \cdot) | \mathcal{Y}_{nh}^h] \\ &= E_{\bar{P}}[Z_{(n-1)h}(\omega_W, \cdot) | \mathcal{Y}_{(n-1)h}^h] E_{\bar{P}}[\exp(\int_{(n-1)h}^{nh} c(s, X_s)^* dY_s - \frac{1}{2} \int_{(n-1)h}^{nh} \|c(s, X_s)\|^2 ds) (\omega_W, \cdot) | \Delta Y_{nh}], \end{aligned}$$

where $\Delta Y_{nh} = Y_{nh} - Y_{(n-1)h}$. The following equation can be proved by replacing (X_t) by a discrete approximation and using the properties of the normal distribution (remember that (Y_t) is a Brownian motion). However, the proof is lengthy and not very interesting. It will not be given here but we would like to point out that this is nothing new, also in [21] this formula is given without proof.

$$\begin{aligned} &E_{\bar{P}}[\exp(\int_{(n-1)h}^{nh} c(s, X_s)^* dY_s - \frac{1}{2} \int_{(n-1)h}^{nh} \|c(s, X_s)\|^2 ds) (\omega_W, \cdot) | \Delta Y_{nh}] \\ &= \exp(\frac{1}{h} (\int_{(n-1)h}^{nh} c(s, X_s) ds)^* (Y_{nh} - Y_{(n-1)h}) - \frac{1}{2h} \|\int_{(n-1)h}^{nh} c(s, X_s) ds\|^2) (\omega_W, \cdot). \end{aligned}$$

Denote this by $\hat{L}_{(n-1)h}^h$. Choose a positive function $f \in C_b(\mathbf{R}^N)$. From Fubini's Theorem it follows that

$$\begin{aligned} E_{\bar{P}}[f(X_{nh}) Z_{nh} | \mathcal{Y}_{nh}^h] (\omega_Y) &= \int_{\Omega_W} f(X_{nh}(\omega_W)) E_{\bar{P}}[Z_{(n-1)h}(\omega_W, \cdot) | \mathcal{Y}_{(n-1)h}^h] (\omega_Y) \hat{L}_{(n-1)h}^h(\omega_W, \omega_Y) P_W(d\omega_W) \\ &= \int_{\Omega_W} f(X_{nh}(\omega_W)) \hat{Z}_{(n-1)h}^h(\omega_W, \omega_Y) \hat{L}_{(n-1)h}^h(\omega_W, \omega_Y) P_W(d\omega_W), \end{aligned}$$

where $\hat{Z}_{(n-1)h}^h(\omega_W, \omega_Y)$ is a shorthand for $E_{\bar{P}}[Z_{(n-1)h}(\omega_W, \cdot) | \mathcal{Y}_{(n-1)h}^h] (\omega_Y)$. Note that because of the first part of the proof there should be no problems with existence of appropriate versions for the conditional expectations. On the other hand, because (X_t) is a Markov process

$$\begin{aligned} &\int_{\Omega_W} f(X_{nh}(\omega_W)) \hat{Z}_{(n-1)h}^h(\omega_W, \omega_Y) \hat{L}_{(n-1)h}^h(\omega_W, \omega_Y) P_W(d\omega_W) \\ &= \int_{\Omega_W} E_{\bar{P}}[f(X_{nh}) \hat{L}_{(n-1)h}^h(\cdot, \omega_Y) | X_{(n-1)h}] (\omega_W) \hat{Z}_{(n-1)h}^h(\omega_W, \omega_Y) P_W(d\omega_W). \end{aligned}$$

This is in fact what we wanted to prove though it is not entirely obvious. Define

$$g(\omega_W, \omega_Y) = E_{\tilde{P}}[f(X_{nh})\hat{L}_{(n-1)h}^h(\cdot, \omega_Y)|X_{(n-1)h}](\omega_W).$$

It is clear that g may be chosen $\sigma(X_{(n-1)h}) \times \sigma(\Delta Y_{nh})$ -measurable (here $\sigma(X_{(n-1)h})$ is considered as a subset of \mathcal{F}_W and $\sigma(\Delta Y_{nh})$ of \mathcal{F}_Y). Thus we may choose positive functions (s_j) of the form

$$s_j = \sum_{i=0}^j a_{ji} I_{A_{ji} \times B_{ji}},$$

where $A_{ji} \in \sigma(X_{(n-1)h})$, $B_{ji} \in \sigma(\Delta Y_{nh})$ and $s_j \uparrow g$. But for these simple functions, it is clear that for any $C \in \mathcal{Y}_{nh}^h$

$$\begin{aligned} & \int_C \int_{\Omega_W} s_j(\omega_W, \omega_Y) \hat{Z}_{(n-1)h}^h(\omega_W, \omega_Y) P_W(d\omega_W) P_Y(d\omega_Y) \\ &= \int_C \int_{\Omega_W} s_j(\omega_W, \omega_Y) Z_{(n-1)h}^h(\omega_W, \omega_Y) P_W(d\omega_W) P_Y(d\omega_Y) \\ &= \int_C \int_{\Omega_W} s_j(\omega_W, \omega_Y) \pi_{(n-1)h}^{h, \omega_Y}(d\omega_W) P_Y(d\omega_Y) \end{aligned}$$

and now the result follows by taking the limit $j \rightarrow \infty$.

Unfortunately these recursive formulas are in most cases not straightforwardly computable and numerical approximations are necessary. We will use the same simple approximation as in our main references here, [21] and [6]. Let $(X_{nh}, n \geq 0)$ be a Markov chain such that $E[\|X_{nh}^h - X_{nh}\|^2] = O(h)$. This may for example be the Euler approximation to (X_t) . Then we define $X_t^h = X_{nh}^h$ if $nh \leq t < (n+1)h$ which is assumed to be adapted to (\mathcal{F}_t^W) . A natural thing to do now is to define the probability measure

$$dP^h = \exp\left(\int_0^T c(s, X_s^h)^* dY_s - \frac{1}{2} \int_0^T \|c(s, X_s^h)\|^2 ds\right) d\tilde{P}$$

on \mathcal{F}_T (assuming that the exponential local martingale is a martingale). Under this new measure we may write

$$Y_t = \int_0^t c(s, X_s^h) ds + V_t^h,$$

where (V_t^h, \mathcal{F}_t) is a Brownian motion independent of (W_t) . The Kallianpur-Striebel formula is of course valid and the unnormalized conditional measure is denoted by $\tilde{\pi}_t^h$.

Theorem 7.2 We have the following equation:

$$\begin{aligned} & \int_{\mathbf{R}^N} f(x) \tilde{\pi}_{nh}^{h, y_0, \dots, y_{nh}}(dx) \\ &= \int_{\mathbf{R}^N} \int_{\mathbf{R}^N} P_{(n-1)h}^h(u, dx) f(x) \exp(c((n-1)h, u)^*(y_{nh} - y_{(n-1)h}) - \frac{h}{2} \|c((n-1)h, u)\|^2) \tilde{\pi}_{(n-1)h}^{h, y_0, \dots, y_{(n-1)h}}(du), \end{aligned} \quad [7.4]$$

where $P_t^h(u, dx)$ is the transition probability for (X_t^h) .

Proof Note first that

$$E_{\tilde{P}}[f(X_{nh}^h) L_{(n-1)h}^h | \mathcal{Y}_{nh}^h, X_{(n-1)h}^h] = E[f(X_{nh}^h) | X_{(n-1)h}^h] \exp\left(\int_{(n-1)h}^{nh} c(s, X_s^h)^* dY_s - \frac{1}{2} \int_{(n-1)h}^{nh} \|c(s, X_s^h)\|^2 ds\right).$$

Now we may use the recursive formula 7.3.

Error bounds for this approximation can be found in [20] and [26]. This method of discretizing the filtering problem is easy to implement in practice and also has good stability properties. This kind of time discretization leads to a discrete time filtering problem. Thus for an engineer who is mainly interested in applications, learning the abstract theory of continuous time filtering is not essential once he knows the basics of the discrete time theory.

7.2 Interacting Particle Systems Approximations

In this section we will derive a powerful numerical method for solving the filtering problem. Once the reader is familiar with the Kallianpur-Striebel formula, the theory in this section should be quite easy to understand. The approach is simple: first the continuous-time filter is approximated by a discrete-time approximation and after that a branching particle algorithm is used to solve it.

In the previous section we derived an approximative recursive formula for the unnormalized conditional distribution. However, solving it in closed form is not an easy task (it is possible in some special cases, see [32]) and more numerical approximations are needed. We suppose for simplicity that a, b and c do not depend on t . Then the time homogenous Markov chain (X_{nh})

has a transition kernel P which has the approximation $P^h \approx P$, where P^h is the kernel for (X_{nh}^h) . Define the $C(\mathbf{R}^N \times \mathbf{R}^M)$ -function

$$g^h(u, \Delta y) = \exp(c(u)^* \Delta y - \frac{h}{2} \|c(u)\|^2).$$

Let $\mathcal{P}(\mathbf{R}^N)$ be the space of probability measures on \mathbf{R}^N and suppose we have observations y_0, \dots, y_N . The mappings $\Phi_{nh}^{h, y_0, \dots, y_{nh}} : \mathcal{P}(\mathbf{R}^N) \rightarrow \mathcal{P}(\mathbf{R}^N)$ ($0 \leq n \leq N$) are defined as

$$\Phi_{nh}^{h, y_0, \dots, y_{nh}}(\pi)(A) = \frac{\int_{\mathbf{R}^N} P^h(u, A) g^h(u, \Delta y_{nh}) \pi(du)}{\int_{\mathbf{R}^N} g^h(u, \Delta y_{nh}) \pi(du)}. \quad [7.5]$$

If $p_{nh}^{h, y_0, \dots, y_{nh}}$ is the conditional probability measure defined by

$$p_{nh}^{h, y_0, \dots, y_{nh}}(A) = E_P[I_{(X_{nh}^h)^{-1}(A)} | Y_0 = y_0, \dots, Y_{nh} = y_{nh}]$$

then by theorem 7.2

$$\Phi_{nh}^{h, y_0, \dots, y_{nh}}(p_{(n-1)h}^{h, y_0, \dots, y_{(n-1)h}}) = p_{nh}^{h, y_0, \dots, y_{nh}}.$$

Thus the recursion determines the distributions completely and the only problem left is calculating the integrals. This is not easy especially if N is relatively big. We will use a recursive particle algorithm which is by no means the only possible choice.

Let (ζ_{nh}) be an $\mathbf{R}^{N \times L}$ -valued Markov chain defined on some auxiliary probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathcal{P}})$. The columns correspond to individual components of a discrete probability measure. The key idea in particle approximations is at each time step to replace $p_{nh}^{h, y_0, \dots, y_{nh}}$ by the particle approximation

$$\frac{1}{L} \sum_{i=1}^L \delta_{\zeta_{nh}^i}.$$

The transition probability for (ζ_{nh}^i) is defined by

$$P(\zeta_{nh} \in dx | \zeta_{(n-1)h}) = \prod_{j=1}^L (\Phi_{nh}^{h, y_0, \dots, y_{nh}}(\frac{1}{L} \sum_{i=1}^L \delta_{\zeta_{(n-1)h}^i}))(dx^j)$$

as in [6]. This is a mathematical model for the algorithm. Now by substituting into 7.5

$$\Phi_{nh}^{h, y_0, \dots, y_{nh}}(\frac{1}{L} \sum_{i=1}^L \delta_{\zeta_{(n-1)h}^i})(dx) = \frac{\sum_{i=1}^L P(\zeta_{(n-1)h}^i, dx) g^h(\zeta_{(n-1)h}^i, \Delta y_{nh})}{\sum_{i=1}^L g^h(\zeta_{(n-1)h}^i, \Delta y_{nh})}.$$

This gives us the following algorithm:

1. Choose $p_0^L(dx) = \frac{1}{L} \sum_{i=0}^L \delta_{\zeta_0^i}$ and set $n = 0$.

2. Calculate weights

$$w_{nh}^i = \frac{g^h(\zeta_{nh}^i, \Delta y_{(n+1)h})}{\sum_{i=0}^L g^h(\zeta_{nh}^i, \Delta y_{(n+1)h})}$$

3. Sample particles $(z_i)_{i=1}^L$ from (ζ_{nh}^i) according to the weights w_{nh}^i and for $i = 1, \dots, L$ simulate $(\zeta_{(n+1)h}^i)$ from $P(z_i, dx)$.

4. Set

$$p_{(n+1)h}^L(dx) = \frac{1}{L} \sum_{i=1}^L \delta_{\zeta_{(n+1)h}^i}.$$

5. Set $n = n + 1$ and move to step 2.

The basic idea is clear: those particles which are close to the right value get relatively high weights. This algorithm is often called the particle filter and it is a sequential Monte Carlo method. There exist also connections to genetic algorithms.

There exist convergence results for the algorithm. For some of these we refer to [6], where error bounds of order $O(1/L)$ are given for $E[|p_{nh}^L(A) - p_{nh}^{h, y_0, \dots, y_{nh}}(A)|^2]$. At the moment theoretical error bounds are quite loose and not very useful in most practical situations.

7.3 Galerkin's Method

In this section we derive Galerkin's method for solving the filtering problem for the model given by equations 7.1 and 7.2. The derivation is relatively heuristic and no proof of convergence is given. This is because the theory is complicated and in fact still quite undeveloped. Again we suppose that the model is time-homogenous. This is because the inhomogenous case may increase the computational complexity of the algorithm.

Galerkin's method is a standard way of solving deterministic PDEs and it can be generalized straightforwardly to this stochastic problem. More on this method can be found in [9] though the text should be understandable also to those who are not familiar with it beforehand. Important references on Galerkin's method applied to nonlinear filtering are [3] and [1].

Let $\phi_t(x, \omega)$ be the conditional density defined in section 6.1. We chose a different notation because it fits here better. Recall the Kushner-Stratonovitch equation

$$d\phi_t = A^* \phi_t dt + (c^* \phi_t + \phi_t \int_{\mathbf{R}^N} c^* \phi_t dx) dv_t. \quad [7.6]$$

Let $(w_i)_{i=1}^{\infty}$ be a linearly independent dense set of functions in $C^2(\mathbf{R}^N)$ which go to zero in infinity fast enough. Then ϕ_t may be approximated by

$$\tilde{\phi}_t = \sum_{i=1}^l b^i(t) w_i. \quad [7.7]$$

Note that the functions $(w_i)_{i=1}^{\infty}$ are not necessarily orthogonal. The inner product in $L^2(\mathbf{R}^N)$ will be denoted by (\cdot, \cdot) . Next the approximation 7.7 is substituted into 7.6 and then the difference between the right hand side and the left hand side is projected onto the subspace spanned by $(w_i)_{i=1}^l$. The result is set to 0 resulting in the following equation for $j = 1, \dots, l$:

$$(w_j, \phi_t) = \sum_{i=1}^l b^i(t) (w_j, w_i) = (w_j, \phi_0) + \sum_{i=1}^l \int_0^t b^i(s) (w_j, A^* w_i) ds \quad [7.8]$$

$$+ \sum_{k=1}^N \sum_{i=1}^l \int_0^t b^i(s) (w_j, w_i c^k) dv_s^k \quad [7.9]$$

$$+ \int_0^t \left(\sum_{k=1}^l b^k(s) (w_j, w_k) \right) \left(\sum_{i=1}^l \int_{\mathbf{R}^N} b^i(s) w_i c^* dx \right) dv_s. \quad [7.10]$$

For writing this in a matrix form, we define the matrices

$$M^{ij} = (w_i, w_j)$$

$$A^{ij} = (w_i, A^* w_j)$$

$$T_{ij}^{(k)} = (w_i, c^j w_k) \quad (k=1, \dots, l)$$

$$H^{ij} = (w_i, c^j).$$

Then 7.10 can be written as

$$db(t) = M^{-1} A b dt + \sum_{i=1}^l b_i M^{-1} T^{(i)} dv_t + b b^* H dv_t.$$

The only problem left is choosing the basis functions, which is an application specific task. A typical choice is a Gaussian basis. In most practical situations this equation should be used in the Stratonovich form. See [15] for more information on this point.

The main problem of this method is the curse of dimensionality. If the dimension of the state space is over, say four, then the method is in trouble. This is typical to all grid-based methods. In contrary Monte Carlo methods like the particle algorithm presented in the previous section do not suffer from this problem nearly as badly.

The biggest motivation for researching Galerkin's method and grid-based methods in general (for example the finite difference method) is that real-world problems are often low dimensional. Typical examples of fields where these methods would be applicable are mathematical finance and analysis of time series. It seems that there remains a lot of work to do for those who are interested in applications.

7.4 Gaussian Filters

In this section we will use the theory and notation introduced in previous sections. The reader should know the basics of linear regression analysis. For simplicity possible time dependencies in the state-space model are omitted. The algorithm which we will derive is recursive and at each time step results in a Gaussian approximation. The derivation is based on theorem 7.2.

Let $\hat{p}_{nh}^{y_0, \dots, y_{nh}}$ be a Gaussian approximation to the conditional probability density $p_{nh}^{y_0, \dots, y_{nh}}$ defined in section 7.2. It has the following form:

$$\hat{p}_{nh}^{y_0, \dots, y_{nh}}(u) = \frac{1}{\sqrt{(2\pi)^d |C_{nh}|}} \exp\left(-\frac{1}{2}(u - m_{nh})^* C_{nh}^{-1} (u - m_{nh})\right).$$

Recall that

$$\Delta y_{(n+1)h} = y_{(n+1)h} - y_{nh} \approx hc(X_{nh}) + V_{(n+1)h} - V_{nh}. \quad [7.11]$$

We will first examine the so-called updating phase. This phase corresponds to multiplying $\hat{p}_{nh}^{h,y_0,\dots,y_{nh}}$ by the exponential function in formula 7.4. Given the observation y_0, \dots, y_{nh} , the random variable in equation 7.11 can be approximated by

$$\tilde{z} = A(X_{nh} - m_{nh}) + hm_c + B^{1/2}v,$$

where $m_c = \int_{\mathbf{R}^N} c(u) \hat{p}_{nh}^{y_0,\dots,y_{nh}}(u) du$ and v is independent and distributed as $N(0, I)$. A natural way to choose the matrices A and B is to use linear regression. This leads straightforwardly to $A = hP_{xc}^* C_{nh}^{-1}$, where

$$P_{xc} = \int_{\mathbf{R}^N} (u - m_{nh})(c(u) - m_c)^* \hat{p}_{nh}^{y_0,\dots,y_{nh}}(u) du.$$

In addition we define

$$P_{cc} = \int_{\mathbf{R}^N} (c(u) - m_u)(c(u) - m_u)^* \hat{p}_{nh}^{y_0,\dots,y_{nh}}(u) du.$$

The following formula is important:

$$\begin{aligned} & \int_{\Omega} \int_{\mathbf{R}^N} (hc(u) + V_{(n+1)h}(\omega) - V_{nh}(\omega) - A(u - m_{nh}) - hm_c) \\ & \cdot (hc(u) + V_{(n+1)h}(\omega) - V_{nh}(\omega) - A(u - m_{nh}) - hm_c)^* \hat{p}_{nh}^{y_0,\dots,y_{nh}}(u) du dP \\ & = h^2 P_{cc} + hI - h^2 P_{xc}^* C_{nh}^{-1} P_{xc}. \end{aligned}$$

Now the choice of B is easy: by choosing $B = h^2 P_{cc} + hI - h^2 P_{xc}^* C_{nh}^{-1} P_{xc}$ the conditional variances of \tilde{z} and $\Delta y_{(n+1)h}$ given the observations y_0, \dots, y_{nh} are approximately equal, of course depending on how good the estimate $\hat{p}_{nh}^{y_0,\dots,y_{nh}}$ is. B corresponds to linearization error. The noise term is important because if the linearization error hadn't been taken into account, the estimate would be inconsistent (inconsistency means that the covariance of the approximation is smaller than that of the real distribution). In the following calculations we will use $\sqrt{h}B^{-1/2}\tilde{z}$ (the scaling is needed because the noise covariance must be I) for updating $p_{nh}^{y_0,\dots,y_{nh}}$. This means that the update is calculated assuming that $\Delta y_{(n+1)h}$ comes from a linear model which approximates the original one. If c is linear enough in the region where $\hat{p}_{nh}^{y_0,\dots,y_{nh}}$ has its probability mass, then this approximation is good. $\hat{p}_{nh}^{y_0,\dots,y_{nh}}$ is not the true conditional probability density and the regression coefficients are approximations to those obtained by using the real conditional probability

density. The probability masses of both distributions should be concentrated sufficiently close to each other.

The updating phase can be solved analytically for the linearized model. The scaling constants are not important in this case and are omitted. This is because we are only interested in probability densities. $c(u)$ is replaced by $\frac{1}{\sqrt{h}}B^{-1/2}A(u - m_{nh}) + \sqrt{h}B^{-1/2}m_c$ and $\Delta y_{(n+1)h}$ by $\sqrt{h}B^{-1/2}\Delta y_{(n+1)h}$. By an elementary manipulation

$$\begin{aligned} & \exp\left(\left(\frac{1}{\sqrt{h}}B^{-1/2}A(u - m_{nh}) + \sqrt{h}B^{-1/2}m_c\right)^* \sqrt{h}B^{-1/2}\Delta y_{(n+1)h}\right. \\ & \left. - \frac{h}{2}\left\|\frac{1}{\sqrt{h}}B^{-1/2}A(u - m_{nh}) + \sqrt{h}B^{-1/2}m_c\right\|^2\right) \\ & \propto \exp\left((Au)^*B^{-1}\Delta y_{(n+1)h} - \frac{1}{2}(Au)^*B^{-1}Au - (Au)^*(hB^{-1}m_c - B^{-1}Am_{nh})\right). \end{aligned}$$

In the next formulas, P_{yy} stands for $h^2P_{cc} + hI$. By a standard calculation for Gaussian distributions

$$\begin{aligned} & \exp\left(-\frac{1}{2}\begin{bmatrix} \Delta y_{(n+1)h} - hm_c \\ u - m_{nh} \end{bmatrix}^* \begin{bmatrix} P_{yy} & hP_{xc}^* \\ hP_{xc} & C_{nh} \end{bmatrix}^{-1} \begin{bmatrix} \Delta y_{(n+1)h} - hm_c \\ u - m_{nh} \end{bmatrix}\right) \\ & \propto \exp\left(-\frac{1}{2}(\Delta y_{(n+1)h} - hm_c - hP_{xc}^*C_{nh}^{-1}(u - m_{nh}))^*\right. \\ & \quad \left.(P_{yy} - h^2P_{xc}^*C_{nh}^{-1}P_{xc})^{-1}(\Delta y_{(n+1)h} - hm_c - hP_{xc}^*C_{nh}^{-1}(u - m_{nh}))\right) \cdot \\ & \quad \exp\left(-\frac{1}{2}(u - m_{nh})^*C_{nh}^{-1}(u - m_{nh})\right) \\ & \propto \exp\left(-\frac{1}{2}(\Delta y_{(n+1)h} - hm_c - A(u - m_{nh}))^*B^{-1}(\Delta y_{(n+1)h} - hm_c - A(u - m_{nh}))\right) \cdot \\ & \quad \exp\left(-\frac{1}{2}(u - m_{nh})^*C_{nh}^{-1}(u - m_{nh})\right) \\ & \propto \exp\left(-\frac{1}{2}(Au)^*B^{-1}Au + (Au)^*B^{-1}\Delta y_{(n+1)h} - (Au)^*(hB^{-1}m_c - B^{-1}Am_{nh})\right)\hat{P}_{nh}^{y_0, \dots, y_{nh}}. \end{aligned}$$

The formulas for variances and means of conditional Gaussian distributions are well known.

By conditioning on $\Delta y_{(n+1)h}$ it follows that

$$\begin{aligned} & \exp\left(\left(\frac{1}{\sqrt{h}}B^{-1/2}A(u - m_{nh}) + \sqrt{h}B^{-1/2}m_c\right)^* \sqrt{h}B^{-1/2}\Delta y_{(n+1)h}\right. \\ & \left. - \frac{h}{2}\left\|\frac{1}{\sqrt{h}}B^{-1/2}A(u - m_{nh}) + \sqrt{h}B^{-1/2}m_c\right\|^2\right)\hat{P}_{nh}^{y_0, \dots, y_{nh}} \\ & \propto N(m_{nh} - hP_{xc}^*P_{yy}^{-1}(\Delta y_{(n+1)h} - hm_c), C_{nh} - h^2P_{xc}^*P_{yy}^{-1}P_{xc}). \end{aligned}$$

This equation tells us exactly how the variance and mean are updated. We still have the prediction phase left (the integral with respect to the transition kernel $P^h(u, dx)$). The transition kernel is not necessarily linear and this means that additional approximations are necessary. One possibility would be to form a linear model for the dependency between $X_{(n+1)h}^h$ and X_{nh}^h in a similar way as was done in the update phase. This would result in a Gaussian approximation $\hat{P}_{(n+1)h}^{y_0, \dots, y_{(n+1)h}}$ with the same variance and mean as $\int_{\mathbf{R}^N} \hat{P}_{nh}^{y_0, \dots, y_{(n+1)h}}(u) P^h(u, dx) du$, where $\hat{P}_{nh}^{y_0, \dots, y_{(n+1)h}}$ is the approximation after the updating phase. Forming a regression model is not necessary, we may simply define the following rules:

$$\int_{\mathbf{R}^N} u \hat{P}_{(n+1)h}^{y_0, \dots, y_{(n+1)h}}(u) du = \int_{\mathbf{R}^N} \int_{\mathbf{R}^N} x \hat{P}_{nh}^{y_0, \dots, y_{(n+1)h}}(u) P^h(u, dx) du \quad [7.12]$$

and

$$\int_{\mathbf{R}^N} uu^* \hat{P}_{(n+1)h}^{y_0, \dots, y_{(n+1)h}}(u) du = \int_{\mathbf{R}^N} \int_{\mathbf{R}^N} xx^* \hat{P}_{nh}^{y_0, \dots, y_{(n+1)h}}(u) P^h(u, dx) du \quad [7.13]$$

For the implementation of the filter, the matrices P_{cc} , P_{xc} and the expectations in equations 7.12 and 7.13 must be calculated at each time step. There are two possibilities to do this. The first is to calculate analytically, which is often possible. The problem with this approach is that the calculations are often quite lengthy depending on the dimension of the state space and the discretization method. But the benefit is also big because analytic calculations are fast and accurate. In the simulation part we will illustrate this by a comparison study. Even though the analytic approach is well-known, surprisingly often numerical methods are used instead.

Another possibility is to use numerical methods like the Gauss-Hermite quadrature, central differences or the filter of Juhlier-Uhlmann. More on these standard techniques can be found in [14]. In lower dimensions the Gauss-Hermite quadrature is a very powerful method.

It may also be possible to combine the two approaches. Often it is easy to simplify the numerical calculations by using the properties of the normal distribution. Characteristic functions and conditionalizing are very useful tools in this regard.

Filters using the technique introduced in this section will be called (Gaussian) linear-regression filters (LRF). They can also be derived by making the assumption that certain densities are Gaussian, see for example [14]. If P_{cc} , P_{xc} and equations 7.12 and 7.13 are solved

by using the first order Taylor approximation, then the resulting filter is called the Extended Kalman filter (EKF). An example of this kind of approximation is

$$\int_{\mathbf{R}^N} c(u) \hat{P}_{nh}^{y_0, \dots, y_{nh}}(u) du \approx c(m_{nh}),$$

which follows from linearizing c at the point m_{nh} . The EKF is easy to implement and widely used, but may be inaccurate. The EKF approximation also tends to be inconsistent. It has been investigated quite a lot and has connections with other stochastic algorithms. For example in parameter estimation problems it corresponds to a modified Newton method (see [25]), where a crude first order approximation to the Hessian of the cost function is made. In the experimental part by linear-regression filter we mean filters which calculate the integrals with high accuracy.

The error in the linearization phase can be estimated straightforwardly. If $\Delta y_{(n+1)h}$ is a scalar, then the residual sum of squares is given by

$$SSE = P_{yy} - h^2 P_{xy}^* C^{-1} P_{xy}$$

and the total sum of squares

$$S_{yy} = P_{yy}.$$

The R-squared value can be calculated from SSE and S_{yy} :

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}} \quad [7.14]$$

These values give some advice in the performance of the filter.

Deriving error bounds for the error of the Gaussian approximation is not an easy thing. Some work on this subject has been done in [14]. There still remains a lot of work to do on this subject.

Here we have adopted a linear regression approach because it gives insight into the properties of the method. A more general method is to apply information geometry for projecting densities onto a finite dimensional manifold. See for example [4] on this subject. Choosing the finite dimensional manifold is not easy. The family of Gaussian distributions which we used here has the advantage that numerical integration can often be avoided.

Another way to improve the linear-regression filter is to use the expectation propagation algorithm. An important and easy to read work on this subject is [24]. This algorithm is relatively new but will certainly gain popularity in the near future. It is based on the assumed density filtering algorithm which is more accurate than the linear-regression approach but is computationally more difficult to implement. The linear-regression filter can be considered as an approximation to the Gaussian assumed-density filter. If the observation equation is linear, then these two filters coincide.

Chapter 8

Applications

8.1 Simulations with Galerkin's Method

In this section we study a simple phase tracking problem. We will study Galerkin's method, Gaussian filters and the particle filter. The main point is to test Galerkin's method on a simple problem and assess its performance. This method for continuous-time problem has been very rarely used in applications.

The problem in a state-space form is

$$X_t = W_t \tag{8.1}$$

$$y_t = \sin(w_0 t + X_t)/r + v_t,$$

where v_t is white noise, W_t is a Brownian motion and $r > 0$ a constant. The task is to calculate the conditional mean of the state given the observation process (Y_t) . This kind of problems are for example studied in [31], where a Gaussian sum filter is used to approximate the optimal filter. Though this problem is simple, it is important because a wide variety of signal processing problems like demodulation and frequency tracking have much in common with this problem and similar ideas can be applied to them as well. A review on frequency estimation and tracking problems can be found in [19].

The simulations are divided into two parts, in the first one the frequency w_0 is known and in the second one we only have an apriori estimate.

8.1.1 Phase Tracking with a Known Frequency

In this section we assume that w_0 is known and θ_0 is distributed as $N(0, C^2)$, where C is the s.d. of the normal distribution. The observation process is given by

$$Y_t = \int_0^t c(s, X_s) ds + V_t,$$

where (V_t) is a Brownian motion and $h(s, x) = \sin(w_0 s + x)/r$ is the measurement function. The infinitesimal generator A for (X_s) is given by

$$A\phi = \frac{1}{2} \frac{d^2 \phi}{dx^2}.$$

We chose $C = 0.1$ and $w_0 = 3$. In section 7.3 we derived formulas for calculating the weights for a Galerkin approximation to the optimal filter. In this example the observation function depends on time. This implies that the matrices H and $T^{(q)}$ depend on time and must be recalculated at every time step. However, all matrices can be calculated analytically and this poses no problems.

For implementing Galerkin's method a basis $(w_i)_{i=0}^N$ must be chosen. We chose

$$w_i = \exp\left(-\frac{1}{2\sigma^2}(x - c_i)^2\right),$$

where $c_0 = -5$ and $c_i = -5 + i\Delta c$. The number of Gaussian functions was chosen so that the last center point is at 5. We made the choice $\Delta c = 0.1$. Now the implementation is straightforward, just form the matrices and solve the resulting stochastic differential equation. We have the following familiar (formal) equation for the weight vector:

$$dw(t) = M^{-1}Aw(t) dt + \left(\sum_{l=1}^N w_l(t)M^{-1}(T^{(l)})_t - w(t)w(t)'H_t \right) (y_t - H_t'w(t)) dt.$$

We have added subscripts to those matrices which depend on time. This equation was discretized at time points $0, h, \dots, T$, where $h = 0.0015$ is the discretization interval. The discretization was done as follows ($t_i \leq t \leq t_{i+1}$):

$$w'(t) = M^{-1}Aw(t) + \left(\sum_{l=1}^N w_l(t_i)M^{-1}(T^{(l)})_{t_i} - w(t_i)w(t_i)'H_{t_i} \right) (y_{t_i} - H_{t_i}'w(t_i)).$$

r	Pfilt., e_1	EKF, e_1	LRF, e_1	Gal., e_1
0.3	0.801	1.817	1.308	0.823
0.5	1.435	1.561	1.400	1.408
0.7	2.360	2.310	2.090	2.200
0.9	2.130	2.390	1.887	1.924
1.1	2.707	3.288	2.600	2.623

Table 8.1 Results of the phase tracking experiment with constant frequency.

This is a linear differential equation with the initial condition $w(t_i)$ at every interval $[t_i, t_{i+1}]$. After this approximation the weights can be solved analytically. The weights were normalized at every time step.

Because the state variable changes quite quickly the grid of Gaussian functions was moved to its center point every 0.05 seconds. This can be done simply by choosing the component with the highest weight and moving the grid so that this component is in the middle.

As a comparison we implemented the EKF and the particle filter with 15000 particles. These are standard tools and the implementation was easy. In addition to these we implemented the linear-regression filter.

The following error measure was calculated:

$$e_1 = \frac{1}{NM} \sum_{j=1}^M \sum_{i=1}^N (\hat{X}_{ih}^j - X_{ih}^j)^2, \quad [8.2]$$

where \hat{X}_{ih}^j is the estimate (conditional mean) and X_{ih}^j is the real state at simulation j . The results are in table 8.1. The number of simulation for each value of r , M , was 200.

The error measure e_1 gives information about the optimality of the filters. The optimal estimate minimizes e_1 . There is randomness in the results, but the big picture should be clear.

It is clear that Galerkin's method performed better than the EKF in this example. This is due to the fact that even though the model for the hidden states is linear, the nonlinear observation equation makes the optimal solution non-Gaussian. Because the dimension of the state-space is low, Galerkin's method is accurate.

The Gaussian linear-regression filter was tested because it is of practical interest in frequency and phase tracking. Typically the EKF or the Gaussian mixture filter have been used in this kind of problems. It is interesting that these problems are often such that the linear-regression filter can be implemented with the same computational complexity as the EKF. For example in this case expectations with respect to Gaussian distributions were calculated using the formula $\sin(z) = \frac{1}{2i}(\exp(iz) - \exp(-iz))$ and the characteristic function of the normal distribution. This analytic approach is promising as the results show, but as will be seen, linear regression is not necessarily a good choice. To our knowledge the exact linear-regression filter has not been applied before to this class of problems.

From table 8.1 we see that the particle filter performed better than the EKF but worse than Galerkin's method in the sense of e_1 . The results with a high noise level are not very good though. It seems that for $r > 0.5$ the measurements are not very informative. With more sophisticated methods or more particles it would probably have performed better.

Galerkin's method offers an alternative to other filtering methods in phase tracking. In this one dimensional case it can be implemented online because all calculations in the implementation are easy to do analytically.

8.1.2 Phase Tracking with an Unknown Frequency

In this section we consider the same problem as in the previous section, but now the frequency w_0 is an unknown constant. We also modify the state-space model :

$$X_t = 0.3W_t.$$

The a priori distribution for (w_0, θ_0) is $N(\hat{X}_0, C)$, where C is a diagonal covariance matrix with $C_{11} = 0.4^2$, $C_{22} = \pi^2/16$ and $\hat{X}_0 = (3.37, 0.5)$. The real initial condition was $X_0 = (2.97, 0.5)$.

As before, for the implementation of Galerkin's method we needed to choose a basis $(w_i)_{i=0}^N$. We again chose a Gaussian basis with means c_i and diagonal covariance matrices σ . The means were placed on a uniform grid on the interval $[2.5, 4.5] \times [-\pi, \pi]$. The distances between two points on the grid on x- and y-directions were $\Delta c_x = 0.1$ and $\Delta c_y = 0.2\pi$. For the variances we chose $\sigma_{11} = 0.1$ and $\sigma_{22} = 0.2\pi$. The implementation was otherwise similar to that in the previous section except that now the grid was not moved during the simulation.

We also implemented the EKF and the particle filter with 20000 particles. The problem here is that the frequency is constant and therefore sample impoverishment is a problem. We decided to use the simple method called roughening introduced in [10]. In this method, before resampling a Gaussian disturbance is added to each particle in the population. In this case only the frequency parameter was perturbed. The s.d. of the perturbation was chosen to be $KEN^{-1/2}$, where E is the length of the interval between the maximum and minimum frequencies in the population and N is the number of particles. K was chosen to be a matrix with $K_{11} = 0.05$ and $K_{22} = 0$.

The step size for time discretization of the filters was 0.001. To compare the filters, the error measures

$$e_1 = \frac{1}{NM} \sum_{j=1}^M \sum_{i=1}^N \|(\hat{X}_{ih}^1)^{(j)} - (X_{ih}^1)^{(j)}\|^2$$

and

$$e_2 = \frac{1}{NM} \sum_{j=1}^M \sum_{i=1}^N \|(\hat{X}_{ih}^2)^{(j)} - (X_{ih}^2)^{(j)}\|^2$$

r	Galerkin, e_1	Galerkin, e_2	EKF, e_1	EKF, e_2	Pfilt., e_1	Pfilt., e_2
0.4	0.090	0.319	0.089	0.329	0.113	0.387
0.5	0.096	0.280	0.099	0.319	0.108	0.327
0.6	0.106	0.340	0.110	0.355	0.116	0.479

Table 8.2 Results of the phase tracking experiment with unknown frequency.

were calculated. We chose $M = 50$. The results of the experiment are in table 8.2. Galerkin's method worked otherwise well but numerical instability was a problem. This is probably because the method which was used for solving (w_t) was not very sophisticated. Probably a good idea would be to use an operator splitting approach for discretizing the Kushner-Stratonovitch equation which would lead to different equations. More on this subject can be found in [13].

Again there is some randomness in the results. The error values have in fact decreased between $r = 0.4$ and $r = 0.5$. All filters were tested with same realizations, so this setting still gives information about the performance of the filters.

The results of Galerkin's method were better than those of the EKF. This was of course expected, as the EKF works well only with low noise covariances. The bad performance of the particle filter came as a slight surprise. The bad result is due to the suboptimal roughening method. A better and more complicated approach could be to use the EM-algorithm for estimating the frequency. More on using Monte Carlo methods for parameter estimation can be found in [8].

In this two dimensional case choosing centers for the basis functions was already not easy. If initial uncertainty is high, then high accuracy is not possible without adapting the grid somehow. This is because the conditional density becomes more peaked when lots of observations are available. A desired solution would be such that at each time step the basis functions would lie on an estimate of the region where the conditional density is significantly above zero. A possible approach, if the one we used in the previous example is not good enough, would be to use an orthogonal basis. In that case forming the matrix M and its inverse would be simple. In [11] a Fourier basis was used resulting in an efficient implementation. Maybe a good idea

would be to discretize the continuous-time problem as was done in section 7.1 and use the ideas presented in that paper. Fourier transform techniques are quite efficient and have been used also in [22]. Also Hermite polynomials could be considered, but then the implementation would be more complicated.

8.1.3 Conclusions

The results clearly show that the performance of Galerkin's method is good compared to other methods. In practice time discretization of the Kushner-Stratonovitch equation should be done using for example operator splitting methods ([13]). This should not pose additional problems as the methods are quite simple to use.

Maybe the biggest problem in applying Galerkin's method is that its implementation may be quite difficult. The basic method can of course be programmed quite easily, but to make it work properly is not always easy. The first difficulty is choosing basis functions. If the state-space is not compact it may be necessary to use a time varying the basis. Otherwise the conditional density moves away from the region where there are basis functions or becomes too peaked. This means that quite a lot of computation needs to be done and it may be necessary to be able to do it analytically. Possible solutions to this problem were described in the previous section.

Another problem of Galerkin's method is its computational complexity, which is $O(N^2)$, where N is the number of basis functions used. If high accuracy is needed, the amount of computation required grows fast. This problem becomes worse once the dimension of the state space grows. However, it is often the case that using Monte Carlo methods like the particle filter lots of particles are needed to achieve similar accuracy.

When the dimension of the state-space is under four, Galerkin's method might be a good choice. However, its usefulness in online state estimation is questionable. This is because of the computational complexity, it is often possible to implement faster and accurate enough methods. If this is not a problem (calculations are made offline for example) and high accuracy is needed, then Galerkin's method might be a good choice.

Before deciding to use Galerkin's method one should consider using grid-based methods like numerical quadratures or finite differences. A grid of points is much easier to set up and modify than for example a Gaussian basis. The implementation is in general easy because no analytic or numerical integration is needed. The curse of dimensionality is of course a problem here also. More information about these methods in a discrete-time setting can be found in [8]. Numerical quadrature based methods are especially interesting because they offer high accuracy.

Other methods also gave interesting results. The results of the particle filter were surprisingly bad, but one could argue that the implementation was not really very efficient. Still it is clear that in many low dimensional problems the particle filter is not the best choice.

The linear-regression filter gave relatively good results without additional computational cost in a problem where traditionally the EKF has been used. This is not a surprise as the linear-regression filter uses a much more accurate linearization method.

8.2 Simulations with Gaussian Filters

In this sections we will test Gaussian filters on two experiments. The results are compared with those of the particle filter. The main point is finding strengths and weaknesses of the filters. We also want to highlight the fact that very often at least part of the integration necessary for implementing a Gaussian linear-regression filter can be done analytically. This point is interesting taking into account the fact that the EKF and the unscented Kalman filter (the unscented Kalman filter is a popular method, see for example [12]) are so commonly used. This analytic approach is applied to the problem of frequency tracking where, as in the phase tracking problem, traditionally methods like the EKF and UKF have been used. As in the phase tracking problem, the possible improvement in performance comes with no additional cost. But we also show that this improvement may often be questionable.

8.2.1 Vertically Falling Body

In this section we will examine the problem of estimating the velocity, altitude and ballistic coefficients of a vertically falling body. This is a well-known problem, see for example [14] and [2]. In [14] this problem was used as a benchmark problem for Gaussian quadrature filters. The techniques used in this section can be applied straightforwardly for example to target tracking with range-only measurements.

The state variables are the altitude (x_t^1), the velocity (x_t^2) and the ballistic coefficient (x_t^3) of the body. The state equations are

$$dx_t^1 = -x_t^2 dt, \quad [8.3]$$

$$dx_t^2 = -\exp(-gx_t^1)(x_t^2)^2 x_t^3 dt \quad [8.4]$$

and

$$\frac{dx_t^3}{dt} = 0. \quad [8.5]$$

More on the physical meaning of these equations can be found in [2]. Equations 8.3, 8.4 and 8.5 are far from linear, which makes the problem challenging. Next we define the measurement model. We assume that the radar is at the point (M, H) and the measurement signal is continuous. Then we have

$$Y_t = \frac{1}{R} \int_0^t h(X_s) ds + V_t, \quad [8.6]$$

where

$$h(x) = \sqrt{M^2 + (x^1 - H)^2}.$$

We chose the parameters $g = 5 \cdot 10^{-5}$, $H = 10^5$, $R = 80$ and $M = 10^5$.

For state estimation we implemented the EKF, the linear-regression filter and the particle filter (with 15000 particles). The Gaussian filters could be implemented without problems. Denote by p_t a Gaussian approximation to the density $p(x|\mathcal{Y}_t)$. The integrals

$$\int_{\mathbf{R}} h(x^1) p_t(x^1) dx^1$$

and

$$\int_{\mathbf{R}} x^1 h(x^1) p_t(x^1) dx^1$$

were calculated using the Gauss-Hermite quadrature rule with 5 points. The integrals

$\int_{\mathbf{R}} \int_{\mathbf{R}} x^i h(x^1) p_i(x^1, x^i) dx^1 dx^i$ ($i = 2, 3$) were calculated by writing

$$\int_{\mathbf{R}} \int_{\mathbf{R}} x^i h(x^1) p_i(x^1, x^i) dx^1 dx^i = \int_{\mathbf{R}} \int_{\mathbf{R}} x^i p_i(x^i | x^1) dx^i h(x^1) p_i(x^1) dx^1.$$

The latter integral can be calculated by integrating out first the inner integral analytically and then using numerical integration in one dimension. The rest of the integrals necessary for implementation of the linear-regression filter were calculated analytically using the characteristic function of the normal distribution and its derivatives.

Discretization of the continuous-time problem was done with step size 0.005. For the Gaussian filters we used the Euler method and for the particle filter an accurate Runge-Kutta discretization.

Because the model for the hidden states is deterministic the particle filter can't be used as such because of the degeneracy problem, unless lots of particles are used. To avoid this problem the resampling phase was modified. Denote by (ζ_i, w_i) a set of N particles and their corresponding weights. In the ordinary branching particle algorithm at the update phase the discrete distribution

$$\sum_{i=1}^K w_i \delta_{\zeta_i} \tag{8.7}$$

is sampled. This resampling operation is now modified so that instead of 8.7 the samples are taken from the distribution

$$\sum_{i=1}^K w_i N(\zeta_i, B_K),$$

where K is the number of particles and $N(\zeta_i, B_K)$ is the density function of the normal distribution with covariance B_K and mean ζ_i . Sampling from this distribution is easy: just draw according to the weights and add noise. The choice of B_K is of course critical. As proposed in [8] we choose

$$B_K = \left(\frac{4}{(n_x + 2)K} \right)^{1/(n_x+4)} E d,$$

where E is the covariance matrix of the discrete distribution 8.7 and in this case $n_x = 3$. We also made the choice $d = 1/50$.

To evaluate the performance of the filters, the root mean square error (RMS) was evaluated over 50 simulations at each time point. This was done by calculating the mean square error at each time point and taking the square root of it. The initial condition for the state process was $X_0 = (3 \cdot 10^5, 2 \cdot 10^4, 10^{-3})$. The initial distribution for the filtering processes was $N(\hat{X}_0, C)$, where C is diagonal and $\hat{X}_0 = (3 \cdot 10^5, 2 \cdot 10^4, 3 \cdot 10^{-5})$. The diagonal of the covariance matrix is given by $C_{11} = 10^6$, $C_{22} = 4 \cdot 10^6$ and $C_{33} = 10^{-4}$. The results are plotted in figure 8.1. In figure 8.2 we have plotted one realization of the filtering processes.

The results show clearly the superior performance of the linear-regression filter over the EKF. This is due to the strong nonlinearity of the state-space model which can be seen from figure 8.2. An especially difficult moment for all methods is the point where the velocity starts to fall quickly. $(x^1 - H)^2$ is relatively small around $t = 10$ and this means that the noise makes it difficult to estimate the state. The error decreases after the object has passed the critical point and differences between the filters grow.

The EKF is able to perform decently and does not diverge. However, as expected the regression approximation is more accurate than the one based on simple first order approximation (although there is no mathematical proof that this should always be the case). The error of the EKF is mainly due to the prediction phase as the mean and variance should not be updated by such a crude way. As will be seen in the next section, when covariances of the estimates are too big problems may arise. In this example this was not the case.

The particle filter performed fine after the critical point. Estimation of the ballistic coefficient caused difficulties but on the other hand the particle filter was superior to the other filters after $t = 10$. It seems that if the model is deterministic, the choice of resampling method should be done carefully. In this case the method was probably not efficient enough and finding good parameter values was not easy. More efficient methods can be found in [8]. The results from the particle filter still show that the linear-regression filter is not quite optimal.

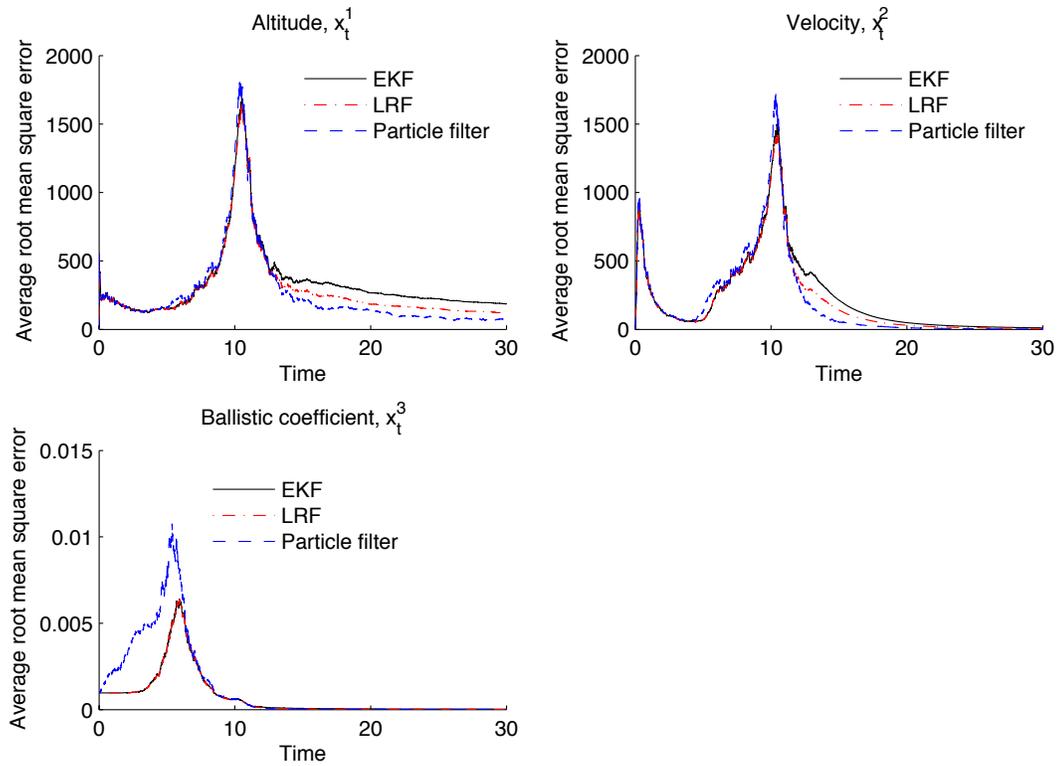


Figure 8.1 RMS of the filters calculated over 50 simulations.

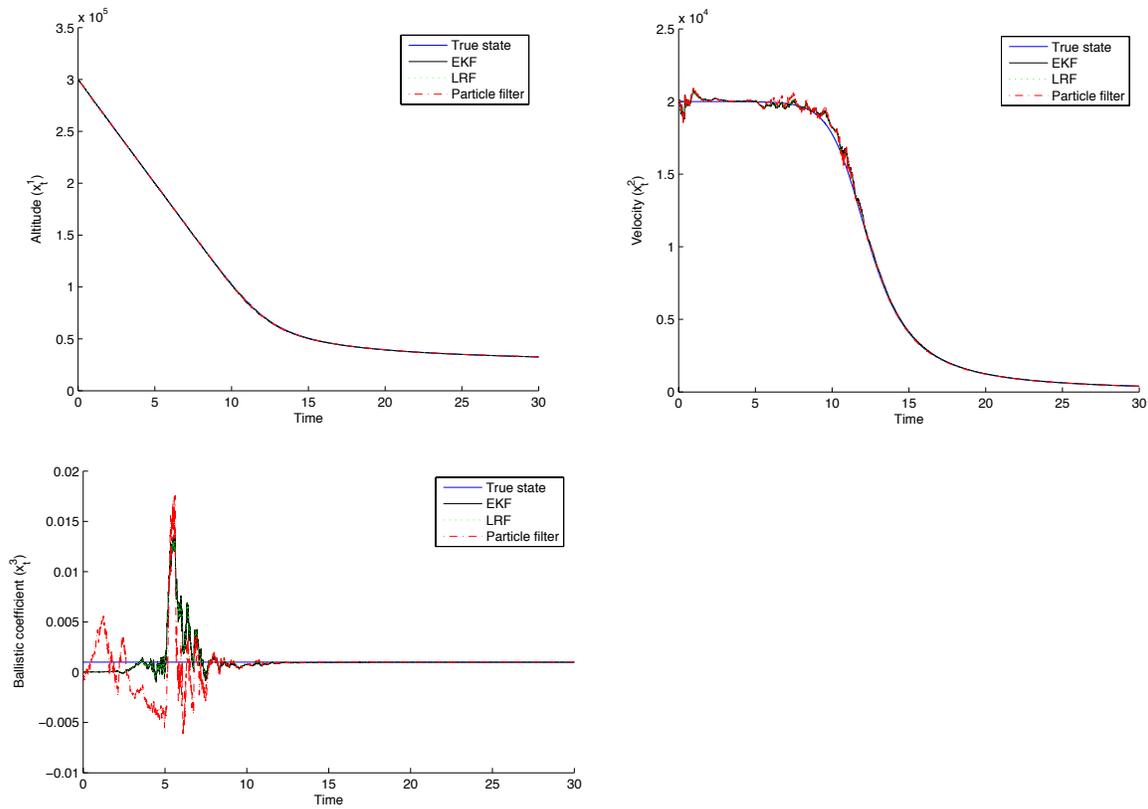


Figure 8.2 A realization of the filtering processes.

It can be concluded that when the problem in question has significant nonlinearities but a Gaussian distribution is still able to approximate the conditional distribution the linear-regression filter may be a good choice. One must also take into account the level of uncertainty. High observation noise leads to high covariances in the Gaussian approximations. The computational cost depends on the dimension of the state space and the numerical integration method which is used to approximate the integrals. The Gauss-Hermite quadrature is a powerful method for calculating integrals numerically when necessary, especially in low dimensional spaces. An additional advantage of quadrature based filters over the EKF is that no derivatives are needed.

In this example we wanted to demonstrate the advantages of analytic calculations: by analytic calculation we were able to avoid numerical integration in three dimensions. This makes the algorithm much faster and more accurate. By numerical integration, the amount of computation needed for the calculations is much higher. In general there exist numerous situations where analytic calculations are possible but the EKF or another numerical approximation is used. In most cases where the model equations are available it is possible to calculate at least part of the integrals for the linear-regression filter analytically and simplify the rest. This is an interesting point and even though it is quite elementary, it is often the case that this fact is not exploited. Typical examples where analytic calculations are helpful are orbit determination (see the simulations in [15]) and training of RBF networks (see [12]). It seems that in many cases vast improvements over the traditional EKF would be easily obtainable.

8.2.2 Frequency Tracking

In this section we will examine the problem of frequency tracking in the case of multiple sinusoidal signals (in this example two). This problem is interesting, because frequency tracking, phase estimation and demodulation are typical problems which occur in practice and have been investigated a lot. The frequency tracking problem is harder to solve than the phase tracking problem in section 8.1, because the dimensionality of the state space is higher and depends

on the number of sinusoidal signals. Otherwise the problems are close to each other and for example the implementation of the linear-regression filter is also here easy.

The state vector is $X_t = (x_t^1, \dots, x_t^6)$. The components x_t^1 , x_t^2 and x_t^3 correspond to the first sin-signal and the rest to the second. The model for the hidden states is linear:

$$dX_t = AX_t dt + B dW_t.$$

The matrix A is given by

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0.05 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.4 \end{bmatrix}.$$

The true initial condition for the simulations was $x_0 = (1, 0.4, 5, 2, 1, 3)$. The function h in the observation model 8.6 is defined as

$$h(x) = x_1 \sin(x_2) + x_4 \sin(x_5).$$

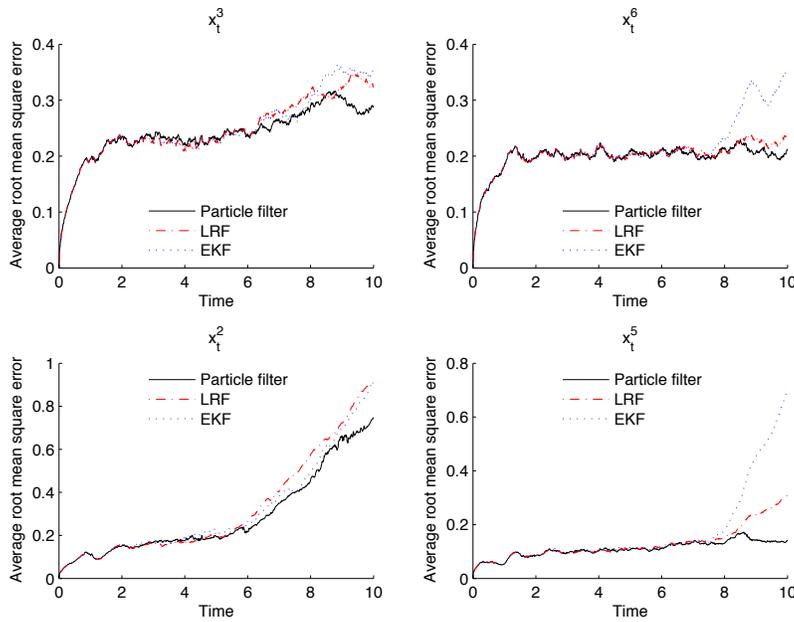


Figure 8.3 Results of experiment 1, RMS of the filters over 1000 simulations.

The a priori state estimate was chosen to be Gaussian with the mean $(1.1, 0.4, 5, 2.1, 1, 3)$ and covariance C_0 . In the first simulation C_0 was chosen to be

$$C_0 = \begin{bmatrix} 0.1^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1^2 \end{bmatrix}.$$

For the noise parameter r we chose the value $1/25$. Time discretization was done with the step size 0.01 . In figure 8.3 we have plotted the RMS over 300 simulations at each time step. The particle filter was implemented with 15000 particles. There was no need for numerical integration in the implementation of the linear-regression filter.

The particle filter should be quite close to optimality. There exist lots of possible improvement strategies. From these we would like to point out Rao-Blackwellisation for calculating the

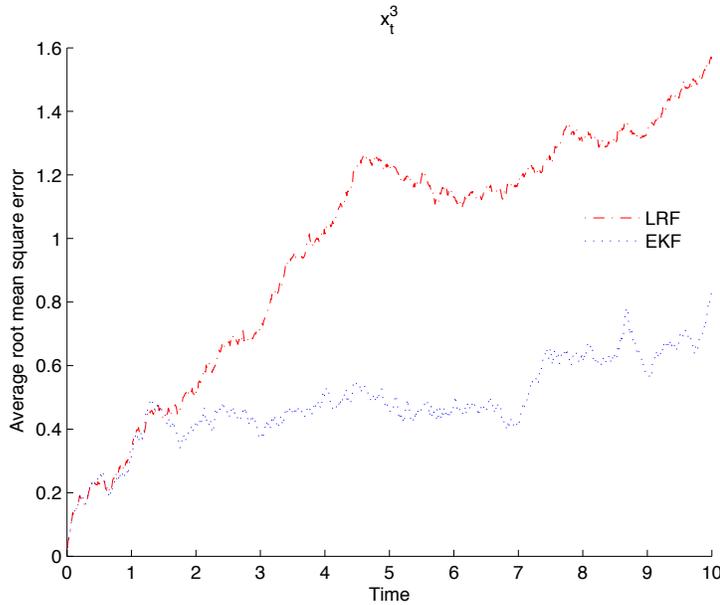


Figure 8.4 Experiment 2, an example of a situation where the linear-regression filter fails.

amplitudes using the Kalman filter. See [5] for more information on this method. The particle filter is a promising approach to this problem as it brings vast improvement over the traditional EKF and there are lots of possibilities for a more efficient implementation.

The linear-regression filter was slightly worse than the EKF at estimating the component x_t^2 but clearly better at estimating the other components. Both filters gave bad estimates towards the end. The extended Kalman filter seems to diverge faster. It is well-known that the stability properties of the EKF are not always very good. In general the linear-regression filter is maybe a bit better on this regard.

Next we will modify the initial condition by setting $C(2, 2) = \pi^2$ and $C(5, 5) = \pi^2$. In addition we chose $r = 1/20$. In figure 8.4 we have plotted the RMS over 30 simulations for this initial condition. The particle filter was left out because we wanted to examine the Gaussian filters only.

Now the linear-regression filter performed very badly. The average value of R^2 (equation 7.14) over 30 simulations with these parameter values was not significantly above 0. The corresponding value calculated with the parameters which were used in the first simulation was

0.101. Both values are low and linear approximations are not very good. In the former case the linear regression approximation is so bad that it is totally useless. This signifies that the noise term in the linearized model is set to a high value and the filter doesn't go anywhere. The EKF on the other hand uses only local information and doesn't suffer from the same problem. In fact in parameter estimation problems the EKF is a modified Newton algorithm (see [25]). The point is that the EKF does not approximate the original model but instead performs optimization steps to minimize a cost function. Without going into details, it is clear that this kind of approximation does often produce better results than linear regression even though the resulting distribution does not necessarily have much to do with the real conditional distribution.

The second experiment was done with parameter values which lead to a total failure of the linear-regression filter. Situations where the EKF was better, but by not as big a margin were also common in our experiments. Thus it is not clear what can be achieved by using the linear-regression filter as such in this problem.

To avoid situations like the one in the second simulation, the linear-regression filter should definitely be combined with a more sophisticated method. One could consider using unscented particle filters or Gaussian mixtures (see [14] and [12]) and replace the UKF or EKF based algorithm by one based on accurate linear regression. The key point is of course that the variance of Gaussian components should stay small enough so that divergence does not happen. The first experiment shows that improvements over the EKF should be possible.

8.2.3 Conclusions

From the simulations it is clear that the EKF is often far from optimal. By using linear-regression filters it may be possible to improve its performance significantly. The computational cost depends on the dimensionality of the state space and on the accuracy of numerical integration. In most cases it is possible to calculate all or some of the integrals analytically and simplify the rest. We demonstrated this in our examples, especially the application to frequency tracking is interesting as the EKF has been a standard tool in frequency and phase tracking.

When the equation for the hidden states is linear and the measurement equation is not too nonlinear, the advantage of the linear-regression filter is not obvious. In section 8.2.2 we demonstrated that the EKF may even perform much better in these kind of situations. Typically the linear-regression filter brings vast improvement over the EKF when the level of noise is not too high and the model is significantly nonlinear. The sizes of the covariances of the Gaussian approximations are of crucial importance: if the uncertainty is too high, then the linear approximation may be bad. The performance of the filter can be evaluated heuristically using standard techniques from regression analysis.

The branching particle algorithm performed well in both examples. Using more sophisticated methods and variance reduction techniques would have improved its performance. It is clear that the basic version of the algorithm is not efficient enough for many problems. For practical applications the computational complexity may be too high and the use of Gaussian filters should at least be considered. In many problems the accuracy obtained with these is good enough with a low computational cost. An additional advantage of Gaussian filters is that they are easy to use and understand. The particle filter is useful as a benchmark for assessing other filters because with enough particles it is nearly optimal.

An interesting approach which was proposed is combining the linear-regression filtering technique with some other filtering technique. This idea has been rarely used. But consider for example the unscented particle filter which puts the unscented Kalman filter and particle filter together. In a wide variety of problems combining the linear-regression filter and the particle filter would be just as easy as implementing the unscented particle filter.

LIST OF REFERENCES

- [1] N. U. Ahmed and S. M. Radaideh. A powerful numerical technique solving zakai equation for nonlinear filtering. *Dynamics and Control*, 7(3):293–308, 1997.
- [2] M. Athans, R. P. Wishner, and A. Bertolini. Suboptimal state estimation for continuous-time nonlinear systems from discrete noise measurements. *IEEE Transaction on Automatic Control*, 13(5):504–514, 1968.
- [3] R. Beard, J. Gunther, T. Oliphant, J. Wilson, and W. Stirling. Galerkin approximations of the kushner equation in nonlinear estimation, 1997.
- [4] D. Brigo, B. Hanzon, and F. Le Gland. A differential geometric approach to nonlinear filtering: the projection filter. Technical report, INRIA-Rennes, 1995.
- [5] G. Casella and C. P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.
- [6] D. Crisan. Interacting particle systems approximations of the kushner-stratonovitch equation. *Adv. in Appl. Probab.*, 31(3):819–838, 1999.
- [7] J. L. Doob. *Classical Potential Theory and Its Probabilistic Counterpart*. A Series of Comprehensive Studies in Mathematics. Springer-Verlag, 1984.
- [8] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Sciences. Springer, 2001.
- [9] C. A. J. Fletcher. *Computational Galerkin Methods*. Springer, 1984.
- [10] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE-Proceedings-F*, 140(2):107–113, 1993.
- [11] J. Gunther, R. Beard, J. Wilson, T. Oliphant, and W. Stirling. Fast nonlinear filtering via galerkin’s method. In *American Control Conference*, 1997.
- [12] S. Haykin, editor. *Kalman Filtering and Neural Networks*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley Sons, 2001.

- [13] K. Ito and B. Rozovskii. Approximation of the kushner equation for nonlinear filtering. *SIAM Journal on Control and Optimization*, 38:893–915, 2000.
- [14] K. Ito and K. Q. Xiong. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927, 2000.
- [15] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, 1970.
- [16] G. Kallianpur. *Stochastic Filtering Theory*. Number 13 in Applications of Mathematics. Springer-Verlag, 1980.
- [17] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82(Series D):35–45, 1960.
- [18] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Number 23 in Applications of Mathematics. Springer-Verlag, 1992.
- [19] P. J. Kootsookos. A review of the frequency estimation and tracking problems. Technical report, The University of Queensland, 1993.
- [20] H. Korezlioglu and G. Maziotto. Modelization and filtering of discrete systems and discrete approximation of continuous systems. In *VI Conférence INRIA, Modélisation et Optimisation des Systèmes*, 1983.
- [21] H. Korezlioglu and G. Mazziotto. Modelization and filtering of discrete systems and approximation of continuous systems. In *System Modelling and Optimization, Proceedings of the 11th IFIP Conference, Copenhagen*, number 59 in Lecture Notes in Control and Information Sciences. Springer-Verlag, 1984.
- [22] M. A. Kouritzin. On exact filters for continuous signals with discrete observations. *IEEE Transactions on Automatic Control*, 1998.
- [23] T. Lefebvre, H. Bruyninckx, and J. De Schutter. Kalman filters for nonlinear systems: a comparison of performance. Technical report, Katholieke Universiteit Leuven, Department of Mechanical Engineering, Belgium, 2001.
- [24] T. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, MIT, 2001.
- [25] A. T. Nelson. *Nonlinear Estimation and Modeling of Noisy Time-series by Dual Kalman Filtering Methods*. PhD thesis, Oregon Graduate Institute, 2000.
- [26] J. Picard. Approximation of non linear filtering problems and order of convergence. In *Filtering and Control of random processes*, number 61 in Lecture Notes in Control and Information Sciences, 1984.

- [27] Da Prato and Giuseppe. *Stochastic Equations in Infinite Dimensions*, volume 45 of *Encyclopaedia of mathematics and its applications*. Cambridge University Press, 1992.
- [28] L. C. G. Rogers. *Diffusions, Markov processes and Martingales*, volume 2. Cambridge University Press, 2nd edition, 2000.
- [29] L. C. G. Rogers and David Williams. *Diffusions, Markov processes and Martingales*, volume 1. Cambridge University Press, 2nd edition, 2000.
- [30] B. L. Rozovskii. On stochastic partial differential equations. *Math. USSR-Sb.*, 25:295–322, 1975.
- [31] P. K. Tam and J. B. Moore. A gaussian sum approach to phase and frequency estimation. *IEEE Transactions on Communications*, 25(9):935–942, 1977.
- [32] S. Yau. Explicit construction of finite dimensional nonlinear filters with state space dimension 2. In *Proceedings of the IEEE Conference on Decision*, 1993.