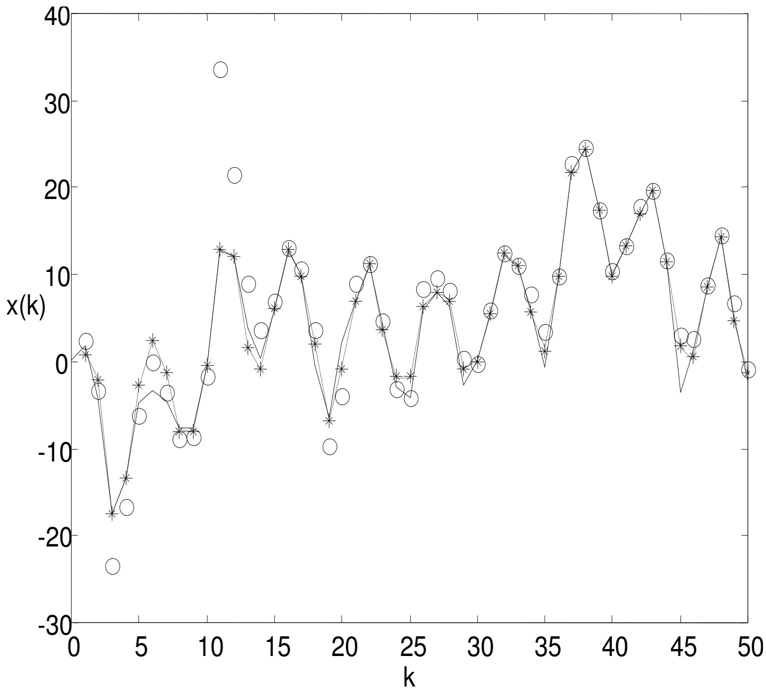
FOI-R--1074--SE
December 2003
1650-1942

Scientific report

Fredrik Berefelt, Johan Hamberg
John W.C. Robinson

Geometric Aspects of Nonlinear Filtering



System Technology Division
SE-172 90 STOCKHOLM
Sweden

System Technology Division
SE-172 90 STOCKHOLM
Sweden

Swedish Defence Research Agency
System Technology Division
SE-172 90 STOCKHOLM
Sweden

FOI-R--1074--SE
December 2003
1650-1942

Scientific report

Fredrik Berefelt, Johan Hamberg
John W.C. Robinson

Geometric Aspects of Nonlinear Filtering

Issuing organization

Swedish Defence Research Agency
System Technology Division
SE-172 90 STOCKHOLM
Sweden

Report number, ISRN: FOI-R--1074--SE

Report type: Scientific report

Research area code: Combat

Month year: December 2003

Project no.: E6004

Customers code: Commissioned Research

Sub area code: Weapons and Protection

Author/s (editor/s)

Fredrik Berefelt, Johan Hamberg
John W.C. Robinson

Project manager: Johan Hamberg

Approved by: Monica Dahlén

Sponsoring agency: Swedish Armed Forces

Scientifically and technically responsible: Lennart Widlind

Report title

Geometric Aspects of Nonlinear Filtering

Abstract

Future military systems must meet new and higher demands on precision, safety, economic and environmental concerns. Vehicle guidance and control clearly illustrate this by having a natural intrinsic geometry that calls for general geometric methods in modeling, filtering and control. This report addresses one particular line of research, differential geometric filtering, giving a picture of the present state of the research efforts in this area. It will be followed by a more complete report. The present work includes a generalization of the Zakai equation, a coordinate free construction of jets of mappings and a completely intrinsic construction of the geodesic spray of an affine connection. Some technical questions concerning the necessity and sufficiency of certain conditions for projections filters are also addressed.

Keywords

Missile, Guidance, Control, Differential Geometry, Modeling

Further bibliographic information

Language: English

ISSN: 1650-1942

Pages: 80

Distribution: By sendlist

Price: Acc. to pricelist

Security classification: Unclassified

Utgivare

Totalförsvarets forskningsinstitut
Avdelningen för Systemteknik
SE-172 90 STOCKHOLM
Sweden

Rapportnummer, ISRN: FOI-R--1074--SE

Klassificering: Vetenskaplig rapport

Forskningsområde: Bekämpning

Månad, år: December 2003

Projektnummer: E6004

Verksamhetsgren: Uppdragsfinansierad verksamhet

Delområde: VVS med styrda vapen

Författare/redaktör

Fredrik Berefelt, Johan Hamberg
John W.C. Robinson

Projektledare: Johan Hamberg

Godkänd av: Monica Dahlén

Uppdragsgivare/kundbeteckning: Försvarsmakten

Tekniskt och/eller vetenskapligt ansvarig: Lennart Widlind

Rapportens titel

Geometriska Aspekter av Olinjär filtrering

Sammanfattning

Framtida militära system måste möta nya och högre krav på precision, säkerhet, ekonomiska samt miljömässiga hänsyn. Problemet med styrning och reglering av fordon illustrerar tydligt detta genom att besitta en naturlig inneboende geometri vilken kräver generella geometriska metoder för fullständigt utnyttjande av möjligheterna till modellering, filtrering och reglering. Dennna rapport behandlar ett speciellt forskningsområde, differentialgeometrisk filtrering, och ger en bild av det aktuella tillståndet i forskningen. Föreliggande arbete innefattar bl. a. en generalisering av Zakaiekvationen, en koordinatfri konstruktion av jets av avbildningar och en fullständigt koordinatfri konstruktion av det geodetiska flöde som hör till en affin konnektion. Några tekniska frågor rörande nödvändigheten och tillräckligheten av vissa villkor för projektionsfilter behandlas också.

Nyckelord

Robot, Styrning, Reglering, Differentialgeometri, Modellering

Övriga bibliografiska uppgifter

Språk: Engelska

ISSN: 1650-1942

Antal sidor: 80

Distribution: Enligt missiv

Pris: Enligt prislista

Sekretess: Öppen

Contents

1 Introduction **1**

2 Nonlinear Stochastic Systems **3**
2.1 Stochastic Processes and Dynamical Systems 3
2.1.1 Preliminaries on Probability and Stochastic Processes 3
2.1.2 A Generic Dynamical System 9

3 Nonlinear Filtering **11**
3.1 A Basic Filtering Problem 11
3.1.1 Change of measure. 11
3.1.2 A Bayes' formula 13
3.2 Recursive Filters 13
3.2.1 The Zakai Equation. 13
3.2.2 Density form of the Zakai equation. 14
3.2.3 Robust Version of the Zakai Equation. 15
3.2.4 Lie Form of the (Robust) Zakai Equation 16
3.3 Finite Dimensional Filters 17
3.4 Observations in additive "white noise" 17
3.4.1 The Wei-Norman Technique 18

4 An algebraic approach to geometry **19**
4.1 Differentiable Manifolds 19
4.2 Basic constructions 19
4.2.1 Reconstruction of the function algebra structure 19
4.2.2 Smooth mappings 19
4.2.3 Product manifolds 20
4.2.4 Tangent vectors 20
4.2.5 Fiber bundles 20
4.2.6 Pullback and push forward 21
4.2.7 Jet bundles 21
4.2.8 Cotangent vectors 21
4.3 Tensor algebra 22
4.3.1 Exterior algebra 22
4.3.2 Fractional density algebras 22
4.3.3 Tensor bundles 23
4.4 Analysis on manifolds 23
4.4.1 Linear differential operators 23
4.4.2 Jet bundle characterization of differential operators 24
4.4.3 Principal symbols 24
4.4.4 Exterior calculus 24
4.5 Flows and Lie derivatives 24
4.5.1 Topology on $C^\infty(M)$ 24
4.5.2 The flow of a vector field 24
4.5.3 Lie derivatives 24
4.5.4 Riemannian and subriemannian structures 25
4.6 Affine Connections 25

4.6.1 Covariant derivatives . . . 25
4.6.2 Geodesic spray . . . 25
4.6.3 Levi-Civita connection . . . 26

5 The coordinate approach to geometry **27**
5.1 Tangent vectors, tensors . . . 27
5.2 Jets and differential operators . . . 27
5.3 Affine connections . . . 28
5.4 Representation of differential operators . . . 30

6 Stochastic calculus on manifolds **31**
6.1 Itô's rule and affine connections . . . 31
6.2 Redundancy in the representation . . . 32
6.3 Particular choices of connections . . . 32
6.3.1 Classical Itô equations . . . 32
6.3.2 Stratonovich equations . . . 32
6.3.3 Geometric diffusion equations . . . 32
6.3.4 Driftless Stratonovich equations . . . 32
6.4 Connections and the definition of SDEs . . . 33
6.5 Geometric Invariants . . . 34
6.6 Other geometric issues . . . 34

7 Statistical Manifolds **35**
7.1 Parametric models without parameters . . . 35
7.1.1 Examples . . . 35
7.1.2 Moser's theorem . . . 36
7.1.3 Invariants of a pair of probability densities . . . 36
7.1.4 Invariants of infinitesimally close pairs . . . 37
7.2 Exponential families . . . 38
7.2.1 Examples . . . 38
7.3 Construction of a statistical manifolds . . . 38
7.3.1 Example . . . 39

8 Differential geometric filtering **41**
8.1 Modeling issues . . . 41
8.2 Geometric filters . . . 41
8.2.1 Exact finite dimensional filters . . . 41
8.2.2 Intrinsic geometric filters . . . 41
8.2.3 Projection filters . . . 42
8.2.4 Extended Kalman filters . . . 42
8.3 Statistical geometry in geometric filtering . . . 42

9 The Projection Filter **43**
9.1 The geometrical setting . . . 43
9.2 A projection map . . . 45
9.2.1 Remarks . . . 45
9.3 The filtering problem . . . 47
9.4 Projection filter, general case . . . 48
9.5 Projection filter for an exponential family . . . 49
9.5.1 Remarks . . . 49
9.6 Numerical implementation . . . 50

10 Filter Implementation **53**
10.1 Particle Projection Filter: Background . . . 53
10.2 Particle Projection Filter: Implementation . . . 54

A Properties of x_0, w, and v under the measures $\mathbb{P}$ and $\mathbb{P}_0$ **59**

B Derivation of the Bayes' formula (3.6) **61**

C Derivation of the Zakai Equation (3.9) **63**
C.1 Itô form 63
C.2 Stratonovich form 65

D Existence of the extended unnormalized density (3.14) **67**

E Derivation of the robust Zakai equation **69**

F Stratonovich form of the Kushner-Stratonovich equation **71**

G Filter toolbox code **73**

1. Introduction

The present report is a revised and enlarged second edition of the report ‘Nonlinear Filtering: Geometric Aspects’ (FOI-R–0766–SE), and it replaces the latter.

The additions in this edition are mainly the entirely new chapters on Statistical Manifolds and on Numerical implementation of the filter techniques together with corrections and enlargements of the other chapters. This work is an ongoing research project, and the final edition of this report is scheduled for 2004.

Future military systems must meet new and higher demands on precision, safety, economic and environmental concerns. The network based defense paradigm is but one aspect of this general tendency towards high-tech and state of the art technology.

The general problem of vehicle guidance and control clearly illustrates this phenomenon. The system is inherently nonlinear and has an intrinsic natural geometry. This calls for general geometric methods in modeling, filtering, signal processing, control and implementation.

A well implemented geometrical description of the relevant subsystems is crucial for those synergy effects necessary for top performance in complex systems. This report addresses one particular line of research, differential geometric filtering.

The report gives a picture of the present state of the research efforts in this area. It will be followed by a more complete report next year.

The present work includes a generalization of the Zakai equation to systems with more general measurement processes than has been considered heretofore. This is an important but largely ignored aspect of modeling. It is however known that the properties of a stochastic filtering problem heavily depend on the details of the noise model. Another reason to generalize the Zakai equation is that this seems to be necessary for a truly geometric theory with nontrivial transformational properties.

The report provides a rather lengthy introduction to differential geometry as such. Several novelties are presented, such as a coordinate free construction of jets of mappings and a completely intrinsic construction of the geodesic spray of an affine connection. Many of the topics covered in this differential geometric survey are not explicitly used in the other chapters, but are included as a preparation for future reports.

The report concludes with a chapter describing the so-called projection filters by Brigo et.al. The presentation provides an alternative setup for these filters, deemphasizing the role of fractional densities. Some technical questions concerning the necessity and sufficiency of certain conditions for these filters are addressed.

2. Nonlinear Stochastic Systems

In this first chapter we will treat stochastic processes in relation to nonlinear dynamical systems. The main objective is to introduce notation and, at the same time, to give a quick introduction to the basic concepts and constructs employed in order to aid readers with a less complete background in probability and stochastic processes. This presentation will be sketchy and we will not explain fully (or even define properly) all the entities used. We motivate this by the fact that the emphasis of the report is geometric filtering, i.e. the geometric aspects of the equations for nonlinear filtering. Indeed, once these equations have been derived there is little need for the probability foundation on which they rest.

Good general references to the material covered here are [32] for the basic probability theory needed (abstract probability and conditional probability) and [18] for the specifics on continuous time Martingales (such as the Wiener process and solutions to stochastic differential equations).

2.1 Stochastic Processes and Dynamical Systems

2.1.1 Preliminaries on Probability and Stochastic Processes We discuss here the basic entities in probability theory and stochastic processes to be used in the following chapters. At the end of this chapter we shall briefly encounter some connections with differential geometry (Lie derivatives). The reference material for these parts is given in Ch. 4.

Probability Space, Stochastic Process In the sequel we will consider jointly random variables and random functions (i.e. stochastic processes) indexed by "time," in the framework of the modern theory of semimartingale stochastic processes [18]. In order to do this, we must assume the existence of an "abstract" probability space [32, Ch. 1–2]

$$\tilde{\Omega} = (\Omega, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathcal{F}, \mathbb{P}) \tag{2.1}$$

where Ω is the basic set of outcomes, $\mathcal{F}$ is the largest sigma-algebra occurring in the discussion; the one on which the probability measure $\mathbb{P}$ is defined and $\{\mathcal{F}_t\}_{t\in[0,T]}$ is the basic filtration (a nested sequence if sigma-algebras). The different parts of this construction will now be explained in a very informal manner.

Since we will only be considering stochastic processes indexed by a time interval $[0, T]$ with values in $\mathbb{R}^k$ and with continuous paths, and random variables with values in $\mathbb{R}^\ell$, for some $k, \ell \in \mathbb{N}$, it is here sufficient to think of Ω in (2.1) as a product space $C([0, T], \mathbb{R}^k) \times \mathbb{R}^\ell$. A *stochastic process* [18, Ch. 1] η can then be thought of as a map

$$\eta : C([0, T], \mathbb{R}^k) \times \mathbb{R}^\ell \to C([0, T], \mathbb{R}^k), \tag{2.2}$$

defined by

$$\eta_t((\omega, \nu)) = \omega_t, \quad \omega \in C([0, T], \mathbb{R}^k),\ \nu \in \mathbb{R}^\ell$$

representing "one coordinate" of Ω, and likewise a *random variable* ξ can be thought of as another map

$$\xi : C([0, T], \mathbb{R}^k) \times \mathbb{R}^\ell \to \mathbb{R}^\ell, \tag{2.3}$$

defined by

$$\xi((\omega,\nu)) = \nu, \quad \omega \in C([0,T],\mathbb{R}^k),\ \nu \in \mathbb{R}^\ell$$

representing another "coordinate" of Ω. For fixed $t \in [0,T]$ the object $\eta_t(\cdot)$ is a function of the "random outcomes" $(\omega,\nu) \in \Omega$ (actually only of ω), hence is a random variable, and is traditionally written simply as η_t. For fixed $(\omega,\nu) \in \Omega$ the object $\eta_{(\cdot)}(\omega,\nu)$ is a continuous function of $t \in [0,T]$, called a *path* of η. In our later applications to dynamical systems the basic stochastic processes will be processes like η which will represent the noise processes of the system and random variables like ξ that will represent the initial conditions. All other random variables and stochastic processes occurring in the discussion, such as the state and observation process, will be be thought of as being derived from η, ξ, as different forms of functions or "functionals."

Sigma-algebra, Filtration A *sigma-algebra* [32, Ch. 1] is a nonempty set of subsets of Ω closed under the operations of complement and countable unions. In probability theory, sigma-algebras in general have the interpretation as the amount of information about the basic outcomes on Ω gained by observing the values of some random variables or processes at certain time points or over time intervals. The *filtration* [18, Ch. 1] $\{\mathcal{F}_t\}_{t\in[0,T]}$ in (2.1) is an increasing sequence of sub-sigma-algebras of the basic sigma-algebra $\mathcal{F}$, i.e. $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2} \subseteq \mathcal{F}$, for $t_t \le t_2$, such that the events (subsets of Ω) determined by observing the values of (η_s, ξ) during $s \in [0,t]$ are always sets in $\mathcal{F}_t$. An example of a set $A \in \mathcal{F}_t$ could be

$$A = \{(\omega,\nu) \in \Omega : \eta^j_{t_1} > 0, -1.3 < \eta^k_{t_2} \le 0.2\} \quad \text{for} \quad t_t, t_2 \in [0,t], \tag{2.4}$$

where the superscripts j, k indicate components of a vector. A random variable ξ for which all the sets A determined in a manner analogous to (2.4) belong to a certain sigma-algebra $\mathcal{G}$ on Ω is said to be *measurable with respect to* $\mathcal{G}$. [1] The smallest sigma-algebra containing two sigma-algebras $\mathcal{F}$ and $\mathcal{G}$ is denoted $\mathcal{F} \vee \mathcal{G}$. If a process η and filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$ are such that η_t is measurable with respect to $\mathcal{F}_t$ for every $t \in [0,T]$ then η is said to be *adapted* to $\{\mathcal{F}_t\}_{t\in[0,T]}$. When observing a particular process $\tilde\eta$ on $\tilde\Omega$ (such as certain components of η and no other variables) the minimal filtration to which $\tilde\eta$ is adapted is called the *filtration generated by* $\tilde\eta$ and is denoted $\{\mathcal{F}^{(\tilde\eta)}_t\}_{t\in[0,T]}$. If another stochastic process $\check\eta$ is adapted to $\{\mathcal{F}^{(\tilde\eta)}_t\}_{t\in[0,T]}$ as well it can (in general) be expressed as a certain form of function (or functional) [32, Sec. 3.13], [18, Sec. 3.4.D] of the variables $\tilde\eta_s, s \in [0,t]$, for each $t \in [0,T]$.

Probability Measure, Conditional Expectation The last component of $\tilde\Omega$, the *probability measure* $\mathbb{P}$ [32, Ch. 1], describes how probable different events (or sets of outcomes) in $\mathcal{F}$ are; it is a map from the members of $\mathcal{F}$ to $[0,1]$. The (mathematical) *expectation* [32, Ch. 6] with respect to $\mathbb{P}$ of a random variable such as η_t is denoted $\mathbb{E}_{\mathbb{P}}(\eta_t)$ and defined as the "abstract" integral

$$\mathbb{E}_{\mathbb{P}}(\eta_t) = \int_\Omega \eta_t\big((\omega,\nu)\big)\, d\mathbb{P}\big((\omega,\nu)\big). \tag{2.5}$$

Given a set $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$, determined for instance as in (2.4) by observing a stochastic process over a time interval, one can construct a new probability measure $\mathbb{P}_A$ on $(\Omega,\mathcal{F})$ by

$$\mathbb{P}_A(B) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}, \quad B \in \mathcal{F}, \tag{2.6}$$

[1] Technically, a map such as (2.3) will only qualify as a random variable if it is measurable with respect to $\mathcal{F}$.

which is the *conditional probability* of B given A. One can then compute expectations as in (2.5) using the probability measure $\mathbb{P}_A$ instead, and obtain a *conditional expectation* [32, Ch. 9]. (Expectations with respect to the basic measure $\mathbb{P}$ is usually written without subindex as $\mathbb{E}(\cdot)$.) For example, if ξ is a random variable taking only a finite number of values on Ω, say

$$\xi = \sum_{j=1}^{N} \beta_j \mathbf{1}_{B_j},$$

where $\beta_j \in \mathbb{R}^k$, the sets B_j are disjoint and $\mathbf{1}_{B_j}$ is the indicator function [2] of B_j, then

$$\mathbb{E}_{\mathbb{P}_A}(\xi) = \sum_{j=1}^{N} \beta_j \mathbb{P}_A(B_j). \tag{2.7}$$

However, the most important forms of conditional probability and expectation are the "differential" forms obtained when (formally) letting $\mathbb{P}(A)$ in (2.6), (2.7) tend to zero. If we were to do this, the left hand side of (2.6) would become a function on Ω (a random variable) $\tilde{\mathbb{P}}_{(\cdot)}(B)$ such that

$$\int_A \tilde{\mathbb{P}}_{\big((\omega,\nu)\big)}(B)\, d\mathbb{P}\big((\omega,\nu)\big)) = \mathbb{P}(B \cap A) \tag{2.8}$$

and similarly the left hand side of (2.7) would become a function $\tilde{\mathbb{E}}_{(\cdot)}$ on Ω such that

$$\int_A \tilde{\mathbb{E}}_{\big((\omega,\nu)\big)}(\xi)\, d\mathbb{P}\big((\omega,\nu)\big) = \sum_{j=1}^{N} \beta_j \mathbb{P}(B_j \cap A) = \int_A \xi\big((\omega,\nu)\big)\, d\mathbb{P}\big((\omega,\nu)\big). \tag{2.9}$$

In fact, for any random variable $\xi \in L_1(\mathbb{P})$ and any sub-sigma-algebra $\mathcal{G} \subseteq \mathcal{F}$ there exists an integrable random variable $\tilde{\mathbb{E}}_{(\cdot)}$, measurable with respect to $\mathcal{G}$, such that a relation like the one (formed by the "outer ends of") (2.9) holds for any $A \in \mathcal{G}$. This random variable, usually denoted $\mathbb{E}_{\mathbb{P}}(\xi|\mathcal{G})$, is the general definition of the *conditional expectation of ξ given $\mathcal{G}$*. (The conditional probability in (2.8) is a special case.) In particular, if ξ is measurable with respect to $\mathcal{G}$ we have $\mathbb{E}_{\mathbb{P}}(\xi|\mathcal{G}) = \xi$. An important property of $\mathbb{E}_{\mathbb{P}}(\xi|\mathcal{G})$ is that if also $\xi \in L_2(\mathbb{P})$ then $\mathbb{E}_{\mathbb{P}}(\xi|\mathcal{G})$ minimizes $\mathbb{E}_{\mathbb{P}}\|z - \xi\|_2$ over all $z \in L_2(\mathbb{P})$ that are measurable with respect to $\mathcal{G}$ [32, Sec. 9.4].

An important class of stochastic processes in the sequel are martingales. A process η is a *martingale* [32, Ch. 10], [18] with respect to a filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$ and probability measure $\mathbb{P}$ if η_t (has finite expectation and) is adapted to $\mathcal{F}_t$ for all t and $\mathbb{E}_{\mathbb{P}}(\eta_t|\mathcal{F}_s) = \eta_s$ for $s \le t$. By the properties of the conditional expectation it follows that $\mathbb{E}_{\mathbb{P}}(\eta_t - \eta_s|\mathcal{F}_s) = 0$ and a martingale η is thus an abstraction of a "fair game" since the future increments of η are "not biased" based on the information in $\mathcal{F}_s$. One of the most important martingales is the Brownian motion, or *Wiener process* [18, Ch. 2], which is (formally) the time integral of "white Gaussian noise."

Stochastic Integrals; Itô vs. Stratonovich In what follows we will study integrals with respect to Wiener processes in two forms: Itô and Stratonovich [18, Ch. 3]. [3] The two forms are equivalent in that there exists a simple ("invertible") transformation between the two (when the involved integrands satisfy conditions such that both forms exist simultaneously) but each have advantages in different applications: The Itô form is the preferred choice in most probabilistic calculations because the

[2] $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ if $x \notin A$.

[3] The reason one why one has to introduce these new notions of the integral is that the paths of a Wiener process are of unbounded variation, hence they cannot be used as "distribution functions" in a Stieltjes integral construction. (They do have bounded second variation, however.)

resulting integral becomes a martingale. The Stratonovich form, on the other hand, is the preferred choice in geometrical arguments because the integral satisfy a relation which typographically looks like the fundamental theorem of (ordinary) calculus (for Stieltjes integrals). Moreover, the Stratonovich form is the one obtained under general conditions if one approximates the Wiener process with a process having paths of bounded variation, which is natural in most real life modeling situations.

Returning to our setup in (2.2), (2.3), we assume that the probability space $\tilde{\Omega}$ in (2.1) is large enough to host a Wiener process w and an auxiliary process z with continuous paths, both with values in $\mathbb{R}$ and adapted to $\{\mathcal{F}_t\}_{t\in[0,T]}$ (we can e.g. take $k \geq 2$ in (2.2) and let w, z be two components of η). We assume also that z is bounded over Ω. The first step in the construction of the stochastic integral [4] is to fix an an arbitrary time point $t \in (0, T)$ and define a piecewise constant (in time) approximation $z^{(N)}$ ($N \in \mathbb{N}$) to z as [5]

$$z_s^{(N)}\big((\omega,\nu)\big) = z_{\tau_N(s)}\big((\omega,\nu)\big), \quad \tau_N(s) = \frac{T}{N}\lfloor\frac{sN}{T}\rfloor,\ s \in [0,t].$$

The Itô stochastic integral $I_t^{(w)}(z^{(N)})$ of the approximating process $z^{(N)}$ with respect to w, written as

$$\int_0^t z_s^{(N)}\, dw_s, \tag{2.10}$$

is defined as a "forward Wiener increment" sum;

$$I_t^{(w)}(z^{(N)}) = \int_0^t z_s^{(N)}\, dw_s = \sum_{j=0}^{N-1} z_{jT/N}(w_{(j+1)T/N} - w_{jT/N}).$$

Because of the way $I_t^{(w)}(z^{(N)})$ is constructed, the sequence $\{I_t^{(w)}(z^{(N)})\}_{N=1}^{\infty}$ converges in $L_2(\mathbb{P})$ ("in mean square") to a random variable $I_t^{(w)}(z)$, which we write also as

$$\int_0^t z_s\, dw_s.$$

The variable $I_t^{(w)}(z)$ is called the *Itô integral* of the process z with respect to w. The Itô integral $I_t^{(w)}(z)$ is $\mathcal{F}_t$ measurable and by letting t vary in $[0,T]$ we obtain a process $I_{(\cdot)}^{(w)}(z)$ which is an $\{\mathcal{F}_t\}_{t\in[0,T]}$ martingale. The Stratonovich integral $\tilde{I}_t^{(w)}(z)$ of z with respect to w is defined similarly but using also backward Wiener increments in the ("symmetrized") approximation;

$$\tilde{I}_t^{(w)}(z^{(N)}) = \sum_{j=0}^{N-1} \frac{1}{2}(z_{(j+1)T/N} + z_{jT/N})(w_{(j+1)T/N} - w_{jT/N}).$$

Also the sequence $\{\tilde{I}_t^{(w)}(z^{(N)})\}_{N=1}^{\infty}$ converges to a limit $\tilde{I}_t^{(w)}(z)$, but the convergence here is "in probability" (which is a weaker topology than "in mean square") i.e. we have $\lim_{N\to\infty} \mathbb{P}(|\tilde{I}_t^{(w)}(z^{(N)}) - \tilde{I}_t^{(w)}(z)| > \varepsilon) = 0$ for any $\varepsilon > 0$. The limit variable $\tilde{I}_t^{(w)}(z)$ is called the *Stratonovich integral* and is denoted

$$\int_0^t z_s \circ dw_s. \tag{2.11}$$

(It does not define a martingale process.) Below we shall give a formula relating the two forms of integrals in the case that the process z is a function of the solution to a stochastic differential equation.

[4] The construction of the integral can be made in a much more general setting and, in particular, the result will not be dependent on the sequence of approximating functions (as it may appear here).

[5] Here $\lfloor\cdot\rfloor$ denotes integer part (truncation).

Since both the Itô and Stratonovich integrals were obtained by a limiting process it is natural to ask what properties the limits have (apart from the martingale property of the Itô integral), in particular do the integrals have continuous paths when regarded as stochastic processes? It turns out that that the answer is affirmative [18, Sec. 3.2B] so that we remain in the realm of continuous path processes also after integration.

It is standard to use the shorthand differential notation

$$z_t \, dw_t, \quad z_t \circ dw_t$$

for the integrals in (2.10) and (2.11), respectively, and we will also use it frequently. Likewise, for (Lebesgue) integrals with respect to time we will use the notation

$$z_t \, dt.$$

In case z is a row vector process the notation $z_t \, dw_t$ for the Itô differential is to be interpreted as a formal vector-vector multiplication (i.e. a linear combination), and this interpretation extends to the case where z is a matrix of such vector processes. These generalizations carry over to the case of Stratonovich differentials $z_t \circ dw_t$ (and ordinary time differentials $z_t \, dt$) as well.

Stochastic Differential Equations Consider two functions $F_0 : \mathbb{R}^n \to \mathbb{R}^n$, and $F = (F_1, \ldots, F_m)$, $F_j : \mathbb{R}^n \to \mathbb{R}^n$, where we assume that F_0, F_j are smooth (C^∞) bounded vector fields, [6] and consider the equation (in the process x)

$$\begin{aligned} dx_t &= F_0(x_t) \, dt + F(x_t) \, dw_t, \\ x_0 &= \xi, \end{aligned} \qquad t \in [0, T], \tag{2.12}$$

where w is a Wiener process with values in $\mathbb{R}^m$ and ξ is a random variable in $\mathbb{R}^n$. An equation such as (2.12) is known as a (Itô) *stochastic differential equation* (SDE) [18, Ch. 5]. The first term on the right on the first line of (2.12) is called the *drift term* and the second the *diffusion term.* Later we shall also need the associated function $a_F : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ defined by

$$a_F(x) = F(x) F^T(x), \quad x \in \mathbb{R}^n, \tag{2.13}$$

called the *diffusion matrix.* If both the initial variable ξ and the Wiener process w are given (relative to $\tilde{\Omega}$), and the object is to find a process x satisfying (2.12) above we talk about a *strong solution* to (2.12). If none of ξ, w or x is given, only the probability distribution of ξ and the specification that w is a Wiener process, we talk about a *weak solution.* The object is then to find a triple (ξ, w, x), defined on *some* probability space $\tilde{\Omega}$, such that (2.12) holds. An important difference between these to forms of a solution is that in the case of a strong solution the process x will be adapted to $\{\mathcal{F}_t^{(\xi,w)}\}_{t \in [0,T]} = \{\mathcal{F}^{(\xi)} \vee \mathcal{F}_t^{(w)}\}_{t \in [0,T]}$ so that, in view of the remarks above, x_t can be thought of as a function (or functional) of the variables ξ, w_s for $s \in [0, t]$. In the case of a weak solution such an interpretation is not possible but weak solutions are nevertheless important since (i) one might only be interested in the set of finite dimensional probability distributions ("the law") of the process x and (ii) they are intimately connected to so-called Girsanov transforms. Both of these aspects will be illustrated below.

Starting with the vector fields F_0, F_j, the Wiener process w and initial condition ξ, one can also seek a (strong) solution $\tilde{x}$ to the SDE on Stratonovich form as

$$\begin{aligned} d\tilde{x}_t &= F_0(\tilde{x}_t) \, dt + F(\tilde{x}_t) \circ dw_t, \\ \tilde{x}_0 &= \xi. \end{aligned} \qquad t \in [0, T], \tag{2.14}$$

[6] In this report we will use the term *vector field* to denote both vector valued functions, such as $F_0(x)$, and the associated "directional" differential operators (differential geometric context), such as $\sum_{j=1}^n F_0^j \partial / \partial x^j$.

A solution to (2.14) is *equivalent* to a solution to a related Itô SDE because if $\tilde{x}$ is a (strong) solution to the Stratonovich (2.14) then $\tilde{x}$ satisfies the Itô SDE

$$\begin{aligned} d\tilde{x}_t &= \big(F_0(\tilde{x}_t) + C_F(\tilde{x}_t)\big)\, dt + F(\tilde{x}_t)\, dw_t, \\ \tilde{x}_0 &= \xi, \end{aligned} \qquad t \in [0, T],$$

where the vector field $C_F : \mathbb{R}^n \to \mathbb{R}^n$, called the *correction term* [18, Sec. 5.2.D] is given by [7]

$$C_F(x) = \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{n} F_j^k(x)\frac{\partial F_j(x)}{\partial x_k} = \frac{1}{2}\sum_{j=1}^{m}\begin{pmatrix} L_{F_j}F_j^1(x) \\ \vdots \\ L_{F_j}F_j^n(x)\end{pmatrix}, \quad x \in \mathbb{R}^n, \tag{2.15}$$

and L denotes the Lie derivative (see Ch. 4). Vice versa, if $\tilde{x}$ satisfies the Itô SDE (2.15) then it satisfies the Stratonovich SDE (2.14). By comparing (2.15) and (2.14) we see that we have

$$F(\tilde{x}_t) \circ dw_t = F(\tilde{x}_t)\, dw + C_F(\tilde{x}_t)\, dt, \qquad t \in [0, T]. \tag{2.16}$$

The above moreover shows that also in the case of a Stratonovich SDE a strong solution $\tilde{x}$ will be adapted to $\{\mathcal{F}_t^{(\xi,w)}\}_{t\in[0,T]}$. The relation (2.16) holds for any smooth vector field F (not just the one occurring in (2.12), (2.14)) and therefore we have also

$$F(x_t) \circ dw_t = F(x_t)\, dw + C_F(x_t)\, dt, \qquad t \in [0, T], \tag{2.17}$$

where x is the solution to the Itô SDE (2.12). To conclude, the difference between the Itô and Stratonovich solutions of an SDE can be expressed as *a change in the drift term.*

Itô's Formula In case a process x satisfies an Itô SDE like (2.12) (weak or strong sense) and $\phi : \mathbb{R}^n \to \mathbb{R}$ is a smooth and bounded function the "transformed" process $\phi(x_{(\cdot)})$ has a stochastic differential given by

$$\begin{aligned} d\phi(x_t) &= (\mathcal{A}\phi)(x_t)\, dt + \big(\nabla\phi(x_t)\big)^T F(x_t)\, dw_t \\ &= (\mathcal{A}\phi)(x_t)\, dt + \big((L_{F_1}\phi)(x_t), \ldots, (L_{F_m}\phi)(x_t)\big)\, dw_t, \quad t \in [0, T], \end{aligned} \tag{2.18}$$

where $\nabla\phi(x) = (\partial\phi(x)/\partial x_1, \ldots, \partial\phi(x)/\partial x_n)^T$, the differential operator $\mathcal{A}$ is defined by

$$\begin{aligned} (\mathcal{A}\phi)(x) &= \big(\nabla\phi(x)\big)^T F_0(x_t) + \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} a_F^{k,i}(x)\frac{\partial^2\phi(x)}{\partial x^k \partial x^i} \\ &= L_{F_0}\phi(x) - L_{C_F}\phi(x) + \frac{1}{2}\sum_{j=1}^{m} L_{F_j}^2\phi(x) \\ &= L_{(F_0 - C_F)}\phi(x) + \frac{1}{2}\sum_{j=1}^{m} L_{F_j}^2\phi(x), \quad x \in \mathbb{R}^n, \end{aligned} \tag{2.19}$$

and $a_F^{k,i}$ is the k:th row, i:th column of the diffusion matrix in (2.13). Relation (2.18) is the celebrated *Itô formula* [18, Ch. 3.3] which in stochastic calculus plays a similar role as the fundamental theorem of calculus does in ordinary calculus.

[7] To show this one needs to introduce the *quadratic variation* of a martingale, but this is one technical point we have chosen to omit in this presentation.

If we use the differential notation from (2.12) we can express (2.18) as

$$d\phi(x_t) = \big(\nabla\phi(x_t)\big)^T\, dx_t + \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{n} a_F^{k,j}(x)\frac{\partial^2\phi(x_t)}{\partial x^k \partial x^j}\, dt, \quad t\in[0,T]. \tag{2.20}$$

We see that the differential for ϕ is not the same as would be obtained from ordinary calculus since it contains also a second order term. The Stratonovich equivalent of (2.18), however, is

$$d\phi(x_t) = \big(\nabla\phi(x_t)\big)^T (F_0(x_t) + C_F(x_t))\, dt + (\nabla\phi(x_t))^T F(x_t)\circ dw_t, \qquad t\in[0,T],$$

which can be symbolically expressed as

$$d\phi(x_t) = (\nabla\phi(x_t))^T \circ dx_t, \qquad t\in[0,T]. \tag{2.21}$$

(where thus the '$\circ$' only effects the diffusion part of dx_t). This is the same form as would be obtained from ordinary calculus. Moreover, this formula holds also if we replace x by $\tilde{x}$, i.e. if we had started with the solution to the Stratonovich equation (2.14) instead (which is clear since (2.21) doesn't depend on the vector field F_0).

Finally, for future reference, we note that if we had started with an SDE on the Stratonovich form (2.14), then Itô's formula (expressed using Itô differentials) will read

$$d\phi(x_t) = (\mathcal{A}_+\phi)(x_t)\, dt + \big(\nabla\phi(x_t)\big)^T F(x_t)\, dw_t, \qquad t\in[0,T], \tag{2.22}$$

where the differential operator $\mathcal{A}_+$ is defined by

$$\mathcal{A}_+ = \mathcal{A} + L_{C_F} = L_{F_0} + \frac{1}{2}\sum_{j=1}^{m} L_{F_j}^2. \tag{2.23}$$

This result is immediate if we recall the transformation rule (2.16) for transformation between Itô and Stratonovich solutions of an SDE; a change in the drift term.

2.1.2 A Generic Dynamical System We will consider stochastic dynamical systems expressed on the Stratonovich differential form

$$\begin{aligned} dx_t &= F_0(x_t)\, dt + F(x_t)\circ dw_t, \quad x_0 = \xi \\ dy_t &= H_0(x_t, y_t)\, dt + H(y_t)\circ dv_t, \quad y_0 = 0, \end{aligned} \qquad t\in[0,T], \tag{2.24}$$

where $F_0 : \mathbb{R}^n \to \mathbb{R}^n$, $F = (F_1,\ldots,F_m)$, $F_j : \mathbb{R}^n \to \mathbb{R}^n$ and $H_0 : \mathbb{R}^n\times\mathbb{R}^p \to \mathbb{R}^p$, $H = (H_1,\ldots,H_p)$, $H_j : \mathbb{R}^p \to \mathbb{R}^p$, and all these functions are smooth and bounded. The diffusion matrix $a_F = FF^T$ of the first equation is assumed to be (strictly) positive definite. The vector valued stochastic processes w and v, with values in $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively, are two independent standard Wiener processes which are independent of the initial condition x_0, all of which are specified with respect to some underlying ('filtered') probability space $(\Omega, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathcal{F}, \mathbb{P})$ where the filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$ satisfies the "usual conditions" [18, p. 10]. [8] We assume that $E\|x_0\|_2 < \infty$ which, together with the conditions on the vector fields F_0, F_j and H_0, H_k is sufficient to guarantee the existence of a unique global (nonexploding) strong solution (x,y) to (2.24) [18, Ch. 5.2] (since the two equations in (2.24) together form a system of the same form as the first equation alone). We moreover assume that $H(y)$ has bounded first order derivatives and has an inverse $H^{-1}(y)$ for all $y\in\mathbb{R}^p$ such that the norm $\|H^{-1}(y)\|_2$ is uniformly bounded in y.

In (2.24), the first equation is to be thought of as representing the "state" x_t which is to be estimated based on noise corrupted "observations" y_s for $s\in[0,t]$,

[8] For all filtrations "generated" by SDEs such as (2.24) we may always assume that the usual conditions are satisfied, see e.g. [18, Sec. 5.2.A].

as described by the second equation. A large class of physical systems arising in the applications can be modeled as a system of the form (2.24) with (intrinsically) Gaussian noise, by an appropriate choice of F_0, F and H_0, H. In particular, systems with explicitly time varying dynamics and colored noise acting on the physical state variables can (by introducing additional "dummy" states) be rewritten on this form.

For later reference we note that the last term in the second equation in (2.24) can be written

$$H(y_t) \circ dv_t = H(y_t)\, dv_t + C_H(y_t)\, dt, \quad t \in [0, T].$$

where C_H is a correction term defined as in (2.15). (This follows from the developments in Sec 5 since the two equations in (2.24) together form a system of the same form as the first equation alone.) We also note that weak solutions [18, Sec. 5.3] to (2.24) are unique "in probability law" and therefore the (finite dimensional) joint probability distributions of (x, y) are unique, in particular they are not dependent on the specific Wiener processes w, v occurring in (2.24).

3. Nonlinear Filtering

It is now time to introduce the filtering problem and present the fundamental equations for its solution. We will focus on the (robust) Zakai equation, in particular the form it takes when expressed in a differential geometric language. This will be the basis for our future investigations on the geometric aspects of the nonlinear filtering problem.

The (most) general formulation of the nonlinear filtering problem is to find the conditional distribution of the state x_t in (2.24) given the (noise corrupted) observations y_s over the time interval $[0, t]$, for $t \in [0, T]$. This is equivalent to determining the conditional expectation

$$\pi_t(\phi) = \mathbb{E}(\phi(x_t)|\mathcal{F}_t^{(y)}) \tag{3.1}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}$ is an arbitrary (Borel measurable) function: [1] For $\phi = \mathbf{1}_A$, the indicator function for some (Borel) set $A \subseteq \mathbb{R}^n$, we obtain $\pi_t(\phi) = \mathbb{P}(A|\mathcal{F}_t^{(y)})$, i.e. the conditional probability of A given $\mathcal{F}_t^{(y)}$. Once the conditional expectation $\pi_t(\phi)$ is known, for arbitrary ϕ, one can compute different estimates of x_t based on $\mathcal{F}_t^{(y)}$ that are optimal in different ways. For example, the *minimum mean square error* (MMSE) estimate $\hat{x}_t$ of x_t given $\mathcal{F}_t^{(y)}$ is, as mentioned in Sec. 1, the conditional expectation $\mathbb{E}(x_t|\mathcal{F}_t^{(y)})$ which is obtained by taking, in succession, $\phi(x) = x^j$ for $j = 1 \ldots n$.

3.1 A Basic Filtering Problem

The solution to the general nonlinear filtering problem, as described by (3.1), that we are going to give is based on a so-called change-of-measure technique. The first step will be to derive a weak solution to the equations in (2.24) (and thus in effect replace the system in (2.24) with an equivalent one, having the same probability distributions). In doing so we shall introduce an auxiliary probability measure which will also be used when we replace the basic relation (3.1) with one that is easier to work with. The new relation is an unnormalized variant of (3.1) for which it is straightforward to derive recursive "updating equations" driven by the observation process y.

3.1.1 Change of measure. To begin with we therefore take one step back and consider (2.24), but for a moment "forget" the basic probability measure $\mathbb{P}$ (and the filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$) mentioned in connection with (2.24). Consider the first equation in (2.24) and assume that in addition to the initial variable x_0 and the process w there is given another process $\tilde{y}$, also defined on Ω. Assume moreover that there exists a probability measure $\mathbb{P}_0$ on $(\Omega, \mathcal{F})$ such that under $\mathbb{P}_0$ the following holds; [2]

[1] In most of what follows, however, ϕ will be a smooth (and bounded) function (to be thought of as a "test" function).

[2] All of this is easy to accomplish by a product space setup as in (2.2), (2.3) using the product of three spaces, one for each of x_0, w and $\tilde{y}$. Since w is Wiener with respect to $\{\mathcal{F}_t^{(w)}\}_{t\in[0,T]}$, and x_0 is independent of w, it follows that w is Wiener also with respect to $\{\mathcal{F}_t^{(w)} \vee \mathcal{F}^{(x_0)}\}_{t\in[0,T]}$. Using this argument one more time we obtain a Wiener process $\tilde{y}$ which is also Wiener with respect to $\{\mathcal{F}_t^{(\tilde{y})} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$.

(i) x_0 has the desired distribution,

(ii) w is a Wiener process with respect to $\{\mathcal{F}_t^{(x_0,w)}\}_{t\in[0,T]}$,

(iii) $\tilde{y}$ is a Wiener process with respect to $\{\mathcal{F}_t^{(\tilde{y})} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$,

(iv) x_0, w and $\tilde{y}$ are independent.

Under these conditions we know ([18, Sec. 5.2]) that (under $\mathbb{P}_0$) there exists a unique strong solution x to the first equation in (2.24) and this solution will be adapted to $\{\mathcal{F}_t^{(x_0,w)}\}_{t\in[0,T]}$. (The process x here can indeed be identified with the one occurring in (2.24).) Note that the conditions (i)–(iv) above imply that the processes x and $\tilde{y}$ are independent (under $\mathbb{P}_0$). Define the process y (at this point *not* to be identified with the one occurring in (2.24)) as the unique strong solution (under $\mathbb{P}_0$) to the Itô SDE

$$\begin{aligned} dy_t &= H(y_t)\, d\tilde{y}_t, \\ y_0 &= 0, \end{aligned} \qquad t \in [0,T]. \tag{3.2}$$

It follows that also the processes x and y are independent. From the definition of y and properties of strong solutions it follows that y is a martingale with respect to $\{\mathcal{F}_t^{(\tilde{y})}\}_{t\in[0,T]}$ and that $\mathcal{F}_t^{(y)} \subseteq \mathcal{F}_t^{(\tilde{y})}$. Hence, y is a martingale also with respect to $\{\mathcal{F}_t^{(y)}\}_{t\in[0,T]}$ and so is the integral [3] with differential $H^{-1}(y_t)\, dy_t$. However, by a step-function approximation argument (similar to the one used in the construction of the Itô integral) we have that $H^{-1}(y_t)\, dy_t = d\tilde{y}_t$ and therefore $\mathcal{F}_t^{(\tilde{y})} \subseteq \mathcal{F}_t^{(y)}$. Therefore we can conclude that

$$\mathcal{F}_t^{(y)} = \mathcal{F}_t^{(\tilde{y})}, \quad t \in [0,T]. \tag{3.3}$$

To proceed, let H_0, H be as in the second equation of (2.24) and define the ("likelihood") process Λ by

$$\Lambda_t = \exp\Big(\int_0^t h^T(x_s, y_s) d\tilde{y}_s - \frac{1}{2} \int_0^t \|h(x_s, y_s)\|_2^2\, ds \Big), \quad t \in [0,T] \tag{3.4}$$

where the first integral is an Itô integral and we have introduced $h : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^p$ as

$$h(x,y) = H^{-1}(y)\big(H_0(x,y) + C_H(y)\big), \quad x \in \mathbb{R}^n, y \in \mathbb{R}^p. \tag{3.5}$$

It is well-known [18, Sec. 3.5] that (since $h(x_s, y_s)$ is bounded) we have $E_{\mathbb{P}_0}\Lambda_t \equiv 1$ and we can define a family of probability measures $\mathbb{P}_t$ on $(\Omega, \mathcal{F})$ by

$$d\mathbb{P}_t = \Lambda_t\, d\mathbb{P}_0, \quad t \in [0,T],$$

i.e. we use Λ_t as a probability *density.* Define then the measure $\mathbb{P}$ by $\mathbb{P} = \mathbb{P}_T$. Under $\mathbb{P}$, the following holds (see appendix A): [4]

- The process v defined by $dv_t = d\tilde{y}_t - h(x_t, y_t)\, dt$ is a standard Wiener process with respect to $\{\mathcal{F}_t^{(\tilde{y})} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$.
- The processes v is independent of x_0, w.
- The pair (x_0, w) has the same (finite dimensional) probability distributions as under $\mathbb{P}_0$ (and thus the process x the same probability distributions as under $\mathbb{P}_0$).

[3] This integral is a generalization of the Itô integral with respect to a Wiener process described in Sec. 2, see [18, Ch. 3].

[4] This is basically an application of Girsanov's Theorem [18, Sec. 3.5]. Girsanov's Theorem can be viewed as an infinite dimensional extension of the change-of-measure by change-of-means property of Gaussian random vectors in $\mathbb{R}^N$.

By a step-function approximation argument it follows that the process y in (3.2) satisfies

$$\begin{aligned} dy_t &= \big(H_0(x_t, y_t) + C_H(y_t)\big)\,dt + H(y_t)\,dv_t, \\ &= H_0(x_t, y_t)\,dt + H(y_t) \circ dv_t, \qquad\qquad t \in [0, T]. \end{aligned}$$

Further, by the remark (3.3) it follows that v is also a Wiener process with respect to $\{\mathcal{F}_t^{(y)} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$. Hence, if we let our basic filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$ be $\{\mathcal{F}_t^{(y)} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$ and take $\mathbb{P}$ as our basic measure we have in (x, y) here a *weak* solution to the overall system in (2.24). Moreover, by the remark in Sec. 2.1.2 about uniqueness in probability law for such solutions we know that the joint probability distributions of (x, y) here (under $\mathbb{P}$) will be the same as those occurring in (2.24). This is all we need in order to compute a solution to our filtering problem. [5]

3.1.2 A Bayes' formula The basic filtering problem of determining $\pi_t(\phi)$ in (3.1) for the system (2.24) can now be addressed using the following Bayes' formula (a version of the so-called *Kallianpur-Striebel* formula)

$$\pi_t(\phi) = \frac{\mathbb{E}_{\mathbb{P}_0}\big(\phi(x_t)\Lambda_t|\mathcal{F}_t^{(y)}\big)}{\mathbb{E}_{\mathbb{P}_0}\big(\Lambda_t|\mathcal{F}_t^{(y)}\big)}, \tag{3.6}$$

which follows straightforwardly from the properties of $x_t, y_t, \mathbb{P}_0$ and $\mathbb{P}$ (see appendix B). If we for smooth bounded $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ define the process $\sigma_t(\tilde{\phi})$ by

$$\sigma_t(\tilde{\phi}) = \mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_t, y_t)\Lambda_t|\mathcal{F}_t^{(y)}\big) \tag{3.7}$$

we see that it is sufficient to compute $\sigma_t(\phi)$ (with ϕ as in (3.1)) in order to find $\pi_t(\phi)$ since we have

$$\pi_t(\phi) = \frac{\sigma_t(\phi)}{\sigma_t(1)}. \tag{3.8}$$

For obvious reasons, $\sigma_t(\phi)$ is known as the *unnormalized estimate* of ϕ. In what follows we shall frequently consider the case where ϕ and $\tilde{\phi}$ are in fact vector valued functions, in which case we interpret the definitions of π and σ as being applied componentwise so that π, σ become vectors of the corresponding dimensions.

3.2 Recursive Filters

The nonlinear filtering problem outlined in the previous section can be considered to be solved if we find an equation (which can be solved by *some* practical means) for the quantity $\sigma(\phi)$ occurring in (3.8). (Indeed, in an abstract sense, the formulas (3.8)–(3.6) taken together *is* a solution to the filtering problem.) If this equation is in the form of a stochastic differential such that the solution $\sigma_t(\phi)$ depends only of $\sigma_s(\phi)$ and y_s for $s \in [0, t]$ we have moreover a *recursive* solution. Next we shall proceed to find such recursive filters.

3.2.1 The Zakai Equation. By expanding the product $\phi(x_t)\Lambda_t$ (where x is the solution to (2.24) and Λ is defined in (3.4)) using Itô's formula and the properties of conditional expectation it is not very hard to arrive (see appendix C) at the *Duncan-Mortensen-Zakai equation,* or *Zakai equation* for short, on Itô form as [6]

$$\left.\begin{aligned} d\sigma_t(\phi) &= \sigma_t(\mathcal{A}_+\phi)\,dt + \sigma_t\big(\phi h^T H^{-1}\big)\,dy_t, \\ \sigma_0(\phi) &= \mathbb{E}\big(\phi(x_0)\big), \end{aligned}\right\} \quad t \in [0, T], \tag{3.9}$$

[5] We can use the pair (x, y) occurring here as a "model" for the corresponding pair in the "system" (2.24); all we need to do is to reproduce the correct probability distributions.

[6] Since $H(y_t)$ is measurable with respect to $\mathcal{F}_t^{(y)}$ we have moreover $\sigma_t\big(\phi h^T H^{-1}\big) = \sigma_t\big(\phi h^T\big) H^{-1}$.

where $\mathcal{A}_+$ is the operator in (2.23) and h is the function in (3.5). The Zakai equation will be the basis for our developments on finite dimensional filters. In the present setting it can be shown that [7]

$$\sigma_t\big(\phi h^T H^{-1}\big) \circ dy_t = \sigma_t\big(\phi h^T H^{-1}\big)\, dy_t + \sigma_t(\gamma\phi)\, dt, \tag{3.10}$$

where $\gamma : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is a certain function of (x_t, y_t) (an explicit expression for γ is given in appendix C). In the important case $H_0(x, y) = H_0(x), H(y) = \mathbf{I}$ ($\mathbf{I}$ being the identity matrix) considered below the function γ becomes

$$\gamma(x, y) = \frac{1}{2}\|H_0(x)\|_2^2, \quad x \in \mathbb{R}^n, y \in \mathbb{R}^p.$$

In all cases we may thus express the Zakai equation on Stratonovich form as

$$\begin{aligned} d\sigma_t(\phi) &= \sigma_t(\hat{\mathcal{A}}\phi)\, dt + \sigma_t\big(\phi h^T H^{-1}\big) \circ dy_t, \\ \sigma_0(\phi) &= \mathbb{E}\big(\phi(x_0)\big), \end{aligned} \quad t \in [0, T], \tag{3.11}$$

where the operator $\hat{\mathcal{A}}$ is given by

$$\hat{\mathcal{A}} = \mathcal{A}_+ - \gamma = L_{F_0} + \frac{1}{2}\sum_{j=1}^{n} L_{F_j}^2 - \gamma,$$

and γ is the function in (3.10). The properties of the operator $\hat{\mathcal{A}}$ and its adjoint $\hat{\mathcal{A}}^*$, given by

$$\begin{aligned} \hat{\mathcal{A}}^* &= \mathcal{A}_+^* - \gamma \\ &= -\mathrm{tr}(\nabla F_0^T) - L_{F_0} - \gamma \\ &\quad + \frac{1}{2}\sum_{j=1}^{m}\Big(\big(\mathrm{tr}(\nabla F_j^T)\big)^2 + 2\mathrm{tr}(\nabla F_j^T) L_{F_j} + L_{F_j}\mathrm{tr}(\nabla F_j^T) + L_{F_j}^2\Big), \end{aligned} \tag{3.12}$$

will be in the center of our subsequent investigations on the geometrical aspects of nonlinear filtering.

3.2.2 Density form of the Zakai equation.

It is a standard fact (cf. [18, Sec. 5.3]) that the map $A \mapsto \mathbb{P}(A|\mathcal{F}_t^{(y)})$ on the Borel sets of $\mathbb{R}^n$ given by (3.1) for $\phi(x) = \mathbf{1}_A(x)$ defines, for almost all (under $\mathbb{P}$) paths $y_{(\cdot)}$, a probability measure. In case this probability measure has a Lebesgue density then so has the measure $A \mapsto \mathbb{E}_{\mathbb{P}_0}(\mathbf{1}_A\Lambda_t|\mathcal{F}_t^{(y)})$ (and vice versa). If the density of the latter measure is denoted $q_t^{(y)}$ we thus have

$$\sigma_t(\phi) = \mathbb{E}_{\mathbb{P}_0}\big(\phi(x_t)\Lambda_t|\mathcal{F}_t^{(y)}\big) = \int_{\mathbb{R}^n} \phi(x) q_t^{(y)}(x)\, dx. \tag{3.13}$$

The function $q_t^{(y)}$ is known as the *unnormalized density* of the filtering problem. By an approximation argument (see appendix D) it follows that if $\tilde{\phi}$ is as in (3.7) we have moreover

$$\mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_t, y_t)\Lambda_t|\mathcal{F}_t^{(y)}\big) = \int_{\mathbb{R}^n} \tilde{\phi}(x, y_t) q_t^{(y)}(x)\, dx \tag{3.14}$$

[7] Note that the Zakai equation is not an SDE of the form (2.12) (so we cannot use the formulas (2.16), (2.17)); it is merely a stochastic differential representation for $\sigma_t(\phi)$.

and (3.9) implies (by "duality") the following Itô stochastic partial differential equation (SPDE)

$$dq_t^{(y)}(x) = \mathcal{A}_+^* q_t^{(y)}(x)\,dt + q_t^{(y)}(x)h^T(x,y_t)H^{-1}(y_t)\,dy_t,$$
$$q_0^{(y)}(x) = q_0(x),$$
$$t \in [0,T], x \in \mathbb{R}^n, \quad (3.15)$$

where q_0 is the initial density ("unconditional" unnormalized probability density for the state x_0, before any observations have been made). Likewise, (3.11) implies the Stratonovich SPDE

$$dq_t^{(y)}(x) = \hat{\mathcal{A}}^* q_t^{(y)}(x)\,dt + q_t^{(y)}(x)h^T(x,y_t)H^{-1}(y_t)\circ dy_t,$$
$$q_0^{(y)}(x) = q_0(x),$$
$$t \in [0,T], x \in \mathbb{R}^n. \quad (3.16)$$

It is well-known (see e.g. [7], [25]) that (at least) for $\tilde{\phi} = \phi$ (with ϕ as in the previous section, i.e. no explicit y dependence in $\tilde{\phi}$) and $q_0 = \delta_{x_0}$ (Dirac delta) the SPDE in (3.16) (or equivalently (3.15)) has a unique solution $q_t^{(y)}$ (nonnegative) for $t \in (0,T]$ which is of class C^2 and rapidly decaying for $\|x\| \to \infty$.

3.2.3 Robust Version of the Zakai Equation. The path $y_{(\cdot)}$ enters into (3.16) in a way which is not continuous in the sup-norm. In real-life time (and space) discretized implementations of the solution to the filtering this might be a problem and therefore it is preferable to express (3.16) on a "robust" form where the dependence on the path is continuous. By introducing a change of variables ("gauge transformation")

$$\varsigma_t^{(y)}(x) = \exp\big(-\tilde{h}(x,y_t)\big)q_t^{(y)}(x)$$

where $\tilde{h} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is function such that

$$\nabla_y \tilde{h}(x,y) = \big(H^{-1}(y)\big)^T h(x,y), \quad y \in \mathbb{R}^p \quad (3.17)$$

and applying Itô's formula, one obtains after some straightforward calculations (see appendix E)

$$d\varsigma_t^{(y)}(x) = \exp\big(-\tilde{h}(x,y_t)\big)\hat{\mathcal{A}}^* \exp\big(\tilde{h}(x,y_t)\big)\varsigma_t^{(y)}(x)\,dt,$$
$$\varsigma_0^{(y)}(x) = \exp\big(-\tilde{h}(x,0)\big)q_0(x),$$
$$t \in [0,T], x \in \mathbb{R}^n,$$

(recall that $y_0 = 0$ in (2.24)) or equivalently

$$\begin{aligned} \big(\frac{\partial}{\partial t} - \tilde{\mathcal{A}}_t\big)\varsigma_t^{(y)} &= 0, \\ \varsigma_0^{(y)}(x) &= \exp\big(-\tilde{h}(x,0)\big)q_0(x), \end{aligned} \qquad t \in [0,T], x \in \mathbb{R}^n, \quad (3.18)$$

where the (time dependent) operator $\tilde{\mathcal{A}}_t$ is given by

$$\tilde{\mathcal{A}}_t\phi(x) = \exp\big(-\tilde{h}(x,y_t)\big)\hat{\mathcal{A}}^* \exp\big(\tilde{h}(x,y_t)\big)\phi(x),$$
$$t \in [0,T], x \in \mathbb{R}^n, \quad (3.19)$$

for smooth $\phi : \mathbb{R}^n \to \mathbb{R}$. The positive definiteness of the diffusion matrix a_F ensures that equation (3.18) is a parabolic partial differential equation (PDE) and thus has all the well-known properties of such equations. [8] The big difference between (3.16) and (3.18), however, is that in the latter equation the path $y_{(\cdot)}$ enters "functionally" (as a "parameter") rather than in terms of its increments (in a stochastic integral). Thus, the solutions to (3.18) depend continuously on the path and therefore (3.18) represents a robust solution to the filtering problem.

3.2.4 Lie Form of the (Robust) Zakai Equation If we define the differential operator $L_t^{(1)}$ by

$$L_t^{(1)}\phi(x) = \tilde{h}(x, y_t)\phi(x), \quad t \in [0, T], \tag{3.20}$$

(where $\tilde{h}$ is the function in (3.17)) and for ease of notation also put

$$L_t^{(0)} = \hat{\mathcal{A}}^* \tag{3.21}$$

(where $\hat{\mathcal{A}}^*$ is the operator in (3.12)) then the operator $\tilde{\mathcal{A}}_t$ in (3.19) can be written

$$\tilde{\mathcal{A}}_t = \exp\big(-L_t^{(1)}\big)L_t^{(0)}\exp\big(L_t^{(1)}\big).$$

Now, if we define the related operator Ψ by

$$\Psi(\tau) = \exp\big(-\tau L_t^{(1)}\big)L_t^{(0)}\exp\big(\tau L_t^{(1)}\big), \quad \tau \in \mathbb{R},$$

we have $\Psi(1) = \tilde{\mathcal{A}}_t$ and by differentiation we obtain

$$\begin{aligned}\frac{d}{d\tau}\Psi(\tau) &= \exp\big(-\tau L_t^{(1)}\big)\big(-L_t^{(1)}L_t^{(0)} + L_t^{(0)}L_t^{(1)}\big)\exp\big(\tau L_t^{(1)}\big)\\ &= \exp\big(-\tau L_t^{(1)}\big)[L_t^{(0)}, L_t^{(1)}]\exp\big(\tau L_t^{(1)}\big).\end{aligned}$$

Repeated differentiation gives

$$\frac{d^k}{d\tau^k}\Psi(\tau) = \exp\big(-\tau L_t^{(1)}\big)\mathrm{ad}^k_{L_t^{(1)}}L_t^{(0)}\exp\big(\tau L_t^{(1)}\big), \quad k \in \mathbb{N},$$

where $\mathrm{ad}^k_{L_t^{(1)}}$ is defined recursively by

$$\mathrm{ad}^{k+1}_{L_t^{(1)}}L_t^{(0)} = [\mathrm{ad}^k_{L_t^{(1)}}L_t^{(0)}, L_t^{(1)}], \quad k \in \mathbb{N},$$

with $\mathrm{ad}^0_{L_t^{(1)}}$ being the identity. This shows that the map $\tau \mapsto \Psi(\tau)$ is C^∞ and since $L_t^{(1)}$ is a differential operator of degree 0, and $L_t^{(0)}$ is a (linear) differential operator of degree 2, it follows that (see e.g. [13, pp. 22–23])

$$\mathrm{ad}^k_{L_t^{(1)}}L_t^{(0)} = 0, \quad k > 2. \tag{3.22}$$

Hence, the Taylor series for $\Psi(\tau)$ contains only three terms and we have

$$\begin{aligned}\tilde{\mathcal{A}}_t &= \Psi(1)\\ &= \Psi(0) + \frac{d}{d\tau}\Psi(\tau)|_{\tau=0} + \frac{1}{2}\frac{d^2}{d\tau^2}\Psi(\tau)|_{\tau=0}\\ &= L_t^{(0)} + [L_t^{(0)}, L_t^{(1)}] + \frac{1}{2}[[L_t^{(0)}, L_t^{(1)}], L_t^{(1)}].\end{aligned}$$

[8] This is the only place where this condition on the diffusion matrix is used. A standard case where the condition would be violated is when some of the vector fields $F_0, F_1, \ldots, F_m$ and $H_0, H_1, \ldots, H_p$ are time-varying, and this is modeled by adding an extra state variable representing time. In this case the diffusion matrix would however have an upper left block satisfying the condition and we would still get a parabolic PDE, albeit with time-varying coefficients.

The robust form (3.18) of the Zakai equation can now be expressed on Lie form as

$$\frac{d\varsigma_t^{(y)}(x)}{dt} = L_t^{(0)}\varsigma_t^{(y)}(x) + [L_t^{(0)}, L_t^{(1)}]\varsigma_t^{(y)}(x) + \frac{1}{2}[[L_t^{(0)}, L_t^{(1)}], L_t^{(1)}]\varsigma_t^{(y)}(x),$$
$$\varsigma_0^{(y)}(x) = \exp\big(-\tilde{h}(x,0)\big)q_0(x),$$
$$t \in [0,T], x \in \mathbb{R}^n. \quad (3.23)$$

It is worth noting here that by the same argument that was alluded to in connection with (3.22) above the operator $[L_t^{(0)}, L_t^{(1)}]$ is a first degree (linear) differential operator and $[[L_t^{(0)}, L_t^{(1)}], L_t^{(1)}]$ is an operator of degree zero, i.e. a (possibly time-varying) smooth function (in x).

3.3 Finite Dimensional Filters

The unnormalized density $q_t^{(y)}$ in (3.16), and its transformed version $\varsigma_t^{(y)}$ in (3.23), are both objects in $L_1(\mathbb{R}^n)$, which is an infinite dimensional vector space. Thus, direct solution of (3.16) or (3.23) is in general not feasible, at least not in real-time. However, if there exists a smooth manifold $\mathbb{M}$ with a differential equation

$$\frac{d\xi(t)}{dt} = b(\xi(t), y_t), \quad t \in [0,T], \qquad \xi(0) = \tilde{\xi}, \quad (3.24)$$

where b is a smooth vector field on $\mathbb{M} \times \mathbb{R}^p$, and a smooth "output" function $\theta : \mathbb{M} \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\theta(\xi(t), t, x) = \varsigma_t^{(y)}(x), \quad t \in [0,T], x \in \mathbb{R}^n,$$

then we say that we have a *robust* [9] *finite dimensional filter* (FDF) for the filtering problem in (3.1). The number of examples for which an FDF is known is relatively small; it includes the *Kalman filter,* the *Beneš filter* and a few other cases.

A special case of (3.24) that will be of particular interest to us in the following is when b can be factored as (for some $k > 0$)

$$b(p, y) = \sum_{j=1}^{k} c_j(y) b_j(p), \quad p \in \mathbb{M}, y \in \mathbb{R}^p,$$

where $b_j : \mathbb{M} \to \mathbb{R}$ are smooth vector fields and $c_j : \mathbb{R}^p \to \mathbb{R}$ are smooth functions. When b is on this form the so-called Wei-Norman technique can (sometimes) be applied to explicitly construct FDFs.

3.4 Observations in additive "white noise"

For the special case

$$H_0(x, y) = H_0(x), \quad H(y) = \mathbf{I}, \quad (3.25)$$

($\mathbf{I}$ being the identity matrix) the question of existence, characterization and conditions for FDFs has been given a fairly complete answer (see e.g. [26] and the references therein). In the case (3.25) we can take $\tilde{h}$ in (3.17) as

$$\tilde{h}(x, y) = H_0^T(x) y, \quad x \in \mathbb{R}^n, y \in \mathbb{R}^p$$

[9] If we instead have an SDE on $\mathbb{M}$ analogous to (3.24) with an output function θ producing the values of $q_t^{(y)}$ we say that we have a *finite dimensional filter*.

and the operator $L_t^{(1)}$ in (3.20) can be represented in terms of a sum

$$L_t^{(1)} = \sum_{j=1}^{p} y_t^j L_j$$

where L_j is defined as

$$L_j \phi(x) = H_0^j(x)\phi(x), \quad x \in \mathbb{R}^n, j = 1, \ldots, p,$$

for smooth $\phi : \mathbb{R}^n \to \mathbb{R}$. The operator $L_t^{(0)}$ in (3.21) also simplifies in the case (3.25); it reduces to the (time invariant) operator $L^{(0)}$ given by

$$L^{(0)} = \hat{\mathcal{A}}^* = \mathcal{A}_+.$$

The Lie form of the (robust) Zakai equation (3.23) can therefore for the case (3.25) be written as

$$\begin{aligned} \frac{d\varsigma_t^{(y)}(x)}{dt} &= L^{(0)}\varsigma_t^{(y)}(x) + \sum_{j=1}^{p} y_t^j [L^{(0)}, L_j]\varsigma_t^{(y)}(x) \\ &\quad + \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p} y_t^j y_t^k [[L^{(0)}, L_j], L_k]\varsigma_t^{(y)}(x), \qquad t \in [0,T], x \in \mathbb{R}^n. \\ \varsigma_0^{(y)}(x) &= q_0(x), \end{aligned} \tag{3.26}$$

Again, since $L_1, \ldots, L_p$ are differential operators of degree zero the operator $[L^{(0)}, L_j]$ is of first degree and the operator $[[L^{(0)}, L_j], L_k]$ is of degree zero. Moreover, all the operators $L^{(0)}, L_1, \ldots, L_p$ here are *deterministic and time invariant;* the time varying and stochastic parts of the right hand side of (3.26) are confined to $\varsigma_t^{(y)}$ and y_t.

For the setting in (3.26), *Brockett's conjecture* is that a necessary condition for the existence of an FDF is that the *estimation algebra* $\mathcal{E}$, defined as the Lie algebra generated by the differential operators $L^{(0)}, L_1, \ldots, L_p$, is finite dimensional. Brockett's conjecture has been verified in a number of cases, including the Kalman and Beneš cases mentioned above.

3.4.1 The Wei-Norman Technique The Wei-Norman technique for constructing explicit solutions to linear differential equations on a manifold can be applied to obtain a concrete representation of an FDF. We shall here briefly sketch how this is done; a more detailed account can be found in [29].

Assume that the estimation algebra $\mathcal{E}$ of the system (2.24), for the special case in (3.26), is finite dimensional and has a basis consisting of the differential operators $E_0 = L_t^{(0)}$ and $E_1, \ldots, E_k$ (for some $k > 0$), the latter of the form

$$\sum_{j=1}^{n} \alpha_{i,j} D_j + \beta_j,$$

where

$$D_j = \frac{\partial}{\partial x^j} - F_0^j,$$

and the $\alpha_{i,j}$'s are constants and the β_j's polynomials in the components of the state variable x, and zero degree differential operators $E_{k+1}, \ldots, E_\ell$ (for some ($\ell > k$)). Moreover, assume that $[E_i, E_j]$ is a constant for $i \geq 1, 1 \leq j \leq k$, and that all the zero degree differential operators in $\mathcal{E}$ are in the span of $E_{k+1}, \ldots, E_\ell$. Then the (robust) Zakai equation on Lie form (3.26) can be represented as

$$\varsigma_t^{(y)}(x) = \exp\big(r_\ell(t)E_\ell\big) \cdots \exp\big(r_1(t)E_1\big)\exp(tE_0)q_0(x), \quad t \in [0,T], x \in \mathbb{R}^n,$$

where $r_j : [0,T] \to \mathbb{R}$ and the r_j's satisfy ordinary differential equations.

4. An algebraic approach to geometry

In this chapter some basic differential geometric concepts and notation are quickly introduced, leaving aside many points of rigor and finer details. For a comprehensive presentation of this material, see [30], [1], [12], [13], [23]. The main point made in this chapter is that all relevant constructions can be made without explicit use of any coordinate system.

4.1 Differentiable Manifolds

An n-dimensional *topological manifold* is a (second countable Hausdorff) topological space M such that every point in M has a neighborhood homeomorphic to an open set in $\mathbb{R}^n$. The family of such local diffemorphisms ('coordinate systems' or 'charts') is called a topological atlas for M. A differentiable manifold is a topological manifold together with a preferred subatlas (a 'differentiable atlas') being such that the transition mapping from one coordinate system to another is smooth (C^∞). A mapping $M \to \mathbb{R}$ is then said to be *smooth* whenever it is represented by a smooth function in every coordinate system belonging to the preferred subatlas. Denote by $C^\infty(M)$ the algebra of smooth mappings $M \to \mathbb{R}$. One refers to the subatlas, or alternatively the algebra $C^\infty(M)$, as a *differential structure* of the topologial manifold M.

There are similar definitions of C^k manifolds and real-analytic (C^ω) manifolds, but much of the algebraic approach given below does not work in the case of $C^k$and C^ω manifolds.

4.2 Basic constructions

One approach to coordinate-free differential geometry consists in regarding the algebra $C^\infty(M)$ as the basic object. Everything else, including the manifold M itself and its topology, may then be reconstructed from the algebra structure of $C^\infty(M)$ alone.

4.2.1 Reconstruction of the function algebra structure The *points* of M are identified with the maximal ideals of $C^\infty(M)$, *viz.* the point $p \in M$ is identified with the maximal ideal $\mathcal{I}_p = \{\varphi \in C^\infty(M) \,|\, \varphi(p) = 0\}$. The maximal ideals all have codimension one, and $C^\infty(M) = \mathcal{I}_p \oplus (C^\infty(M)/\mathcal{I}_p)$ where $C^\infty(M)/\mathcal{I}_p = \mathbb{R}$ is identified with the constant functions, *i.e.* the subalgebra of $C^\infty(M)$ generated by its identity element.

The *value* $\varphi(p)$ of a function φ is equal to the element $\varphi/\mathcal{I}_p$ in $\mathbb{R}$. The 'abstract algebra' $C^\infty(M)$ is thereby realized as an algebra of functions, and the topology of M is recovered as the topology generated by those functions' preimages of open intervals. In particular, the subalgebra $C_0^\infty(M)$ of functions with compact support is well defined from the algebra structure of $C^\infty(M)$ alone. (The *support* suppϕ, is the closure of the set of points where the value of ϕ differs from 0.)

4.2.2 Smooth mappings A mapping $\Theta : M \to N$ is said to be smooth ($\Theta \in C^\infty(M, N)$) if the composition $\phi \circ \Theta \in C^\infty(M)$ for every $\phi \in C^\infty(N)$. For a smooth mapping $\Theta : M \to N$, the *pull-back* mapping $\Theta^* : C^\infty(N) \to C^\infty(M)$ defined by composition is an algebra homomorphism, and conversely any algebra homomorphism

$C^\infty(N) \to C^\infty(M)$ is the pull-back of a smooth mapping. The smooth mapping $\Theta : M \to N$ is called a *diffeomorphism* if it is invertible and if the inverse is smooth. This is the case if and only if Θ^* is an algebra isomorphism.

4.2.3 Product manifolds The *product manifold* differential structure $C^\infty(M \times N)$ can be characterized in terms of the differential structures $C^\infty(M)$ and $C^\infty(N)$ through the condition that $C^\infty(M \times N)$ is the largest function algebra on $M \times N$ such that the product mapping $\Theta_M \times \Theta_N : P \to M \times N$ is smooth whenever its components $\Theta_M : P \to M$ and $\Theta_N : P \to N$ are smooth.

4.2.4 Tangent vectors A *tangent vector* at $p \in M$ is an linear operator $X_p : C^\infty(M) \to \mathbb{R}$ satisfying

$$X_p(\varphi\psi) = \varphi(p)X_p(\psi) + \psi(p)X_p(\varphi) \tag{4.1}$$

The set of such tangent vectors forms a vector subspace of the space of real valued linear operators on $C^\infty(M)$. This subspace is called the tangent space of M at p, and is is denoted by T_pM. The (disjoint) union of all tangent spaces of M is denoted by TM. The projection mapping $\pi_{TM} : TM \to M$ associates elements of the different T_pM to their respective base points p.

A vector field on M is a derivation X of the algebra $C^\infty(M)$, *i.e.* a mapping $X : C^\infty(M) \to C^\infty(M)$ such that

$$X(\varphi\psi) = \varphi X(\psi) + \psi X(\varphi) \tag{4.2}$$

The space of vector fields on M is denoted by $\mathfrak{X}(M)$. If X and Y are vector fields, then their commutator

$$[X, Y] = X \circ Y - Y \circ X \tag{4.3}$$

is also a vector field. (The fact that the commutator of two derivations is a derivation holds for any commutative algebra.)

Let $X \in \mathfrak{X}(M)$. For every $p \in M$, the mapping $\varphi \mapsto X(\varphi)(p)$ satisfies (4.1) and thus defines an element in $X_p \in T_pM$. In this way the vector field X defines (and may in fact be identified with) a mapping $X : M \to TM$ such that $\pi_{TM} \circ X = id_M$. Now define $C^\infty(TM)$ as the algebra of real valued functions on TM such that $\psi \in C^\infty(TM)$ if and only if $\psi \circ X \in C^\infty(M)$ for all $X \in \mathfrak{X}(M)$. This provides *'the tangent bundle' TM* with a differential structure.

4.2.5 Fiber bundles A *fiber bundle* $\pi_P : P \to M$ over a manifold M is a manifold P together with a smooth surjective mapping $\pi_P : P \to M$ such that P is locally diffeomorphic to a product manifold, i.e. every $p \in M$ has an open neighborhood U such that $\pi_P^{-1}(U)$ is diffeomorphic to $U \times \pi_P^{-1}(p)$. In particular, the *fibers* $\pi_P^{-1}(p)$ are diffeomorphic manifolds.

Let $\pi_P : P \to M$ be a fiber bundle. A smooth mapping $\sigma : M \to P$ such that $\pi_P \circ \sigma = id_M$ is called a (smooth) *section* of the bundle. The set of such is denoted by $\Gamma(M, P)$. Note that $C^\infty(M) = C^\infty(M, \mathbb{R}) = \Gamma(M, M \times \mathbb{R})$. The differential structure $C^\infty(P)$ may also be characterized by the space of smooth sections $\Gamma(M, P)$ through the condition that $\psi \in C^\infty(P)$ if and only if $\psi \circ \phi \in C^\infty(M)$ for all $\phi \in \Gamma(M, P)$.

If the fibers are (finite dimensional) vector spaces, the fibre bundle is called a *vector bundle*. The tangent bundle is a vector bundle.

If $\pi_E : E \to M$ is a vector bundle, the *dual vector bundle* $\pi_{E^*} : E^* \to M$ is the (disjoint) union of the dual vector spaces together with the obvious projection mapping. The differential structure of E^* is implicitly defined by demanding that for any $a \in \Gamma(M, E)$ and any $\alpha \in \Gamma(M, E^*)$, the evaluation mapping $\alpha(a) \in C^\infty(M)$. (The evaluation mapping maps the point $p \in M$ to the number $\alpha_p(a_p)$, where α_p and a_p belong to the dual vector spaces that are fibers of E and E^* over the point p.)

4.2.6 Pullback and push forward Let $\Theta : M \to N$ be smooth. Recall the *pull-back* mapping $\Theta^* : C^\infty(N) \to C^\infty(M)$ defined by composition. The *tangent mapping* ('push-forward' or 'differential') of Θ is the smooth mapping $\Theta_* : TM \to TN$ characterized by

$$\Theta_* X_p(\phi) = X_p(\Theta^* \phi)$$

identically in $\phi \in C^\infty(N)$.*

Note that the pullback mapping acts on sections while the push-forward mapping acts on the fiber bundle itself. The pullback and push-forward mappings also behave differently with respect to composition: if $\Theta = \Theta_2 \circ \Theta_1$, then $\Theta_* = \Theta_{2*} \circ \Theta_{1*}$ but $\Theta^* = \Theta_1^* \circ \Theta_2^*$.

4.2.7 Jet bundles The *k-jet* of a function $\phi \in C^\infty(M)$ *at the point* $p \in M$ is the element $j_p^k \phi = \phi / \mathcal{I}_p^{k+1}$ in the *jet space* $\mathcal{J}_p^k(M) = C^\infty(M)/\mathcal{I}_p^{k+1}$. The jet space over p, being the quotient of an algebra by an ideal, has a natural algebra structure and in particular it has the structure of a real vector space. The (disjoint) union $\mathcal{J}^k(M)$ of the $\mathcal{J}_p^k(M)$, $p \in M$ has a natural projection mapping $\pi_{J^k} : \mathcal{J}^k(M) \to M$, which maps the elements of the different $\mathcal{J}_p^k(M)$ on the corresponding p.

The *k-prolongation* of a function $\phi \in C^\infty(M)$ is the mapping $j^k\phi : M \to \mathcal{J}^k(M)$ which maps p on $j_p^k \phi$.

A mapping $\Psi : \mathcal{J}^k(M) \to \mathbb{R}$ is said to be *smooth* whenever all its composition with k-prolongations belong to $C^\infty(M)$. This provides $\mathcal{J}^k(M)$ with a differential structure such that the projection π_{J^k} is smooth, and hence $\pi_{J^k} : \mathcal{J}^k(M) \to M$ is a vector bundle over M, which is called the *k-jet bundle*.

It follows from the construction that there are surjective mappings $\mathcal{J}^\ell(M) \to \mathcal{J}^k(M)$, whenever $\ell \geq k$.

Dual jet bundles Let $\Theta : M \to N$ and be smooth and $\phi \in C^\infty(N)$. It follows directly from the definition of the k-prolongation, that $j_p^k \Theta^* \phi$ at $p \in M$ depends on ϕ only through the value of $j_{\Theta(p)}^k \phi$, and hence that there is a well-defined pullback mapping $\Theta^* : \Gamma(N, \mathcal{J}^k(N)) \to \Gamma(M, \mathcal{J}^k(M))$ such that

$$\Theta^* \circ j^k = j^k \circ \Theta^*$$

holds as an identity of operators on $C^\infty(N)$.

Dually, there is a push-forward mapping $\Theta_* : \mathcal{J}^{k\,*}(M) \to \mathcal{J}^{k\,*}(N)$ on the dual bundles $\mathcal{J}^{k\,*}(M)$ of the jet bundles given by

$$\Theta_* L_p(K) = L_p(\Theta^* K)$$

as an identity in $p \in M$, $L_p \in \mathcal{J}_p^{k\,*}(M)$ and $K \in \Gamma(N, \mathcal{J}^k(N))$. The restriction of this push-forward mapping Θ_* to $\mathcal{J}_p^{k\,*}(M)$ is called the *k-jet of the mapping* Θ and is also denoted by $j_p^k \Theta$. As usual for push-forward mappings $(\Theta_2 \circ \Theta_1)_* = \Theta_{2*} \circ \Theta_{1*}$ or in another notation

$$j^k(\Theta_2 \circ \Theta_1) = j^k \Theta_2 \circ j^k \Theta_1$$

4.2.8 Cotangent vectors For a function $\phi \in C^\infty(M)$, the element $\phi - \phi(p) \in C^\infty(M)$ sits in $\mathcal{I}_p$ and defines uniquely an element $(d\phi)_p$ of $\mathcal{I}_p/\mathcal{I}_p^2$, called *the (exterior) derivative* of ϕ (at p). It follows from (4.1) that $\mathcal{I}_p/\mathcal{I}_p^2$ is the dual space of T_pM and that $X_p(\phi) = (d\phi)_p(X_p)$. The space $\mathcal{I}_p/\mathcal{I}_p^2$ is called the *cotangent space* of M at p and is denoted by T_p^*M. The cotangent spaces are naturally identified with the fibers of the cotangent bundle $\pi_{T^*M} : T^*M \to M$, defined as the the bundle dual to the tangent bundle.

4.3 Tensor algebra

As will be verified in the next chapter, the dimensions of the tangent spaces and the cotangent spaces coincide with the dimension of the manifold itself. From the tangent and cotangent vector spaces (T_pM and T_p^*M respectively) at the point $p \in M$, the space $T_p^{(k,\ell)}M$ of (k,ℓ)-tensors p may be constructed as

$$T_p^{(k,\ell)}M = T_pM \otimes .. \otimes T_pM \otimes T_p^*M \otimes .. \otimes T_p^*M$$

where there are k factors T_pM and ℓ factors T_p^*M. These spaces form the fibers of the corresponding tensor bundles. Elements of $T^{(k,0)}M$ are pushed forward by mappings, while sections of $T^{(0,\ell)}M$ are pulled back.

Completely skew-symmetric $(0,\ell)$-tensors form a subbundle (later to be denoted by $\Lambda^\ell M$) of $T^{(0,\ell)}M$. The sections of this subbundle may be identified with the differential forms defined in another way below.

4.3.1 Exterior algebra For a general n-dimensional real vector space V, we define $\Lambda^\ell V$ as the space of skew symmetric multilinear mappings

$$V \times .. \times V \to \mathbb{R}$$

(k factors V). There is a bilinear operation $\wedge : \Lambda^a V \times \Lambda^b V \to \Lambda^{a+b} V$ defined by

$$\alpha \wedge \beta \left(v_1, .., v_{a+b}\right) = \frac{1}{a!b!} \sum_{\sigma \in S_{a+b}} sgn\left(\sigma\right) \ \alpha\left(v_{\sigma(1)}, .., v_{\sigma(a)}\right) \ \beta\left(v_{\sigma(a+1)}, .., v_{\sigma(a+b)}\right)$$

where $v_1, .., v_{a+b} \in V$ and S_{a+b} is the symmetric group (*i.e.* the permutation group).

With this definition the wedge product $\wedge$ becomes

- associative, $\alpha \wedge (\beta \wedge \gamma) = (\alpha \wedge \beta) \wedge \gamma$,
- graded commutative, $\alpha \wedge \beta = (-1)^{ab} \beta \wedge \alpha$

In particular, associativity is due to our choice of normalization factors.

It is readily seen that the dimension of the space $\Lambda^a V$ is the binomial coefficient $\frac{n!}{a!(n-a)!}$, so that $\Lambda^n V$ is 1-dimensional and the $\Lambda^m V$ are trivial for $m > n$.

There is also an algebraic operation, the interior product, $i_A \alpha \in \Lambda^{a-1} V$, between an element $A \in V$ and an element $\alpha \in \Lambda^a V$. This operation is uniquely characterized by the following properties

$$\begin{aligned} i_A\alpha &= 0 \text{ when } a = 0 \\ i_A\alpha &= \alpha\left(A\right) \text{ when } a = 1 \\ i_A\left(\alpha \wedge \beta\right) &= i_A\alpha \wedge \beta + (-1)^a \alpha \wedge i_A\beta \\ i_A \circ i_A &= 0 \end{aligned}$$

We will use exterior algebra only in the case $V = T_pM$, the sections of the corresponding bundle $\Lambda^\ell M$ then being referred to as differential forms.

4.3.2 Fractional density algebras Consider a 1-dimensional vector space V (later $V = \Lambda_p^n M$). From this space, two important families of 1-dimensional vector spaces may be formed,

- V^k with $k \in \mathbb{Z}$ and
- $|V|^\kappa$ with $\kappa \in \mathbb{R}$.

Since the trace mapping

$$tr : V \otimes V^* \to \mathbb{R}$$

is an isomorphism in the case of a 1-dimensional V, the tensor space $V^{(k+1,\ell+1)}$ is naturally isomorphic with the tensor space $V^{(k,\ell)}$, and hence we may unambiguously write V^{k-l} for $V^{(k,\ell)}$. With this convention $V^a \otimes V^b = V^{a+b}$ holds for integer a and b of arbitrary signs.

The 1-dimensional vector space V^a has a natural orientation when $a = 2k$ is even; a nonzero element α of V^{2k} constitutes a positively oriented basis whenever it can be represented as $\beta \otimes \beta$ with β in V^k.

Let κ be real and denote by $|V|^\kappa$ the linear space of mappings $\psi : V^* \to \mathbb{R}$ satisfying $\psi(z\, u) = |z|^\kappa \psi(\, u)$ identically in $z \in \mathbb{R}$ and $u \in V^*$. This family is closed under tensor products if we identify the element $\psi_1 \otimes \psi_2$ of $|V|^{\kappa_1} \otimes |V|^{\kappa_2}$ with the product of ψ_1 and ψ_2 as functions, which is a function in $|V|^{\kappa_1+\kappa_2}$. We refer to the family $|V|^\kappa$ ($\kappa \in \mathbb{R}$) as the *fractional density algebra* generated by V, but the exponents κ need by no means be rational numbers. The vector spaces $|V|^\kappa$ all come with a natural orientation; a nonzero element α of $|V|^\kappa$ constitutes a positively oriented basis whenever it is a positive function on V^*. The space of *oriented scalars* constructed from V is the space $\mathbb{R}_V = V \otimes |V|^{-1}$.

There are then natural identifications

$$\begin{aligned} V^{2k} &= |V|^{2k} \\ V^{2k+1} &= \mathbb{R}_V \otimes |V|^{2k+1} \end{aligned}$$

These constructions are of interest in differential geometry with the particular 1-dimensional vector space $V = \Lambda^n_p M$.

4.3.3 Tensor bundles Via the constructions above, vector bundles like $T^{(k,\ell)}M$, $\Lambda^\ell M$ and $\mathbb{R}_{\Lambda^m M} M$ are formed by fiberwise algebraic operations (tensor products etc.). Alternatively one may consider the corresponding algebraic operations on the spaces of sections. These two approaches give the same resulting bundles, as has been verified in [14].

4.4 Analysis on manifolds

4.4.1 Linear differential operators A linear differential operator of order 0 on $C^\infty(M)$ is by definition (multiplication by) an element of $C^\infty(M)$. Differential operators of higher order are defined recursively: a linear mapping $L : C^\infty(M) \to C^\infty(M)$ is a differential operator of order at most k if the commutator $[L, \chi]$ is a differential operator of order at most $k-1$ for every differential operator χ of order 0. A linear differential operator is said to be 'pure' if it vanishes on constant functions. Thus, a pure first order linear differential operator is the same thing as a vector field, and a general first order linear differential operator is uniquely decomposed as the sum of a vector field and a zeroth order operator. For linear differential operators of order $k > 1$, there is however no natural decomposition into terms of 'exact order' $\ell = 0, 1, ..k$. (Such decompositions will depend on the choice of a coordinate system or, more generally, on the choice of an affine connection.)

Linear differential operators are *local* in the sense that if L is a linear differential operator on $C^\infty(M)$, then

$$\mathrm{supp} L\phi \subseteq \mathrm{supp}\phi$$

for every $\phi \in C^\infty(M)$. (This property in fact characterizes linear partial differential operators on a compact manifold. On a non-compact manifold it is still true that a local operator is locally given by a differential operator, but there may be no global bound on the order.)

The support, $\mathrm{supp} L$, of the linear differential operator L is defined as the closure of $\cup_{\phi \in C^\infty(M)} \mathrm{supp} L\phi$.

4.4.2 Jet bundle characterization of differential operators A linear differential operator $L : C^\infty(M) \to C^\infty(M)$ of order k also has the interpretation as a fiberwise linear form on $\mathcal{J}^k(M)$, and is hence a section in the dual jet bundle $\mathcal{J}^{k\,*}(M)$. It is easily verified that the support, supp$L$, as defined above coincides with the support of L considered as a section of $\mathcal{J}^{k\,*}(M)$.

4.4.3 Principal symbols Let χ be a zeroth order differential operator. The commutator mapping $ad_{-\chi} = [\cdot, \chi]$ maps order $\leq k$ differential operators L on order $\leq k-1$ differential operators $[L, \chi]$ and consequently $ad^k_{-\chi}L$ is a zeroth order operator. Due to the Jacobi identity, $ad_{-\chi_1}\, ad_{-\chi_2} ..\, ad_{-\chi_k} L$ is completely symmetric in χ_1, χ_2, .. χ_k . It is also clear that it depends on χ_1 only via its differential $d\chi_1$. From this follows that $ad^k_{-\chi}L = \sigma_L(d\chi)$, where σ_L, the principal symbol of L, is a k:th order homogeneous polynomial function on T^*M. The differential operator L is said to be *elliptic* if $\sigma_L \neq 0$ away from the zero section of T^*M.

4.4.4 Exterior calculus Differential operators between vector bundles are defined analogously to scalar differential operators, in terms of commutators. There is a natural linear differential operator d on differential forms, mapping sections of $\Lambda^k M$ to sections of $\Lambda^{k+1}M$ (for every k) which is characterized by the conditions that $d_{|\Lambda^0 M}$ =ordinary differential and

$$\begin{aligned} d\,(\alpha \wedge \beta) &= d\alpha \wedge \beta + (-1)^a\, \alpha \wedge d\beta \\ d \circ d &= 0 \end{aligned}$$

(where a is the order of α). The operator d is known as the exterior differential.

4.5 Flows and Lie derivatives

4.5.1 Topology on $C^\infty(M)$ The algebra $C^\infty(M)$ has a natural topology given by the seminorms $\phi \mapsto \sup|L\phi|$, where L ranges over linear differential operators with compact support.

This topology is complete, and all linear differential operators are continuous with respect to it, as are pull-backs etc.

4.5.2 The flow of a vector field Let X be a vector field. A one-parameter family of diffeomorphisms $\Phi^t : M \to M$, $(t \in \mathbb{R})$, with $\Phi^0 = \mathrm{id}_M$, is said to be a (global) *flow* of X if the following equation holds

$$\frac{d\Phi^{t*}}{dt} = X \circ \Phi^{t*}$$

The derivative on the left hand side of this equation is defined in terms of the $C^\infty(M)-$topology defined above. Whenever a vector field has a flow, it is unique, and the existence of a flow is guaranteed if X has compact support. The flow of X is denoted by Φ^t_X (another popular notation is $exp(Xt)$)

4.5.3 Lie derivatives The pullback operator Φ^{t*}_X is well defined for other objects than scalar functions, e.g. for vector and tensor fields, jet bundle sections etc. The Lie derivative along X is the generic differential operator L_X acting on such objects such that

$$\frac{d\Phi^{t*}}{dt} = L_X \circ \Phi^{t*}$$

holds identically. In particular

$$L_X Y = [X, Y]$$

for a vector field Y and

$$L_X \alpha = d i_X \alpha + i_X d\alpha$$

for a differential form α.

4.5.4 Riemannian and subriemannian structures A subriemannian structure on a manifold M is a section g of $T^{(2,0)}M$ such that g considered as a bilinear function on T^*M is symmetric and nonnegative. The subriemannian structure is riemannian if it is positive definite.

4.6 Affine Connections

An affine connection on a manifold M is a vector field valued bilinear operator ∇ on the space of vector fields $\mathfrak{X}(M)$ satisfying

$$\begin{aligned} \nabla_{fX+Y} Z &= f\nabla_X Z + \nabla_Y Z \\ \nabla_Z (fX+Y) &= f\nabla_Z X + \nabla_Z Y + X\ i_Z df \end{aligned}$$

identically in $X, Y, Z \in \mathfrak{X}(M)$ and $f \in C^\infty(M)$.

Let ∇ be an affine connection. Then the expressions

$$T^\nabla(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]$$

and

$$K^\nabla(X,Y)\,Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

are both $C^\infty(M)$-linear in each argument $X, Y, Z \in \mathfrak{X}(M)$ and thereby define tensorial quantities T^∇ and K^∇, the torsion and curvature, respectively, of the affine connection ∇.

4.6.1 Covariant derivatives An affine connection ∇ and a vector field X thus give rise to a differential operator ∇_X on (the sections of) TM, the covariant derivative in the direction X. This differential operator is extended to vector bundles other than TM by postulating the following product rules

$$\begin{aligned} \nabla_X (S \otimes T) &= \nabla_X S \otimes T + S \otimes \nabla_X T \\ i_X d(i_Y \alpha) &= i_{\nabla_X Y}\alpha + i_Y(\nabla_X \alpha) \end{aligned}$$

In particular this gives a meaning to the Hessian $H^\nabla(f)$ satisfying

$$i_X H^\nabla(f) = \nabla_X df$$

which is a bilinear form on TM. The Hessian $H^\nabla(f)$ is symmetric for all f if and only if the connection is torsion free.

4.6.2 Geodesic spray Let ∇ be an affine connection. There is then a naturally defined vector field Z^∇ defined on TM (considered as a $2n$-dimensional manifold). The vector field Z^∇ is characterized by the following properties

$$\begin{aligned} L_{Z^\nabla}(\pi_{TM}^* f) &= df \\ L_{Z^\nabla}(df) &= H^\nabla(f) \end{aligned}$$

for every $f \in C^\infty(M)$. In these formulas, df and $H^\nabla(f)$ are identified with the corresponding polynomial functions on TM. The vector field Z^∇ is called the geodesic spray of ∇, if it has a globally defined flow, the connection is said to be geodesically complete. The projections on M of the integral curves of Z^∇ are called the geodesics of the connection.

4.6.3 Levi-Civita connection Let g be a Riemannian structure on M. There is then a unique torsionfree connection $\overset{(g)}{\nabla}$, called the Levi-Civita connection, that satisfies $\overset{(g)}{\nabla} g = 0$.

5. The coordinate approach to geometry

In this chapter we collect coordinate expression for the different constructs of the preceding chapter.

5.1 Tangent vectors, tensors

Locally the manifold M is described in terms of coordinates $x^1, ..x^n$. By Taylors theorem, a function $f \in C^\infty(M)$ may be written

$$f\left(x^i\right) = f\left(0\right) + x^i h_i\left(x\right)$$

where from now on the summation convention is followed: $x^i h_i\left(x\right)$ is short hand for $\sum_{i=1}^{n} x^i h_i\left(x\right)$, summation being tacitly understood over all repeated indices. From the Taylor theorem representation it is readily seen that the tangent vectors at the origin are the operators of the form $X^i \frac{\partial}{\partial x^i}$. Similarly, a vector field has the general form

$$X = X^i\left(x\right)\frac{\partial}{\partial x^i}$$

The dimension of the tangent spaces therefore coincide with the dimension of the manifold itself. The component functions $X^i\left(x\right)$ are simply the result of letting the vector field act as a differential operator on the coordinate functions. The coordinate vector fields $\frac{\partial}{\partial x^i}$ constitute a basis for the tangent vectors and the differentials $dx^i$of the coordinate functions constitute the dual basis for the cotangent vectors. Under a change of coordinates

$$x^i = x^i\left(\bar{x}\right)$$

the components transform as

$$X^j = \frac{\partial x^j}{\partial \bar{x}^i}\bar{X}$$

and similarly the components of a tensor

$$t = t^{i_1..i_I}_{j_1..j_J}\frac{\partial}{\partial x^{i_1}} \otimes .. \otimes \frac{\partial}{\partial x^{i_1}} \otimes dx^{j_1} \otimes .. \otimes dx^{j_J}$$

transform as

$$t^{k_1..k_I}_{l_1..l_J} = \bar{t}^{i_1..i_I}_{j_1..j_J}\frac{\partial x^{k_1}}{\partial \bar{x}^{i_1}}..\frac{\partial x^{k_I}}{\partial \bar{x}^{i_I}}\frac{\partial \bar{x}^{j_1}}{\partial x^{l_1}}..\frac{\partial \bar{x}^{j_J}}{\partial x^{l_J}}$$

5.2 Jets and differential operators

The 1-jet of a mapping $\Theta : M \to N$

$$y^\alpha = \Theta^\alpha(x)$$

is simply the list

$$\left(x^i, y^\alpha(x), \frac{\partial y^\alpha}{\partial x^i}\right)$$

and similarly for higher jets.

The flow of a vector field $X^j(x)$ is the solution to the system of ordinary differential equations

$$\frac{dx^i(t)}{dt} = X^i(x(t))$$

A linear scalar differential operator of order m has the representation

$$L = \sum_{k=0}^{m} a^{i_1..i_k} \frac{\partial^k}{\partial x^{i_1}..\partial x^{i_k}}$$

and its principal symbol is the function $\sigma_L = a^{i_1..i_m}\xi_{i_1}..\xi_{i_m}$, where (x,ξ) are the induced coordinates on T^*M.

5.3 Affine connections

An affine connection ∇ is characterized by its action the coordinate fields.We introduce the notation $\partial_k = \frac{\partial}{\partial x^k}$ and write

$$\nabla_{\partial_k}\partial_j = \Gamma^i_{jk}\partial_i$$

The coefficients Γ^i_{jk} are known as the connection coefficients or Christoffel symbols of the connection. Generally it then holds that

$$\begin{aligned} \nabla_X Y &= X^i\left(\frac{\partial Y^j}{\partial x^i} + \Gamma^j_{li}Y^l\right)\partial_j \\ H^{\nabla}(f) &= \left(\frac{\partial^2 f}{\partial x^i \partial x^j} - \Gamma^k_{ij}\frac{\partial f}{\partial x^k}\right)dx^i \otimes dx^j \end{aligned}$$

The difference between two connections $\overset{(2)}{\nabla}$ and $\overset{(1)}{\nabla}$ is a tensor t

$$\overset{(2)}{\Gamma^k}_{ij} - \overset{(1)}{\Gamma^k}_{ij} = t^k_{ij}$$

and conversely any connection $\overset{(2)}{\nabla}$ may be expressed as the sum of any other connection $\overset{(1)}{\nabla}$ and their difference tensor t.

The torsion and curvature tensors are given by

$$\begin{aligned} \left(T^{\nabla}\right)^i_{jk} &= \Gamma^i_{jk} - \Gamma^i_{kj} \\ \left(K^{\nabla}\right)^i_{j\ kl} &= \frac{\partial \Gamma^i_{jk}}{\partial x^l} - \frac{\partial \Gamma^i_{jl}}{\partial x^k} + \Gamma^i_{ml}\Gamma^m_{jk} - \Gamma^i_{mk}\Gamma^m_{jl} \end{aligned}$$

respectively.

The geodesic spray of ∇ is the vector field Z^{∇}

$$Z^{\nabla} = \xi^l \frac{\partial}{\partial x^l} - \xi^j \xi^k \Gamma^l_{jk} \frac{\partial}{\partial \xi^l}$$

defined on TM. Here (x,ξ) are the the coordinates on TM, *induced* by the coordinates x on M, so that a tangent vector at a point of M with coordinates x and components ξ w.r.t. the induced basis $\frac{\partial}{\partial x}$, is the point in TM represented by the coordinates (x,ξ).

The geodesics of ∇ are the projections onto M of the integral curves of Z^{∇}.

In induced coordinates, the integral curves of Z^{∇} are the solutions of the system of differential equations

$$\begin{aligned}\frac{dx^i(t)}{dt} &= \xi^i(t)\\ \frac{d\xi^i(t)}{dt} &= -\Gamma^i_{jk}\left(x(t)\right)\xi^j(t)\xi^k(t)\end{aligned}$$

Eliminating the $\xi^i(t)$, we see that the geodesics are the solution to the system of second order differential equations

$$\frac{d^2x^i(t)}{dt^2} + \Gamma^i_{jk}\frac{dx^j(t)}{dt}\frac{dx^k(t)}{dt} = 0$$

An affine connection is said to be *flat* if the torsion and curvature tensors both vanish. This is the case if and only if there is a coordinate system such that the corresponding connection coefficients identically vanish.

Most affine connections in stochastic differential geometry are either torsion-free, but having nonvanishing curvature or *vice versa*, curvature-free, but having nonvanishing torsion. The latter kind of connections are closely related to the concept of a *moving frame.*

A moving frame is an ordered set of pointwise independent vector fields $F_1,.., F_n$ spanning TM. Any moving frame F uniquely defines an affine connection $\overset{(F)}{\nabla}$ through the conditions

$$\overset{(F)}{\nabla} F_A = 0$$

Another moving frame G defines the same connection if and only if $G_A = C^B_A F_B$ with *constant* coefficients C^B_A. Affine connections arising in this manner have vanishing curvature but in general nonvanishing torsion. The torsion of $\overset{(F)}{\nabla}$ is characterized by

$$\left(T^{\nabla}\right)^i_{jk} F^j_A F^k_B = -\left[F_A, F_B\right]^i$$

which follows directly from the definition of torsion $T^{\nabla}(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]$ and the condition that $\overset{(F)}{\nabla} F_A = 0$.

The Levi-Civita connection of a riemannian structure $g = g^{ij}\partial_i \otimes \partial_j$ is given by the formula

$$\Gamma^i_{jk} = \frac{1}{2}g^{il}\left(\frac{\partial g_{lj}}{\partial x^k} + \frac{\partial g_{kl}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^l}\right)$$

where $g_{kl}\ dx^k \otimes dx^l$ is the riemannian metric dual to the riemannian structure $g = g^{ij}\partial_i \otimes \partial_j$. The Levi-Civita connection is torsion-free, but in general has nonvanishing curvature.

Several different affine connections are of interest in the differential geometric filtering theory:

- Flat connections associated with coordinate systems (in classical Itô theory)
- Curvature-free connections related to moving frames (in noise modeling)
- The symmetric part of a curvature-free connection. (This part has curvature but is torsion-free)
- The Levi-Civita connection corresponding to the noise covariance metric.
- The Levi-Civita connection corresponding to the Fisher information metric on a statistical manifold.
- The Amari α-connections on a statistical manifold.

Curvature and torsion both have simple geometric interpretations. Consider the following control system

$$\dot{x}^i(t) = F_A^i(x(t))\,\dot{v}^A(t)$$

where $F_A^i(x)$ are the components of a moving frame and the *controls*, $\dot{v}^A(t)$, are represented as the time derivative of a small closed trajectory in $\mathbb{R}_v^n$. The resulting trajectory $x^i(t)$ in the manifold is in in general *not* closed, and to first order in the enclosed v-area dA^{AB}, the end point differs from the starting point according to the formula

$$x^i(t_{end}) = x^i(0) - \frac{1}{2}\left(T^{\nabla}\right)^i_{jk} F_A^j F_B^k dA^{AB} + o\left(dA^{AB}\right)$$

Consider now instead a small closed curve $\gamma : t \mapsto x^i(t)$ in the manifold, and an initial frame with components $F_A^j(x(0))$. Each of the basis vectors F_A is *parallel transported* along γ according to the formula

$$\nabla_{\dot{x}^i\partial_i} F_A = 0$$

that is,

$$\frac{dF_A^i}{dt} + \Gamma^i_{jk} F_A^j \dot{x}^k = 0$$

so that $F_A(t)$ becomes a frame at the point $x(t)$. At the endpoint of the closed curve, the parallel transported frame $F_A(t_{end})$ in general differs from the initial frame, and to first order in the enclosed x-area dA^{kl}

$$F_A^i(t_{end}) = F_A^i(0) + \frac{1}{2}\left(K^{\nabla}\right)^i_{j\ kl} F_A^j(0)\, dA^{kl} + o\left(dA^{kl}\right)$$

These torsion and curvature formulas hold to the first order in the enclosed area for closed curves of infinitesimal diameter. From them, the corresponding exact formulas for finite closed curves may be obtained, but this requires the notion of multiplicative integration.

5.4 Representation of differential operators

An affine connection may be considered as a differential operator

$$\nabla : \Gamma\left(M, T^{(k,l)}M\right) \to \Gamma\left(M, T^{(k,l+1)}M\right)$$

according to the identification

$$i_X(\nabla A) = \nabla_X A$$

A general differential operator L of order k between the ‘geometric vector bundles’ $\pi_E : E \to M$ and $\pi_F : F \to M$ may be uniquely represented in the following form

$$L\psi = \sum_{j\leq k}\left(L_j, \nabla^j\psi\right)$$

Here, $\nabla^j\psi$ is the result of applying the connection j times to ψ, the L_j is an $F \otimes E^*$-valued symmetric $T^{(0,j)}$-tensor, $(\cdot,\cdot)$ is the natural duality pairing, and finally, a ‘geometric vector bundle’ is one in which the affine connection makes natural sense (namely direct sums of fractional density tensor bundles).

In this way a differential operator is split into terms of homogeneous order, each term defining its own coefficient tensor via its principal symbol. The 1-1 correspondence between the differential operator and the list of coefficient tensors depends on the connection used. This is the underlying reason for the strange transformational properties of the ‘drift vector field of an Itô diffusion’, when the concept of connections is not properly acknowledged.

6. Stochastic calculus on manifolds

In this chapter the notion of stochastic differential equations on manifolds is addressed.

6.1 Itô's rule and affine connections

The data for a manifold valued SDE consist of

- A drift vector field $F_0 = F_0^i \frac{\partial}{\partial x^l}$
- An $\mathbb{R}^m$ valued standard Brownian motion w^A
- a set of noise vector fields $F_A = F_A^i \frac{\partial}{\partial x^l}$
- a geodesically complete affine connection ∇

We intend to give give meaning to a formal SDE of the form

$$d^{\nabla} x^i = F_0^i \left(x\left(t\right)\right) dt + F_A^i \left(x\left(t\right)\right) dw^A$$

as an Itô SDE.

Put $g = \sum_{A=1}^{m} F_A \otimes F_A = F_A^i \delta^{AB} F_B^j \partial_i \otimes \partial_j$. This is the subriemannian structure defined by the noise. By the left hand side of the equation is loosely speaking meant the expression $d^{\nabla} x^i = dx^i + \frac{1}{2} g^{jk} \Gamma_{jk}^i dt$.

More precisely, the following property is postulated for the expression $d^{\nabla} x^i$

$$d^{(\nabla + T)} x^i = d^{\nabla} x^i + \frac{1}{2} g^{jk} T_{jk}^i dt$$

If, furthermore, it is agreed to identify $d^{\nabla} x^i$ with the Itô differential dx^i in the case when Γ_{jk}^i vanish identically in the coordinate system employed, the above formulas are related to Itô's rule in the following sense

- Itô's rule shows that the equation association from the quadruple $\left(F_0, w^A, F_A, \nabla\right)$ to the equation $d^{\nabla} x^i = F_0^i \left(x\left(t\right)\right) dt + F_A^i \left(x\left(t\right)\right) dw^A$ (more precisely the Itô equation $dx^i = \left(F_0^i \left(x\left(t\right)\right) - \frac{1}{2} g^{jk} \Gamma_{jk}^i\right) dt + F_A^i \left(x\left(t\right)\right) dw^A$) is coordinate independent.
- Itô's rule itself is encoded in the formula $d^{(\nabla + T)} x^i = d^{\nabla} x^i + \frac{1}{2} g^{jk} T_{jk}^i dt$

In this formulation, the 'strange' coordinate transformation rules of a traditional Itô SDE is due to a simultaneous tacit change of connection, from the flat connection of the first coordinate system to that of the second.

6.2 Redundancy in the representation

It is clear from the above that the quadruple $\left(F_0, w^A, F_A, \nabla\right)$ uniquely determines the SDE. The representation is however redundant, since for any section T of $T^{(1,2)}M$ and any mapping $Q : M \to SO\left(\mathbb{R}^m\right)$, the quadruple

$$\left(F_0^i - \frac{1}{2} g^{jk} T^i_{jk}, w^A, Q^B_A F_B, \nabla + T\right)$$

defines the '*same* SDE' (strongly equivalent if $Q^A_B \equiv \delta^A_B$, otherwise weakly equivalent).

6.3 Particular choices of connections

Certain choices of connections give rise to particularly interesting representations

6.3.1 Classical Itô equations The use of a particular coordinate system and its corresponding flat connection gives us back the classical Itô SDE as it is usually presented. When performing a nonlinear change of coordinates, *together with* the ensuing change of corresponding flat connection, the drift vector field has to be corrected according to the redundancy formula above.

6.3.2 Stratonovich equations Consider for simplicity the case when g^{jk} is nondegenerate (and $m = n$), the general case being only slightly more involved. By using the curvature-free connection $\overset{(F)}{\nabla}$ defined by the moving frame F_A, one arrives at the Stratonovich equation (written as an Itô equation). In other words, by the convention of always using $\nabla = \overset{(F)}{\nabla}$ in $\left(F_0, w^A, F_A, \nabla\right)$ we arrive at the Stratonovich interpretation of a formal SDE in terms of $\left(F_0, w^A, F_A\right)$. Recall that $\overset{(F)}{\nabla}$ in general has nonvanishing torsion.

When performing a nonconstant orthonormal change of the moving frame, *together with* the ensuing change of corresponding curvature-free connection, the drift vector field has to be corrected according to the redundancy formula above. Observe also that only the torsion-free part of the moving frame connection affects the SDE.

6.3.3 Geometric diffusion equations Consider again the case when g^{jk} is nondegenerate (and $m = n$). This time the condition is essential. To the riemannian structure g^{jk} is associated its Levi-Civita connection $\overset{(g)}{\nabla}$, which is torsion-free, but in general has nonvanishing curvature. By the convention of always using $\nabla = \overset{(g)}{\nabla}$ in $\left(F_0, w^A, F_A, \nabla\right)$ we obtain an SDE which in the case of vanishing drift, $F_0 = 0$, reflects properties of the riemannian manifold (M, g) only. Its corresponding forward Kolmogorov equation is the famous intrinsic heat equation of riemannian geometry. This formulation is particularly useful, when the riemannian metric has simple properties, which is true in the case of Lie groups. Navigation filtering problems usually involve Lie groups ($SO\,(n)$ or $SE\,(n)$ in dimensions $n = 2,\ 3$).

6.3.4 Driftless Stratonovich equations According to the redundancy formula, the drift vector field is changed when the one moving frame connection is replaced by another. It is natural to ask whether it is possible to select a moving frame, such that the drift vector field is completely absorbed. This is *impossible* for generic drift fields (in sharp contrast to the erroneous claim in Elworthy et al [20], Theorem 2.1.1.

p 31). The precise integrability conditions are particularly simple in two dimensions, and for the flat case

$$\begin{aligned} dx^1 &= f^1(x^1,x^2)dt + dw^1 \\ dx^2 &= f^2(x^1,x^2)dt + dw^2 \end{aligned}$$

we may use the moving frame

$$\begin{aligned} F_1 &= \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix} \\ F_2 &= \begin{pmatrix} -\sin\alpha \\ \cos\alpha \end{pmatrix} \end{aligned}$$

whereby the drift is modified by the term

$$\begin{pmatrix} \frac{\partial\alpha}{\partial x^2} \\ -\frac{\partial\alpha}{\partial x^1} \end{pmatrix}$$

from which it is clear that the drift vector field can be completely absorbed if and only if it is divergence free

$$\frac{\partial f^1}{\partial x^1} + \frac{\partial f^2}{\partial x^2} = 0$$

In higher dimensions, the integrability conditions on the drift vector field becomes nonlinear.

6.4 Connections and the definition of SDEs

In the preceding section, it was noted that by adjoining the choice of an affine connection to the list of data of an Itô SDE, we automatically incorporate the correct transformational properties of such equations. However, no real insight into how the connection enters the definition of the SDE was given. One way of achieving this is to abandon additive integrals in the integral formulation of Itô SDEs. We may formulate a multiplicative integral Itô equation thus

$$x(t) = \prod_{\tau=0}^{t=\tau} \exp^{\nabla}\left(F_0\left(x\left(\tau\right)\right)dt + F_A\left(x\left(\tau\right)\right)dw^A\right)x(0)$$

Here the exponential mapping $\exp^{\nabla}$ is the composition of the flow of the geodesic spray and the projection from the tangent bundle and the Ito product integral is defined as the limit of finite products. The notation $\prod_{\tau=0}^{t=\tau}$ (with the initial time value, 0, at the right of the equality sign, and the final time value, t, at the left of the equality sign) is meant to mean that later factors are put to the left of earlier factors, and in particular that

$$\prod_{\tau=0}^{t=\tau} \exp^{\nabla}(...) = \prod_{\tau=t_1}^{t=\tau} \exp^{\nabla}(...) \prod_{\tau=0}^{t_1=\tau} \exp^{\nabla}(...)$$

In the case of Stratonovich equations the the exponential mapping coincides with the vector field flow. It is our intention to return to this formulation in a later report.

6.5 Geometric Invariants

Due to the high degree of redundancy in our formulation of SDEs, it is natural ask which of its features are invariant, i.e. true properties of the process itself. From the redundancy formula it is readily seen that the subriemannian structure is a candidate. This is indeed so, and it can in fact be shown that, with unit probability, the subriemannian structure along a single trajectory is determined by the 'size of its wiggles'. Hence, in the riemannian (= nondegenerate) case, the Levi-Civita connection is well-defined standard choice of connection, and therefore an invariant. Finally, the drift vector field, as defined using the Levi-Civita connection, is an invariant. These invariants, in turn, determines the process up to weak equivalence, since all SDE data are determined, except for an immaterial (for weak equivalence) choice of orthonormal frame.

Another approach to reducing the redundancy is the concept of an Itô bundle. In our terminology, the Itô bundle may be defined as the bundle of SDE data modulo the group of redundancy transformations. This quotient bundle may alternatively be constructed by means of explicit transition formulas between bundle charts. Such a presentation, however, leaves obscure the close relation between SDEs and affine connections.

6.6 Other geometric issues

In the literature, there are several differential geometric constructions aiming at the definition of SDEs on manifolds. Due to Nash' riemannian imbedding theorem, SDEs on manifolds may be obtained as special cases of SDEs on euclidean spaces. In this approach one has to check that the resulting SDE in no way depends on the imbedding chosen, but on the intrinsic properties of the manifold (or rather its SDE data, as above) only.

There are constructions making use of frame bundles. These formulations are in fact close to ours.

7. Statistical Manifolds

In this chapter, a differential geometric oriented presentation of parametric models is given, *cf.* [10], [27].

7.1 Parametric models without parameters

A statistical model may be described as follows. Given the n-dimensional manifold M, a subset, S of the smooth nowhere vanishing probability densities on M (*i.e.* elements ϕ of $\Gamma(M, |\Lambda^n| M)$ satisfying $\phi > 0$ everywhere and $\int_M \phi = 1$) is a 'parametric model'. If S is given a manifold structure (of dimension N, say) such that the pointwise evaluations mapping ρ_{SM}

$$\rho_{SM} : S \times M \to |\Lambda^n| M$$

is jointly smooth, then S is a *statistical manifold*, and the triple (M, S, ρ_{SM}) is a *smooth parametric model.* (The 'parameters' of the model are the local coordinate functions on S).

Our aim is to understand such smooth 'parametric' models in a coordinate-free fashion, so as to identify genuine properties of the triple (M, S, ρ_{SM}), irrespective of any choice of coordinates on either M or S. In more technical terms, we look for invariants of (M, S, ρ_{SM}) under the actions of independent simultaneous diffeomorphisms of S and M.

It turns out that the structure of smooth parametric models is so rich, that (generically) more or less unique standard parameterizations can be defined. Nevertheless, it is often useful to restrict attention to smaller sets of invariants (of certain forms), that do not suffice to determine canonical coordinate systems.

7.1.1 Examples 1) An *ad hoc* model on the unit circle $\mathbb{S}$ is given by

$$\begin{aligned} M &= \mathbb{S}_x (= \mathbb{R}_x / 2\pi\mathbb{Z}) \\ S &= \mathbb{S}_{x_0} \\ \rho(x_0, x) &= \frac{2 + \sin(x - x_0)}{4\pi} |dx| \end{aligned}$$

We see that a simultaneous rotation of both circles M and S leaves the model invariant. The formulas suggest that S may be identified with M via $x \leftrightarrow x_0$, and that this would provide a preferred point estimate of $x \in M$, given $x_0 \in S$. On the other hand, by Moser's theorem below, there exists a diffeomorphism of S whose pullback of the density $x_0 \in S$ is $\frac{1}{2\pi} |dx|$, which does not seem to single out any natural point estimate. Hence, if the identification $x \leftrightarrow x_0$ (or the corresponding point estimate) has any invariant meaning at all, it is a property of the geometry of the whole model (M, S, ρ_{SM}) and not of any single density x_0.

2) The one-dimensional normal family is given by

$$\begin{aligned} M &= \mathbb{R}_x \\ S &= \mathbb{R}_\xi \times \mathbb{R}^+_\sigma \\ \rho(\xi, \sigma, x) &= \frac{e^{\frac{-1}{2\sigma^2}(x-\xi)^2}}{\sigma\sqrt{2\pi}} |dx| \end{aligned}$$

Most of the discussion of the preceding example carries over to this one. This model is an example of an exponential family, which implies a benign behavior. The property of being exponential is invariant, and is related to the vanishing of the curvature of an appropriate Amari connection. We shall return to this example in order to compute, among other things, its Fisher information metric.

7.1.2 Moser's theorem The following is a well-known theorem of differential geometry: two nowhere vanishing volume forms on a compact, oriented manifold M are equivalent via the pullback of a diffeomorphism, if and only if they have the same total volume. This can be proved by a Lie transform method.

Apart from technicalities (compact and oriented manifold), this theorem tells us that, generally, single probability densities on M have no individual properties. As the second simplest invariants would be properties of *pairs of densities*, we turn our attention to these.

7.1.3 Invariants of a pair of probability densities Let ρ_1 and ρ_2 be two nowhere vanishing probability densities on the manifold M. Their quotient q

$$\rho_2 = q\ \rho_1$$

is a mapping $q : M \to \mathbb{R}^+$. The push-forward of ρ_1 (as a measure), $q_*\rho_1$, is a probability measure on $\mathbb{R}^+$, whose probability function, F,

$$F : \mathbb{R}^+ \to [0, 1]$$

is a function space invariant of the pair ρ_1 and ρ_2. It is our conjecture that there are no further pair invariants. This would imply that any scalar pair invariant may be written in the form

$$\int_M f(q)\, \rho_1$$

and it is clear that for every function f, this expression is an invariant.

For the special choice

$$f(q) = -\ln q$$

this is the famous Kullback-Leibler divergence and for the choice

$$f(q) = \sqrt{q}$$

this is the scalar product of the fractional densities $\sqrt{\rho_1}$ and $\sqrt{\rho_2}$ as unit vectors in the natural Hilbert space of square-roots of probability densities. The squared Hilbert distance between $\sqrt{\rho_1}$ and $\sqrt{\rho_2}$, a.k.a. the *Hellinger metric* is

$$2 - 2\int_M \sqrt{q}\rho_1$$

7.1.4 Invariants of infinitesimally close pairs Let $\varepsilon \mapsto \rho_\varepsilon$ be a smooth one-dimensional family of densities, and consider the Taylor coefficients of

$$I_\varepsilon = \int_M f(q_\varepsilon)\,\rho_0$$

It holds that

$$\begin{aligned} I_\varepsilon &= f(1) + \frac{\varepsilon^2}{2} f''(1) \int_M \left(\frac{\rho_0'}{\rho_0}\right)^2 \rho_0 + \\ &+ \frac{\varepsilon^3}{6}\left(f'''(1) \int_M \left(\frac{\rho_0'}{\rho_0}\right)^3 \rho_0 + 3f''(1) \int_M \left(\frac{\rho_0'}{\rho_0}\right)\left(\frac{\rho_0''}{\rho_0}\right)\rho_0 \right) + O\left(\varepsilon^4\right) \end{aligned}$$

The family $\varepsilon \mapsto \rho_\varepsilon$ is a curve on S. In terms of local coordinates $\theta^A, (A = 1..N)$, the curve is given by $\varepsilon \mapsto \theta^A(\varepsilon)$. Introducing the *Fisher information metric*

$$g_{AB} = \int_M \frac{\rho_A'}{\rho}\frac{\rho_B'}{\rho}\rho$$

and the *skewness tensor*

$$T_{ABC} = \int_M \frac{\rho_A'}{\rho}\frac{\rho_B'}{\rho}\frac{\rho_C'}{\rho}\rho$$

the above formula takes the form

$$\begin{aligned} I_\varepsilon &= f(1) + \frac{\varepsilon^2}{2} f''(1)\, g_{AB}\dot\theta^A\dot\theta^B + \frac{\varepsilon^3}{2} f''(1)\, g_{AB}\dot\theta^A \frac{\overset{0}{D}\,\dot\theta^B}{dt} + \\ &+ \frac{\varepsilon^3}{6}\left(f'''(1) + \frac{3}{2} f''(1)\right) T_{ABC}\dot\theta^A\dot\theta^B\dot\theta^C + O\left(\varepsilon^4\right) \end{aligned}$$

where $\frac{\overset{0}{D}\dot\theta^B}{dt}$ is the acceleration w.r.t. the Levi-Civita connection of the Fisher metric.

With the introduction of Amari's α-connections

$$\begin{aligned} \overset{\alpha}{\Gamma}{}^A_{BC} &= \overset{0}{\Gamma}{}^A_{BC} - \frac{\alpha}{2} g^{AD} T_{DBC} \\ (\alpha &\in \mathbb{R}) \end{aligned}$$

where $\overset{0}{\Gamma}{}^A_{BC}$ are the Christoffel symbols for the Levi-Civita connection, we may also write

$$\begin{aligned} I_\varepsilon &= f(1) + \frac{\varepsilon^2}{2} f''(1)\, g_{AB}\dot\theta^A\dot\theta^B + \frac{\varepsilon^3}{2} f''(1)\, g_{AB}\dot\theta^A \frac{\overset{-1}{D}\,\dot\theta^B}{dt} + \\ &+ \frac{\varepsilon^3}{6} f'''(1)\, T_{ABC}\dot\theta^A\dot\theta^B\dot\theta^C + O\left(\varepsilon^4\right) \end{aligned}$$

where $\frac{\overset{-1}{D}\dot\theta^B}{dt}$ is the acceleration w.r.t. Amari's -1-connection.

Summing up, we see that *any* invariant measure of divergence, has an $O\left(\varepsilon^4\right)$-expansion in terms of the Fisher metric and the skewness tensor.

7.2 Exponential families

In the important case, when the smooth parametric model (M, S, ρ_{SM}) has the following properties

- S is an open subset of $\mathbb{R}^N_\theta$
- $\rho_{SM}(\theta, x) = e^{\theta^A c_A(x) - \psi(\theta)} |d^n x|$ for some functions c_A on M.

it is said to constitute an *exponential family.*

It is obvious that the this functional form is invariant under an *affine* change of θ^A, and it is in fact elementary to show that no other reparametrization can preserve this form. Put in other words, to the exponential family is associated a *flat affine connection.* It turns out that this connection is the +1-Amari connection, which for this reason is also called *the exponential connection.*

7.2.1 Examples Returning to our earlier examples, for the family we may

$$\rho(x_0, x) = \frac{2 + \sin(x - x_0)}{4\pi} |dx|$$

compute the Fisher metric $g = \left(1 - \frac{\sqrt{3}}{2}\right) dx_0 \otimes dx_0$ and the skewness $T = 0$. The identification $x \leftrightarrow x_0$ is the maximum likelihood point estimate. The $\overset{\alpha}{\Gamma}$-geodesic coordinate x_0 on S hence translates into a preferred coordinate x on M, so the given parameterizations (x and x_0) of M and S may be reconstructed from properties of the model itself, up to a common additive term.

For the family

$$\rho(\xi, \sigma, x) = \frac{e^{\frac{-1}{2\sigma^2}(x-\xi)^2}}{\sigma\sqrt{2\pi}} |dx|$$

we have

$$\begin{aligned} g &= \frac{1}{\sigma^2}(d\xi \otimes d\xi + 2d\sigma \otimes d\sigma) \\ T &= \frac{2}{\sigma^3}(4d\sigma \otimes d\sigma \otimes d\sigma + d\xi \otimes d\xi \otimes d\sigma + d\sigma \otimes d\xi \otimes d\xi + d\xi \otimes d\sigma \otimes d\xi) \end{aligned}$$

The Fisher metric is that of the standard hyperbolic plane, while the nontrivial skewness tensor in this case leads to a flat exponential connection, confirming that the normal family is an exponential family. Affine coordinates for this family are

$$\begin{aligned} \theta^1 &= \frac{-1}{2\sigma^2} \\ \theta^2 &= \frac{\xi}{\sigma^2} \end{aligned}$$

The Fisher metric is invariant under the full hyperbolic group, while only ξ-translations, simultaneous rescaling of σ and ξ and combinations of these are symmetries for both g and T.

7.3 Construction of a statistical manifolds

The Kalman filter is a finite dimensional exact filter. Its statistical manifold S is the space of normal probability densities on the state space M, which in this case is an affine space (with translation vector space V, say). The elements of this S are identified by the location of their maximum $\xi \in M$ (maximum likelihood estimate w.r.t. the Lebesgue measure) and their covariance tensor $P \in V \otimes V$. From this, it is seen that the manifold S of normal densities may be identified with the convex conical subbundle of symmetric, nonnegative elements of $T^{(2,0)}M$.

We now intend to reverse this construction, and associate to any manifold M, endowed with an geodesically complete affine connection ∇, a statistical manifold $S_{(M,\nabla)}$ of densities that generalizes the normal densities. As a manifold, $S_{(M,\nabla)}$ is a copy of the bundle of symmetric, nonnegative elements of $T^{(2,0)}M$. Each element (ξ, P) of this bundle may be identified with the unique normal density on $T_\xi M$, that is centered around the origin and has its covariance matrix equal to P. This normal density, considered as a probability measure on $T_\xi M$, is pushed forward onto the manifold M by means of the ∇-exponential mapping (the flow of the geodesic spray). This gives a well-defined probability measure on M. There is, however, no guarantee that these probability measures have smooth, nowhere vanishing densities.

7.3.1 Example Let M be the unit circle $\mathbb{S}_x$, with the standard connection. Then $S_{(M,\nabla)}$ will be the space of densities obtained by 'winding' a scalar normal density around $\mathbb{S}$. It may be parametrized by $\mathbb{S}_{x_0} \times R^+_\sigma$ and the pointwise evaluation mapping ρ becomes

$$\rho\left(x_0, \sigma, x\right) = \frac{1}{2\pi}\vartheta_3\left(\frac{\left(x - x_0\right) \mathrm{mod} 2\pi}{2}, e^{\frac{-\sigma^2}{2}}\right)$$

where

$$\vartheta_3\left(u, q\right) = \sum_{n=-\infty}^{\infty} q^{n^2} \cos\left(2nu\right)$$

is one of the elliptic theta functions.

8. Differential geometric filtering

The subjects touched upon in this chapter will be elaborated in later reports. Here only some definitions, ideas and programme statements will be given.

There are several good reasons to keep any geometric properties of a problem intact throughout its solution. For one thing, if the solution is unique, then it will enjoy any symmetry etc. of the problem.

8.1 Modeling issues

As has been discussed in earlier chapters, one and the same SDE may be written in several ways using different connections and drift vector fields. For a given SDE, the choice of representation in further investigations is entirely a matter of taste and convenience. However, in theoretical modeling, noise is often added to a Siffre model, and then the question arises, not only what covariance tensor the noise has, but also by means of what connection the resulting SDE should be defined. When a detailed noise model is given, and noise enters separately though different channels, it is natural to express this by means of a corresponding moving frame, and to use the frame's own connection. This is tantamount to using the frame together with the Stratonovich calculus, which is known to have good robustness properties w.r.t. noise coloring. On the other hand, in a situation where only the net result of the noise is known, it seems wise to express the noise in terms of the Levi-Civita connection. In this case, it is straight forward to impose any possible symmetry requirements on the noise model, which is hard if an *ad hoc* frame formulation is used. In any case, our geometric formulation of SDEs is helpful in pinpointing exactly what assumptions are hidden in any noise model.

8.2 Geometric filters

There exist a few differential geometrically motivated nonlinear filters

8.2.1 Exact finite dimensional filters These are the finite dimensional solutions to the Zakai equation studied in earlier chapters. For such filters to exist, some nongeneric integrability conditions have to be fulfilled. The case with a flat riemannian metric has been thoroughly investigated by [33] Yau and others. We intend to consider more general cases for the future report, particularly symmetric and homogeneous spaces. It is, however, important to realize that the integrability conditions imply both local *and global* restrictions on the situations, where finite dimensional exact filters are possible. Dynamics on the circle $\mathbb{S}$ may locally look like affine dynamics and thereby suggest the existence of a Kalman like filter. Nevertheless, there is a global obstruction to the integrability conditions, so no finite dimensional filter seems to exist even in this simple-looking case.

8.2.2 Intrinsic geometric filters These are the nonlinear filters of Darling. They are defined in terms of an affine connection the corresponding notion of an intrinsic location parameter. In this case, the system dynamics is time continuous, but the observations are time discrete. These filters produce point estimates, and (*a priori*) no densities.

8.2.3 Projection filters These filters will be reviewed in the next chapter. The basic idea is to project the Zakai equation onto a selected statistical manifold. By identifying a probability density with the corresponding half-density (a fractional density) which is an element of the unit sphere of a naturally defined L^2-space, it is clear that orthogonal projection (of the half-density) is a well-defined invariant operation. Once this observation is made, there is no real need for half-densities in the further development of the projection filter: the 'projected Zakai equation' is nonlinear, regardless of whether written as an equation for the density or its square root.

A promising line of development is the combination of geometric ideas from the projection filter and modern Monte Carlo methods (particle filters).

8.2.4 Extended Kalman filters The popular *extended Kalman filter* (EKF), as it is traditionally presented, depends in an *ad hoc* manner on the coordinate systems chosen. The system dynamics is 'linearized' at each point, which for the nonequilibrium points requires the choice of a connection. It is wise to consider EKF as something produced by a dynamical system with a preferred or selected affine connection. This opens up a possibility to construct natural and tailor made EKF filters in geometric situations. This possibility will be investigated in later research.

The EKF yields a point estimate ξ together with a formal 'measure of dispersion' P, playing a role similar to the conditional covariance matrix of the genuine Kalman filter. However, ξ and P might be considered as producing a probability density via the naturally defined statistical manifold $S_{(M,\nabla)}$. By this construction, the EKF can be compared to other filters taking values on statistical manifolds, and in particular the quality of an EKF could be judged against how close its produced densities are (expected to be) to the ones solving the Zakai equation. In this sense one might hope for an optimal choice of connection, and in favorable cases, this optimal connection might be determined from symmetry properties alone, but this is so far only speculation.

8.3 Statistical geometry in geometric filtering

The filters discussed above provide stochastic processes on statistical manifolds. The statistical manifolds are provided with (several) natural affine connections, and one might expect that the filter SDE has particularly transparent properties when expressed in some of these natural connections.

9. The Projection Filter

We give here a short review of Brigo et. al. [8], [9] on the projection filter and its geometrical significance. We also review some results from Amari [2]. We will try to clarify the basic assumptions and to tighten up the presentation somewhat.

The Kushner-Stratonovich equation defines a vector field in a space of probability densities. By approximating the initial condition with a density in some given parameterized family regarded as a finite dimensional submanifold, and projecting the vector field at each point of the manifold on the corresponding tangent space, one ends up with a stochastic differential equation in the finite dimensional parameter space of the given family — the projection filter.

9.1 The geometrical setting

We will be concerned with parameterized families of probability density functions on $\mathbb{R}^n$ with respect to the Lebesgue measure dx that may be regarded as finite dimensional differentiable manifolds (see [34] or [15] for standard definitions). The densities in a family are all supposed to be strictly positive (almost everywhere) on $\mathbb{R}^n$, that is to say, the corresponding measures are mutually absolutely continuous (equivalent).

We will only consider local aspects and assume for simplicity that such a family $S = \{p(\cdot, \theta)\}$ admits an atlas consisting of a single chart (S, φ_Θ), where φ_Θ is a bijection from S onto an open set Θ in $\mathbb{R}^m$ such that $\varphi_\Theta(p(\cdot, \theta)) = \theta$, i.e. the parameters $\theta = (\theta_1, \dots, \theta_m)$ constitutes a global coordinate system in S. From this point of view, S is just an ordinary finite dimensional manifold, with coordinate vectors $\{\frac{\partial}{\partial \theta_i}\}_{i=1}^m$ spanning the tangent space T_pS at each point $p \in S$ etc.

In order to make contact with the properties of S as a set of functions, we would like to view S as embedded in some space of functions. An infinite dimensional manifold is defined in complete analogy with the finite dimensional case as a set M covered by an atlas of compatible charts (U_i, φ_i) taking their values in a Banach space E, $\varphi_i : U_i \to E$ ([1] or [34] give precise definitions). E is called the model space, and the given set M is called a Banach manifold modeled on E. We need only consider the trivial case where the set M is the Banach space E itself, covered by the single chart $(E, identity)$.

Recall further that the relevant concept of differentiability of a map f from one Banach space E to another F is the Fréchet derivative Df; the derivative $Df|_e$ at a point $e \in E$ is the unique bounded linear map $E \to F$ that approximates f in a neighborhood of e. This defines also a formal derivative Df of a function $f : M \to N$ between two manifolds modeled on E and F respectively in the following way. If (U, φ) is a chart at $m \in M$ and (W, ψ) a chart at $f(m)$, then the Fréchet derivative $D(\psi \circ f \circ \varphi^{-1})|_{\varphi(m)}$ is called the representative of $Df|_m$ in the given charts.

A tangent vector v at a point m in an infinite dimensional manifold M is usually defined formally as an equivalence class of curves γ through m (maps from an interval $I \subset \mathbb{R}$, $0 \in I$ and $\gamma(0) = m$) such that the derivatives $D(\varphi \circ \gamma)|_{t=0}$ exist and coincide in some chart (U, φ). The object $V = D(\varphi \circ \gamma)|_{t=0}$ is by definition (of differentiability) a vector belonging to the model space E and is called the representative of v in the chart (U, φ). [1] In less fancy terms, V is simply the "velocity" vector at $\varphi(m)$ of the curve $(\varphi \circ \gamma)(t)$ in E. Conversely, each element of E is the representative of a tangent vector at m to some curve γ in M through m, so the tangent space (the collection of formal tangent vectors) $T_m M$ at m is isomorphic to E itself. With this isomorphism between the tangent space and the model space we see that the formal derivative Df defined above of a map f from one manifold M to another N, may be interpreted as a map $Df|_m : T_m M \to T_{f(m)} N$, and is often referred to as the tangent map. [2] Let $\bar{S}$ denote the set $S = \{p(\cdot, \theta)\}$ regarded as a subset of the Banach space L^1 (measurable functions with norm $\|f\|_1 = \int |f(x)|dx < \infty$) and let ι be the corresponding inclusion map $S \to \bar{S} \subset L^1$. Recall that we assumed S to be covered by a single chart (S, φ_Θ) and that L^1 considered as a manifold is trivial in this sense too. The tangent map $D\iota|_p$ at $p \in S$ maps the coordinate vectors $\frac{\partial}{\partial \theta_i} \in T_p S$ onto tangent vectors $\frac{\partial p(\cdot,\theta)}{\partial \theta_i} \in T_p L^1 = L^1$. Its representation in the chart (S, φ_Θ) is given by the Fréchet derivative $D(\iota \circ \varphi_\Theta^{-1})|_\theta = (\frac{\partial p(\cdot,\theta)}{\partial \theta_i})$ which, regarded as a row vector, maps (velocity) vectors in $\mathbb{R}^m$ onto (velocity) vectors in L^1. This map exists if we assume the functions $p(\cdot, \theta)$ to be smooth in θ. In order for $\bar{S}$ to be a submanifold of L^1, ι must be injective and the tangent map $D\iota|_p$ injective for all $p \in S$, which will be the case if the functions $\{\frac{\partial p(\cdot,\theta)}{\partial \theta_i}\}$ are linearly independent for all $\theta \in \Theta$ and thus span a m-dimensional subspace $T_p \bar{S}$ of L^1 at each $p \in \bar{S}$. [3] Note that the full tangent space $T_p L^1$ at a point $p \in L^1$ which is a probability density ($\|p\|_1 = 1$) may be thought of as a space of random variables (measurable functions on $\mathbb{R}^n$). Amari [2] therefore calls $T_p \bar{S}$ the "random variable representation" of $T_p S$, which is a nice way of putting things.

Next, turn $\bar{S}$ into a Riemannian manifold by defining a metric tensor $g_p(\cdot, \cdot)$, i.e. an inner product on the tangent space $T_p \bar{S}$ at each point $p \in \bar{S}$, by the following action on the coordinate vectors, i.e. components in θ coordinates

$$g_{ij}(\theta) = g_{p_\theta}(\frac{\partial p_\theta}{\partial \theta_i}, \frac{\partial p_\theta}{\partial \theta_j}) = \int \frac{\partial p(x,\theta)}{\partial \theta_i} \frac{\partial p(x,\theta)}{\partial \theta_j} \frac{dx}{p(x,\theta)}. \tag{9.1}$$

Here, we demand that the right hand side exists for each $p_\theta := p(\cdot, \theta) \in \bar{S}$. This integral may also be written as an expectation $\mathrm{E}_p\{\cdot\}$ with respect to a density p, $\mathrm{E}_{p_\theta}\{\frac{\partial \log p_\theta}{\partial \theta_i} \frac{\partial \log p_\theta}{\partial \theta_j}\}$, which is well known in statistics as the Fisher information matrix. g_p is obviously symmetric, and from the assumptions that $p_\theta > 0$ and $\{\frac{\partial p_\theta}{\partial \theta_i}\}$ are linearly independent follows that g_p is positive definite. The Fisher information matrix plays a fundamental role in statistics, e.g. its inverse gives a lower bound of the covariance of any unbiased estimator $\hat{\theta}$ of the parameter θ (the Cramér-Rao theorem [28]). It is obviously invariant under coordinate transformations in the sample space $\mathbb{R}^n$. Amari [2, section 3.8] gives further interesting comments on the importance of the Fisher information matrix from a statistics perspective.

In the statistical literature a parameterized family S of densities satisfying the conditions stated so far is called a regular statistical model and is used for statistical inference problems [2], [3]. In that context the main concern seems to be the nature of S as a finite dimensional manifold. For the purpose of defining the projection filter however, there is a point in viewing S as embedded in a larger space of functions. And to emphasize this we are using the rather pedantic separate notation "$\bar{S}$".

[1] This definition is more concrete than the elegant definition of a tangent vector as a derivation of functions at m, which is usually employed in the finite dimensional case. This latter definition is unfortunately problematic in infinite dimensions [1, p. 292 ff.]. In [34, chapter III and VII] a nice summary of different ways of defining tangent vectors can be found.

[2] Although we will not do so here, it is good to use a separate notation "f_*" for this tangent map and reserve "Df" for the Fréchet derivative.

[3] With terminology from [1] $\bar{S}$ is called an immersed submanifold. If, in addition, the topology of $\bar{S}$ induced from $\mathbb{R}^m$ is the relative L^1-topology, then the map $\iota \circ \varphi_\Theta^{-1} : \Theta \to L^1$ is a homeomorphism and is referred to as an embedding. This assumption will not be needed in our context however.

9.2 A projection map

Let $p \in \bar{S}$. Even though $g_p(u,v)$ is defined only for vectors u and v in the subspace $T_p\bar{S}$ of L^1, the defining integral in (9.1) may exist as well for pairs of vectors z and v, where v belongs to $T_p\bar{S}$ but z does not. Let us continue the pedantry and denote by $\bar{g}_p$ the tensor g_p with the domain of the *first* argument extended to all $z \in L^1$ such that the defining integral exists with the second argument still in $T_p\bar{S}$. This extended domain $\mathcal{D}_p$ is a linear subspace of L^1 and may be characterized by $\mathcal{D}_p = \{z \in L^1 | \frac{z}{p_\theta}\frac{\partial p_\theta}{\partial \theta} \in L^1\}$.

Let $g_p^{-1}(\cdot,\cdot)$ be the 2-contravariant inverse of g_p, i.e. $g_p^{-1}(\cdot, g_p(v,\cdot)) = v$ is the identity map on $T_p\bar{S}$. Denote by $g^{ij}(\theta)$ its components in θ coordinates, thus (g^{ij}) is the inverse of the Fisher information matrix. By combining g_p^{-1} with $\bar{g}_p$ we get a map $z \longmapsto \Pi_p(z) := g_p^{-1}(\cdot, \bar{g}_p(z,\cdot))$, or explicitly in terms of θ coordinate vectors

$$\begin{aligned}\Pi_\theta(z) &= \sum_{i,j=1}^{m} g^{ij}(\theta)\, \bar{g}_{p_\theta}(z, \frac{\partial p_\theta}{\partial \theta_j})\, \frac{\partial p_\theta}{\partial \theta_i} \\ &= \sum_{i,j=1}^{m} g^{ij}(\theta)\, [\int z(x)\, \frac{\partial p(x,\theta)}{\partial \theta_j}\, \frac{dx}{p(x,\theta)}]\, \frac{\partial p_\theta}{\partial \theta_i} \\ &= \sum_{i,j=1}^{m} g^{ij}(\theta)\, \mathrm{E}_{p_\theta}\{\frac{z}{p_\theta}\frac{\partial \log p_\theta}{\partial \theta_j}\}\, \frac{\partial p_\theta}{\partial \theta_i},\end{aligned} \tag{9.2}$$

which may be interpreted as an orthogonal projection $\mathcal{D}_p \to T_p\bar{S}$. This is the map which is used to define the projection filter.

The requirement for the right hand side in (9.1) to exist may be stated as $\frac{1}{\sqrt{p_\theta}}\frac{\partial p_\theta}{\partial \theta} \in L^2$. Brigo et al. assume $\mathrm{E}_{p_\theta}\{(\frac{z}{p_\theta})^2\} < \infty$, i.e. $\frac{z}{\sqrt{p_\theta}} \in L^2$, as a condition for the applicability of the map Π_p (assumption "D" in [9, p. 506]). By the Cauchy-Schwarz inequality this implies $\frac{z}{p_\theta}\frac{\partial p_\theta}{\partial \theta} \in L^1$, i.e. $z \in \mathcal{D}_p$, so this condition is sufficient but, presumably, not necessary. In remark 4.3 of [8, p. 15] Brigo et al. raises the question (in the context of an exponential family of densities) whether the geometrical interpretation of Π_p still holds under the weaker condition $z \in \mathcal{D}_p$. Our presentation shows that this question is answered in the affirmative.

9.2.1 Remarks In our application, we are primarily interested in applying the above projection map to vectors tangent to curves made up of probability densities. Therefore, we would have liked to regard the set $\mathcal{M}$ of *all* probability densities on $\mathbb{R}^n$ as a manifold per se and to have an intrinsic characterization of the tangent spaces $T_p\mathcal{M}$. The finite dimensional family S would then be defined as a submanifold of $\mathcal{M}$. Unfortunately, it seems difficult to define a topology on $\mathcal{M}$ in order to get the notion of differentiable curves in $\mathcal{M}$. (This problem is touched upon in [2, p. 93], [9, p. 498] and [27, p. 76], see also the comment in [19, p. 8].) For this reason, one is forced to regard $\mathcal{M}$ as a subset of a larger space, as L^1 above, and the full tangent space at a point $p \in \mathcal{M}$ will then contain "to many" vectors. Remember that a vector tangent to a curve γ in $\mathcal{M} \subset L^1$ belongs to L^1 by the very definition of differentiability of the map $\gamma : I \subset \mathbb{R} \to L^1$.

Another possibility to embed $\mathcal{M}$ in a larger space is given by the fact that the square root $\sqrt{p}$ of a density is an element of L^2 (square integrable functions on $\mathbb{R}^n$). $\mathcal{M}$ is thus represented by $\mathcal{R} := \{2\sqrt{p}\,|\,p \in \mathcal{M}\}$ as a subset of L^2 (the factor 2 is inserted for later convenience). A vector tangent to a curve in $\mathcal{R} \subset L^2$ is now by definition (of differentiability) an element in L^2.

If $t \mapsto 2\sqrt{p_t}$ is such a curve in $\mathcal{R}$, differentiable at some t with tangent vector $v = \frac{\partial 2\sqrt{p_t}}{\partial t} = \frac{1}{\sqrt{p_t}}\frac{\partial p_t}{\partial t} \in L^2$, the Cauchy-Schwartz inequality imply that $p_t \in \mathcal{M}$ is a curve differentiable in L^1 with tangent vector $v\sqrt{p_t}$. However, the reverse implication is not true: p_t differentiable in L^1 does not imply that $\sqrt{p_t}$ is differentiable in L^2.[4] Thus, all vectors in L^2 tangent to $\mathcal{R}$ are in this sense also tangent vectors to $\mathcal{M}$ in L^1, but not the converse.

Consider again a family S of densities strictly positive a.e. Let $p \in \bar{S} \subset L^1$ and denote by $\mathcal{B}_p$ the set of measurable functions u such that $\frac{u}{\sqrt{p}} \in L^2$. As before, Cauchy-Schwartz gives that $\mathcal{B}_p \subset L^1$. $\mathcal{B}_p$ is the maximal subspace of T_pL^1 $(= L^1)$ to which the metric tensor g_p defined in (9.1) may naturally be extended (in *both* its arguments). Since $p > 0$, the map $u \mapsto v = \frac{u}{\sqrt{p}}$ is a bijection $\mathcal{B}_p \to L^2$, and the extension $\tilde{g}_p$ of g_p from $T_p\bar{S} \times T_p\bar{S}$ to $\mathcal{B}_p \times \mathcal{B}_p$ is simply the ordinary inner product $(\cdot|\cdot)$ turning L^2 into a Hilbert space

$$\begin{aligned} \tilde{g}_p(u_1, u_2) &= \int \frac{u_1(x)}{\sqrt{p(x)}} \frac{u_2(x)}{\sqrt{p(x)}}\, dx \\ &= \int v_1(x)\, v_2(x)\, dx = (v_1|v_2), \end{aligned} \tag{9.3}$$

where $v_i = \frac{u_i}{\sqrt{p}} \in L^2$. Note that $\mathcal{B}_p \subset \mathcal{D}_p$, the domain of the projection map in (9.2), but that $\mathcal{B}_p$ is defined without reference to the specific family S (apart from the single point p).

The family $S = \{p_\theta\}_{\theta\in\Theta}$ may be embedded as a submanifold $\bar{S}^{1/2} := \{2\sqrt{p_\theta}\}_{\theta\in\Theta}$ of L^2, if the derivatives $\frac{\partial 2\sqrt{p_\theta}}{\partial\theta}$ exist in L^2 for all $\theta \in \Theta$. The metric tensor induced on S by the inner product in L^2 is of course identical to g_p. Considering this embedding is effectively a handy way of restricting ones attention to vectors in the subspace $\mathcal{B}_p$ of L^1, and the resulting natural access to the Hilbert space structure of L^2 is conceptually nice. Note that the previously explicitly made assumption, that the Fisher information matrix should exist, is now "hidden" in the assumption of differentiability of the map $\theta \mapsto 2\sqrt{p_\theta}$. The representation of the projection map Π_p in (9.2) (restricted to $\mathcal{B}_p$) is obviously the orthogonal projection $L^2 \to T_p\bar{S}^{1/2}$ given by the inner product in L^2. This embedding is employed by Brigo et al. [8], [9] for their derivation of the projection filter. The result is of course the same in terms of θ coordinates of the image vector, irrespective of the chosen embedding.

Amari [2] has introduced a whole family of so called α-representations of densities strictly positive on $\mathbb{R}^n$. Define a one parameter family of functions

$$F_\alpha(p) = \begin{cases} \frac{2}{1-\alpha} p^{(1-\alpha)/2}, & \alpha \neq 1 \\ \log p, & \alpha = 1 \end{cases} \tag{9.4}$$

and consider the sets $\mathcal{R}_\alpha := \{F_\alpha(p)\,|\,p > 0, p \in \mathcal{M}\}$. (It is possible to consider other equivalence classes of measures, i.e. densities with an arbitrary common support $X \subset \mathbb{R}^n$.) For $\alpha \in [-1, 1)$, $\mathcal{R}_\alpha$ may be regarded as a subset of the Banach space L^q (measurable functions with norm $\|f\|_q = (\int |f(x)|^q dx)^{\frac{1}{q}} < \infty$) with $q \equiv \frac{2}{1-\alpha} \geq 1$. A differentiable curve $F_\alpha(p_t)$ in $\mathcal{R}_\alpha \subset L^q$, with tangent vector $v = \frac{\partial F_\alpha(p_t)}{\partial t} = p_t^{-(1+\alpha)/2} \frac{\partial p_t}{\partial t} \in L^q$, is by the Hölder inequality also a differentiable curve p_t in $\mathcal{M} \subset L^1$ (since $\frac{1}{q} + \frac{1+\alpha}{2} = 1$). It is possible to define an inner product $< v, w >_\alpha := \int v(x)w(x)p^\alpha(x)\, dx$ in some subspace of the tangent space T_fL^q $(= L^q)$, $f = F_\alpha(p) \in \mathcal{R}_\alpha$, and this subspace is isomorphic to $\mathcal{B}_p \subset L^1$. As before, the metric tensor induced by this inner product on a finite dimensional family S embedded via $\mathcal{R}_\alpha$, is identical to the metric given by the Fisher information matrix.

[4] If the support of p_t changes along the curve (p_t *becomes* zero on a set of non-vanishing measure) then this is not true for obvious reasons. In the case when the support *is* fixed along p_t we are not aware of any argument allowing us to reverse the implication. It is easy to give an example of a function $u \in L^1$ *not* tangent to $\mathcal{M}$ and a $p \in \mathcal{M}$ $(p > 0)$ such that $\frac{u}{\sqrt{p}}$ does not belong to L^2, but we need such a function u *tangent* to $\mathcal{M}$ in order to have a counter-example to the reversed implication.

Amari uses these representations to define a family of connections (covariant derivatives) on S, the α-connections, which is the starting point for endowing a statistical model with some geometric structure [2]. Consider a 2-parameter set $\mathcal{A}$ of functions $f_{s,t}$ in L^q, defined for s,t in some open neighborhood of 0 in $\mathbb{R}^2$. If $s,t \mapsto f_{s,t}$ is appropriately smooth, the derivatives $\frac{\partial f_{s,t}}{\partial t}$ and $\frac{\partial f_{s,t}}{\partial s}$ may be interpreted as vector fields Y and Z respectively, with domain $\mathcal{A}$. The second order derivative $\frac{\partial^2 f_{s,t}}{\partial s \partial t}$ may then be naturally regarded as the derivative of the vector field Y along the curves $s \mapsto f_{s,t}$ (whose tangent vectors are given by the field Z). This defines a covariant derivative $\nabla_Z Y := \frac{\partial^2 f_{s,t}}{\partial s \partial t}$ of Y in the direction Z on $\mathcal{A}$ (or the other way round) and a natural flat connection on L^q (analogous to component-wise derivation of vectors in $\mathbb{R}^n$).

Let $\bar{S}_\alpha$ denote the embedding of $S = \{p_\theta\}$ in $\mathcal{R}_\alpha \subset L^q$ and let $e_i := \frac{\partial F_\alpha(p_\theta)}{\partial \theta_i}$ be coordinate vector fields on $\bar{S}_\alpha$. The representation of the projection map in (9.2) is an orthogonal projection $\Pi_{F_\alpha(p_\theta)}(z) = \sum_{k,l} g^{kl}(\theta) < z, e_l >_\alpha e_k$ onto the tangent space of $\bar{S}_\alpha$ at $F_\alpha(p_\theta)$, assuming that z, e_l belong to the domain of $< \cdot, \cdot >_\alpha$. Applying this projection to the vector field $\nabla_{e_i} e_j = \frac{\partial^2 F_\alpha(p_\theta)}{\partial \theta_i \partial \theta_j}$ defines a covariant derivative on $\bar{S}_\alpha$ and thus on S. The e_k-coefficients of the image are by definition the connection coefficients Γ_{ij}^k

$$\begin{aligned} \Gamma_{ij}^k(\theta; \alpha) &= \sum_{l=1}^m g^{kl}(\theta) < \nabla_{e_i} e_j, e_l >_\alpha \\ &= \sum_{l=1}^m g^{kl}(\theta) \int \frac{\partial^2 F_\alpha(p_\theta)}{\partial \theta_i \partial \theta_j} \frac{\partial F_\alpha(p_\theta)}{\partial \theta_l} p_\theta^\alpha \, dx. \end{aligned} \tag{9.5}$$

This is the family of α-connections on S introduced by Amari. They are defined whenever the integral in the right hand side exists (and may thus be contemplated without indulging in the L^q spaces and their topologies). The case $\alpha = 0$, i.e. the embedding in L^2 considered earlier, is again special since the 0-connection is the Levi-Civitá connection corresponding to the Fisher information matrix (the Riemannian connection given by the metric g_p).

9.3 The filtering problem

Let $\{X_t\}_{0 \le t}$ be a diffusion process in $\mathbb{R}^n$, with drift vector $f_t(x)$ and diffusion matrix $a_t(x)$, which is partially observed through a process $\{Y_t\}_{0 \le t}$ in $\mathbb{R}^d$ given by

$$dY_t = h_t(X_t)\, dt + \rho_t \, dW_t \tag{9.6}$$

where $\{W_t\}_{0 \le t}$ is a standard ("unit" variance) Brownian motion in $\mathbb{R}^q$.

Assume that $R_t := \rho_t \rho_t^{\mathrm{T}}$ is invertible for $t \ge 0$, that $\{X_t\}$ and $\{W_t\}$ are independent and $\mathrm{E}\{\int_0^t |h_s(X_s)|^2_{R_s^{-1}} \, ds\} < \infty$ for all $t \ge 0$, where we are using the notation $|h|^2_{R^{-1}} := h^{\mathrm{T}} R^{-1} h$. The filtering problem consists in integrating, for given initial condition $p = p_0$, the Kushner-Stratonovich equation for the conditional density $p_t := p_t(\cdot|\mathcal{Y}_t)$ of the state X_t given observations $\mathcal{Y}_t = \sigma\{Y_s, 0 \le s \le t\}$,

$$dp_t = \mathcal{L}_t^* p_t \, dt + p_t [h_t - \mathrm{E}_{p_t}\{h_t\}]^{\mathrm{T}} R_t^{-1} [dY_t - \mathrm{E}_{p_t}\{h_t\} \, dt], \tag{9.7}$$

where $\mathrm{E}_{p_t}\{\cdot\}$ as before denotes expectation with respect to p_t, $h_t := h_t(\cdot)$ and the forward diffusion operator $\mathcal{L}_t^*$ is given by

$$\mathcal{L}_t^* \phi = -\sum_{i=1}^n \frac{\partial}{\partial x_i}(f_t^i \phi) + \tfrac{1}{2} \sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \partial x_j}(a_t^{ij} \phi). \tag{9.8}$$

By exercising the Itô-Kunita-Watanabe's theorem [17, theorem 17.11] we get (see appendix F) the following Stratonovich version of equation (9.7)

$$\begin{aligned} dp_t = \mathcal{L}_t^* p_t\, dt - p_t \tfrac{1}{2}(|h_t|^2_{R_t^{-1}} - \mathrm{E}_{p_t}\{|h_t|^2_{R_t^{-1}}\})\, dt \\ + \sum_{k=1}^{d} p_t[(R_t^{-1}h_t)^k - \mathrm{E}_{p_t}\{(R_t^{-1}h_t)^k\}] \circ dY_t^k. \end{aligned} \tag{9.9}$$

Since the Stratonovich differential obeys the usual Newton-Leibniz calculus it is the appropriate formulation in a geometric setting where we would like to interpret the differential dp formally as a "tangent vector", or "velocity", and be able to define the usual tangent map (see [11, chapter 8] or [21]). The right hand side of (9.9) thus defines, in some sense, a time dependent vector field in a space of probability densities on $\mathbb{R}^n$. By restricting this vector field to a given family of densities and apply the projection map Π_p defined in the last section, we will get a finite dimensional filtering problem.

9.4 Projection filter, general case

Let $S = \{p(\cdot, \theta), \theta \in \Theta\}$ be a family of probability densities, where Θ is an open subset of $\mathbb{R}^m$. Impose on $p_\theta := p(\cdot, \theta)$ and h_t the conditions

$$\begin{aligned} \frac{\mathcal{L}_t^* p_\theta}{p_\theta} \frac{\partial p_\theta}{\partial \theta} \in L^1 \\ |h_t|^2_{R_t^{-1}} \frac{\partial p_\theta}{\partial \theta} \in L^1 \end{aligned} \tag{9.10}$$

for all $\theta \in \Theta$ and all $t \geq 0$. They imply that the coefficients of dt and dY_t in the right hand side of the Kushner-Stratonovich equation (9.9), evaluated for an arbitrary member $p = p_\theta \in S$, may be interpreted as vectors belonging to the domain $\mathcal{D}_\theta$ of the projection map Π_θ defined in (9.2). [5] The image of $z \in \mathcal{D}_\theta$,

$$\Pi_\theta(z) = \sum_{i,j=1}^{m} g^{ij}(\theta)\, [\int z(x)\, \frac{\partial p(x,\theta)}{\partial \theta_j}\, \frac{dx}{p(x,\theta)}]\, \frac{\partial p_\theta}{\partial \theta_i}, \tag{9.11}$$

is a tangent vector to $\bar{S}$ at p_θ. Since $\{\frac{\partial p_\theta}{\partial \theta_i}\}$ are the coordinate vectors, $dp_\theta = \sum_i \frac{\partial p_\theta}{\partial \theta_i} \circ d\theta_i$, their coefficients in (9.11) give the representative in $\mathbb{R}^m$ of this image vector. Thus, by plugging the right hand side of (9.9) with $p_t = p_{\theta_t}$ into (9.11) and noting that $\mathrm{E}_{p_\theta}\{\frac{\partial p_\theta}{\partial \theta_i}\} = 0$, we get the finite dimensional projection filter for the given family S in vectorized notation as

$$\begin{aligned} d\theta_t = g^{-1}(\theta_t)\, [\int \mathcal{L}_t^* p(x,\theta_t)\, \frac{\partial p(x,\theta_t)}{\partial \theta}\, \frac{dx}{p(x,\theta_t)}]\, dt \\ - g^{-1}(\theta_t)\, [\int \tfrac{1}{2}\, |h_t(x)|^2_{R_t^{-1}}\, \frac{\partial p(x,\theta_t)}{\partial \theta}\, dx]\, dt \\ + g^{-1}(\theta_t) \sum_{k=1}^{d} [\int (R_t^{-1} h_t(x))^k\, \frac{\partial p(x,\theta_t)}{\partial \theta}\, dx] \circ dY_t^k \end{aligned} \tag{9.12}$$

where θ and $\frac{\partial p_\theta}{\partial \theta} := (\frac{\partial p_\theta}{\partial \theta_i})$ are regarded as column vectors and $g^{-1} = (g^{ij})$ is a matrix.

There remains the question of how to map the given initial condition p_0 onto a starting point θ_0 for the projection filter. Brigo et al. [8] suggests using the device of minimizing the Kullback-Leibler information $\int \log[\frac{p_0(x)}{p(x,\theta)}]\, p_0(x)\, dx$ with respect to

[5] The second condition in (9.10) covers also the last terms in (9.9) as is shown by the following schematic argument: $\int \left| h \frac{\partial p_\theta}{\partial \theta} \right| dx \leq \int h^2 \left| \frac{\partial p_\theta}{\partial \theta} \right| dx + \int \left| \frac{\partial p_\theta}{\partial \theta} \right| dx < \infty$, since both $h^2 \frac{\partial p_\theta}{\partial \theta}, \frac{\partial p_\theta}{\partial \theta} \in L^1$.

$\theta \in \Theta$. How this approximation is done does not seem to be very important however. But on the other hand, how sensitive are the solutions of (9.12) to small changes in the initial conditions? The best thing is, of course, if S is chosen in such a way that p_0 already belongs to S.

9.5 Projection filter for an exponential family

For a given function $c(x) \in \mathbb{R}^m$ on $\mathbb{R}^n$ define probability densities

$$p(x, \theta) = \exp[\theta^{\mathrm{T}} c(x) - \psi(\theta)] \tag{9.13}$$

with normalization factor

$$\psi(\theta) = \log \int \exp[\theta^{\mathrm{T}} c(x)]\, dx \tag{9.14}$$

and domain $\Theta_0 = \{\theta \in \mathbb{R}^m : \psi(\theta) < \infty\}$.

For any open $\Theta \subset \Theta_0$, $S = \{p(\cdot, \theta), \theta \in \Theta\}$ is called an *exponential family* of probability densities. In order to interpret S as a m-dimensional manifold with coordinates θ, the coordinate vectors $\{\frac{\partial p_\theta}{\partial \theta_i}\}$ must be linearly independent. From $\frac{\partial p(x,\theta)}{\partial \theta_i} = p(x, \theta)(c_i(x) - \frac{\partial \psi(\theta)}{\partial \theta_i})$ follows that a sufficient condition for this is that the collection of functions $\{1, c_1, ..., c_m\}$ are linearly independent.

Assume that the conditions in (9.10) apply. By noting the relation $\int \mathcal{L}^* p(x, \theta) \frac{\partial p(x,\theta)}{\partial \theta} \frac{dx}{p(x,\theta)} = \int \mathcal{L}^* p(x, \theta)(c_i(x) - \frac{\partial \psi(\theta)}{\partial \theta_i}) dx = \int p(x, \theta) \mathcal{L} c_i(x) dx$, where $\mathcal{L}$ is the adjoint of $\mathcal{L}^*$ (the backward diffusion operator), the projection filter equation (9.12) for the exponential family S becomes

$$\begin{aligned} d\theta_t = {} & g^{-1}(\theta_t)\, \mathrm{E}_{p_{\theta_t}}\{\mathcal{L}_t c\}\, dt \\ & - g^{-1}(\theta_t)\, \mathrm{E}_{p_{\theta_t}}\{\tfrac{1}{2}\, |h_t|^2_{R_t^{-1}}\, (c - \frac{\partial \psi(\theta_t)}{\partial \theta})\}\, dt \\ & + g^{-1}(\theta_t) \sum_{k=1}^{d} \mathrm{E}_{p_{\theta_t}}\{(R_t^{-1} h_t)^k (c - \frac{\partial \psi(\theta_t)}{\partial \theta})\} \circ dY_t^k \end{aligned} \tag{9.15}$$

where $\frac{\partial \psi(\theta)}{\partial \theta} := (\frac{\partial \psi(\theta)}{\partial \theta_i})$ is regarded as a column vector.

The initial conditions for this equation may easily be obtained by minimizing the Kullback-Leibler information as suggested in [8]; namely, for a given initial density p_0 find $\theta_0 \in \Theta$ such that

$$\frac{\partial \psi(\theta_0)}{\partial \theta} = \int c(x)\, p_0(x)\, dx. \tag{9.16}$$

9.5.1 Remarks Exponential families ("EF" in the following) are of special importance in statistics, both as a theoretical tool ([5], [6]) and for modeling purposes in statistical inference and estimation ([2], [28, section IV.C]). The ubiquitous family of normal distributions is an example of an EF. EFs have some blessed computational properties. Brigo et al. introduce, for example, an EF for which the coefficients of the diffusion part in (9.15) (the coefficients of dY) are independent of θ_t and thus deterministic, which is a nice feature when solving the equation numerically.

Let us give some examples of the interesting analytical and geometrical properties of EFs. In the statistical literature one defines the *Laplace transform* $M(\theta)$ for $\theta \in \mathbb{R}^m$ of a random vector $c \in \mathbb{R}^m$ on the probability space $(\Omega, \mathcal{F}, P)$ as

$$M(\theta) = \mathrm{E}\{\exp[\theta^{\mathrm{T}} c]\} = \int \exp[\theta^{\mathrm{T}} c(\omega)] P(d\omega) \tag{9.17}$$

with domain $\Theta_0 = \{\theta \in \mathbb{R}^m : M(\theta) < \infty\}$. The *cumulant transform* is then defined by $\psi(\theta) = \log M(\theta)$ which is a closed convex function on $\mathbb{R}^m$, and if the covariance

$\text{Cov}(c)$ is positive definite, $\psi(\theta)$ is strictly convex on Θ_0 [5]. Moreover, $M(\theta)$ is a real analytical function in the interior of Θ_0 (=: $\text{int}\Theta_0$) and is the moment generating function of c (strictly speaking, only if $0 \in \text{int}\Theta_0$). $\psi(\theta)$ is the generating function of the cumulants of c [16].

The probability measures $\{P_\theta, \theta \in \Theta_0\}$ given by densities p_θ

$$p_\theta(\omega) = \frac{dP_\theta}{dP}(\omega) = M(\theta)^{-1} \exp[\theta^{\mathrm{T}} c(\omega)] = \exp[\theta^{\mathrm{T}} c(\omega) - \psi(\theta)] \tag{9.18}$$

are called the exponential family generated by c and P. The Laplace transform of c under P_θ is $\mathrm{E}_{p_\theta}\{\exp[\xi^{\mathrm{T}} c]\} = M(\xi + \theta)/M(\theta)$ which thus acts as the generating function of the c-moments under P_θ.

By differentiating the identity $\int p_\theta dP = 1$ once and twice with respect to θ, together with the definition of the Fisher information matrix $g_{ij}(\theta) = \mathrm{E}_{p_\theta}\{\frac{\partial \log p_\theta}{\partial \theta_i} \frac{\partial \log p_\theta}{\partial \theta_j}\}$, we get the useful relations

$$\mathrm{E}_{p_\theta}\{c\} = \frac{\partial \psi(\theta)}{\partial \theta} \tag{9.19}$$

$$\mathrm{Cov}_{p_\theta}(c) = (g_{ij}(\theta)) = (\frac{\partial^2 \psi(\theta)}{\partial \theta_i \partial \theta_j}). \tag{9.20}$$

When $\psi(\theta)$ is strictly convex we can ask for its Legendre transform $L(\eta) = \theta^{\mathrm{T}} \eta - \psi(\theta)$, where $\theta(\eta)$ is uniquely given by the equations

$$\eta = \frac{\partial \psi(\theta)}{\partial \theta}. \tag{9.21}$$

The parameters $\eta \in \mathbb{R}^m$ are in view of (9.19) called the *expectation parameters* of the EF and $L(\eta) = \mathrm{E}_{p_\theta}\{\theta^{\mathrm{T}} c - \psi(\theta)\} = \int p_\theta \log p_\theta dP$ shows that $L(\eta)$ is the negative entropy of the density $p_{\theta(\eta)}$ [2].

If the EF is regarded as a manifold S, the transformation $\theta \mapsto \eta$ is just a change of coordinates whose Jacobian matrix is given by $\frac{\partial \eta_i}{\partial \theta_j} = \frac{\partial^2 \psi(\theta)}{\partial \theta_i \partial \theta_j} = g_{ij}(\theta)$. The Fisher information defines a metric tensor on S according to $< \frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j} >= g_{ij}(\theta)$. The relation between the coordinate vector fields of the two systems θ and η, written as $\frac{\partial}{\partial \eta_i} = \frac{\partial \theta_j}{\partial \eta_i} \frac{\partial}{\partial \theta_j} = g^{ij}(\theta) \frac{\partial}{\partial \theta_j}$, shows that the inner product between two such coordinate vectors is $< \frac{\partial}{\partial \eta_i}, \frac{\partial}{\partial \theta_j} >= g^{ik} < \frac{\partial}{\partial \theta_k}, \frac{\partial}{\partial \theta_j} >= g^{ik} g_{kj} = \delta^i_j$, i.e. the coordinate systems θ and η are mutually dual [2].

The projection filter equation (9.15) expressed in terms of the coordinates η becomes

$$\begin{aligned} d\eta_t = g(\theta_t) \circ d\theta_t = {} & \mathrm{E}_{p_{\theta(\eta_t)}}\{\mathcal{L}_t c\}\, dt \\ & - \mathrm{E}_{p_{\theta(\eta_t)}}\{\tfrac{1}{2} |h_t|^2_{R_t^{-1}} (c - \eta_t)\}\, dt \\ & + \sum_{k=1}^{d} \mathrm{E}_{p_{\theta(\eta_t)}}\{(R_t^{-1} h_t)^k (c - \eta_t)\} \circ dY_t^k. \end{aligned} \tag{9.22}$$

By noting that $\eta = \mathrm{E}_{p_\theta}\{c\}$ and writing all expectations $\mathrm{E}_{p_\theta}\{...\}$ in (9.22) as $\widehat{(...)}$, one immediately recognizes the resulting expression as the Fujisaki-Kallianpur-Kunita equation for the moments $\widehat{c}$ (see equation (F.5) in the appendix, with $\phi = c$). The additional assumption in (9.22) is of course that the conditional probability density a fortiori belongs to the expectation family S. As [8] points out, this brings about a connection with the Stratonovich-based assumed density filter.

9.6 Numerical implementation

It is not clear how the exponential family S should be chosen. The choice is certainly problem specific. Brigo et al. [8] presents the following simplifying situation. Suppose it is possible to write $|h_t(x)|^2_{R_t^{-1}}$ and the components of $h_t(x)$ as (time dependent) linear combinations of the components of $c(x)$ defining the exponential family, i.e. suppose there are vectors $\lambda_t^0, \lambda_t^1, ..., \lambda_t^d \in \mathbb{R}^m$such that

$$\begin{aligned} \tfrac{1}{2}\,|h_t(x)|^2_{R_t^{-1}} &= (\lambda_t^0)^{\mathrm{T}} c(x) \\ h_t^k(x) &= (\lambda_t^k)^{\mathrm{T}} c(x), \qquad k = 1, ..., d. \end{aligned} \tag{9.23}$$

From (9.19) and (9.20) then follows that the filter equation (9.15) takes the form

$$d\theta_t = \; g^{-1}(\theta_t)\, \mathrm{E}_{p_{\theta_t}} \{\mathcal{L}_t c\}\, dt - \lambda_t^0 dt + \Lambda_t R_t^{-1} dY_t. \tag{9.24}$$

where we have collected the vectors $\lambda_t^1, ..., \lambda_t^d$ in a matrix $\Lambda_t = [\lambda_t^1 ... \lambda_t^d]$. The coefficients of dY_t in the right side are now deterministic, which makes the equation easier to solve numerically since the ordinary Euler scheme coincides with the Milshtein scheme [22]. However, even in the simple case when the components of $c(x)$ are polynomials, the first term in the right hand side of (9.24) is exceedingly expensive to calculate. Another drawback is the lack of any error bounds.

10. Filter Implementation

10.1 Particle Projection Filter: Background

For discrete-time observations

$$y_n = h_n(x_{t_n}) + \varepsilon_n, \tag{10.1}$$

where $0 \leq t_1 < ... < t_n < ...$ and $\{\varepsilon_n\}_{n\geq 1}$ is a white random sequence in $\mathbb{R}^d$ independent of $\{x_t\}$ with probability density $q_n(\varepsilon)$, the filtering problem falls into two parts. *Prediction*: between two observations at t_{n-1} and t_n, the conditional density $p_t = p_t(\cdot|Y_{n-1})$ of x_t given the sequence of past observations $Y_{n-1} = \{y_1, ..., y_{n-1}\}$ evolves according to the Fokker-Planck equation

$$\frac{\partial p_t}{\partial t} = \mathcal{L}_t^* p_t, \qquad t_{n-1} \leq t < t_n.$$

Correction: the new observation at t_n is incorporated by Bayes rule

$$p_{t_n}(x|Y_n) = \frac{\Psi_n(x) p_{t_n}(x|Y_{n-1})}{\int \Psi_n(x) p_{t_n}(x|Y_{n-1}) dx} =: (\Psi_n \cdot p_{t_n})(x) \tag{10.2}$$

where $\Psi_n(x)$ is the likelihood function, i.e. the conditional probability of observing y_n given x_t, $p(y_n|x_t = x) = q_n(y_n - h_n(x))$. The operator in the right hand side is the *projective product*; note that the normalization of Ψ_n is arbitrary. Here we assume that $p(y_n|x_t, Y_{n-1}) = p(y_n|x_t)$, which follows from the mutually independence of $\{\varepsilon_n\}$. In the Bayesian paradigm, $p_{t_n}(x|Y_{n-1})$ is referred to as the *prior distribution* and $p_{t_n}(x|Y_n)$ as the *posterior distribution*.

If we apply the projection filter to this case, the prediction step consists in solving the ODE

$$d\theta_t = g^{-1}(\theta_t)\, \mathrm{E}_{p_{\theta_t}}\{\mathcal{L}_t c\}\, dt \tag{10.3}$$

for $t_{n-1} \leq t < t_n$ with $\theta_{t-1} = \theta_{n-1}$. If the observation noise is Gaussian, $\varepsilon_n \sim \mathcal{N}(0, R_n)$, and the exponential family S chosen such that the components of h_n are in the span of the components of c in the sense of (9.23), the correction step becomes just a matter of updating the parameters θ_{t_n}

$$\theta_n = \theta_{t_n} - \lambda_n^0 + \Lambda_n R_n^{-1} y_n \tag{10.4}$$

because Ψ_n is (with proper normalization) a member of S and an exponential family of densities is closed under projective multiplication. Note the obvious connection with the last terms in the right hand side of (9.24) (we will soon return to this fact).

This is the starting point for a paper by Azimi-Sadjadi and Krishnaprasad [4], where they propose a combination of the classical particle filter with aspects from the projection filter of Brigo et al. [8], the *projection particle filter*.

10.2 Particle Projection Filter: Implementation

Figure 10.1 to 10.4 shows the result of a simulation of a strongly nonlinear system

$$x_k = 0.5x_{k-1} + 25\frac{x_{k-1}}{1+(x_{k-1})^2} + 8\cos(1.2k) + v_k \tag{10.5}$$

$$y_k = \frac{(x_k)^2}{20} + \varepsilon_k. \tag{10.6}$$

The initial state is $x_0 = 0.1$ and the noises given by $v_k \sim \mathcal{N}(0, 50)$ and $\varepsilon_k \sim \mathcal{N}(0, 1)$. The initial prior distribution for the filters is $x_0 \sim \mathcal{N}(0, 5)$. The generated true trajectory and the observations are shown in figure 10.1. Figure 10.2 shows the EKF estimate of the posterior mean. We can see that EKF has great difficulties tracking the true state. Figure 10.3 shows the posterior mean estimated by the bootstrap and the particle projection filter. The number of particles is 500 and the exponential family S is simply the family of Gaussians. Both filters track the system pretty well. Figure 10.4 shows an example of the empirical PDF generated by the bootstrap filter at $k = 8$. Note the strong bimodal character of the distribution.

We have also simulated a linear system observed by a cubic sensor

$$x_k = x_{k-1} + 8\cos(1.2k) + v_k \tag{10.7}$$

$$y_k = \frac{(x_k)^3}{60} + \varepsilon_k. \tag{10.8}$$

Here, $v_k \sim \mathcal{N}(0, 10)$, otherwise the same parameters have been used. Figure 10.5 shows the generated true trajectory and the observations. Figure 10.6 shows that even the EKF behaves quite well in this case. Figure 10.7 shows a typical empirical PDF ($k = 20$), which is more Gaussian like in this case.

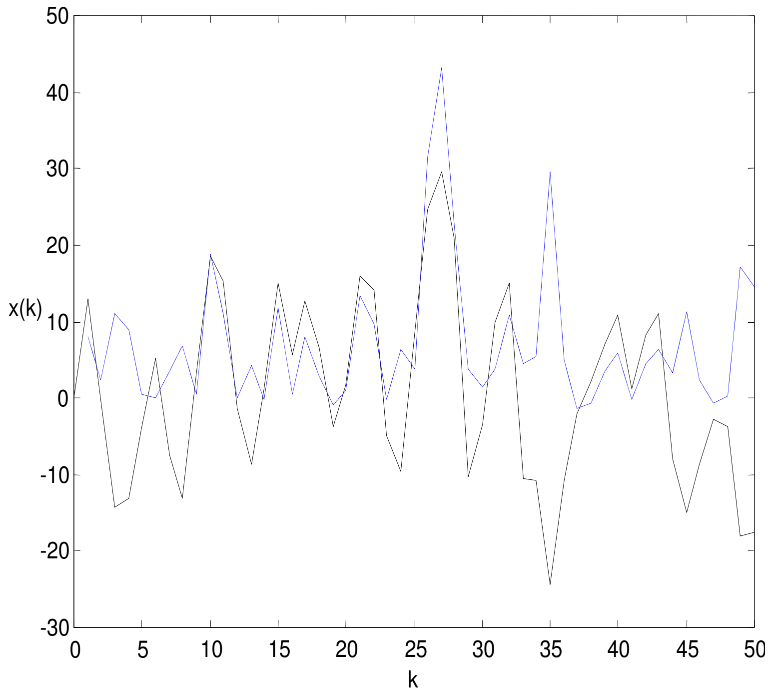


Figure 10.1: 50 point realization of the nonlinear reference model (black) and quadratic observations (blue).

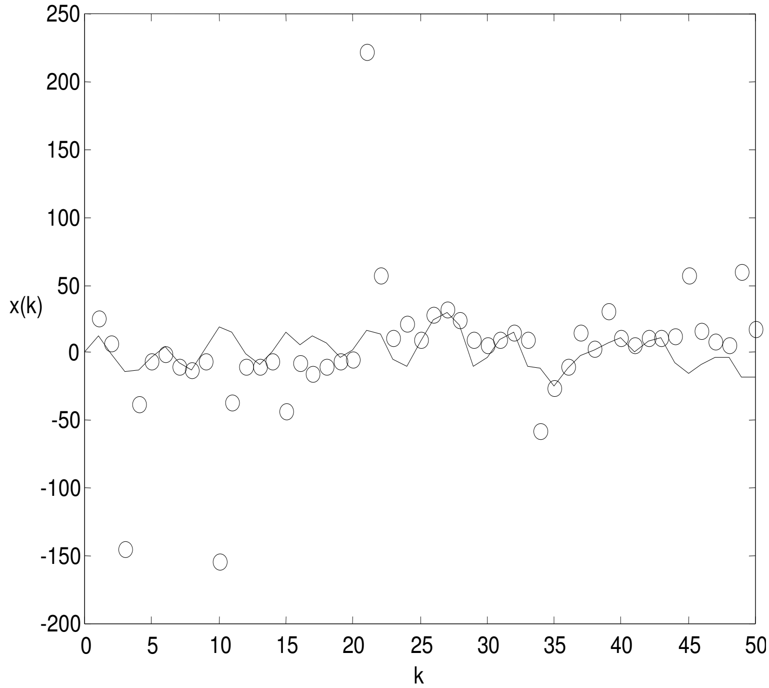


Figure 10.2: EKF estimate of the posterior mean (circles), true state (solid line).

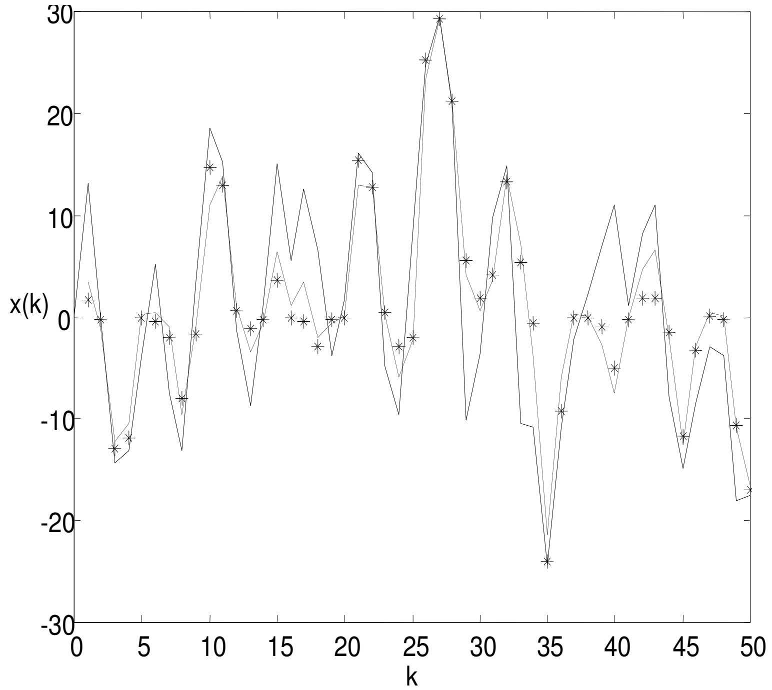


Figure 10.3: Estimate of the posterior mean using the projection particle filter (dotted line) and the bootstrap filter (stars), true state (solid line).

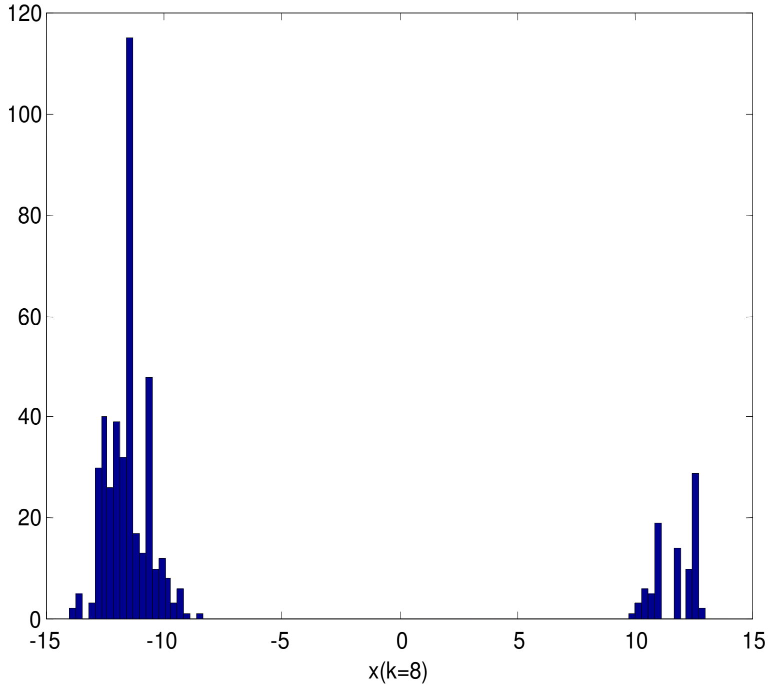


Figure 10.4: The 500 particles of the bootstrap filter sorted in 100 equally spaced bins giving an approximate representation of the conditional probability density at $k = 8$.

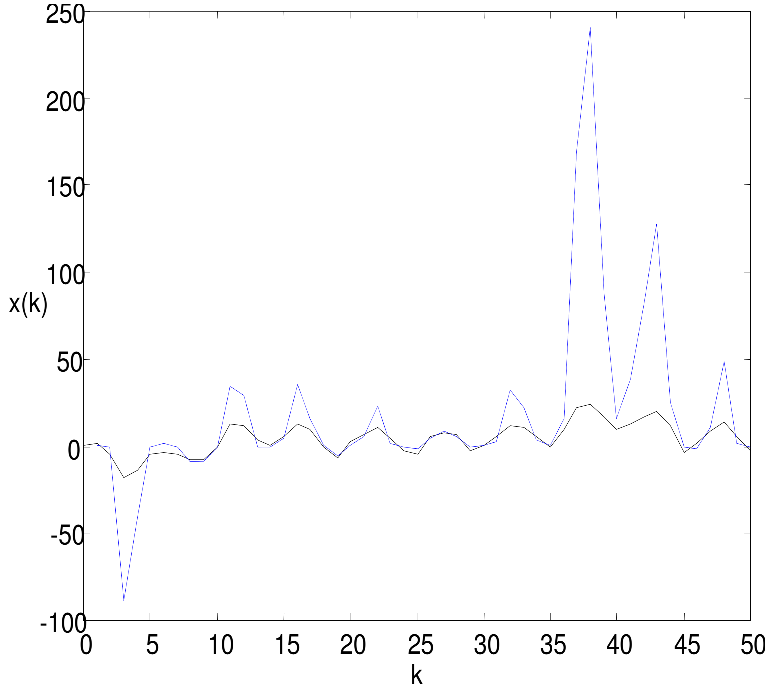


Figure 10.5: 50 point realization of a linear model (black) and cubic observations (blue).

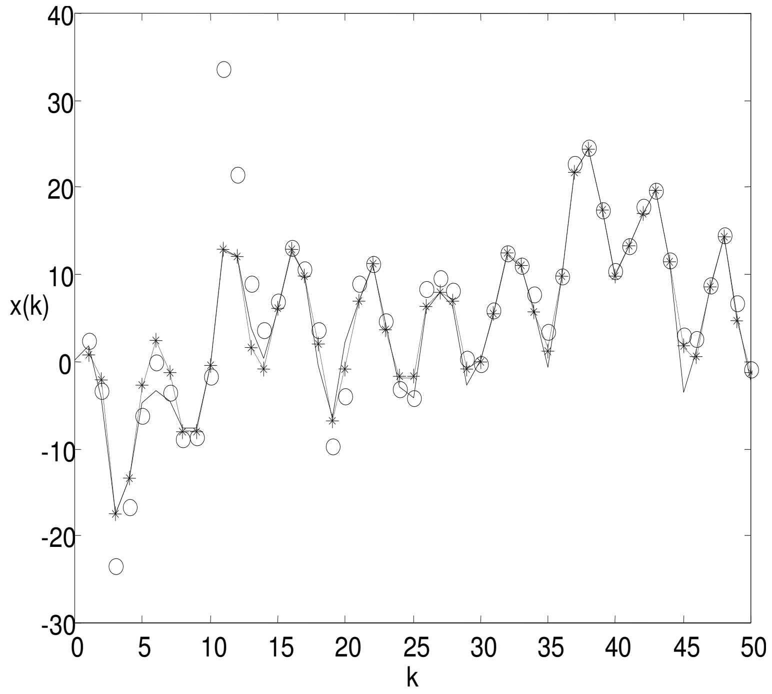


Figure 10.6: Estimate of the posterior mean using EKF (circles), the projection particle filter (dotted line) and the bootstrap filter (stars), true state (solid line).

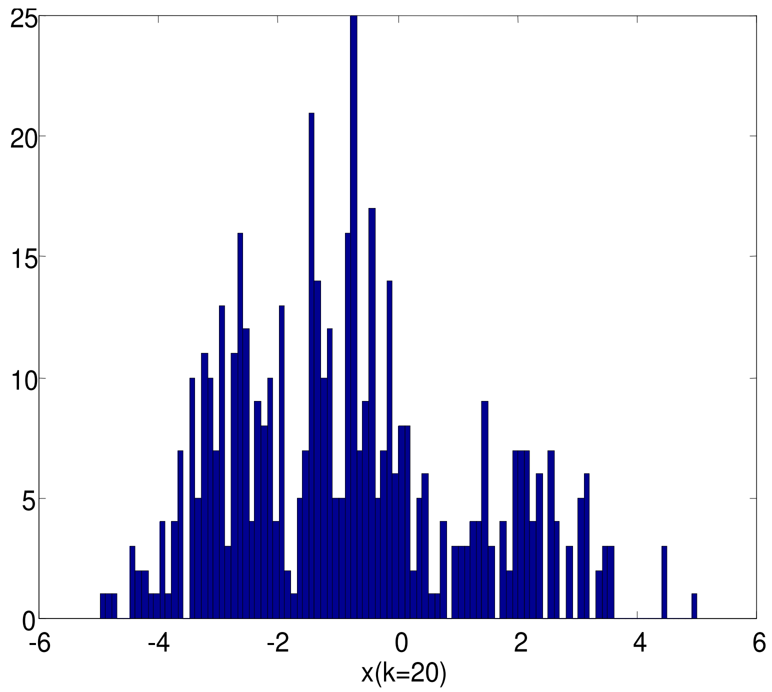


Figure 10.7: The 500 particles of the bootstrap filter sorted in 100 equally spaced bins giving an approximate representation of the conditional probability density at $k = 20$.

A. Properties of x_0, w, and v under the measures $\mathbb{P}$ and $\mathbb{P}_0$

Define the filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$ by $\mathcal{F}_t = \mathcal{F}_t^{\tilde{y}} \vee \mathcal{F}_T^{(x_0,w)}$. We can then apply Girsanov's formula [18, Sec. 5.2] to obtain that under $\mathbb{P}$ the process v defined by $dv_t = d\tilde{y}_t - h(x_t, y_t)\, dt$ is a Wiener process with respect to $\{\mathcal{F}_t\}_{t\in[0,T]}$. Since v is a Wiener process, v_t is independent of $\mathcal{F}_0^{\tilde{y}} \vee \mathcal{F}_T^{(x_0,w)} = \mathcal{F}_T^{(x_0,w)}$, so v is independent of x_0, w. This proves the first two properties of x_0, w, and v. As for the last, by the equality (3.3) and the computations in (B.1) it follows that when restricted to $\mathcal{F}_T^{(x_0,w)}$ the measures $\mathbb{P}$ and $\mathbb{P}_0$ coincide, which means that the pair (x_0, w) has the same probability distributions under $\mathbb{P}$ and $\mathbb{P}_0$.

B. Derivation of the Bayes' formula (3.6)

To begin with we note that since $\mathbb{E}_{\mathbb{P}_0}\Lambda_t \equiv 1$ the process Λ is a martingale under $\mathbb{P}_0$ with respect to $\{\mathcal{F}_t^{(\tilde{y})} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$ (c.f [18, Sec. 3.5.D]), and by (3.3) it is a martingale with respect to $\{\mathcal{F}_t^{(y)} \vee \mathcal{F}_T^{(x_0,w)}\}_{t\in[0,T]}$ as well. The martingale property implies

$$\forall A \in \mathcal{F}_t^{(y)} \vee \mathcal{F}_T^{(x_0,w)} : \quad \mathbb{P}(A) = \int_A d\mathbb{P} = \int_A \Lambda_T \, d\mathbb{P}_0 = \int_A \mathbb{E}_{\mathbb{P}_0}(\Lambda_T | \mathcal{F}_t^{(y)} \vee \mathcal{F}_T^{(x_0,w)}) \, d\mathbb{P}_0 = \int_A \Lambda_t \, d\mathbb{P}_0 = \int_A d\mathbb{P}_t = \mathbb{P}_t(A), \quad \text{(B.1)}$$

and it follows therefore by a standard approximation argument that

$$\forall B \in \mathcal{F}_t^{(y)} : \quad \int_B \mathbb{E}(\phi(x_t)|\mathcal{F}_t^{(y)}) \, d\mathbb{P} = \int_B \mathbb{E}(\phi(x_t)|\mathcal{F}_t^{(y)}) \, d\mathbb{P}_t = \int_B \mathbb{E}(\phi(x_t)|\mathcal{F}_t^{(y)})\Lambda_t \, d\mathbb{P}_0 = \int_B \mathbb{E}(\phi(x_t)|\mathcal{F}_t^{(y)})\mathbb{E}_{\mathbb{P}_0}(\Lambda_t|\mathcal{F}_t^{(y)}) \, d\mathbb{P}_0.$$

This taken together with the relations

$$\forall B \in \mathcal{F}_t^{(y)} : \quad \int_B \mathbb{E}(\phi(x_t)|\mathcal{F}_t^{(y)}) \, d\mathbb{P} = \int_B \phi(x_t) \, d\mathbb{P} = \int_B \phi(x_t) \, d\mathbb{P}_t = \int_B \phi(x_t)\Lambda_t \, d\mathbb{P}_0 = \int_B \mathbb{E}_{\mathbb{P}_0}(\phi(x_t)\Lambda_t|\mathcal{F}_t^{(y)}) \, d\mathbb{P}_0$$

yields the formula in (3.6).

C. Derivation of the Zakai Equation (3.9)

We shall here derive a slightly more general version of the Zakai equation in (3.9). (The added degree of generality is needed in order to obtain the Stratonovich form in (3.11).) All computations will be performed under measure $\mathbb{P}_0$.

C.1 Itô form

Let h be defined as in (3.5) and consider the vector stochastic process $(x^T, y^T, z^T)^T$ where x is the solution to the first equation in (2.24), y is the solution to (3.2) and z is given by

$$dz_t = h^T(x_t, y_t)\, d\tilde{y}_t - \frac{1}{2}\|h(x_t, y_t)\|_2^2\, dt, \quad z_0 = 0, \quad t \in [0, T].$$

Note that all three processes x, y, z are semimartingales driven by the two (independent) $\mathbb{P}_0$-Wiener processes $w, \tilde{y}$. Let further $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be smooth and bounded. If we (after rewriting the differential for x on Itô form using the correction (2.16)) apply Itô's formula (2.18) for semimartingales having Itô differentials to the product $\tilde{\phi}(x_t, y_t)\Lambda_t = \tilde{\phi}(x_t, y_t)\exp(z_t)$ we obtain

$$\begin{aligned}
d\big(\tilde{\phi}(x_t, y_t)\Lambda_t\big) &= \exp(z_t)(\mathcal{A}_+\tilde{\phi})(x_t, y_t)\, dt + \exp(z_t)\big(\nabla_x\tilde{\phi}(x_t, y_t)\big)^T F(x_t)\, dw_t \\
&\quad + \exp(z_t)\Big(\frac{1}{2}\sum_{j,k=1}^{p} d_H^{j,k}(y_t)\frac{\partial^2\tilde{\phi}(x_t, y_t)}{\partial y^j \partial y^k}\, dt\Big) \\
&\quad + \exp(z_t)\Big(\nabla_y\tilde{\phi}(x_t, y_t)\Big)^T H(y_t)h(x_t, y_t)\, dt\Big) \\
&\quad + \exp(z_t)\Big(\tilde{\phi}(x_t, y_t)h^T(x_t, y_t)\, d\tilde{y}_t + \big(\nabla_y\tilde{\phi}(x_t, y_t)\big)^T H(y_t)\, d\tilde{y}_t\Big) \\
&\quad + \exp(z_t)\tilde{\phi}(x_t, y_t)\Big(-\frac{1}{2}\|h(x_t, y_t)\|_2^2\, dt + \frac{1}{2}\|h(x_t, y_t)\|_2^2\, dt\Big) \\
&= \Lambda_t\Big((\mathcal{A}_+\tilde{\phi})(x_t, y_t) + \nabla_y\tilde{\phi}(x_t, y_t)\big)^T H(y_t)h(x_t, y_t)\, dt\Big) \\
&\quad + \Lambda_t\Big(\frac{1}{2}\sum_{j,k=1}^{p} d_H^{j,k}(y_t)\frac{\partial^2\tilde{\phi}(x_t, y_t)}{\partial y^j \partial y^k}\Big)\, dt \\
&\quad + \Lambda_t\big(\nabla_x\tilde{\phi}(x_t, y_t)\big)^T F(x_t)\, dw_t \\
&\quad + \Lambda_t\Big(\big(\nabla_y\tilde{\phi}(x_t, y_t)\big)^T H(y_t) + \tilde{\phi}(x_t, y_t)h^T(x_t, y_t)\Big)\, d\tilde{y}_t,
\end{aligned} \tag{C.1}$$

where $d_H^{j,k}$ is the j:th row, k:th column of the diffusion matrix

$$d_H(y) = H(y)H^T(y), \quad y \in \mathbb{R}^p$$

and $\mathcal{A}_+$ is the operator in (2.23). (In (C.1) and henceforth we use the obvious extension of $\mathcal{A}_+$ to smooth $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ by letting $\mathcal{A}_+$ act only on the first n arguments of $\tilde{\phi}$.) It follows that

$$\tilde{\phi}(x_t, y_t)\Lambda_t = \tilde{\phi}(x_0, 0) + \int_0^t \Lambda_s \mathcal{L}\tilde{\phi}(x_s, y_s)\, ds + \int_0^t \Lambda_s \Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T H(y_s) + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) \Big)\, d\tilde{y}_s + \int_0^t \Lambda_s \big(\nabla_x \tilde{\phi}(x_s, y_s)\big)^T F(x_s)\, dw_s, \quad t \in [0, T], \tag{C.2}$$

where $\mathcal{L}$ is the operator

$$\mathcal{L}\tilde{\phi}(x, y) = (\mathcal{A}_+\tilde{\phi})(x, y) + \big(\nabla_y \tilde{\phi}(x, y)\big)^T H(y) h(x, y) + \frac{1}{2}\sum_{j,k=1}^{p} d_H^{j,k}(y) \frac{\partial^2 \tilde{\phi}(x, y)}{\partial y^j \partial y^k}. \tag{C.3}$$

Since $\tilde{y}$ is Wiener and independent of x_0 under $\mathbb{P}_0$, and x has the same probability distribution under $\mathbb{P}_0$ as under $\mathbb{P}$, we have using (3.3), that

$$\mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_0, 0) | \mathcal{F}_t^{(y)}\big) = \mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_0, 0) | \mathcal{F}_t^{(\tilde{y})}\big) = \mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_0, 0)\big) = \mathbb{E}\big(\tilde{\phi}(x_0, 0)\big). \tag{C.4}$$

Similarly, since $\tilde{y}$ is Wiener (independent increments) and independent of x_0, w it follows using piecewise constant (in time) approximations of integrands (as in the construction of the Itô integral in Sec. 2) that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_0}\Big(\int_0^t \Lambda_s \mathcal{L}\tilde{\phi}(x_s, y_s)\, ds \,\big|\, \mathcal{F}_t^{(y)} \Big) &= \mathbb{E}_{\mathbb{P}_0}\Big(\int_0^t \Lambda_s \mathcal{L}\tilde{\phi}(x_s, y_s)\, ds \,\big|\, \mathcal{F}_t^{(\tilde{y})} \Big) \\
&= \int_0^t \mathbb{E}_{\mathbb{P}_0}\Big(\Lambda_s \mathcal{L}\tilde{\phi}(x_s, y_s) \,\big|\, \mathcal{F}_t^{(\tilde{y})} \Big)\, ds \\
&= \int_0^t \mathbb{E}_{\mathbb{P}_0}\Big(\Lambda_s \mathcal{L}\tilde{\phi}(x_s, y_s) \,\big|\, \mathcal{F}_s^{(\tilde{y})} \Big)\, ds \\
&= \int_0^t \sigma_s\big(\mathcal{L}\tilde{\phi}(x_s, y_s)\big)\, ds,
\end{aligned} \tag{C.5}$$

where we have again used (3.3) for the first and last equalities. Likewise,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_0}\Big(\int_0^t \Lambda_s \Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T H(y_s) + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) \Big)\, d\tilde{y}_s \,\big|\, \mathcal{F}_t^{(y)} \Big) \\
&= \mathbb{E}_{\mathbb{P}_0}\Big(\int_0^t \Lambda_s \Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T H(y_s) + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) \Big)\, d\tilde{y}_s \,\big|\, \mathcal{F}_t^{(\tilde{y})} \Big) \\
&= \int_0^t \mathbb{E}_{\mathbb{P}_0}\Big(\Lambda_s \Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T H(y_s) + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) \Big) \big| \mathcal{F}_t^{(\tilde{y})} \Big)\, d\tilde{y}_s \\
&= \int_0^t \mathbb{E}_{\mathbb{P}_0}\Big(\Lambda_s \Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T H(y_s) + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) \Big) \big| \mathcal{F}_s^{(\tilde{y})} \Big)\, d\tilde{y}_s \\
&= \int_0^t \mathbb{E}_{\mathbb{P}_0}\Big(\Lambda_s \Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) H^{-1}(y_s) \Big) \big| \mathcal{F}_s^{(y)} \Big) H(y_s)\, d\tilde{y}_s = \\
&= \int_0^t \sigma_s\Big(\big(\nabla_y \tilde{\phi}(x_s, y_s)\big)^T + \tilde{\phi}(x_s, y_s) h^T(x_s, y_s) H^{-1}(y_s) \Big)\, dy_s.
\end{aligned} \tag{C.6}$$

Finally, if $s, \tau \in \mathbb{R}_+$ are arbitrary subject to $s, s+\tau \in [0, t]$ we have, by the (in)dependence structure of $\tilde{y}, w, x_0$ (and (3.3) again), that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_0}\big(\Lambda_s \big(\nabla_x \tilde{\phi}(x_s, y_s)\big)^T F(x_s)(w_{s+\tau} - w_s)\big) \,|\, \mathcal{F}_t^{(\tilde{y})}\big) = \\
&\mathbb{E}_{\mathbb{P}_0}\Big(\mathbb{E}_{\mathbb{P}_0}\big(\Lambda_s \big(\nabla_x \tilde{\phi}(x_s, y_s)\big)^T F(x_s)(w_{s+\tau} - w_s)\big) \,|\, \mathcal{F}_t^{(\tilde{y})} \vee \mathcal{F}_s^{(x_0, w)}\big) | \, \mathcal{F}_t^{(\tilde{y})}\Big) = \\
&\mathbb{E}_{\mathbb{P}_0}\big(\Lambda_s \big(\nabla_x \tilde{\phi}(x_s, y_s)\big)^T F(x_s)\, \mathbb{E}_{\mathbb{P}_0}\big(w_{s+\tau} - w_s \,|\, \mathcal{F}_t^{(\tilde{y})} \vee \mathcal{F}_s^{(x_0, w)}\big) \,|\, \mathcal{F}_t^{(\tilde{y})}\big) = \\
&\mathbb{E}_{\mathbb{P}_0}\big(\Lambda_s \big(\nabla_x \tilde{\phi}(x_s, y_s)\big)^T F(x_s)\, \mathbb{E}_{\mathbb{P}_0}\big(w_{s+\tau} - w_s \,|\, \mathcal{F}_s^{(w)}\big) \,|\, \mathcal{F}_t^{(\tilde{y})}\big) = 0.
\end{aligned}$$

By a simple approximation argument we therefore obtain

$$\mathbb{E}_{\mathbb{P}_0}\Big(\int_0^t \Lambda_s\big(\nabla_x\tilde{\phi}(x_s,y_s)\big)^T F(x_s)\,dw_s \,|\, \mathcal{F}_t^{(y)}\Big) =$$
$$\mathbb{E}_{\mathbb{P}_0}\Big(\int_0^t \Lambda_s\big(\nabla_x\tilde{\phi}(x_s,y_s)\big)^T F(x_s)\,dw_s \,|\, \mathcal{F}_t^{(\tilde{y})}\Big) = 0. \tag{C.7}$$

Combining (C.2)–(C.7) yields the result

$$d\sigma_t(\tilde{\phi}) = \sigma_t\big(\mathcal{L}\tilde{\phi}(x_s,y_s)\big)\,dt + \sigma_t\Big(\big(\nabla_y\tilde{\phi}\big)^T + \tilde{\phi}h^TH^{-1}\Big)\,dy_t, \quad t\in[0,T], \tag{C.8}$$

which, in case $\tilde{\phi} = \phi$, (i.e. no y dependence) collapses to (3.9). This concludes the derivation of the Zakai equation on Itô form.

C.2 Stratonovich form

We now turn to the Stratonovich form of the Zakai equation. Consider the martingale part of (3.9). By Itô's product rule [18, p. 155] we have [1]

$$\sigma_t\big(\phi h^TH^{-1}\big)\,dy_t = \sigma_t\big(\phi h^TH^{-1}\big)\circ dy_t - \frac{1}{2}\sum_{j=1}^p d\langle\sigma_{(\cdot)}\big(\phi(H^{-1})^Th\big)^j, y^j_{(\cdot)}\rangle_t,$$
$$t\in[0,T]. \tag{C.9}$$

To evaluate the last term on the right hand side we put $\tilde{\phi} = \phi(H^{-1})^Th$ and apply (C.8) to obtain the differential for $\sigma_t\big(\phi(H^{-1})^Th\big)$ (only the martingale part is of interest). It follows that

$$\sum_{j=1}^p d\langle\sigma_{(\cdot)}\big(\phi(H^{-1})^Th\big)^j, y^j_{(\cdot)}\rangle_t =$$
$$\mathrm{tr}\Big(\Big(\sigma_t\big(\phi\nabla_y^T((H^{-1})^Th)\big) + \sigma_t\big(\phi(H^{-1})^Thh^TH^{-1}\big)\Big)HH^T\Big)\,dt =$$
$$\mathrm{tr}\Big(\sigma_t\big(\phi\nabla_y^T((H^{-1})^Th)HH^T\big) + \sigma_t\big(\phi(H^{-1})^Thh^TH^T\big)\Big)\,dt$$
$$\sigma_t\Big(\mathrm{tr}\big(\phi\nabla_y^T((H^{-1})^Th)HH^T + \phi(H^{-1})^Thh^TH^T\big)\Big), \quad t\in[0,T], \tag{C.10}$$

and the function γ in (3.10) is thus given by

$$\gamma = \frac{1}{2}\mathrm{tr}\big(\nabla_y^T((H^{-1})^Th)HH^T + (H^{-1})^Thh^TH^T\big)$$
$$= \frac{1}{2}\mathrm{tr}\big(\nabla_y^T((H^{-1})^Th)HH^T\big) + \frac{1}{2}\|h\|_2^2,$$

which, in the special case $H_0(x,y) = H_0(x), H(y) = \mathbf{I}$ (so that $h(x,y) = H_0(x)$), becomes

$$\gamma(x,y) = \frac{1}{2}\|H_0(x)\|_2^2, \quad x\in\mathbb{R}^n, y\in\mathbb{R}^p.$$

Combining (C.9), (C.10) and inserting the result in (C.8) gives the Stratonovich form of the Zakai equation.

[1] The brackets $\langle\cdot,\cdot\cdot\rangle$ denote (quadratic) covariation. The quadratic variation between two semi-martingales is defined as the quadratic variation between their martingale parts, cf. [24, Sec. 2.2].

D. Existence of the extended unnormalized density (3.14)

Let $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ be smooth and bounded, say $|\tilde{\phi}| \leq C$. From basic calculus we know that for any $\varepsilon > 0$ there exists a "piecewise constant" function $\varphi : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ of the form

$$\varphi(x, y) = \sum_{j=1}^{N} f_j(x) g_j(y), \quad x \in \mathbb{R}^n, y \in \mathbb{R}^p, \tag{D.1}$$

where $f_j : \mathbb{R}^n \to \mathbb{R}, g_j : \mathbb{R}^p \to \mathbb{R}$ are indicator functions of "rectangles," such that

$$\sup_{x \in \mathbb{R}^n, y \in \mathbb{R}^p} |\tilde{\phi}(x, y) - \varphi(x, y)| < \varepsilon.$$

Using (3.13) we see that the function φ moreover satisfies

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_0}\big(\varphi(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) &= \mathbb{E}_{\mathbb{P}_0}\big(\sum_{j=1}^{N} f_j(x_t) g_j(y_t) \Lambda_t | \mathcal{F}_t^{(y)}\big) \\
&= \sum_{j=1}^{N} g_j(y_t) \mathbb{E}_{\mathbb{P}_0}\big(f_j(x_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) \\
&= \sum_{j=1}^{N} g_j(y_t) \int_{\mathbb{R}^n} f_j(x) q_t^{(y)}(x)\, dx \\
&= \int_{\mathbb{R}^n} \varphi(x, y_t) q_t^{(y)}(x)\, dx.
\end{aligned} \tag{D.2}$$

It follows that for any $A \in \mathcal{F}_t^{(y)}$ we have the following two bounds

$$\begin{aligned}
&\Big| \int_A \Big(\mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) - \mathbb{E}_{\mathbb{P}_0}\big(\varphi(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) \Big)\, d\mathbb{P}_0 \Big| = \\
&\Big| \int_A \big(\tilde{\phi}(x_t, y_t) - \varphi(x_t, y_t)\big)\Lambda_t\, d\mathbb{P}_0 \Big| \leq \\
&\int_\Omega |\tilde{\phi}(x_t, y_t) - \varphi(x_t, y_t)|\, d\mathbb{P}_t \leq \varepsilon
\end{aligned} \tag{D.3}$$

and

$$\begin{aligned}
&\Big| \int_A \Big(\mathbb{E}_{\mathbb{P}_0}\big(\varphi(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) - \int_{\mathbb{R}^n} \tilde{\phi}(x, y_t) q_t^{(y)}(x)\, dx \Big)\, d\mathbb{P}_0 \Big| = \\
&\Big| \int_A \int_{\mathbb{R}^n} \big(\varphi(x, y_t) - \tilde{\phi}(x, y_t)\big) q_t^{(y)}(x)\, dx\, d\mathbb{P}_0 \Big| \leq \\
&\int_\Omega \int_{\mathbb{R}^n} |\varphi(x, y_t) - \tilde{\phi}(x, y_t)|\, q_t^{(y)}(x)\, dx\, d\mathbb{P}_0 \leq \varepsilon \int_{\mathbb{R}^n} q_t^{(y)}(x)\, dx.
\end{aligned} \tag{D.4}$$

If we combine (D.2)–(D.4) we obtain

$$\begin{aligned}
\forall A \in \mathcal{F}_t^{(y)} : \quad & \Big| \int_A \Big(\mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) - \int_{\mathbb{R}^n} \tilde{\phi}(x, y_t) q_t^{(y)}(x)\, dx \Big)\, d\mathbb{P}_0 \Big| \\
& \leq \Big| \int_A \Big(\mathbb{E}_{\mathbb{P}_0}\big(\tilde{\phi}(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) - \mathbb{E}_{\mathbb{P}_0}\big(\varphi(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) \Big)\, d\mathbb{P}_0 \Big| \\
& \quad + \Big| \int_A \Big(\mathbb{E}_{\mathbb{P}_0}\big(\varphi(x_t, y_t)\Lambda_t | \mathcal{F}_t^{(y)}\big) - \int_{\mathbb{R}^n} \tilde{\phi}(x, y_t) q_t^{(y)}(x)\, dx \Big)\, d\mathbb{P}_0 \Big| \\
& \leq \varepsilon\big(1 + \int_{\mathbb{R}^n} q_t^{(y)}(x)\, dx\big).
\end{aligned}$$

Since A and ε are arbitrary, the result (3.14) follows.

E. Derivation of the robust Zakai equation

For fixed but arbitrary $x \in \mathbb{R}^n$ define the process $\zeta^{(y)}(x)$ by

$$\zeta_t^{(y)}(x) = \exp\big(-\tilde{h}(x, y_t)\big) q_t^{(y)}(x), \quad t \in [0, T], \tag{E.1}$$

where $\tilde{h} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is a function such that

$$\nabla_y \tilde{h}(x, y) = \big(H^{-1}(y)\big)^T h(x, y), \quad y \in \mathbb{R}^p, \tag{E.2}$$

and h is defined in (3.5). By Itô's product rule we have (computing under $\mathbb{P}_0$)

$$\begin{aligned} d\zeta_t^{(y)}(x) = \exp\big(-\tilde{h}(x, y_t)\big) \circ dq_t^{(y)}(x) \\ + q_t^{(y)}(x) \circ d\big(\exp\big(-\tilde{h}(x, y_t)\big)\big). \end{aligned} \tag{E.3}$$

The rightmost factor of the last term on the right can be expanded using Itô's "C^3-formula" (i.e. (2.21)) as

$$d\big(\exp\big(-\tilde{h}(x, y_t)\big)\big) = -\exp\big(-\tilde{h}(x, y_t)\big)\big(\nabla_y^T \tilde{h}(x, y_t)\big) \circ dy_t. \tag{E.4}$$

Combining (3.16) with (E.2)–(E.4) now yields

$$\begin{aligned} d\zeta_t^{(y)}(x) &= \exp\big(-\tilde{h}(x, y_t)\big)\hat{\mathcal{A}}^* q_t^{(y)}(x)\, dt \\ &\quad + \exp\big(-\tilde{h}(x, y_t)\big) q_t^{(y)}(x) h^T(x, y_t) H^{-1}(y_t) \circ dy_t \\ &\quad - \exp\big(-\tilde{h}(x, y_t)\big) q_t^{(y)}(x)\big(\nabla_y^T \tilde{h}(x, y_t)\big) \circ dy_t \\ &= \exp\big(-\tilde{h}(x, y_t)\big)\hat{\mathcal{A}}^* q_t^{(y)}(x)\, dt \\ &= \exp\big(-\tilde{h}(x, y_t)\big)\hat{\mathcal{A}}^* \exp\big(\tilde{h}(x, y_t)\big)\zeta_t^{(y)}(x)\, dt. \end{aligned}$$

F. Stratonovich form of the Kushner-Stratonovich equation

We will perform the exercise of transforming the Itô version (9.7) of the Kushner-Stratonovich equation into its Stratonovich form (9.9). A very nice summary of the stochastic calculus needed is given in [31].

Let ϕ be some test function defined on $\mathbb{R}^n$. With the notation $\widehat{(...)}_t := \mathrm{E}_{p_t}\{...\}$, the equation (9.7) formulated in terms of expectations (moments) $\widehat{\phi}_t$ reads

$$\widehat{\phi}_t = \widehat{\phi}_0 + \int_0^t \widehat{\mathcal{L}\phi}_s \, ds + \int_0^t (\widehat{\phi}_s \widehat{h}_s^{\mathrm{T}} R_s^{-1} \widehat{h}_s - \widehat{\phi h}_s^{\mathrm{T}} R_s^{-1} \widehat{h}_s) \, ds$$
$$+ \int_0^t (\widehat{\phi h}_s - \widehat{\phi}_s \widehat{h}_s)^{\mathrm{T}} R_s^{-1} \, dY_s. \tag{F.1}$$

This is sometimes referred to as the Fujisaki-Kallianpur-Kunita equation, often with the notation $\pi_t(...) := \mathrm{E}_{p_t}\{...\}$ used instead. It is the last term (the local martingale) that we will put on S-form. R is symmetric by assumption. Define a vector $r := R^{-1} h$, so $\sum_i r^i R^{ik} = h^k$. The last term in (F.1) may be written

$$\sum_i \int_0^t (\widehat{\phi r^i}_s - \widehat{\phi}_s \widehat{r^i}_s) \, dY_s^i. \tag{F.2}$$

In order to convert this to S-form we need to calculate the covariation process [17, p. 332]

$$[(\widehat{\phi r^k} - \widehat{\phi}\, \widehat{r^k}), Y^k]_t. \tag{F.3}$$

This is done by repeatedly using (F.1) and the Itô-Kunita-Watanabe's theorem [17, theorem 17.11], each time noting that terms of locally finite variation do not contribute to the covariation.

Begin by calculating from (9.6)

$$[Y^i, Y^k]_t = \sum_{j,l} [\int^t \rho_s^{ij} \, dW_s^j, \int^t \rho_s^{kl} \, dW_s^l]$$
$$= \sum_{j,l} \int^t \rho_s^{ij} \rho_s^{kl} \, d[W^j, W^l]_s$$
$$= \sum_{j,l} \int^t \rho_s^{ij} \rho_s^{kl} \delta^{jl} \, ds = \int^t R_s^{ik} \, ds$$
$$\Longrightarrow$$
$$d[Y^i, Y^k]_t = R_t^{ik} \, dt. \tag{F.4}$$

In the following we will not write out time indices for simplicity. Integration by parts [31, eq. (9.3)] gives

$$\widehat{\phi}\, \widehat{r^k} = [\widehat{\phi}, \widehat{r^k}] + \int \widehat{\phi} \, d\widehat{r^k} + \int \widehat{r^k} \, d\widehat{\phi}$$

and the second term in (F.3) becomes

$$[\widehat{\phi}\,\widehat{r^k}, Y^k] = \int \widehat{\phi}\, d[\widehat{r^k}, Y^k] + \int \widehat{r^k}\, d[\widehat{\phi}, Y^k].$$

Using (F.1), (F.2) and (F.4), the two differentials in the right hand side are given by

$$\begin{aligned} d[\widehat{r^k}, Y^k] &= \sum_i (\widehat{r^k r^i} - \widehat{r^k}\,\widehat{r^i})\, d[Y^i, Y^k] = (\widehat{r^k h^k} - \widehat{r^k}\,\widehat{h^k})\, dt \\ d[\widehat{\phi}, Y^k] &= \sum_i (\widehat{\phi r^i} - \widehat{\phi}\,\widehat{r^i})\, d[Y^i, Y^k] = (\widehat{\phi h^k} - \widehat{\phi}\,\widehat{h^k})\, dt. \end{aligned}$$

The first term in (F.3) is calculated in the same manner as

$$[\widehat{\phi r^k}, Y^k] = \sum_i \int (\widehat{\phi r^k r^i} - \widehat{\phi r^k}\,\widehat{r^i})\, d[Y^i, Y^k] = \int (\widehat{\phi r^k h^k} - \widehat{\phi r^k}\,\widehat{h^k})\, ds.$$

The term to be added to (F.2) for its transformation to S-form is

$$\begin{aligned} &\tfrac{1}{2}\sum_k [(\widehat{\phi r^k} - \widehat{\phi}\,\widehat{r^k}), Y^k]_t = \tfrac{1}{2}\int_0^t (\phi\,\widehat{|h|^2_{R^{-1}}} - \widehat{\phi h}^{\mathrm{T}} R^{-1}\widehat{h})\, ds \\ &-\tfrac{1}{2}\int_0^t (\widehat{\phi}\,\widehat{|h|^2_{R^{-1}}} - \widehat{\phi}\,\widehat{h}^{\mathrm{T}} R^{-1}\widehat{h})\, ds - \tfrac{1}{2}\int_0^t (\widehat{h}^{\mathrm{T}} R^{-1}\widehat{\phi h} - \widehat{\phi}\,\widehat{h}^{\mathrm{T}} R^{-1}\widehat{h})\, ds \\ &= \tfrac{1}{2}\int_0^t (\widehat{\phi\,|h|^2_{R^{-1}}} - \widehat{\phi}\,\widehat{|h|^2_{R^{-1}}})\, ds + \int_0^t (\widehat{\phi}\,\widehat{h}^{\mathrm{T}} R^{-1}\widehat{h} - \widehat{\phi h}^{\mathrm{T}} R^{-1}\widehat{h})\, ds. \end{aligned}$$

The last term here is already present in (F.1) so the resulting Stratonovich version of this "FKK"-equation becomes

$$\begin{aligned} \widehat{\phi}_t = \widehat{\phi}_0 &+ \int_0^t \widehat{\mathcal{L}\phi}_s\, ds - \tfrac{1}{2}\int_0^t (\widehat{\phi\,|h|^2_{R^{-1}}}_s - \widehat{\phi}_s \widehat{|h|^2_{R^{-1}}}_s)\, ds \\ &+ \int_0^t (\widehat{\phi h}_s - \widehat{\phi}_s \widehat{h}_s)^{\mathrm{T}} R_s^{-1} \circ dY_s. \end{aligned} \tag{F.5}$$

Under appropriate differentiability assumptions this may be written as the stochastic partial differential equation (9.9) for the conditional density function p_t.

G. Filter toolbox code

In this appendix we present implementations (as Matlab code) of three of the main nonlinear filter techniques; the extended Kalman filter (EKF), the classical Bootstrap (particle) filter and a variant of the Bootstrap filter using an an approximation of the optimal importance function (due to Doucet). Moreover, we present an implementation of the projection particle filter.

```
%%%% MAIN

global STATEMODEL OBSMODEL

%============================
STATEMODEL = 1;
OBSMODEL   = 1;

vsigma = sqrt(10);
nsigma = sqrt(1);

T = 50;
N = 250;
Nthres = 2*N/3;

X0mean = 0;
X0sigma = sqrt(5);
x0 = X0mean + X0sigma*randn;     %0.1;

khist = {0};
%============================



%=== Simulate trajectory ====
x = zeros(T,1);
y = zeros(T,1);

x(1) = statedyn(x0,1) - vsigma*randn;
for k = 2:T
   x(k) = statedyn(x(k-1),k) - vsigma*randn;
end
y = observ(x) + nsigma*randn(size(x));


%=== EKF ===================
xE = zeros(T,1);
PE = zeros(T,1);
```

```
x1 = X0mean;
P1 = X0sigma^2;
for k = 1:T
   [x2 F] = statedyn(x1,k);
   [h H] = observ(x2);
   P2 = P1*F^2 + vsigma^2;
   K = P2*H/(P2*H^2 + nsigma^2);
   x1 = x2 + K*(y(k) - h);
   P1 = P2*(1-K*H)^2 + (nsigma^2)*K^2;
   xE(k) = x1;
   PE(k) = P1;
end


%=== Bootstrap ==============
xB = zeros(T,1);
Xboot = X0mean + X0sigma*randn(1,N);

for k = 1:T
   Xboot = statedyn(Xboot,k) + vsigma*randn(1,N);
   h = observ(Xboot);
   W = gauss(y(k)-h,nsigma);
   if sum(W)<=eps
      disp('Bootstrap spårar ur vid:'), k, j, break
   end
   W = W/sum(W);
   xB(k) = W*Xboot';
   Xboot = mnomres(Xboot,W,N);
   switch k case khist, figure, hist(Xboot, 100), end
end


%=== Doucet linearisation ===
xD = zeros(T,1);
XDouc = X0mean + X0sigma*randn(1,N);
W = ones(1,N)/N;

ii = 0;
for k = 1:T
   f = statedyn(XDouc,k);
   [XDouc pop] = optimport(f,vsigma,nsigma,y(k));
   h = observ(XDouc);
   W = W.*gauss(y(k)-h,nsigma).*gauss(XDouc-f,vsigma)./pop;
   if sum(W)<=eps
      disp('Doucet spårar ur vid:'), k, j, break
   end
   W = W/sum(W);
   xD(k) = W*XDouc';
   if 1/(W*W')<Nthres
      XDouc = mnomres(XDouc,W,N);
      W = ones(1,N)/N;
      ii = ii + 1;
   end
end
```

```
%procentDoucetSIR = 100*ii/T


%=== Proj ==================
xP = zeros(T,1);
XProj = X0mean + X0sigma*randn(1,N);

for k = 1:T
   XProj = statedyn(XProj,k) + vsigma*randn(1,N);
   pm = mean(XProj);
   ps = sqrt(mean(XProj.^2) - pm^2);
   XProj = rejnorm(pm,ps,y(k),nsigma,N);
   xP(k) = mean(XProj);
end


%=== Output =================
figure
t = 1:T;
plot([0 t],[x0; x],'k',t,y,'y',t,xE,'mo',t,xB,'b:',t,xD,'g:',t,xP,'r:')



%%%%%%%%%%%%%%%%%%%%%%%%%
function [f, F] = statedyn(x, k)
%STATEDYN

global STATEMODEL

switch STATEMODEL
case 1
   f = 0.5*x + 25*x./(1 + x.^2) + 8*cos(1.2*k);
   if nargout==2
      F = 0.5 + 25./(1 + f.^2) - 50*(f.^2)./((1 + f.^2).^2);
   end

case 2
   f = x + 8*cos(1.2*k);
   if nargout==2
      F = 1;
   end

end



%%%%%%%%%%%%%%%%%%%%%%%%%
function [h, H] = observ(x)
%OBSERV

global OBSMODEL

switch OBSMODEL
case 1
```

```
   h = x.^2/20;
   H = x/10;

case 2
   h = x;
   H = ones(size(x));

case 3
   h = x.^3/60;
   H = x.^2/20;

end



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function xs = mnomres(x, w, N)
%MNOMRES

xs = zeros(1,N);
R = length(w);
t = cumsum(-log(rand(1,N+1)));
q = cumsum(w);
t = t/t(N+1);
q = q/q(R);

i = 1;
for j = 1:R
   while i<=N & q(j)>t(i)
      xs(i) = x(j);
      i = i+1;
   end
end



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function xs = rejnorm(priorMean, priorSigma, y, nsigma, N)
%REJNORM

xs = zeros(1,N);
Max = gauss(0,nsigma);

for i=1:N
   while 1
      x = priorMean + priorSigma*randn;
      if rand<=gauss(y-observ(x),nsigma)/Max, break, end
   end
   xs(i) = x;
end
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [X, pop] = optimport(f, vsigma, nsigma, y)
%OPTIMPORT

vv = vsigma^2;
nv = nsigma^2;

[h H] = observ(f);

v = 1./(1/vv + H.^2/nv);
m = v.*(f/vv + H/nv.*(y - h + H.*f));

s = sqrt(v);
X = m + s.*randn(size(f));
pop = gauss(X-m,s);



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function y = gauss(x, sigma)
%GAUSS

y = exp(-0.5*(x./sigma).^2)./sigma/sqrt(2*pi);
```

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics.* Benjamin/Cummings Publishing Company, second edition edition, 1978.

[2] S. Amari. *Differential-Geometrical Methods in Statistics.* Lecture Notes in Statistics, Vol. 28. Springer-Verlag, Berlin, 1985.

[3] S. Amari. *Differential geometrical theory of statistics.* Volume 10 of *Lecture Notes – Monograph Series*, **(author?)** [10], 1987.

[4] B. Azimi-Sadjadi and P.S. Krishnaprasad. Approximate nonlinear filtering and its application for gps. Technical Report TR 2001-26, University of Maryland, 2001.

[5] O.E. Barndorff-Nielsen. *Information and Exponential Families.* Wiley, Chichester, 1978.

[6] O.E. Barndorff-Nielsen and D.R. Cox. *Asymptotic Techniques for Use in Statistics.* Chapman & Hall, London, 1989.

[7] Václav E. Benes and Ioannis Karatzas. On the relation of zakai's and mortensen's equations. *SIAM Journal Control on Control and Optimization*, 21:472–489, 1983.

[8] B. Hanzon D. Brigo and F. Le Gland. A differential geometric approach to nonlinear filtering: The projection filter. Technical Report Publication Interne 914, IRISA, 1995.

[9] B. Hanzon D. Brigo and F. Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5(2):495–534, 1999.

[10] S. Gupta (ed). *Differential Geometry in Statistical Inference*, volume 10 of *Lecture Notes – Monograph Series.* Inst. of Math. Stat., 1987.

[11] K.D. Elworthy. *Stochastic Differential Equations on Manifolds*, volume 70 of *Lecture Notes of London Mathematical Society.* Cambridge University Press, Cambridge, 1982.

[12] R. V. Gamkrelidze, editor. *Geometry I. Basic Ideas and Concepts of Differential geometry.* Encylopaedia of Mathematical Sciences. Springer-Verlag, 1991.

[13] Victor Guillemin and Shlomo Sternberg. *Geometric Asymptotics.* Mathematical Surveys. American Mathematical Society, Providence, Rhode Island, 1977.

[14] Sigurdur Helgason. *Differential Geometry, Lie Groups and Symmetric Spaces.* Academic Press, 1978.

[15] N.J. Hicks. *Notes on Differential Geometry.* Van Nostrand, Princeton, 1965.

[16] J.L. Jensen. *Saddlepoint Approximations.* Oxford University Press, Oxford, 1995.

[17] O. Kallenberg. *Foundations of Modern Probability, 2nd ed.* Springer, New York, 2002.

[18] I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus.* Graduate Texts in Mathematics. Springer-Verlag, New York, 1988.

[19] R.E. Kass. *Introduction*, volume 10 of *Lecture Notes – Monograph Series.* Inst. of Math. Stat., Hayward, California, 1987.

[20] Y. Le Jan K.D. Elworthy and X.-M Li. *On the Geometry of Diffusion Operators and Stochastic Flows.* Number 1720 in Lecture Notes in Mathematics. Springer, 1999.

[21] W.S. Kendall. Stochastic differential geometry: An introduction. *Acta Applicandae Mathematicae*, 9:29–60, 1987.

[22] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations.* Springer, Berlin, 1992.

[23] Ivan Kolar, Peter W. Michor, and Jan Slovak. *Natural Operations in Differential Geometry.* Springer-Verlag, 1993.

[24] H. Kunita. *Stochastic Flows and Stochastic Differential Equations.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, UK, 1990.

[25] S. Kusuoka and D. Stroock. The partial malliavin calclulus and its application to non-linear filtering. *Stochastics*, 12:83–142, 1984.

[26] M. Cohen De Lara. Finite-dimensional filters. part i: The wei-norman technique. *SIAM J. Control Optim*, 35(3):980–1001, 1997.

[27] Michael K. Murray and Joha W. Rice. *Differential Geometry and Statistics.* Monographs on Statistics and Applied Probability. Chapman & Hall, 1993.

[28] H.V. Poor. *An Introduction to Signal Detection and Estimation.* Springer, New York, 1994.

[29] Luen-Fai Tam, Wing Shing Wong, and Stephen S. T. Yau. On a necessary and sufficient condition for finite dimensionality of estimation algebras. *SIAM J. Control and Optimization*, 28(1):173–185, 1990.

[30] Frank W. Warner. *Foundations of Differentiable Manifolds and Lie Groups.* Scott, Foresman and Company, 1971.

[31] D. Williams. *To begin at the beginning*, volume 851 of *Lecture Notes in Mathematics*, pages 1–55. Springer, Berlin, 1981.

[32] D. Williams. *Probability with Martingales.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, UK, 1991.

[33] Wing Shing Wong and Stephen S. T. Yau. The estimation algebra of nonlinear filtering systems. In J Baillieul and J. C. Willems, editors, *Mathematical Control Theory*, pages 33–65. Springer-Verlag, 1999.

[34] C. DeWitt-Morette Y. Choquet-Bruhat and M. Dillard-Bleick. *Analysis, Manifolds and Physics.* North-Holland, Amsterdam, 1982.