

# Optimal approximation of SDEs on submanifolds: the Itô-vector and Itô-jet projections

John Armstrong  
Dept. of Mathematics  
King's College London  
john.1.armstrong@kcl.ac.uk

Damiano Brigo  
Dept. of Mathematics  
Imperial College London  
damiano.brigo@imperial.ac.uk

First version: Sept. 12, 2015. This version: July 10, 2017

## Abstract

We define two new notions of projection of a stochastic differential equation (SDE) onto a submanifold: the Itô-vector and Itô-jet projections. This allows one to systematically develop low dimensional approximations to high dimensional SDEs using differential geometric techniques. The approach generalizes the notion of projecting a vector field onto a submanifold in order to derive approximations to ordinary differential equations, and improves the previous Stratonovich projection method by adding optimality analysis and results. Indeed, just as in the case of ordinary projection, our definitions of projection are based on optimality arguments and give in a well-defined sense “optimal” approximations to the original SDE in the mean-square sense. We also show that the Stratonovich projection satisfies an optimality criterion that is more ad hoc and less appealing than the criteria satisfied by the Itô projections we introduce.

As an application we consider approximating the solution of the non-linear filtering problem with a Gaussian distribution and show how the newly introduced Itô projections lead to optimal approximations in the Gaussian family and briefly discuss the optimal approximation for more general families of distribution. We perform a numerical comparison of our optimally approximated filter with the classical Extended Kalman Filter to demonstrate the efficacy of the approach.

**Keywords:** Stochastic differential equations, Jets, SDEs as jets, SDEs projection on a manifold, SDEs on submanifolds, Stratonovich projection, Itô-vector projection, Itô-jet projection, Optimal projection, Gaussian Itô-jet filter, Gaussian Itô-vector filter.

AMS classification codes: 58A20, 39A50, 58J65, 60H10, 60J60, 65D18

# 1 Introduction

In this paper we define three notions of projecting a stochastic differential equation (SDE) onto a (sub)manifold  $M$ . Our aim is to derive practical numerical methods for solving SDEs and we will illustrate our theory with an example drawn from signal processing.

To explain the general idea, let us first consider projecting an ordinary differential equation (ODE) from the Euclidean space  $\mathbb{R}^r$  onto an  $n$ -dimensional manifold  $M \subseteq \mathbb{R}^r$ . An ODE in  $\mathbb{R}^r$  can be thought of as defining a vector field in  $\mathbb{R}^r$ . At every point  $x \in M$  we can use the Euclidean metric to project the vector at  $x$  onto the tangent space  $T_x M$ . In this way one obtains a vector field on  $M$  which can be thought of as a new ODE on  $M$  that approximates the full ODE in  $\mathbb{R}^r$ . This is illustrated in Figure 1. It is easy to prove that this will be the best way of approximating the ODE in  $\mathbb{R}^r$  with an ODE on  $M$ . To be precise, if the initial condition for an ODE is a point  $x$  on the manifold, then any curve on  $M$  with tangent not equal to the projected vector field will diverge from the solution to the ODE faster than a curve which is tangent to the projected vector field. In this sense, the projected ODE is the only ODE which is “optimal” at each point. This paper addresses the question of how projection can be generalized from ODEs to SDEs. After some brief preliminaries on Itô–Taylor series in Section 2, we answer this question by describing three possible generalizations to SDEs in Section 3.

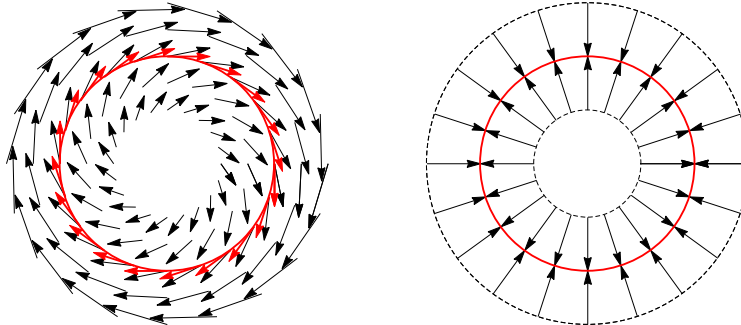


Figure 1: Left: A pictorial representation of the projection of an ODE defined on  $\mathbb{R}^2$  to an ODE defined on the circle. Right: Metric projection from  $\mathbb{R}^2$  to the circle.

The first generalization of projection to SDEs has been proposed previously: what we shall call the Stratonovich projection. The Stratonovich projection is obtained by simply applying the projection operator to the coefficients of the SDE written in Stratonovich form. No optimality result has been derived for the Stratonovich projection. This projection has simply been derived heuristically from the deterministic case. Nevertheless, it appears to be a good approximation in practice and it has been used to find good quality numerical solutions to the non-linear filtering problem (See [9], [10], [3]). The Stratonovich projection is

a natural first choice from the following point of view. As is obvious to anyone with experience of stochastic differential equations on manifolds, and we refer to the monographs and articles [15],[14],[17], [13], [19], [2], simply applying the projection operator to the coefficients of the SDE written in Itô form will not work. This is because solutions to the projected equation don't stay on the manifold, contrary to the Stratonovich case. Nevertheless, we will be able to obtain two modifications of this idea, which we will call the *Itô-vector projection* and the *Itô-jet projection*. These both give well-defined SDEs on the manifold.

We derive the Itô-vector projection by seeking an SDE on the manifold which optimally approximates the original SDE on the manifold when the size of the errors are measured in the mean square ambient metric of  $\mathbb{R}^r$ . The mean squared error between a trajectory following the original SDE and following an SDE on the manifold will typically grow at a rate  $O(t^{\frac{1}{2}})$ . The diffusion term of the projected SDE is determined by minimizing the coefficient of  $t^{\frac{1}{2}}$  in this growth estimate. Choosing the drift term is more delicate, but we give two minimization arguments that indicate that the optimal choice of drift term is given by what we call the Itô-vector projection. The first argument identifies the drift by minimizing the coefficient of the  $O(t)$  term in the estimate of the error, notwithstanding the fact that there is also an  $O(t^{\frac{1}{2}})$  term. The second argument is to find an SDE on the manifold such that the difference between the means of the solutions to the original SDE and to the SDE on the manifold are minimized.

Both of these arguments are somewhat unsatisfying. As an alternative approach we consider finding the SDE on the manifold that most closely tracks the *metric projection* of the solution to the original SDE. The metric projection is the map that sends a point in the ambient space to the closest point of an embedded manifold  $M$ . It is well known to be well-defined and smooth on a tubular neighbourhood. The metric projection is illustrated in Figure 1. It is possible to find an SDE on the manifold such that the mean squared distance between the solutions on the manifold and the metric projection of the solution to the original SDE grows at a rate  $O(t)$ . This requirement determines the diffusion term of the SDE on the manifold and makes the  $O(t^{\frac{1}{2}})$  term coefficient vanish, rather than merely minimize it. Minimizing the coefficient of the order  $t$  term in this estimate determines the drift. We call the SDE determined in this way the *Itô-jet projection*.

It is natural to ask if the Stratonovich projection can also be derived from an optimality argument. We will show that the Stratonovich projection is optimal when using a time-reflection-symmetric optimality criterion anchored to the deterministic initial condition of the process as a special state. We will clarify this notion of optimality in the paper. However, as we will see, for our applications to filtering, the form of optimality achieved by the Stratonovich projection is not particularly useful. This is because the filtering problem is inherently asymmetric in time, as indeed are most applications of SDEs. Nevertheless, it is conceivable that in some applications of SDEs to physics, time reversal symmetry may be a paramount concern. In this case the Stratonovich

projection may be preferred.

Surprisingly the Itô-vector projection, the Itô-jet projection and the Stratonovich projection are all distinct. All of them reduce to classical projection in the case of ODEs. Thus, while optimality arguments lead to a single best method for projecting ODEs, the situation is more complex for SDEs. Since both Itô projections are derived from optimality arguments that are much less ad hoc than the argument for optimality of the Stratonovich projection, there is a clear sense in which they are an improvement upon the Stratonovich projection—both theoretically and in practice.

However, it is not immediately clear whether one should prefer the Itô-vector or the Itô-jet projection. We investigate this question in Section 4.

We consider a simple toy example in Section 4 which we believe strongly suggests that the Itô-jet projection is the better approximation. We also prove a simple theorem that shows how this example can be generalized.

We use this same toy example to illustrate another (entirely non-rigorous) reason for preferring the Itô-jet projection: mathematical aesthetics. As we shall see, each of the different notions of projection is best understood using different formulations of SDEs on manifolds. As its name suggests, the Stratonovich projection is most readily understood using Stratonovich calculus. The Itô-vector projection is most readily understood using the formulation of SDEs on manifolds in terms of Itô calculus first introduced by Itô in [20]. Finally, the Itô-jet projection is most readily understood using the 2-jet formulation of [2]. As we will see, the Itô-jet projection has a very elegant formulation in the language of 2-jets. It is even possible to draw a diagram that allows one to interpret the Itô-jet projection visually. We will present a diagram that visually represents the Itô-jet projection of our toy example. In fact, the development of the 2-jet formulation of SDEs in [2] was originally motivated by the development of these projection methods. It is for this reason that we have called the projections the Itô-vector and Itô-jet projections respectively.

Section 5 is devoted to a detailed calculation of the Itô-jet projection in local coordinates. This calculation amounts to computing the Taylor series for the metric projection map up to second order. This calculation is essential to using the projection for applications.

Section 6 demonstrates how the notion of projection can be applied in practice. In particular, we will apply it to the non-linear filtering problem. We will derive general projection formulae for the non-linear filtering problem. We will then apply this to the problem of approximating a non-linear filter using a Gaussian distribution. A reader who is unfamiliar with non-linear filtering will want to consult Section 6.1 for a brief review.

Gaussian approximations to non-linear filters are widely used in practice (see for example [21, 6]). In particular, the Extended Kalman Filter (EKF) is a popular approximation technique. Other Gaussian approximations exist such as Assumed Density Filters (ADF) and filters derived from the Stratonovich projection. Our theory indicates that all these classical techniques can be improved upon by using the Itô projections (at least over small time intervals). We confirm this with a numerical example.

The utility of the projection method is by no means restricted to the filtering problem nor to such simple approximations as Gaussian filters. Our previous work [5] shows how the Stratonovich projection can be used to generate far more sophisticated filters and it is clear that the idea of projection should be widely applicable in the study of ODEs, SDEs, PDEs and SPDEs. Nevertheless by focussing on Gaussian filters we can examine in detail the idea that there may be many useful ways of approximating an SDE on a submanifold, but that the Itô projections are in some sense optimal amongst these approximation methods. The point we wish to emphasize is that the Itô projections are able to tell us something new even about the well-worn topic of approximating the non-linear filtering problem using Gaussian distributions.

Finally in Section 7 we summarize our findings.

## 2 Stochastic Taylor Series

The main technical tool we will use are stochastic Taylor series. These are described in detail in [22]. In this section we will recall the main definitions and results. We will make some minor notational changes so that we can use the Einstein summation convention.

Let  $X_t$  satisfy a  $d$ -dimensional stochastic differential equation driven by  $m$  independent Brownian motions  $W_t^\alpha$ ,  $\alpha = 1, 2, \dots, m$ . We write

$$dX_t = a(X, t)dt + b_\alpha(X, t)dW_t^\alpha \quad (1)$$

where  $X_t$  is a random process taking values on  $\mathbb{R}^d$ .  $a$  and  $b_\alpha$  are also  $\mathbb{R}^d$  valued for each  $\alpha$ . We are using the Einstein summation convention that when there are matching indices in an expression one should take the sum over the given index. Thus (1) is an abbreviation for:

$$dX_t = a(X, t)dt + \sum_{\alpha=1}^m b_\alpha(X, t)dW_t^\alpha.$$

The advantage of the Einstein summation convention is not simply that it makes formulae shorter. The convention also makes it easier to spot incorrect formulae. This is because, in formulae that are valid in all coordinate systems, the summed indices should always consist of one upper and one lower index.

In this section we will use Greek indices to index the different Brownian motions and Roman indices to index components of vectors in  $\mathbb{R}^d$ . This additional convention is not strictly necessary as the range of the index can be deduced from the position of the index alone.

A multi-index  $\xi$  is defined to be a finite list of integer numbers between 0 and  $m$  and this definition includes the empty list  $()$ . Let  $l(\xi)$  denote the length of  $\xi$ . Let  $n(\xi)$  denote the number of zeros in  $\xi$ . For  $\xi$  with length greater than 0, we define:  $-\xi$  to be the result of removing the first element from  $\xi$ ;  $\xi-$  for the result of removing the last element;  $\xi_1$  for the first element; and  $\xi_{-1}$  for the last element.

Multi-indices enumerate stochastic integrals with respect to the Brownian motions  $W_t^\alpha$  and time. The following definitions are related to those on page 169 of [22]. We define  $W_t^0 := t$  so that the indices equal to 0 correspond to time. We define the multi-integral associated with  $\xi$  by:

$$I_{t_1, t_2}^\xi(f) = \begin{cases} f(t_2) & \text{if } l(\xi) = 0 \\ \int_{t_1}^{t_2} I_{t_1, s}^{\xi^-} dW_s^{\xi-1} & \text{otherwise.} \end{cases}$$

For example the multi-index  $(0, 1, 2)$  is associated with integrating with respect first to time, then  $W^1$ , then  $W^2$ .

$$I_{t_1, t_2}^{(0,1,2)}(f) = \int_{t_1}^{t_2} \int_{t_1}^u \int_{t_1}^v f(w) dw dW_v^1 dW_u^2.$$

We re-express the notation in [22] (page 177, Eqs. 3.1–3.3) by defining differential operators  $L_\xi$  associated to a multi index as:

$$L_\xi f = \begin{cases} f & \text{if } l(\xi) = 0 \\ \frac{\partial f}{\partial t} + a^i \frac{\partial f}{\partial x^i} + \frac{1}{2} b_\alpha^i b_\beta^j g_E^{\alpha\beta} \frac{\partial^2 f}{\partial x^i \partial x^j} & \text{if } \xi = (0) \\ b_{\xi_1}^i \frac{\partial f}{\partial x^i} & \text{if } l(\xi) = 1 \text{ and } \xi \neq (0) \\ L_{\xi_1}(L_{-\xi} f) & \text{otherwise.} \end{cases}$$

Here  $g_E^{\alpha\beta}$  denotes the covariance matrix of the  $d$  Brownian motions  $W_t^\alpha$ . Since we have assumed that the Brownian motions are independent, this will equal the identity matrix. We choose to write  $g_E$  instead of using the Kronecker delta because it transforms as a tensor of type  $(2, 0)$ . In addition, one can simply replace  $g_E$  with the quadratic co-variation tensor if one wishes to consider SDEs driven by more general continuous semi-martingales.

Since  $L_\xi$  contains a total of  $l(\xi) + n(\xi)$  derivatives,  $L_\xi$  acts on functions in  $C^{l(\xi)+n(\xi)}(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R})$ .

The following definition is related to Eq. (9.1) page 206 in [22].

**Definition 1.** The *Itô–Taylor expansion* of order  $\gamma = 0, \frac{1}{2}, 1, \dots$  is given by

$$X_t^\gamma = \sum_{l(\xi)+n(\xi) \leq 2\gamma} L_\xi(x)|_{(t_0, X_{t_0})} I_{t_0, t}^\xi(1)$$

where  $x$  denotes the function  $x(t, X) = X$ . When we speak of the expansion of a given order, we will assume that all the necessary derivatives exist.

The Itô–Taylor expansion allows one to approximate  $X_t$  using  $X_t^\gamma$ . Loosely speaking, this approximation will be accurate in mean squared up to order  $\gamma$ . A precise statement is given in Proposition 1.

**Definition 2.** The *weak Itô–Taylor expansion* of order  $\beta = 0, 1, 2, \dots$  is given by

$$\eta_\beta(t) = \sum_{l(\xi) \leq \beta} L_\xi(x)|_{(t_0, X_{t_0})} I_{t_0, t}^\xi(1)$$

where  $x$  denotes the function  $x(t, X) = X$ . When we speak of the expansion of a given order, we will assume that all the necessary derivatives exist.

The weak Itô–Taylor expansion is of interest if one measures the error using the size of the expectation of the error, rather than the expectation of the size of the error. We will give a precise statement in Proposition 2.

Given a smooth vector valued function  $f$  defined on  $\mathbb{R}^d$  we have by Itô’s lemma that

$$df(X_t) = L_0(f)dt + L_\alpha(f)dW_t^\alpha. \quad (2)$$

The system of equations (1) and (2) define a higher dimensional SDE. We can use this to compute Itô–Taylor expansions for this higher dimensional system and hence compute approximations to  $f(X_t)$ . This calculation gives rise to the following more general definition.

**Definition 3.** The *Itô–Taylor expansion* of order  $\gamma = 0, \frac{1}{2}, 1, \dots$  for  $f(X_t)$  is given by

$$f_t^\gamma = \sum_{l(\xi)+n(\xi)\leq 2\gamma} L_\xi(f)|_{(t_0, X_{t_0})} I_{t_0, t}^\xi(1)$$

When we speak of the expansion of a given order, we will assume that all the necessary derivatives exist. The weak Itô–Taylor expansion for  $f(X_t)$  is defined similarly.

**Lemma 1.** *We suppose that for all  $i$ ,  $W_0^i = 0$ . Given a time  $t$ , and  $i, j \in \{1, \dots, m\}$ , the integrals*

$$\begin{aligned} I_{0, t}^{(0)}(1) &= t \\ I_{0, t}^{(i)}(1) &= W_t^i \\ I_{0, t}^{(i, j)}(1) &= \int_0^t W_s^i dW_s^j \end{aligned}$$

*are orthogonal in expectation.*

*Proof.* We first show that

$$E\left(I_{0, t}^{(i)}(1)I_{0, t}^{(j, k)}(1)\right) = E\left(W_t^i \int_0^t W_s^j dW_s^k\right) = 0 \quad (3)$$

for all  $i, j, k \in \{1, \dots, m\}$ . If  $i \neq j$  then we see, by reversing the sign of  $W^j$ , that

$$E\left(W_t^i \int_0^t W_s^j dW_s^k\right) = -E\left(W_t^i \int_0^t W_s^j dW_s^k\right)$$

Hence (3) is zero unless  $i = j$ . The same argument shows (3) is zero unless  $i = k$ . Finally we note that when  $i = j = k$ , (3) simplifies to

$$E\left(W_t^i \int_0^t W_s^i dW_s^i\right) = E\left(\int_0^t dW_s^i \int_0^t W_s^i dW_s^i\right) = E\left(\int_0^t W_s^i ds\right) = 0$$

by the Itô isometry.

We also need to show that if  $i \neq k$  or  $j \neq l$

$$E \left( I_{0,t}^{(i,j)}(1) I_{0,t}^{(k,l)}(1) \right) = 0.$$

This follows from Lemma 5.7.2 on page 191 of [22].

The other cases are trivial.  $\square$

For completeness, we wish to state some results on the convergence of Itô–Taylor series. We will first need a few more definitions.

First we define spaces  $\mathcal{H}_\xi$  associated with multi-indices  $\xi$ . Associated to the empty index  $()$  we have the set  $\mathcal{H}_()$  of adapted cadlag processes  $f_t$  with

$$|f(t, \omega)| < \infty$$

with probability one for each  $t \geq 0$ .  $\mathcal{H}_()$  consists of the adapted cadlag processes with

$$\int_0^t |f(s, \omega)| ds < \infty$$

with probability one for each  $t \geq 0$ .  $\mathcal{H}_(\alpha)$  has the same definition for any positive  $\alpha$ : it is the set of adapted cadlag processes with

$$\int_0^t |f(s, \omega)|^2 ds < \infty$$

with probability one for each  $t \geq 0$ . We now recursively define  $\mathcal{H}_\xi$  for  $\xi$  of length greater than 1 to be the set of adapted cadlag processes such that the integral process  $I_{0,t}^{\xi^-}(f)$  when viewed as a function of  $t$  lies in  $\mathcal{H}_{\xi_1}$ .

We define  $\mathcal{M}$  to be the set of all multi-indices.

Given a subset  $\mathcal{A} \subseteq \mathcal{M}$  we define the *remainder set*  $\mathcal{B}(\mathcal{A})$  to be the set

$$\mathcal{B}(\mathcal{A}) = \{\xi \in \mathcal{M} \setminus \mathcal{A} : -\xi \in \mathcal{A}\}$$

Thus the remainder set contains all the indices immediately following the indices in  $\mathcal{A}$ . By estimating integrals in the remainder set, one can bound the error of the Itô–Taylor series as we will see below.

We define

$$\Lambda_k = \{\xi \in \mathcal{M} : l(\xi) + n(\xi) \leq k\}.$$

Thus the order  $\gamma$  Itô–Taylor series is a sum over multi-indexes in  $\Lambda_{2\gamma}$ .

We can now state a result on the convergence of the Itô–Taylor series. The following result is a simplified version of Proposition 5.9.1 in [22].

**Proposition 1.** *Suppose that  $L_\xi x|_{X_{t_0}, t_0} \in \mathcal{H}_\xi$  for all  $\xi \in \Lambda_k$ . Suppose that  $L_\xi x|_{X_t, t} \in \mathcal{H}_\xi$  with*

$$\sup_{0 \leq t \leq T} E \left( |(L_\xi x|_{X_t, t})|^2 \right) \leq C_1$$



for all  $\xi \in \mathcal{B}(\Lambda_k)$  and some constant  $C_1$ . Then

$$E \left( |X_t - X_t^{\frac{k}{2}}|^2 \right) \leq C_2(t - t_0)^{k+1}$$

for some constant  $C_2$ . Here  $X_t^{\frac{k}{2}}$  is the order  $\frac{k}{2}$  Itô-Taylor expansion with  $k = 0, 1, \dots$

This next result on the convergence of weak Itô-Taylor series is a restatement of Proposition 5.11.1 in [22].

**Proposition 2.** *Let  $\beta \in \{1, 2, \dots\}$  and  $T \in (0, \infty)$  be given. Let  $C_P^l(\mathbb{R}^d, \mathbb{R})$  denote the space of  $l$  times continuously differentiable functions whose derivatives of order up to and including  $l$  have polynomial growth. Suppose that  $a^k$  and  $b^{kj}$  are time-independent and satisfy Lipschitz conditions, linear growth bounds and belong to  $C_P^{2(\beta+1)}(\mathbb{R}^d, \mathbb{R})$ . Then for each  $g \in C_P^{2(\beta+1)}(\mathbb{R}^d, \mathbb{R})$  there exist constants  $K \in (0, \infty)$  and  $r \in \{1, 2, \dots\}$  such that*

$$\sup_{0 \leq t \leq T} |E(g(X_t) - g(\eta_\beta(t)))| \leq K(1 + |X_0|^{2r})T^{\beta+1}$$

where  $\eta_\beta(t)$  is the weak Itô-Taylor series and the expectation is taken conditional on the information at time 0.

A quick note on how we plan to use the above results to obtain optimal approximations is in order. Take the strong Taylor series to make the point. In our applications the  $X$  we will expand in strong Taylor series will be the difference between the true solution of a SDE and its approximation on a submanifold. Knowing that the error in the Taylor series is bounded as per Propositions 1 (and 2 for the weak case), we will then concentrate on minimizing the mean square of the truncated Taylor series. Minimizing the mean square for the truncated expansion of the difference rather than the mean square of the difference itself will work in view of the convergence guaranteed by the above proposition. In this sense when we talk about “minizing the error” or “difference” later on in the paper we always mean minimizing the truncated expansion mean square.

### 3 Projecting stochastic differential equations

Let  $M$  be an  $n$ -dimensional submanifold of  $\mathbb{R}^r$  with chart  $\psi : U \rightarrow \mathbb{R}^n$  for some open neighbourhood  $U$  in  $M$ . The inverse  $\phi = \psi^{-1}$  gives an embedding of  $\text{Im } \psi$  into  $\mathbb{R}^r$ . The setup is illustrated in Figure 2.

Suppose we are given an Itô SDE on  $\mathbb{R}^r$ ,  $dX_t = a(X_t, t) dt + b_\alpha(X_t, t) dW_t^\alpha$ , that we write in concise form as

$$dX = a dt + b_\alpha dW_t^\alpha, \quad X_0 \tag{4}$$

with  $X_0 \in M$ .

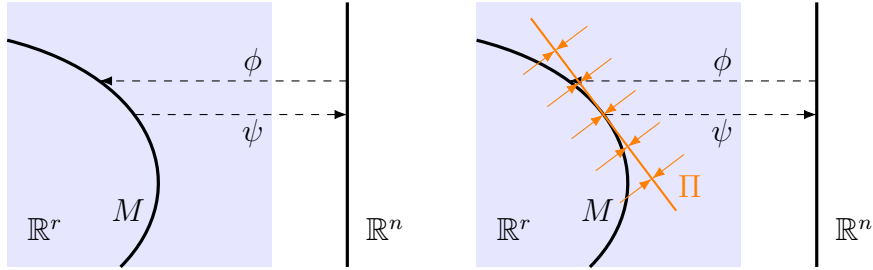


Figure 2: Left: An  $n$ -dimensional manifold  $M$  in  $\mathbb{R}^r$ ,  $r > n$ . Right: Tangent space linear projection used in the Stratonovich and Itô-vector projections

We wish to find an SDE on  $\mathbb{R}^n$  of the form  $dY_t = A(Y_t, t) dt + B_\alpha(Y_t, t) dW_t^\alpha$ , again written concisely as

$$dY = A dt + B_\alpha dW_t^\alpha, \quad Y_0 = \psi^{-1}(X_0), \quad (5)$$

whose mapped solution  $\phi(Y)$  in some sense approximates the solution  $X$  of the original equation on  $\mathbb{R}^r$ . We will consider three approaches.

### 3.1 Stratonovich Projection

**Definition 4.** Let  $W_t$  be an  $\mathbb{R}^m$  valued Brownian motion. Given a Stratonovich SDE on  $\mathbb{R}^r$

$$dX = \bar{a} dt + b_\alpha \circ dW_t^\alpha$$

and a chart  $\psi : U \rightarrow \mathbb{R}^n$  for some neighbourhood in  $M$  we define the *Stratonovich projection* of the SDE to be:

$$dY = \bar{A} dt + B_\alpha \circ dW_t^\alpha$$

where:

$$\bar{A}(Y_t, t) = (\psi_*) \Pi_{\phi(Y_t)}(\bar{a}(\phi(Y_t), t)) \quad (6)$$

$$B_\alpha(Y_t, t) = (\psi_*) \Pi_{\phi(Y_t)}(b_\alpha(\phi(Y_t), t)) \quad (7)$$

where  $\Pi$  is the projection of  $\mathbb{R}^r$  onto  $\phi_*(\mathbb{R}^n)$  defined by the Euclidean metric.

Because we know that projection of vector fields can be defined similarly, and because we know that the coefficients of Stratonovich SDEs transform like vector fields, we see that the definition above defines a Stratonovich SDE on  $M$ . Indeed, if one is willing to accept that projection of vector fields onto a submanifold is well-defined, then one could define the projection of a Stratonovich SDE as the projection of the coefficient functions.

Trying the same method for an Itô SDE does not work. One cannot simply apply projection to the coefficient functions of an Itô SDE because the coefficients of an Itô SDE on a manifold do not transform like vector fields.

The *Stratonovich projection* of an Itô SDE is trivially defined by the recipe:

- (i) rewrite the Itô SDE as a Stratonovich SDE;
- (ii) apply the Stratonovich projection as defined above;
- (iii) rewrite the resulting Stratonovich SDE as an Itô SDE.

In other words, while the definition of Stratonovich projection is most conveniently expressed using Stratonovich calculus, the notion of projection is independent of the calculus used to write down the differential equations.

Linear projection provides the best possible way to approximate vectors in  $\mathbb{R}^r$  with vectors in  $T_X M$ . For ODEs, this implies that the projected ODE is the best possible approximation in  $M$  of the original ODE. However, the situation is different for SDEs. It is not immediately clear how good an approximation the projected Stratonovich SDE solution  $\phi(Y)$  is for the original SDE  $X$  solution. For example, we cannot immediately extend the optimality argument for ODEs to Stratonovich SDEs pathwise, because of the rough paths property of SDEs solutions. In this sense, with the information we have given so far, the definition of the Stratonovich projection is motivated by purely heuristic considerations. Nevertheless, the Stratonovich projection gives good results when applied to approximation of non-linear filtering problems (see [9], [10], [3]) and we will discuss optimality arguments later on, when discussing the Itô-vector projection, and illustrate the time-symmetric optimality of the Stratonovich projection in detail.

In the next sections we will use optimality arguments to derive two alternative notions of projection.

### 3.2 Itô-vector projection

We wish to consider the minimization problem of finding coefficients  $A$  and  $B$  such that the solution of the SDE (5) has the property that  $\phi(Y_t)$  is, in some sense, as close to the solution  $X_t$  of (4) as possible.

The next proposition shows how to give a precise meaning to this notion using the Itô-Taylor expansion.

**Proposition 3.** *Let  $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^d$  and  $F : \mathbb{R}^{d_y} \rightarrow \mathbb{R}^d$  be smooth maps. Let  $x$  be a process on  $\mathbb{R}^{d_x}$  and  $y$  be a process on  $\mathbb{R}^{d_y}$  given by:*

$$\begin{aligned} dx_t &= a(x_t, t) dt + b_\alpha(x_t, t) dW_t^\alpha, & x_0 \\ dy_t &= A(y_t, t) dt + B_\alpha(y_t, t) dW_t^\alpha, & y_0 \end{aligned} \tag{8}$$

with  $f(x_0) = F(y_0)$ . Define

$$z_t = f(x_t) - F(y_t).$$

Let  $z_t^i$  denote the components of the order  $i$  Itô-Taylor expansion for  $z$ . We

have that:

$$\begin{aligned}
E(|z_t^{\frac{1}{2}}|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
E(|z_t^1|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
&\quad + \left( \left| f_*(a(x_0, 0)) - F_*(A(y_0, 0)) \right. \right. \\
&\quad \left. \left. + \frac{1}{2} (\nabla_{b_{\alpha}(x_0, 0)} f_*) b_{\beta}(x_0, 0) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_0, 0)} F_*) B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right|^2 \right. \\
&\quad \left. + \mathcal{R}(f, F, b, B)^2 \right) t^2
\end{aligned} \tag{9}$$

where  $\mathcal{R}(f, F, b, B)$  is a term independent of  $a$ ,  $A$  and  $t$ .

*Proof.* As an example of how to compute the operators  $L_{\mathcal{E}}$  for the system of equations (8), we write down  $L_{(\alpha)}$ .

$$L_{(\alpha)} f = b_{\alpha}^i \frac{\partial f}{\partial x^i} + B_{\alpha}^i \frac{\partial f}{\partial y^i}.$$

Let us now the first few terms of the Itô-Taylor expansion for  $z = f(x) - F(y)$ .

$$\begin{aligned}
L_{(0)}(z) &= f_*(a(x_t, t)) - F_*(A(y_t, t)) \\
&\quad + \frac{1}{2} (\nabla_{b_{\alpha}(x_t, t)} f_*) b_{\beta}(x_t, t) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_t, t)} F_*) B_{\beta}(y_t, t) g_E^{\alpha\beta}. \\
L_{(\alpha)}(z) &= f_*(b_{\alpha}(x_t, t)) - F_*(B_{\alpha}(y_t, t)). \\
L_{(\alpha, \beta)}(z) &= L_{(\alpha)} L_{(\beta)}(z) \\
&= b_{\alpha}^i(x_t, t) \frac{\partial}{\partial x^i} f_*(b_{\beta}(x_t, t)) - B_{\alpha}^i(y_t, t) \frac{\partial}{\partial y^i} F_*(B_{\beta}(y_t, t)).
\end{aligned}$$

We can now write down the order 1 Itô-Taylor expansion  $z_t^1$ . It is

$$\begin{aligned}
z_t^1 &= \left( f_*(a(x_0, 0)) - F_*(A(y_0, 0)) \right. \\
&\quad \left. + \frac{1}{2} (\nabla_{b_{\alpha}(x_0, 0)} f_*) b_{\beta}(x_0, 0) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_0, 0)} F_*) B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right) I_{0,t}^{(0)} \\
&\quad + \left( f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0)) \right) I_{0,t}^{(\alpha)} \\
&\quad + \left( b_{\alpha}^i(x_0, 0) \frac{\partial}{\partial x^i} f_*(b_{\beta}(x_0, 0)) - B_{\alpha}^i(y_0, 0) \frac{\partial}{\partial y^i} F_*(B_{\beta}(y_0, 0)) \right) I_{0,t}^{(\alpha, \beta)}
\end{aligned}$$

We can now use Lemma 1 to calculate  $E(|z_t^1|^2)$ . This gives the desired result.  $\square$

**Remark 1.** For readers familiar with the traditional Itô formula in Euclidean spaces, the term  $(\nabla_{b_{\alpha}(x_0, 0)} f_{i,*}) b_{\beta}(x_0, 0) g_E^{\alpha\beta}$  for the  $i$ -th component of  $f$  might be more familiar when written as

$$(\nabla_{b_{\alpha}} f_{i,*}) b_{\beta} g_E^{\alpha\beta} = \text{Tr} [b^T (H f_i) b]$$

where  $Tr$  is the trace operator and  $H$  is the Hessian operator.

**Theorem 1** (Itô–Taylor series and Itô–vector projection). *Given any time  $t > 0$ , if we wish to find the coefficients  $A$  and  $B$  at time 0 for which the solution to Equation (5) is as close as possible to the solution to Equation (4) in the sense that the mean square ( $L^2$ ) norm of the order  $\frac{1}{2}$  Itô–Taylor series for  $X_t - \phi(Y_t)$  is minimized, we must take*

$$B_\alpha(Y_0, 0) = (\psi_*)_{X_0} \Pi_{X_0} b_\alpha(X_0, 0)$$

where  $\Pi_{X_0}$  is the projection map onto the tangent space of  $M$  at  $X_0$ . If we now suppose that  $B$  is chosen so that this minimum is achieved at all points of  $U$ , a neighborhood of  $X_0$  in  $M$ , then the mean square  $L^2$  norm of the order 1 Itô–Taylor series is minimized by taking

$$A(Y_0, 0) = (\psi_*)_{X_0} \Pi_{X_0} \left( a(X_0, 0) - \frac{1}{2} (\nabla_{B_\alpha(Y_0, 0)} \phi_*) B_\beta(Y_0, 0) g_E^{\alpha\beta} \right).$$

*Proof.* We apply Proposition 3 taking  $f$  equal to the identity,  $F$  equal to  $\phi$ ,  $x_t = X_t$  and  $y_t = Y_t$ . To minimize the order  $\frac{1}{2}$  Itô–Taylor series for  $X_t - \phi(Y_t)$  we must solve the problem:

$$\text{Find } B_\alpha(Y_0, 0) \text{ minimizing } \sum_\alpha |\phi_*(B_\alpha(Y_0, 0)) - b_\alpha(X_0, 0)|^2.$$

The solution to this is given by  $B_\alpha(Y_0, 0) = \psi_*(V_\alpha)$  where the vectors  $V_\alpha$  give a solution to the problem:

$$\text{Find } V_\alpha \in \text{Im } \phi_* \text{ minimizing } \sum_\alpha |V - b_\alpha(X_0, 0)|^2.$$

The standard properties of the projection map tell us that  $V_\alpha = \Pi_{X_0} b_\alpha(X_0, 0)$ .

The same argument is used to find the formula for the coefficient  $A$  that minimizes the order 1 Itô–Taylor expansion.  $\square$

This theorem motivates the following definition.

**Definition 5.** The Itô–vector projection of the SDE (4) onto the manifold  $M$  is given in the chart  $\psi$  by the SDE (5) with

$$\begin{aligned} \phi &:= \psi^{-1} \\ B_\alpha(Y_t, t) &:= (\psi_*)_{\phi(Y_t)} \Pi_{\phi(Y_t)} b_\alpha(\phi(Y_t), t) \\ A(Y_t, t) &:= (\psi_*)_{\phi(Y_t)} \Pi_{\phi(Y_t)} \left( a(\phi(Y_t), t) - \frac{1}{2} (\nabla_{B_\alpha(Y_t, t)} \phi_*) B_\beta(\phi(Y_t), t) g_E^{\alpha\beta} \right) \end{aligned} \tag{10}$$

**Remark 2.** *The optimal  $B$  in the above definition is the same we had in the Stratonovich projection in Eq. (7). The optimal  $A$  is different.*

**Corollary 1.** *The Itô-vector projection defines an SDE on the manifold  $M$ . By this we mean that SDE defined on the manifold  $M$  transforms according to Itô's lemma as we change chart  $\psi$ . See [2] for a more detailed discussion of the Itô formulation of SDEs on manifolds.*

*Proof.* The criteria we are using for finding the optimal coefficients of the SDE is given in terms of an estimate of the growth of the difference between the solution to the SDE in  $\mathbb{R}^r$  and the solution to the SDE on the manifold. Since it is expressed in terms of the solutions to the SDE rather than the coefficients of the SDE, the criterion is independent of the choice of chart  $\psi$ .

It follows that the condition we have derived on the coefficients will transform according to Itô's lemma as we change the choice of chart. For an alternative proof by brute-force calculation see [4].  $\square$

We will demonstrate that the Itô-vector projection is distinct from the Stratonovich projection by calculating an explicit example in Section 6.

One criticism of our derivation of the Itô-vector projection is that it is peculiar to worry about minimizing a term of order 1 when we cannot even ensure that the projection is accurate to order  $\frac{1}{2}$ . It seems uncontroversial that choosing the diffusion coefficient  $B$  by the prescription above will yield the best approximation, but will it make much difference to choose  $A$  in the same way?

The reason that choosing  $A$  is important is that the errors of order  $\frac{1}{2}$  due to the approximation of  $b$  will cancel on average. The correct choice of  $A$  yields the optimal average value for the approximation. This is made precise by the next result.

**Theorem 2** (Itô-vector projection and weak Itô-Taylor expansion). *If we wish to find the coefficient  $A$  at time 0 for which the solution to Equation (5) is as close as possible to the solution to Equation (4) in the sense that the norm of the expectation of the order 1 weak Itô-Taylor series for  $X_t - \phi(Y_t)$  is minimized, we must take:*

$$A(Y_0, 0) = (\psi_*)_{X_0} \Pi_{X_0} \left( a(X_0, 0) - \frac{1}{2} (\nabla_{B_\alpha(Y_0, 0)} \phi_*) B_\beta(Y_0, 0) g_E^{\alpha\beta} \right).$$

where  $\phi = \psi^{-1}$ .

*Proof.* The expectation of the weak Itô Taylor expansion of  $X_t - \phi(Y_t)$  is

$$\left( a(X_0, 0) - \phi_*(A(Y_0, 0)) - \frac{1}{2} (\nabla_{B_\alpha(Y_0, 0)} \phi_*) B_\beta(Y_0, 0) g_E^{\alpha\beta} \right) t.$$

The result now follows immediately from the properties of  $\Pi$ .  $\square$

Thus the Itô-vector projection is the choice of  $A$  and  $B$  that simultaneously minimizes the expectation of the error to order  $\frac{1}{2}$  and the error of the expectation to order 1.

Given the Stratonovich–Taylor expansions described in [22], one might wonder if there are versions of Theorems 1 and 2 using Stratonovich–Taylor series in place of Itô–Taylor series? Might these provide a justification for the Stratonovich projection? The answer is negative. The order 1 Stratonovich–Taylor expansion of [22] is in fact equal to the order 1 Itô–Taylor expansion. The difference is simply that the Stratonovich–Taylor expansion is expressed in terms of Stratonovich coefficients and Stratonovich integrals rather than Itô coefficients and integrals. Thus there is no different “Stratonovich” version of Theorem 1.

However, there is a sense in which the Stratonovich projection is optimal in relation with time symmetry. We will address this optimality after introducing two different optimal approximations, the Itô vector and Itô jet projections.

### 3.3 Itô-jet projection

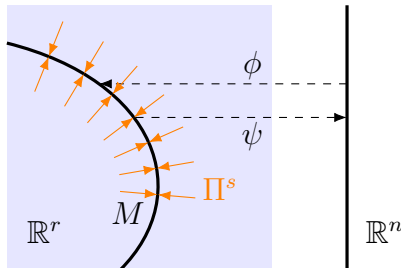


Figure 3: Metric projection  $\Pi^s$  of a tubular neighbourhood of  $M$  in  $\mathbb{R}^r$  onto a neighbourhood  $U$  in  $M$ . This is used to define the Itô-jet projection.

We now suppose that the open set  $U$  inside our manifold  $M$  has been chosen so that we can find a tubular neighbourhood  $N$  of  $U$  such that the metric projection  $\Pi^s$  is smoothly defined on  $N$ . The metric projection is the map sending a point  $x \in \mathbb{R}^r$  to the nearest point in  $N$ . The standard theory of tubular neighbourhoods tells us that if we choose  $U$  small enough, these conditions will apply.

Note that the superscript  $s$  in  $\Pi^s$  is short for smooth and is intended to distinguish this map from the linear projection operator  $\Pi_{\phi(x)}$  onto the tangent space at  $\phi(x)$  ( $x \in \mathbb{R}^n$ ). Since the metric on  $M$  is induced by the  $\mathbb{R}^r$  Euclidean metric, we will have that the tangent-space linear projection,  $\Pi_{\phi(x)}$ , will be the first-order-component or best-linear-approximation of the metric projection,  $\Pi^s$ . See also our explicit calculation in Equation (23) later on.

For ODEs only the first order linear component of the metric projection is necessary to define the projection. However, Itô SDEs involve explicit second order effects, so that there is an actual difference in applying the tangent vector projection or the full metric projection, going beyond the linear term, in approximating a SDE on a submanifold. As we pointed out in [2], an Itô SDE can be interpreted as a 2-jet. It is then not completely surprising that the second

order terms of the metric projection play an important role in understanding the projection of SDEs.

More specifically, in this section we will solve the problem of finding an SDE on the manifold  $M$ ,  $Y_t$  in  $\psi$  coordinates, which minimizes the mean square of the truncated Taylor expansion of the  $M$  geodesic distance between  $\Pi^s(X_t)$  and  $\phi(Y_t)$ , or ambient  $\mathbb{R}^r$  distance between these two points of  $M$ . The two distances will lead to the same result. We call this solution the Itô-jet projection. By contrast, the Itô-vector projection focuses on the  $\mathbb{R}^r$  distance between  $\phi(Y_t)$  and  $X_t$ . Thus the Itô-jet projection uses the metric projection of  $X$  as a benchmark to obtain an optimal approximation  $\phi(Y)$ , whereas the Itô-vector projection uses directly the original  $X$  as a benchmark.

The Itô-jet projection is most neatly defined using the correspondence between 2-jets and SDEs described in [2]. We recall this correspondence now.

Suppose that at each point,  $x$ , of a manifold,  $M$ , we are given a smooth map  $\gamma_x : \mathbb{R}^m \rightarrow M$  with  $\gamma_x(0) = x$ . Suppose also that  $\gamma_x$  depends smoothly on  $x$ . We can then define an SDE on  $M$  driven by  $m$  dimensional Brownian motion by:

$$dX_t^i = \frac{1}{2} g_E^{\alpha\beta} \frac{\partial^2 \gamma^i}{\partial x^\alpha \partial x^\beta}(X_t) dt + \frac{\partial \gamma^i}{\partial x^\alpha}(X_t) dW_t^\alpha.$$

Here  $X^i$  and  $\gamma^i$  denote the components in some coordinate chart on  $M$ . It follows by Itô's Lemma that this SDE is independent of the choice of charts for  $M$ . Since we only use the first two derivatives of  $\gamma$  in this definition, we say that the SDE depends only on the 2-jet of  $\gamma$ .

We can now write down the definition of the Itô-jet projection.

**Definition 6.** Let  $W_t^\alpha$  be independent Brownian motions with  $1 \leq \alpha \leq k$ . Let  $\gamma_x : \mathbb{R}^k \rightarrow \mathbb{R}^r$  be a smoothly varying family of maps satisfying  $\gamma_x(0) = x$  for all  $x \in \mathbb{R}^r$ . We interpret  $\gamma$  as defining an Itô SDE. We define the Itô-jet projection to be the SDE associated with  $\Pi^s \circ \gamma_y : \mathbb{R}^k \rightarrow M$ .

Since this definition only depends upon germs of  $\Pi^s$  and  $\gamma$ , the Itô-jet projection does not depend upon issues such as the tubular neighbourhood used to define  $\Pi^s$ .

We wish to show that the Itô-jet projection solves the problem of finding the best approximation to the SDE on the manifold, if one measures the quality of the approximation using the truncated Itô-Taylor expansion of either the geodesic distance or the distance in the ambient space  $\mathbb{R}^r$ .

**Theorem 3** (Itô-jet projection as optimal approximation). *Let  $\lambda(x, y)$  denote the square of the geodesic distance between two points on  $M$ . Let  $|x - y|_r^2$  denote the square of the distance between two points in the ambient space. If we wish to find the coefficients  $A$  and  $B$  at time 0 for which the solution to (5) is as close as possible to the image on  $M$  of the solution of (4) under  $\Pi^s$  in the sense that the expectation of the square of the order 1 Itô-Taylor expansion of  $\lambda$  or of  $|\cdot|_r^2$  is minimized, we must take*

$$B_\alpha(Y_0, 0) = (\psi_*)_{X_0} \Pi_{X_0} b_\alpha(X_0, 0).$$



If we use this to define  $B$  at all points of  $M$ , we have that the expectation of the square of the order 2 Itô-Taylor expansion of  $\lambda$  or  $|\cdot|_r^2$  is minimized by ensuring that the 2-jet associated with (5) at  $(Y_0, 0)$  is given by  $\tilde{\Pi}^s \circ \gamma_{X_0}$  where  $\gamma_x$  is the 2-jet associated with (4). This results in the following drift for (5):

$$A(Y_0, 0) = \tilde{\Pi}_*^s(a(X_0, 0)) + \frac{1}{2}(\nabla_{b_\alpha(X_0, 0)} \tilde{\Pi}_*^s) b_\beta(X_0, 0) g_E^{\alpha\beta}, \quad (11)$$

where we define  $\tilde{\Pi}^s = \psi \circ \Pi^s$ .

*Proof.* We will first prove the result for the geodesic distance.

It will suffice to prove the result in a single chart. Hence we may assume that our coordinates are normal coordinates based at  $X_0$ .

We have the following Taylor series expansion for the square of the geodesic distance (see for example formula 3.4.3 in [7]):

$$\begin{aligned} \lambda(x, y) &= g_{ij}^E(x^i - y^i)(x^j - y^j) \\ &\quad - \frac{1}{12} R_{ikjl}(x^i + y^i)(x^k - y^k)(x^j + y^j)(x^l - y^l) + O((|x| + |y|)^5). \end{aligned}$$

The first term is just the Euclidean metric on  $\mathbb{R}^n$ , the term  $R_{ikjl}$  denotes the Riemann curvature tensor of  $M$  at the origin.

We can write down the expectation of the order 2 Itô-Taylor expansion of  $|\tilde{\Pi}^s(X_t) - Y_t|^2$  using Proposition 3, taking  $f = \tilde{\Pi}^s$  and  $F$  to be the identity. It is:

$$\begin{aligned} &\sum_{\alpha} |\tilde{\Pi}_*^s(b_\alpha(X_0, 0)) - B_\alpha(Y_0, 0)|^2 t \\ &\quad + \left( \left| \tilde{\Pi}_*^s(a(X_0, 0)) + \frac{1}{2}(\nabla_{b_\alpha(X_0, 0)} \tilde{\Pi}_*^s) b_\beta(X_0, 0) g_E^{\alpha\beta} - A(Y_0, 0) \right|^2 \right. \\ &\quad \left. + \mathcal{R}(\tilde{\Pi}^s, b, B)^2 \right) t^2 \end{aligned} \quad (12)$$

where  $\mathcal{R}(\tilde{\Pi}^s, b, B)$  is a term independent of  $a$  and  $A$ . Our reasoning is that we know that this formula for the expectation of  $|\tilde{\Pi}^s(X_t) - Y_t|^2$  is accurate up to order  $t^2$ , therefore it must equal the expectation of the order 2 Itô-Taylor expansion. This allows us to avoid computing the order 2 Itô-Taylor expansion directly.

The curvature term is fourth order, so it will not influence the order 1 Itô-Taylor expansion for  $\lambda$ . This is because the differential operators  $L_\xi$  in this expansion are all order 2 or less. We deduce that the expectation of the order 1 Itô-Taylor expansion of  $\lambda$  is

$$\sum_{\alpha} |\tilde{\Pi}_*^s(b_\alpha(X_0, 0)) - B_\alpha(Y_0, 0)|^2 t.$$

This is minimized by taking  $B$  as described in the statement of the theorem.

The expectation of an integral  $I_\xi(1)$  is zero if  $\xi$  contains any non-zero entries. This follows by the Martingale property of the Itô integral. Thus the non-zero terms in the expectation of the order 2 Itô expansion for  $\lambda$  correspond to the multi-indices  $(\cdot)$ ,  $(0)$  and  $(0,0)$ . Since the curvature term is fourth order, the only term that will contain a curvature term corresponds to the index  $(0,0)$ . Moreover, only the highest order term of the operator  $L_{0,0}$  is influenced by the curvature. The coefficient of this highest order term may involve only  $b$  and  $B$  but will not involve  $a$  or  $A$ .

Thus the expectation of the order 2 Itô–Taylor expansion is of the form (12) since any curvature correction can be absorbed into the term  $\mathcal{R}(\tilde{\Pi}^s, b, B)^2$ . We deduce that the order 2 Itô–Taylor series is minimized by taking  $A$  as in Equation (11). When these conditions are rewritten in the language of 2-jets, we get the desired result for the metric  $\lambda$ .

The proof for the metric  $|\cdot|$  follows from Lemma 2 given below, and is otherwise essentially identical to that for  $\lambda$ .  $\square$

Note that in this argument we can ensure that the order 1 expansion of  $\lambda$  actually vanishes. By contrast, recall that the corresponding term did not vanish in the derivation of the Itô-vector projection which lead us to give an alternative derivation using the weak Itô–Taylor expansion.

**Lemma 2.** *Let  $U$  be a neighbourhood of the origin in  $\mathbb{R}^n$  and let  $\phi : U \rightarrow \mathbb{R}^r$  be normal coordinates for the Riemannian manifold  $\phi(U)$  centred at the origin, then*

$$|\phi(x) - \phi(y)|_r^2 = |x - y|_n^2 + O((|x|_n + |y|_n)^4).$$

Here  $|\cdot|_n$  is the norm on  $\mathbb{R}^n$ .

*Proof.* Without loss of generality we may assume that the origin is mapped to the origin and the coordinate axes in  $\mathbb{R}^n$  are mapped to the corresponding axes in  $\mathbb{R}^r$ . Given a point  $y \in U$ , we can write the Taylor expansion for the component  $\phi(y)^a$  in the following form:

$$\phi(y)^a = (\delta_n^r)_i^a y^i + A^a_{jk} y^j y^k + O(|y|_n^3). \quad (13)$$

Here  $(\delta_n^r)$  is the tensor representing the projection of  $\mathbb{R}^r$  onto  $\mathbb{R}^n$ . The upper indices of  $(\delta_n^r)$  range from 1 to  $r$  and the lower from 1 to  $n$ .  $(\delta_n^r)_j^i$  is equal to 1 if  $i = j$  and 0 otherwise.  $A^a_{jk}$  is a tensor with upper index  $a$  ranging from 1 to  $r$  and lower indices  $j$  and  $k$  ranging from 1 to  $n$  and which satisfies  $A^a_{jk} = A^a_{kj}$ .

The components of the metric tensor on  $U$  can now be computed as follows:

$$\begin{aligned} g_{ij} &= \left\langle \frac{\partial \phi}{\partial y^i}, \frac{\partial \phi}{\partial y^j} \right\rangle_r \\ &= ((\delta_n^r)_i^a + A^a_{ik} y^k) ((\delta_n^r)_j^b + A^b_{jl} y^l) (g^r)_{ab} + O(|y|_n^2) \end{aligned}$$

Here  $g^r$  is the metric tensor of  $\mathbb{R}^r$ . Our expression for  $g_{ij}$  simplifies to give:

$$g_{ij} = (g^n)_{ij} + A^j_{ik} y^k + A^i_{jk} y^k + O(|y|_n^2).$$

It is well known that in Riemannian normal coordinates the partial derivatives of the metric tensor vanish at the origin. We compute that

$$\partial_k g_{ij}|_0 = A^j_{ik} + A^i_{jk}.$$

So we have  $A^j_{ik} = -A^i_{jk}$ . However, recall that  $A^i_{jk}$  is symmetric in the indices  $j$  and  $k$ . We see that:

$$A^i_{jk} = -A^j_{ik} = -A^j_{ki} = A^k_{ji} = A^k_{ij} = -A^i_{kj} = -A^i_{jk}.$$

So all the components of  $A$  vanish.

We can now use (13) to compute:

$$\begin{aligned} |\phi(x) - \phi(y)|_r^2 &= |(\delta_n^r)_i^a x^i - (\delta_n^r)_i^a y^i|_r^2 + O((|x|_n + |y|_n)^4) \\ &= |x - y|_n^2 + O((|x|_n + |y|_n)^4). \end{aligned}$$

□

### 3.4 Time-symmetric optimality of the Stratonovich projection

Having introduced the Itô vector and Itô jet projections, we are now in a position where we can clarify that also the Stratonovich projection is optimal in a time symmetric sense, even if this optimality is somewhat ad hoc.

Consider the  $x$  SDE (8), in Stratonovich form:

$$dx_t = \bar{a}(x, t)dt + b_\alpha(x, t) \circ dW_t^\alpha, \quad X_0. \quad (14)$$

Recall the Itô Stratonovich transformation

$$\bar{a}_i = a_i - \frac{1}{2} \sum_{j=1}^{d_x} \sum_{\alpha} b_\alpha^j \frac{\partial b_\alpha^i}{\partial x_j}.$$

More generally, by a bar over the drift of an Itô SDE we will mean the drift of the equivalent Stratonovich SDE.

We now extend the SDE to negative time as follows. Define

$$d\xi_t = -\bar{a}(\xi, t)dt - b_\alpha(\xi, t) \circ d\hat{W}_t^\alpha, \quad \xi_0 = X_0 \quad (15)$$

where  $\hat{W}$  is a second standard Brownian motion, independent of  $W$ . Given the symmetric nature of the Stratonovich integral underlying the above SDE and given that formally the chain rule holds, it makes sense to define  $x$  for  $t < 0$  by setting

$$x_{-t} := \xi_t.$$

We now wonder whether the Stratonovich projection could be indeed optimal at time 0 for this SDE extended to negative time at time 0. Suppose that we wish to find the SDE on  $M$

$$dy = \bar{A} dt + B_\alpha \circ dW_t^\alpha, \quad y_0 = x_0, \quad (16)$$

extended similarly to negative time (giving  $y_{-t}$ ), that minimizes the mean square of the truncated Taylor expansion of the vector  $(f(x_t) - F(y_t), f(x_{-t}) - F(y_{-t}))$ . Here  $f$  and  $F$  are functions as defined in Proposition 3.

This optimality criterion is symmetric under time reversal around an anchor state given by a deterministic initial condition. For most applications, for example, when we apply the projection method to filtering, there will be a clear time asymmetry in the problem setting. In these cases, a time-symmetric optimality criterion would not be appropriate. However, in applications to physics one may possibly seek to approximate SDEs that are symmetric under time-reversal in a manner that preserves this symmetry. In this case this criterion would be a natural choice.

Write  $z_t^i$  for the terms in the Itô Taylor expansion of  $z_t = f(x_t) - F(y_t)$ . Proposition 3 states that for positive  $t$ :

$$\begin{aligned}
E(|z_t^{\frac{1}{2}}|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
E(|z_t^1|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
&\quad + \left( \left| f_*(a(x_0, 0)) - F_*(A(y_0, 0)) \right. \right. \\
&\quad \left. \left. + \frac{1}{2} (\nabla_{b_{\alpha}(x_0, 0)} f_*) b_{\beta}(x_0, 0) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_0, 0)} F_*) B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right|^2 \right. \\
&\quad \left. + \mathcal{R}(f, F, b, B)^2 \right) t^2
\end{aligned} \tag{17}$$

The interesting point is that for negative  $t$  the correction terms in this second equation are very different. Recall  $\xi_t = x_{-t}$  and define  $\eta_t = y_{-t}$  so we can write conventional Itô SDEs for  $\xi_t$  and  $\eta_t$ .

$$\begin{aligned}
d\xi_t &= -(a(\xi_t, -t) - (\nabla_{b_{\alpha}(\xi_t, -t)} b_{\beta})(\xi_t, -t) g_E^{\alpha\beta}) dt - b_{\alpha}(\xi_t, -t) d\hat{W}_t^{\alpha} \\
d\eta_t &= -(A(\eta_t, -t) - (\nabla_{B_{\alpha}(\eta_t, -t)} B_{\beta})(\eta_t, -t) g_E^{\alpha\beta}) dt - B_{\alpha}(\eta_t, -t) d\hat{W}_t^{\alpha}.
\end{aligned}$$

By proposition 3 we therefore have the following expressions for  $z_t^i$  for negative

times  $t$ .

$$\begin{aligned}
E(|z_t^{\frac{1}{2}}|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
E(|z_t^1|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
&\quad + \left( \left| f_* \left( -a(x_0, 0) + \nabla_{b_{\alpha}(x_0, 0)} b_{\beta}(x_0, 0) g_E^{\alpha\beta} \right) \right. \right. \\
&\quad \left. \left. - F_* \left( -A(y_0, 0) + \nabla_{B_{\alpha}(y_0, 0)} B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right) \right. \right. \\
&\quad \left. \left. + \frac{1}{2} (\nabla_{b_{\alpha}(x_0, 0)} f_*) b_{\beta}(x_0, 0) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_0, 0)} F_*) B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right|^2 \right. \\
&\quad \left. + \mathcal{R}'(f, F, b, B)^2 \right) t^2
\end{aligned} \tag{18}$$

Combining our results we have for any  $t$ :

$$\begin{aligned}
\frac{1}{2} E(|z_t^{\frac{1}{2}}|^2) + \frac{1}{2} E(|z_{-t}^{\frac{1}{2}}|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
\frac{1}{2} E(|z_t^1|^2) + \frac{1}{2} E(|z_{-t}^1|^2) &= \sum_{\alpha} |f_*(b_{\alpha}(x_0, 0)) - F_*(B_{\alpha}(y_0, 0))|^2 t \\
&\quad + \left( \frac{1}{2} \left| f_* \left( -a(x_0, 0) + \nabla_{b_{\alpha}(x_0, 0)} b_{\beta}(x_0, 0) g_E^{\alpha\beta} \right) \right. \right. \\
&\quad \left. \left. - F_* \left( -A(y_0, 0) + \nabla_{B_{\alpha}(y_0, 0)} B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right) \right. \right. \\
&\quad \left. \left. + \frac{1}{2} (\nabla_{b_{\alpha}(x_0, 0)} f_*) b_{\beta}(x_0, 0) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_0, 0)} F_*) B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right|^2 \right. \\
&\quad \left. + \frac{1}{2} \left| f_* a(x_0, 0) - F_* A(y_0, 0) \right. \right. \\
&\quad \left. \left. + \frac{1}{2} (\nabla_{b_{\alpha}(x_0, 0)} f_*) b_{\beta}(x_0, 0) g_E^{\alpha\beta} - \frac{1}{2} (\nabla_{B_{\alpha}(y_0, 0)} F_*) B_{\beta}(y_0, 0) g_E^{\alpha\beta} \right|^2 \right. \\
&\quad \left. + \mathcal{R}(f, F, b, B)^2 + \mathcal{R}'(f, F, b, B)^2 \right) t^2
\end{aligned} \tag{19}$$

We can now find the optimal projection in a time symmetric sense by mimicking our previous arguments. We will need a lemma that allows us to write down the minimizing  $A$ .

**Lemma 3.** *Let  $(V, g)$  be a Hilbert space containing a closed subspace  $W$ . Let  $v_1$  and  $v_2$  be two vectors in  $V$ . Then the optimization problem*

$$\underset{w \in W}{\text{minimize}} |w - v_1|^2 + |w - v_2|^2$$

has a unique minimizer given by  $w = \Pi_W(\frac{1}{2}v_1 + \frac{1}{2}v_2)$  where  $\Pi_W$  is orthogonal projection onto  $W$ .

*Proof.* By translating the problem by  $\Pi_W(\frac{1}{2}v_1 + \frac{1}{2}v_2)$  we may assume wlog that  $\Pi_W(v_1) = -\Pi_W(v_2)$ . Then for  $w \in W$

$$\begin{aligned} |w - v_1|^2 + |w - v_2|^2 &= |w - \Pi_W v_1|^2 + |w - \Pi_W v_2|^2 + |\Pi_W v_1 - v_1|^2 + |\Pi_W v_2 - v_2|^2 \\ &= |w - \Pi_W v_1|^2 + |w + \Pi_W v_1|^2 + |\Pi_W v_1 - v_1|^2 + |\Pi_W v_2 - v_2|^2 \\ &= 2|w|^2 + 2|\Pi_W v_1|^2 + |\Pi_W v_1 - v_1|^2 + |\Pi_W v_2 - v_2|^2 \end{aligned}$$

This has a unique minimizer,  $w = 0$ .  $\square$

Let us begin with the case  $f = \text{id}$  and  $F = \phi$ . Minimizing the order  $\frac{1}{2}$  expansion requires us to choose  $B$  such that  $\Pi_{\phi(y_t)} b_\alpha(\phi(y_t), t) = \phi_*(B_\alpha(y_t), t)$  at every point. By Lemma 3 and (19), minimizing the order 1 expansion then requires us to choose  $A$  such that

$$\phi_*(A(y_t, t)) = \Pi_{\phi(y_t)} \left( a(\phi(y), t) - \frac{1}{2} \nabla_{b_\alpha(\phi(y_t), t)} b_\beta(\phi(y_t), t) g_E^{\alpha\beta} + \frac{1}{2} \phi_* \nabla_{B_\alpha(\phi(y_t), t)} B_\beta(\phi(y_t), t) g_E^{\alpha\beta} \right).$$

Equivalently

$$\phi_*(\bar{A}(y_t, t)) = \Pi_{\phi(y_t)} (\bar{a}(y, t)).$$

Hence we obtain the Stratonovich projection as the time symmetric analogue of the vector projection.

Now consider the case where  $f = \Pi_s$  and  $F = \text{id}$ . Minimizing the order  $\frac{1}{2}$  expansion requires us to choose  $B$  by  $(\Pi_s)_* b_\alpha = B_\alpha$ . By Lemma 3 and (19), minimizing the order 1 expansion then requires us to choose  $A$  such that

$$A(y_t, t) = (\Pi_s)_* \left( a(\phi(y), t) - \frac{1}{2} \nabla_{b_\alpha(\phi(y_t), t)} b_\beta(\phi(y_t), t) g_E^{\alpha\beta} \right) + \frac{1}{2} \nabla_{B_\alpha(\phi(y_t), t)} B_\beta(\phi(y_t), t) g_E^{\alpha\beta}.$$

Equivalently

$$\bar{A}(y_t, t) = (\Pi_s)_* (\bar{a}(\phi(y), t)).$$

Again, this is the Stratonovich projection.

## 4 A low dimensional example: cross diffusion on a unit circle

We now look at a concrete example which shows the difference between the Itô-vector and Itô-jet projections. Consider the SDE in  $\mathbb{R}^2$  given by

$$\begin{aligned} dX_t &= \sigma Y_t dW_t, \\ dY_t &= \sigma X_t dW_t, \end{aligned} \tag{20}$$

with deterministic initial condition  $(X_0, Y_0)$ . We call this a cross diffusion, since each state crosses over as diffusion coefficient of the other state and the paths

tend to lie on a St Andrew cross, see Appendix A for more details on this process. We wish to project this process equation onto the unit circle given by  $X^2 + Y^2 = 1$ . It is easy to check using Itô's Lemma that if we write  $(X_t, Y_t)$  in polar coordinates as  $(r_t \cos(\theta_t), r_t \sin(\theta_t))$  then  $\theta_t = \arctan(Y_t/X_t)$  satisfies the following exact angular position process equation:

$$d\theta_t = -\frac{1}{2}\sigma^2 \sin(4\theta_t)dt + \sigma \cos(2\theta_t)dW_t, \quad \text{or} \quad d\theta_t = \sigma \cos(2\theta_t) \circ dW_t. \quad (21)$$

Thanks to the special structure of the cross-diffusion, the equation above is already a closed SDE for  $\theta$  without needing to apply any of our projection methods. In this sense we already have the exact angular position SDE and we do not need to project the original  $\mathbb{R}^2$  SDE on the circle  $M$  to approximate the exact angular position with a SDE on the circle. However, we might want to check whether one of our projection methods is consistent with the exact angular position SDE. Let us check how the different projections behave. If we use the same polar coordinate  $\theta$  for the unit circle, we find that the Stratonovich projection and the Itô-jet projection for the  $(X, Y)$  SDE are also given by (21), and are thus consistent with the exact  $\theta$ . However the Itô-vector projection is different and results in:

$$d\theta_t = \sigma \cos(2\theta_t)dW_t.$$

For this example at least, the Itô-jet projection and the Stratonovich projections track the angular position of  $(X_t, Y_t)$  perfectly. Intuitively one might therefore feel that the Stratonovich and Itô-jet projections are “better” approximations to the SDE despite the short time optimality arguments given earlier. It turns out this is a special case of a more general situation, summarized in the following

**Definition 7** (SDE that fibres over a map between manifolds). Let  $f : M \rightarrow N$  be a smooth map between two manifolds. Let  $S$  be an SDE on  $M$  determined by the 2-jets  $\gamma_x : \mathbb{R}^m \rightarrow M$  given at each point  $x \in M$ . We say that  $S$  fibres over  $f$  if  $j_2(f \circ \gamma_{x_1}) = j_2(f \circ \gamma_{x_2})$  whenever  $f(x_1) = f(x_2)$ . This implies that we can define an SDE on the image of  $f$  using the 2-jets  $j_2(f \circ \gamma_x)$  at  $f(x)$ . We call this the *SDE induced by  $f$* .

Returning to projection, we see that we have the following

**Theorem 4** (If SDE fibres over  $\Pi^s$  then Stratonovich = Itô-jet proj.). *If an SDE fibres over the smooth projection map  $\Pi^s$  then the Stratonovich and Itô-jet projection will both be equal to the SDE induced by  $\Pi^s$ .*

*Proof.* This is an immediate consequence of the Stratonovich chain rule in the first case. It is a trivial consequence of the definition of the Itô-jet projection in the second case.  $\square$

Our two-dimensional example of the cross-diffusion on the circle is simply a special case of this more general phenomenon.

It is interesting to note that one can draw a diagram to show the Itô-jet projection. In [2] it is discussed how the jet formulation of SDEs makes it possible to draw pictures of SDEs that transform according to Itô's lemma. For processes driven by one dimensional Brownian motion, one simply finds functions  $\gamma_x$  whose 2-jet represents the SDE and then draws the image of an interval  $[-\epsilon, \epsilon]$  under the map  $\gamma_x$  at each point  $x$ . A picture of this type is shown in Figure 4. It shows how the 2-jets determining the SDE (20) can be projected onto the unit circle simply by composition with  $\Pi^s$ .

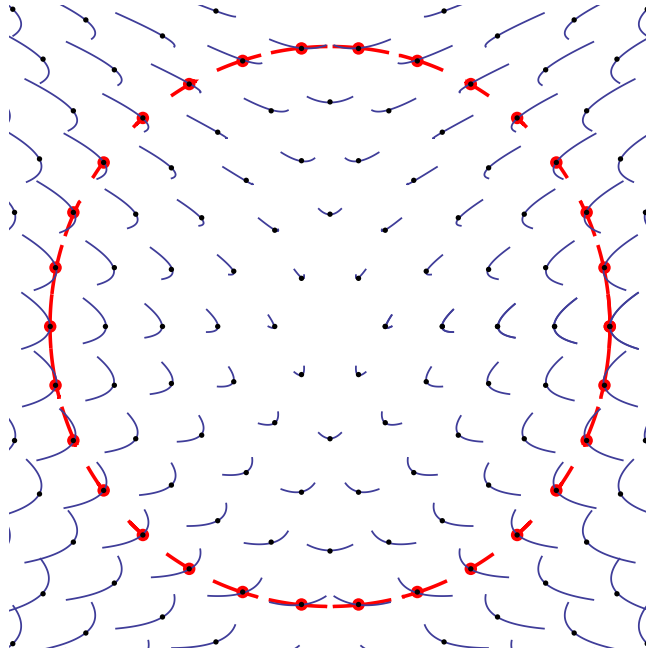


Figure 4: An SDE in  $\mathbb{R}^2$  and its Itô-jet projection onto the unit circle

It seems paradoxical that we derived the Itô-vector projection using optimality arguments that seem to be less ad hoc than for the Stratonovich projection, and yet, for this example, the Itô-vector projection appears manifestly suboptimal.

One possible resolution to this paradox is to say that our notions of tracking  $X_t$  optimally are flawed. Theorem 1 has the weakness that we attempt to minimize a term of order 1 when our approximation is not accurate at order  $\frac{1}{2}$ . Indeed, looking at Equation (17) we see that when we try to minimize the relevant expectation we minimize a combination of terms of order  $t$  and  $t^2$  for the square. Moreover, Theorem 2 has the weakness that we are using the error in the mean to measure the accuracy of our solution. By contrast, the Itô-jet projection has a fully convincing derivation as the optimal approximation of  $\Pi^s(X_t)$  up to order 1.

We will see numerical evidence later that suggests that the Itô-jet projection



performs better in the long term than the Itô-vector projection which lends some support to the idea that the Itô-jet projection is the “right” choice.

We summarize the different projections and the optimality criteria used to determine their drifts in Table 1. The diffusion coefficient is identical for all three projections.

Projection	Properties of drift term
Itô-vector	Minimizes norm of the expectation of the order 1 weak Itô–Taylor expansion between $X$ & $\phi(Y)$ . (ii) Given $B$ minimizing mean square of the order 1/2 strong Itô–Taylor expansion of the difference $X_t - \phi(Y_t)$ , minimizes mean square of order 1 strong Itô–Taylor expansion of the difference.
Itô-jet	Minimizes mean square of 2 Strong Itô–Taylor expansion for $\mathbb{R}^r$ or $M$ distance between $\Pi_s(X)$ & $\phi(Y)$
Stratonovich	Similar to Itô vector below but for the Taylor series of the differences vector $[X_t - \phi(Y_t), X_{-t} - \phi(Y_{-t})]$ at positive and negative time, where negative time processes are defined ad hoc by propagating a second input Brownian motion backward in time.

Table 1: Projections and the associated optimality criteria

## 5 The Itô-jet projection in local coordinates

Our definition of the Itô-jet projection is coordinate free and simple. However, to calculate it in practice we will need an explicit coordinate representation.

We therefore wish to calculate the metric projection map  $\tilde{\Pi}^s = \psi \circ \Pi^s$  up to second order. Then using Itô’s formula for 2-jets we will be able to calculate the Itô-jet projection associated to  $\tilde{\Pi}^s$ .

Most of our calculation involves the deterministic map  $\tilde{\Pi}^s$ . Thus in this section we will drop the convention of using Greek indices exclusively for components of the Brownian motion. In this section we will also use Greek indices to highlight indices over which we are summing. This makes the formulae a little easier to read.

We define the metric tensor on  $U$  by:

$$h_{ab} = \frac{\partial \phi^\alpha}{\partial x^a} \frac{\partial \phi^\alpha}{\partial x^b} \quad (22)$$

The differential  $\tilde{\Pi}_*^s$  of  $\tilde{\Pi}^s$  is well known to be given by the linear projection onto  $\text{Im } \phi_*$  composed with the map  $\phi_*^{-1}$ . Hence  $\tilde{\Pi}_*^s$  is the unique linear map with

$\tilde{\Pi}_*^s \circ \phi_*$  equal to the identity and with kernel equal to the orthogonal complement of  $\text{Im } \phi_*$ . We deduce that  $\tilde{\Pi}_*^s$  has the following components:

$$\Pi_b^a := (\tilde{\Pi}_*^s)_b^a = \frac{\partial \phi^b}{\partial x^a} h^{a\alpha}, \quad a \leq n, \alpha \leq n, b \leq r. \quad (23)$$

We note that the differential or tangent map  $\tilde{\Pi}_*^s$  is the best linear approximation of the metric projection  $\tilde{\Pi}^s$  around the relevant point  $x = \phi(y) \in M$ , and it coincides with the classic linear projection  $\Pi_{\phi(y)}$  on the tangent space of  $M$ . Indeed, Equation (23) shows the classic components of the projection on the tangent space of an  $n$ -dimensional manifold  $M$  embedded in  $\mathbb{R}^r$  and realized as  $\phi$ -image of a subset of  $\mathbb{R}^n$ .

**Lemma 4.** *Suppose for simplicity that  $\phi(0) = 0$  and*

$$(\phi_*)_b^a := \frac{\partial \phi^a}{\partial x^b} = D_b^a := \begin{cases} 1 & a = b \text{ and } a \leq n \\ 0 & \text{otherwise} \end{cases}$$

then  $\Pi^s$  is given up to second order by

$$\tilde{\Pi}^s(y)^a = y^a - \frac{1}{2} \frac{\partial^2 \phi^a}{\partial x^\alpha \partial x^\beta} y^\alpha y^\beta + \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} (\perp)_{\gamma\alpha} y^\alpha y^\beta + O(|y|^3)$$

where we define

$$(\perp)_{ab} = \begin{cases} 1 & a = b \text{ and } a > n \\ 0 & \text{otherwise.} \end{cases}$$

Note that we are using an extension of the Einstein summation convention to cover tensors where some indices range from 1 to  $n$  and some from 1 to  $r$ . Where an index appears twice, we sum over the smaller range. Note also that we are working in a restricted set of coordinate systems, so it no longer holds that all summed pairs of indices will consist of an upper and a lower index.

*Proof.* By our simplifying assumption we may write:

$$\tilde{\Pi}^s(y)^a = y^a + A_{\alpha\beta}^a y^\alpha y^\beta + B_{\alpha\beta\gamma}^a y^\alpha y^\beta y^\gamma + O(|y|^4) \quad (24)$$

where  $A_{\alpha\beta}^a$  is symmetric in  $\alpha$  and  $\beta$  and  $B$  is symmetric in  $\alpha, \beta$  and  $\gamma$ . The Taylor series expansion for  $\phi$  now allows us to compute the components of  $(y - \phi(\tilde{\Pi}^s(y)))$ .

$$\begin{aligned} (y - \phi(\tilde{\Pi}^s(y)))^a &= y^a - D_\alpha^a (y^\alpha + A_{\beta\gamma}^\alpha y^\beta y^\gamma + B_{\beta\gamma\delta}^\alpha y^\beta y^\gamma y^\delta) \\ &\quad - \frac{1}{2} \frac{\partial^2 \phi^a}{\partial x^\alpha \partial x^\beta} (y^\alpha + A_{\delta\epsilon}^\alpha y^\delta y^\epsilon) (y^\beta + A_{\zeta\eta}^\beta y^\zeta y^\eta) + O(|y|^4) \\ &= y^a - D_\alpha^a y^\alpha \\ &\quad - D_\alpha^a A_{\beta\gamma}^\alpha y^\beta y^\gamma - \frac{1}{2} \frac{\partial^2 \phi^a}{\partial x^\alpha \partial x^\beta} y^\alpha y^\beta \\ &\quad - \frac{\partial^2 \phi^a}{\partial x^\alpha \partial x^\beta} A_{\delta\epsilon}^\alpha y^\delta y^\epsilon y^\beta - D_\alpha^a B_{\beta\gamma\delta}^\alpha y^\beta y^\gamma y^\delta + O(|y|^4) \end{aligned}$$

We take the partial derivative of this with respect to  $A_{qr}^p$  to get:

$$\frac{\partial}{\partial A_{qr}^p} (y - \phi(\tilde{\Pi}^s(y)))^a = -D_p^a y^q y^r - \frac{\partial^2 \phi^a}{\partial x^p \partial x^\beta} y^q y^r y^\beta + O(|y|^4).$$

Because of the distance minimizing property of  $\Pi^s$  we know that for all  $p, q, r$  and sufficiently small  $y$  we have:

$$\frac{\partial}{\partial A_{qr}^p} |(y - \phi(\tilde{\Pi}^s(y)))|^2 = 0$$

The left hand side of this expression is equal to:

$$2 \left( \frac{\partial}{\partial A_{qr}^p} (y - \phi(\tilde{\Pi}^s(y)))^a \right) (y - \phi(\tilde{\Pi}^s(y)))^a.$$

We have written down explicit expressions for each term in this product. This enables us to write down the fourth order terms of  $\frac{\partial}{\partial A_{qr}^p} |(y - \phi(\tilde{\Pi}^s(y)))|^2$ . They are given by:

$$\begin{aligned} & 2D_p^a A_{\alpha\beta}^a y^\alpha y^\beta y^q y^r + D_p^a \frac{\partial^2 \phi^a}{\partial x^\alpha \partial x^\beta} y^\alpha y^\beta y^q y^r - 2 \frac{\partial^2 \phi^a}{\partial x^p \partial x^\beta} (y^a - D_\gamma^a y^\gamma) y^\beta y^q y^r \\ &= \left( 2A_{\alpha\beta}^p y^\alpha y^\beta + \frac{\partial^2 \phi^p}{\partial x^\alpha \partial x^\beta} y^\alpha y^\beta - 2 \frac{\partial^2 \phi^a}{\partial x^p \partial x^\beta} (y^a - D_\delta^a y^\delta) y^\beta \right) y^q y^r \\ &= \left( 2A_{\alpha\beta}^p y^\alpha y^\beta + \frac{\partial^2 \phi^p}{\partial x^\alpha \partial x^\beta} y^\alpha y^\beta - 2 \frac{\partial^2 \phi^a}{\partial x^p \partial x^\beta} (\perp)_{\alpha a} y^\alpha y^\beta \right) y^q y^r \end{aligned}$$

We know that this must vanish for all sufficiently small  $y$ . We deduce that

$$2A_{\alpha\beta}^p y^\alpha y^\beta + \frac{\partial^2 \phi^p}{\partial x^\alpha \partial x^\beta} y^\alpha y^\beta - 2 \frac{\partial^2 \phi^a}{\partial x^p \partial x^\beta} (\perp)_{\alpha a} y^\alpha y^\beta = 0.$$

for all sufficiently small  $y$ . This gives us an expression for  $A_{\alpha\beta}^p y^\alpha y^\beta$  which combines with equation (24) to prove the result.  $\square$

We now use the lemma coupled with some coordinate transformations to compute a second order expression for the metric projection in the general case.

**Proposition 4.** *Let  $g_\perp^\phi$  be the symmetric two form on  $\mathbb{R}^r$  defined by:*

$$g_\perp^\phi(X + X^\perp, Y + Y^\perp) = g(X^\perp, Y^\perp) \quad X, Y \in \text{Im } \phi \text{ and } X^\perp, Y^\perp \in (\text{Im } \phi)^\perp$$

where  $g$  is the Euclidean metric on  $\mathbb{R}^r$ . Define coordinates  $\tilde{y}$  centered on  $0 \in R^r$  by  $\tilde{y}^a = y^a - y_0^a$ . Then to second order the metric projection is given by

$$\begin{aligned} \tilde{\Pi}^s(y)^a &= x_0 + \Pi_\alpha^a \tilde{y}^\alpha - \frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\gamma^a \Pi_\delta^\alpha \Pi_\epsilon^\beta \tilde{y}^\delta \tilde{y}^\epsilon \\ &\quad + \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\delta^\beta h^{\alpha\gamma} \tilde{y}^\delta \tilde{y}^\epsilon - \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\epsilon^\beta \Pi_\gamma^\alpha \Pi_\delta^\zeta h_{\eta\zeta} h^{\alpha\gamma} \tilde{y}^\delta \tilde{y}^\epsilon \\ &\quad + O(|y|^3) \end{aligned}$$

and where  $\Pi$  is given by equation 23.

*Proof.* We assume without loss of generality that  $x_0 = 0$  and  $y_0 = \phi(x_0) = 0$ .

We can find a coordinate transformation  $J$  of  $\mathbb{R}^n$  which maps an orthonormal basis of  $\mathbb{R}^n$  to the standard basis vectors. We take  $x$  to be our original coordinates and  $X$  to be the coordinates obtained by applying  $J^{-1}$ . So we have:

$$x^a = J_b^a X^b$$

To satisfy our requirements  $J$  must satisfy:

$$h_{\alpha\beta} J_a^\alpha J_b^\beta = \delta_{ab}$$

Equivalently:

$$h_{ab} = (J^{-1})_a^\alpha (J^{-1})_b^\alpha.$$

So any pseudo square root of  $h_{ab}$  will give an appropriate choice for  $J^{-1}$ . Taking the matrix inverse of the above expression we have:

$$h^{ab} = J_\alpha^a J_\alpha^b \quad (25)$$

We can now find an orthogonal transformation  $T$  of  $\mathbb{R}^r$  mapping  $\text{Im } \phi_*$  to  $\mathbb{R}^n \subseteq \mathbb{R}^r$ . Hence  $\Phi = T \circ \phi \circ J$  satisfies  $(\Phi_*)_b^a = D_b^a$ . We will write  $x$  for the original coordinates on  $\mathbb{R}^n$  and define transformed coordinates  $X$  by:

$$X^a = T_\alpha^a x^\alpha.$$

Let us write  $\Pi'$  for the metric projection associated with the map  $\Phi$ . The various maps we have just defined are summarized in the commutative diagram below:

$$\begin{array}{ccc} \mathbb{R}^n & \begin{array}{c} \xrightarrow{\Phi} \\ \xleftarrow{\Pi'} \end{array} & \mathbb{R}^r \\ J \downarrow & & \uparrow T \\ \mathbb{R}^n & \begin{array}{c} \xrightarrow{\phi} \\ \xleftarrow{\tilde{\Pi}^s} \end{array} & \mathbb{R}^r \end{array}$$

From Lemma 4 we have:

$$\begin{aligned} \Pi'(Y)^a &= \frac{\partial \Phi^a}{\partial X^\alpha} Y^\alpha - \frac{1}{2} \frac{\partial^2 \Phi^a}{\partial X^\alpha \partial X^\beta} Y^\alpha Y^\beta \\ &\quad + \frac{\partial^2 \Phi^\gamma}{\partial X^a \partial X^\beta} (\perp)_{\gamma\alpha} Y^\alpha Y^\beta + O(|Y|^3) \\ &= \frac{\partial (T^{-1} \circ \Phi)^\beta}{\partial X^\alpha} T_\beta^a Y^\alpha - \frac{1}{2} \frac{\partial^2 (T^{-1} \circ \Phi)^\gamma}{\partial X^\alpha \partial X^\beta} T_\gamma^a Y^\alpha Y^\beta \\ &\quad + \frac{\partial^2 (T^{-1} \circ \Phi)^\delta}{\partial X^a \partial X^\beta} T_\delta^\gamma (\perp)_{\gamma\alpha} Y^\alpha Y^\beta + O(|Y|^3). \end{aligned}$$

Hence

$$\begin{aligned}
\Pi'(Ty)^a &= \frac{\partial(T^{-1} \circ \Phi)^\beta}{\partial X^\alpha} T_\beta^a T_\gamma^\alpha y^\gamma - \frac{1}{2} \frac{\partial^2(T^{-1} \circ \Phi)^\gamma}{\partial X^\alpha \partial X^\beta} T_\gamma^a T_\zeta^\alpha T_\eta^\beta y^\zeta y^\eta \\
&\quad + \frac{\partial^2(T^{-1} \circ \Phi)^\delta}{\partial X^a \partial X^\beta} T_\delta^\gamma (\perp)_{\gamma\alpha} T_\zeta^\alpha T_\epsilon^\eta y^\zeta y^\eta + O(|y|^3) \\
&= \frac{\partial(T^{-1} \circ \Phi)^\beta}{\partial x^\gamma} J_\alpha^\gamma T_\beta^a T_\gamma^\alpha y^\gamma - \frac{1}{2} \frac{\partial^2(T^{-1} \circ \Phi)^\gamma}{\partial x^\iota \partial x^\kappa} J_\alpha^\iota J_\beta^\kappa T_\gamma^a T_\zeta^\alpha T_\eta^\beta y^\zeta y^\eta \\
&\quad + \frac{\partial^2(T^{-1} \circ \Phi)^\delta}{\partial x^\iota \partial x^\kappa} J_\alpha^\iota J_\beta^\kappa T_\delta^\gamma (\perp)_{\gamma\alpha} T_\zeta^\alpha T_\eta^\beta y^\zeta y^\eta + O(|y|^3).
\end{aligned}$$

We deduce that:

$$\begin{aligned}
\tilde{\Pi}^s(y)^a &= ((J \circ \pi' \circ T)(y))^a \\
&= \frac{\partial \phi^\beta}{\partial x^\delta} J_\lambda^a J_\alpha^\delta T_\beta^\lambda T_\gamma^\alpha y^\gamma - \frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\iota \partial x^\kappa} J_\lambda^\iota J_\alpha^\kappa J_\beta^\lambda T_\gamma^\alpha T_\zeta^\beta T_\eta^\gamma y^\zeta y^\eta \\
&\quad + \frac{\partial^2 \phi^\delta}{\partial x^\iota \partial x^\kappa} J_\lambda^\iota J_\beta^\kappa J_\delta^\gamma (\perp)_{\gamma\alpha} T_\zeta^\alpha T_\eta^\beta y^\zeta y^\eta + O(|y|^3).
\end{aligned}$$

We now note that:

$$\Pi_b^a = J_\alpha^a D_\beta^\alpha T_b^\beta = J_\alpha^a T_b^\alpha.$$

This allows us to simplify our expression for  $\Pi^s(y)$  to:

$$\begin{aligned}
\tilde{\Pi}^s(y)^a &= \frac{\partial \phi^\beta}{\partial x^\delta} \Pi_\beta^a \Pi_\gamma^\delta y^\gamma - \frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\iota \partial x^\kappa} \Pi_\gamma^a \Pi_\zeta^\iota \Pi_\eta^\kappa y^\zeta y^\eta \\
&\quad + \frac{\partial^2 \phi^\delta}{\partial x^\iota \partial x^\kappa} \Pi_\eta^\kappa J_\lambda^\iota J_\lambda^\delta T_\delta^\gamma (\perp)_{\gamma\alpha} T_\zeta^\alpha y^\zeta y^\eta + O(|y|^3).
\end{aligned}$$

The tensor  $\perp_{ab}$  is equal to  $(g_\perp^\Phi)_{ab}$ . So since  $T$  is an isometry, we may write

$$(\perp)_{\alpha\beta} T_a^\alpha T_b^\beta = (g_\perp^\phi)_{ab}.$$

Using this together with equation (25) we may write:

$$\begin{aligned}
\tilde{\Pi}^s(y)^a &= \frac{\partial \phi^\beta}{\partial x^\delta} \Pi_\beta^a \Pi_\gamma^\delta y^\gamma - \frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\iota \partial x^\kappa} \Pi_\gamma^a \Pi_\zeta^\iota \Pi_\eta^\kappa y^\zeta y^\eta \\
&\quad + \frac{\partial^2 \phi^\delta}{\partial x^\iota \partial x^\kappa} \Pi_\eta^\kappa h^{a\iota} (g_\perp^\phi)_{\delta\zeta} y^\zeta y^\eta + O(|y|^3)
\end{aligned}$$

The first term can be simplified by repeated applications of equations (22) and (23):

$$\begin{aligned}
\frac{\partial \phi^\beta}{\partial x^\delta} \Pi_\beta^a \Pi_\gamma^\delta y^\gamma &= \frac{\partial \phi^\beta}{\partial x^\delta} \frac{\partial \phi^\beta}{\partial x^\gamma} h^{a\gamma} \frac{\partial \phi^\gamma}{\partial x^\epsilon} h^{\epsilon\delta} y^\gamma \\
&= h_{\delta\gamma} h^{a\gamma} \frac{\partial \phi^\gamma}{\partial x^\epsilon} h^{\epsilon\delta} y^\gamma = \frac{\partial \phi^\gamma}{\partial x^\epsilon} h^{\epsilon a} y^\gamma = \Pi_\beta^a y^\beta.
\end{aligned}$$

It is a tautology that the first order term is given by  $\Pi$ , nevertheless this calculation is a reassuring check on our working. Renaming the dummy variables we now have that:

$$\begin{aligned}\tilde{\Pi}^s(y)^a &= x_0 + \Pi_\alpha^a y^\alpha - \frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\gamma^a \Pi_\delta^\alpha \Pi_\epsilon^\beta y^\delta y^\epsilon \\ &\quad + \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\epsilon^\beta h^{a\alpha} (g_\perp^\phi)_{\gamma\delta} y^\delta y^\epsilon + O(|y|^3)\end{aligned}$$

We would like a formula that can be computed efficiently when  $n \ll r$ , so we wish to eliminate the term  $g_\perp^\phi$ . By splitting vectors  $V$  and  $W$  in  $\mathbb{R}^r$  into components in  $\text{Im } \phi_*$  and its orthogonal complement, we see that the Euclidean metric on  $\mathbb{R}^r$  satisfies the decomposition:

$$g_{ab} V^a W^b = (g_\perp^\phi)_{ab} V^a W^b + h_{ab} \Pi_\alpha^a V^\alpha \Pi_\alpha^b W^\beta.$$

Using this formula we obtain:

$$\begin{aligned}\tilde{\Pi}^s(y)^a &= x_0 + \Pi_\alpha^a y^\alpha - \frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\gamma^a \Pi_\delta^\alpha \Pi_\epsilon^\beta y^\delta y^\epsilon \\ &\quad + \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\delta^\beta h^{a\alpha} y^\gamma y^\delta - \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\epsilon^\beta \Pi_\gamma^\eta \Pi_\delta^\zeta h_{\eta\zeta} h^{a\alpha} y^\delta y^\epsilon \\ &\quad + O(|y|^3)\end{aligned}$$

□

We can immediately conclude:

**Theorem 5** (Itô-jet projection in coordinates). *Let  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^r$  be an embedding with  $\phi(x_0) = y_0$  then the Itô-jet projection of the SDE:*

$$dy = a dt + b_\alpha dW_t^\alpha, \quad y_0$$

is

$$dx = A dt + B_\alpha dW_t^\alpha, \quad x_0$$

where:

$$B_\alpha^i = \Pi_\beta^i b_\alpha^\beta$$

and:

$$\begin{aligned}A^i &= \Pi_\alpha^i a^\alpha + \\ &\quad \left( -\frac{1}{2} \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\gamma^i \Pi_\delta^\alpha \Pi_\epsilon^\beta \right. \\ &\quad \left. + \frac{\partial^2 \phi^\epsilon}{\partial x^\alpha \partial x^\beta} \Pi_\delta^\beta h^{i\alpha} - \frac{\partial^2 \phi^\gamma}{\partial x^\alpha \partial x^\beta} \Pi_\epsilon^\beta \Pi_\gamma^\eta \Pi_\delta^\zeta h_{\eta\zeta} h^{i\alpha} \right) \\ &\quad \times b_\kappa^\delta b_t^\epsilon [W^\kappa, W^t]_t.\end{aligned}$$

$\Pi$  is given by (23).  $h_{ab}$  is given by (22).  $h^{ab}$  is the inverse of  $h_{ab}$ .

It is reassuring to check that this formula gives the same result as we found in Section 4 for projection of a particular SDE onto a circle where the projection map was known exactly. In fact, we can find an explicit expression for the Itô-jet projection of any bivariate SDE driven by a single Brownian motion on the plane on the unit circle.

**Example 5.1** (Itô-jet projection of a bivariate SDE on the unit circle). Suppose that our diffusion process in  $\mathbb{R}^r = \mathbb{R}^2$ , driven by a one-dimensional Brownian motion  $W = W^1$ , is

$$\begin{aligned} dX &= a_1(X, Y)dt + b_1^1(X, Y)dW^1, \quad X_0 \\ dY &= a_2(X, Y)dt + b_1^2(X, Y)dW^1, \quad Y_0 \end{aligned}$$

and suppose we wish to approximate this process in the unit circle. If we define  $\theta = \arctan(Y_t/X_t)$ , and compute  $d\theta_t$  via Itô's formula, this won't be in general a closed SDE for  $\theta$ , contrary to the special example of the cross diffusion above. To obtain a closed SDE in  $\theta$  we have to project. One can check that for the one-dimensional manifold given by the unit circle, expressed as

$$M = \{(\cos(\theta), \sin(\theta)), \theta \in [0, 2\pi)\}$$

with coordinates  $Y = \theta$  in  $\mathbb{R}^n = \mathbb{R}^1$ , one has

$$h = 1, \quad h^{-1} = 1, \quad \Pi_1 = -\sin(\theta), \quad \Pi_2 = \cos(\theta), \quad \partial_\theta^2 \phi^1 = -\cos(\theta), \quad \partial_\theta^2 \phi^2 = -\sin(\theta),$$

which allows us to apply Theorem 5 to this system. We obtain (coefficients  $a$  and  $b$  are computed in  $X = \cos(\theta), Y = \sin(\theta)$ )

$$\begin{aligned} A(\theta) &= -a_1 \sin(\theta) + a_2 \cos(\theta) + \frac{1}{2} \sin(2\theta)((b_1^1)^2 - (b_1^2)^2) - \cos(2\theta)b_1^1 b_1^2, \\ B(\theta) &= -\sin(\theta)b_1^1 + \cos(\theta)b_1^2. \end{aligned}$$

In the special case of  $a_1 = a_2 = 0$  and  $b_1^1 = \sigma \sin(\theta), b_1^2 = \sigma \cos(\theta)$  this confirms our previous calculations for the cross-diffusion example.

## 6 Application of the Projection to Non-linear Filtering

As a fundamental application of our new projection methods we consider an area from signal processing, stochastic filtering. This extends our previous work in [5].

In stochastic filtering one has a signal  $X$  that evolves according to a SDE, and observes a process  $Y$  which is a function of this signal plus noise. This is standard notation, but these  $X$  and  $Y$  are not to be confused with the processes we used earlier in the paper, in that they are not the  $\mathbb{R}^r$  process to be approximated and its  $\mathbb{R}^n$  approximation.

The filtering problem consists in estimating the signal  $X$  given the present and past observations  $Y$ . If  $t$  is the current time, the solution of the filtering problem is the probability density of the state  $X_t$  conditional on the observations from time 0 to time  $t$ , call it  $p_t$ . The density  $p_t$  follows the Kushner-Stratonovich (or Zakai) stochastic partial differential equation (SPDE) that, under some technical assumptions, can be seen as a stochastic differential equation in the infinite dimensional  $L^2$  space of square roots of densities (Hellinger metric) or of densities themselves (direct  $L^2$  metric).

The process we wish to approximate on a low dimensional manifold is  $p_t$ , which represents the  $X_t$  of our earlier sections. The  $\mathbb{R}^r$  space of our earlier sections is the  $L^2$  infinite dimensional space, while the submanifold  $M$  is a finite dimensional family of probability densities parametrized by  $\theta$ , acting as coordinates:  $\{p(\cdot, \theta), \theta \in \Theta \subset \mathbb{R}^n\}$ .  $\theta_t$  plays the role of what we were calling  $Y_t$  earlier in the paper. We aim at finding a SDE for  $\theta$  such that  $p(\cdot, \theta_t)$  approximates  $p_t(\cdot)$  in an optimal way. Note that in the previous part of the paper we had a dimensionality reduction from  $r$  to  $n$ , whereas now we go from infinite dimensional  $p_t$  to  $n$ -dimensional  $\theta_t$ .

One may be concerned about taking our finite dimensional results and applying them in an infinite dimensional setting. However, we have stated our results in terms of approximating one Ito-Taylor series of a given order with another Ito-Taylor series. This allows us to avoid the analytical issues that might conceivably arise in considering the convergence of these series. Therefore our results generalize straightforwardly to the Hilbert space setting. As an example, the minimization argument used to prove Theorem 1 relies only on properties of the linear projection operator that remain true in a Hilbert space setting.

In addition the explicit calculation of Section 5 can be generalized unproblematically to the case of a finite dimensional manifold embedded in a Hilbert space. To see this simply note that the vector space spanned by the first two derivatives of the map  $\phi$  at  $p$  gives a finite dimensional space  $V$  and so one can simply apply the result for embedding into the space  $V$ .

The point where complexities might conceivably arise in the infinite dimensional setting is in the generalizations of Proposition 1 and Proposition 2. Folk wisdom suggests that such results can be generalized to Hilbert spaces without difficulty, so we will not attempt to prove that here.

## 6.1 The Kushner Stratonovich equation

We suppose that the state  $X_t \in \mathbb{R}^n$  of a system evolves according to the equation:

$$dX_t = f(X_t, t) dt + \sigma(X_t, t) dW_t$$

where  $f$  and  $\sigma$  are smooth  $\mathbb{R}^n$  valued functions and  $W_t$  is a Brownian motion. One typically adds growth conditions to ensure a global existence and uniqueness result for the signal equation, see for example [5] and references therein for the details.



We suppose that an associated process, the observation process,  $Y_t \in \mathbb{R}^d$  evolves according to the equation:

$$dY_t = b(X_t, t) dt + dV_t$$

where  $b$  is a smooth  $\mathbb{R}^d$  valued function and  $V_t$  is a Brownian motion independent of  $W_t$ . Note that the filtering problem is often formulated with an additional constant in terms of the observation noise. For simplicity we have assumed that the system is scaled so that this can be omitted.

The filtering problem is to compute the conditional distribution of  $X_t$  given a prior distribution for  $X_0$  and the values of  $Y$  for all times up to and including  $t$ .

Subject to various bounds on the growth of the coefficients of this equation, the assumption that the distribution has a density  $p_t$  and suitable bounds on the growth of  $p_t$  one can show that  $p_t$  satisfies the Kushner–Stratonovich SPDE:

$$dp = \mathcal{L}^* p dt + p[b - E_p(b)]^T [dY - E_p(b)dt] \quad (26)$$

where  $E_p$  denotes the expectation with respect to the density  $p$ ,

$E_p[f] = \int f(x)p(x)dx$ , and the forward diffusion operator  $\mathcal{L}_t^*$  is defined by:

$$\mathcal{L}_t^* \phi = -\frac{\partial}{\partial x^i} [f_i(x, t)\phi] + \frac{1}{2} \frac{\partial^2}{\partial x^i \partial x^j} [a_{ij}(x, t)\phi] \quad (27)$$

where  $a = \sigma\sigma^T$ . Note that we are using the Einstein summation convention in this expression.

In the event that the coefficient functions  $f$  and  $b$  are all linear and  $\sigma$  is a deterministic function of time one can show that so long as the prior distribution for  $X$  is Gaussian, or deterministic, the density  $p$  will be Gaussian at all subsequent times. This allows one to reduce the infinite dimensional equation (26) to a finite dimensional stochastic differential equation for the mean and covariance matrix of this normal distribution. This finite dimensional problem solution is known as the Kalman filter.

For more general coefficient functions, however, equation (26) cannot be reduced to a finite dimensional problem [18]. Instead one might seek approximate solutions of (26) that belong to some given statistical family of densities. This is a very general setup and includes, for example, approximating the density using piecewise linear functions to derive a finite difference approximation or approximating the density with Hermite polynomials to derive a spectral method. Other examples include exponential families (considered in [10, 9]) and mixture families (considered in [3, 5]).

Our projection theory tells us how one can find good approximations on a given statistical family with respect to a given metric on the space of distributions. We illustrate this by writing down the Itô-vector and Itô-jet projection of (26) for the  $L^2$  and Hellinger metrics onto a general manifold<sup>1</sup>.

---

<sup>1</sup>Note that it is also possible to consider projecting the Zakai equation. However, as explained in [5], one expects that projecting the Kushner–Stratonovich will lead to smaller error terms.

We will then examine some numerical results regarding the very specific case of seeking approximate solutions using Gaussian distributions. The idea of approximating the solution to the filtering problem using a Gaussian distribution has been considered by numerous authors who have derived variously, the extended Kalman filter [27], assumed density filters [23] and Stratonovich projection filters [9]. Some of these are related, for example the assumed density filters and Stratonovich projection filters in Hellinger metrics for Gaussian (and more generally exponential) families coincide [10]. Using our new projection methods, we will be able to derive projection filters which outperform all these other filters (assuming performance is measured over small time intervals using the appropriate Hilbert space metric).

We note that (26) is an infinite dimensional SDE driven by a continuous semimartingale. The definitions and results given in Section 2 were only stated in the finite dimensional case for SDEs driven by Brownian motion. The definition of Itô–Taylor series can be generalized straightforwardly to this situation and hence the definition of the Itô projections can be applied in this context also.

More generally, for the the geometry of infinite dimensional filtering problems based on  $L^2$  or Orlicz charts and for the related differential geometric approach to statistics with recent advances we refer for example to [29, 25, 26, 16, 10, 5, 11, 12].

## 6.2 Itô-vector projections

### 6.2.1 The Itô-vector projection filter in the $L^2$ direct metric

Let us suppose that the density  $p$  lies in  $L^2$  and so we can use the  $L^2$  norm to measure the accuracy of an approximate solution to equation (26). For a discussion on conditions under which a unnormalized version of  $p$  is in  $L^2$  (Zakai Equation) see for example [1].

We wish to consider an  $m$ -dimensional family of distributions  $p$  parameterized by  $m$  real valued parameters  $\theta^1, \theta^2, \dots, \theta^m$ . For example we will consider the 2 dimensional Gaussian family:

$$p(x) = \frac{1}{(\theta^2)\sqrt{2\pi}} \exp\left(-\frac{(x - (\theta^1))^2}{2(\theta^2)^2}\right). \quad (28)$$

Note that we have chosen to follow differential geometry convention and use upper indices for the coordinate functions  $\theta^i$  so we have been careful to distinguish powers from indices using brackets.

More formally, an  $m$ -dimensional family is given by a smooth embedding  $\phi : \mathbb{R}^m \rightarrow L^2(\mathbb{R}^n)$ . The tangent vectors  $\phi_* \frac{\partial}{\partial \theta^i} \in L^2(\mathbb{R}^n)$  are simply the partial derivatives

$$\frac{\partial p}{\partial \theta^i}.$$

Let us write:

$$g_{ij} = \int_{\mathbb{R}} \frac{\partial p}{\partial \theta^i} \frac{\partial p}{\partial \theta^j} dx.$$

This defines the induced metric tensor on the manifold  $\phi(\mathbb{R}^m)$ . We will write  $g^{ij}$  for the inverse of the matrix  $g_{ij}$ . The projection operator  $\Pi_{\phi(\theta)}$  is then given by

$$\begin{aligned}\Pi_{\phi(\theta)}(v) &= \sum_{i,j=1}^m g^{ij} \left\langle v, \phi_* \frac{\partial}{\partial \theta^i} \right\rangle_{L^2} \phi_* \frac{\partial}{\partial \theta^j} \\ &= \sum_{i,j=1}^m g^{ij} \left( \int_{\mathbb{R}^n} v(x) \frac{\partial p}{\partial \theta^i} dx \right) \phi_* \frac{\partial}{\partial \theta^j}.\end{aligned}$$

Thus

$$\phi_*^{-1} \Pi_{\phi(\theta)}(v) = \sum_{i,j=1}^m g^{ij} \left( \int_{\mathbb{R}^n} v(x) \frac{\partial p}{\partial \theta^i} dx \right) \frac{\partial}{\partial \theta^j}.$$

We can now write down the Itô-vector projection of (26) with respect to the  $L^2$  metric. It is:

$$d\theta^i = A^i dt + B^i dY_t$$

where:

$$B^i = \sum_{j=1}^m g^{ij} \left( \int_{\mathbb{R}} (p(b - E_{p(\theta)}(b)))^T \frac{\partial p}{\partial \theta^j} dx \right)$$

and

$$A^i = \sum_{j=1}^m g^{ij} \left( \int_{\mathbb{R}^n} \left( \mathcal{L}^* p - p(b - E_{p(\theta)}(b))^T E_{p(\theta)}(b) - \frac{1}{2} \sum_{k=1}^m \frac{\partial^2 p}{\partial \theta^j \partial \theta^k} B^k \right) \frac{\partial p}{\partial \theta^j} dx \right).$$

**Example 6.1** (Itô-vector projection filter for cubic sensor in direct metric). Consider as a test case the 1-dimensional problem with  $f(x, t) = 0$ ,  $\sigma(x, t) = 1$  and  $b(x, t) = x + \epsilon x^3$  for some small constant  $\epsilon$ . This problem is a perturbation of a linear filter so one might expect that a Gaussian approximation will perform reasonably well at least for small times. Thus we will use the 2 dimensional manifold of Gaussian distributions given in equation (28).

We first calculate the metric tensor  $g_{ij}$  which is diagonal in this case:

$$g_{ij} = \frac{1}{4\sqrt{\pi}(\theta^2)^3} \begin{pmatrix} 1 & 0 \\ 0 & \frac{3}{2} \end{pmatrix}.$$

This is easily inverted to compute  $g^{ij}$ . We compute the expectation  $E_p(b)$ :

$$E_p(b) = \frac{\epsilon (\sqrt{2\pi}(\theta^1)^3(\theta^2) + 3\sqrt{2\pi}(\theta^1)(\theta^2)^3)}{\sqrt{2\pi}(\theta^2)} + (\theta^1).$$

One can now see that computing the projection equation will simply involve integrating a number of terms of the form a polynomial in  $x$  times a Gaussian.

The end result is:

$$\begin{aligned}
d\theta^1 &= \left( -\frac{1}{4}\theta^1 (\theta^2)^2 \left( 3\epsilon^2 \left( 4(\theta^1)^4 - 4(\theta^2)^2 (\theta^1)^2 - 3(\theta^2)^4 \right) + 16\epsilon (\theta^1)^2 + 4 \right) \right) dt \\
&\quad + \left( \frac{1}{2} (\theta^2)^2 \left( 3\epsilon \left( 2(\theta^1)^2 + (\theta^2)^2 \right) + 2 \right) \right) dY_t \\
d\theta^2 &= \left( -\frac{9\epsilon^2 (\theta^2)^8 + (\theta^2)^4 \left( 60\epsilon^2 (\theta^1)^4 + 48\epsilon (\theta^1)^2 + 4 \right) + 6\epsilon (\theta^2)^6 \left( 9\epsilon (\theta^1)^2 + 2 \right) - 4}{8\theta^2} \right) dt \\
&\quad + \left( 3\epsilon\theta^1 (\theta^2)^3 \right) dY_t
\end{aligned}$$

### 6.2.2 The Itô-vector projection filter in the Hellinger metric

The Hellinger metric is a metric on probability measures. In the case of two probability density functions  $p(x)$  and  $q(x)$  on  $\mathbb{R}^n$ , that now need only be in  $L^1$ , the Hellinger distance is given by the square root of:

$$\frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx.$$

In other words, up to the constant factor of  $\frac{1}{2}$  the Hellinger metric corresponds to the  $L^2$  norm on the square root of the density function rather than on the density itself (as in the previous subsection). The Hellinger metric has the important advantage of making the metric independent of the particular background density that is used to express measures as densities. The  $L^2$  direct distance introduced earlier does not satisfy this background independence.

Now, to compute the Itô-vector projection with respect to the Hellinger metric we first want to write down an Itô equation for the evolution on  $\sqrt{p}$ .

Applying Itô's lemma to equation (26) we formally obtain:

$$\begin{aligned}
d\sqrt{p} &= \left( \frac{\mathcal{L}^* p - p(b - E_p(b))^T E_p(b)}{2\sqrt{p}} - \frac{p^2 (b - E_p(b))^T (b - E_p(b))}{8p\sqrt{p}} \right) dt \\
&\quad + \left( \frac{p(b - E_p(b))^T}{2\sqrt{p}} \right) dY_t \\
&= \left( \frac{\mathcal{L}^* p}{2\sqrt{p}} - \frac{1}{8}\sqrt{p}(b - E_p(b))^T (b + 3E_p(b)) \right) dt \\
&\quad + \left( \frac{1}{2}\sqrt{p}(b - E_p(b))^T \right) dY_t.
\end{aligned}$$

A family of distributions now corresponds to an embedding  $\phi$  from  $\mathbb{R}^m$  to  $L^2(\mathbb{R}^n)$  but now  $p = \phi(\theta)^2$ . The tangent space is spanned by the vectors:

$$\phi_* \frac{\partial}{\partial \theta^i} = \frac{\partial \sqrt{p}}{\partial \theta^i}.$$

We define a metric on the tangent space by:

$$h_{ij} = \int_{\mathbb{R}^n} \frac{\partial \sqrt{p}}{\partial \theta^i} \frac{\partial \sqrt{p}}{\partial \theta^j} dx.$$

We write  $h^{ij}$  for the inverse matrix of  $h_{ij}$ . The projection operator with respect to the Hellinger metric is:

$$\Pi_{\phi(\theta)}(v) = \sum_{i,j=1}^m h^{ij} \left( \int_{\mathbb{R}^n} v(x) \frac{\partial \sqrt{p}}{\partial \theta^i} dx \right) \phi_* \frac{\partial}{\partial \theta^j}.$$

We can now write down the Itô-vector projection of (26) with respect to the Hellinger metric. It is:

$$d\theta^i = A^i dt + B^i dY_t$$

where:

$$B^i = \sum_{j=1}^m h^{ij} \left( \int_{\mathbb{R}^n} \frac{1}{2} \sqrt{p} (b - E_{p(\theta)}(b))^T \frac{\partial \sqrt{p}}{\partial \theta^j} dx. \right)$$

and

$$A^i = \sum_{j=1}^m h^{ij} \left( \int_{\mathbb{R}^n} \left( \frac{\mathcal{L}^* p}{2\sqrt{p}} - \frac{1}{8} \sqrt{p} (b - E_{p(\theta)}(b))^T (b + 3E_{p(\theta)}(b)) - \frac{1}{2} \sum_{k=1}^m \frac{\partial^2 \sqrt{p}}{\partial \theta^j \partial \theta^k} B^k \right) \frac{\partial \sqrt{p}}{\partial \theta^j} dx. \right).$$

**Example 6.2** (Itô-vector projection filter for cubic sensor: Hellinger metric). We may repeat example 6.1 but projecting using the Hellinger metric. We first calculate the metric tensor  $h_{ij}$  which is diagonal also in this case:

$$h_{ij} = \frac{1}{4\theta_2^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

This is easily inverted to compute  $h^{ij}$ . We obtain the following SDEs:

$$\begin{aligned} d\theta^1 &= \left( -\theta^1 (\theta^2)^2 \left( 3\epsilon^2 \left( (\theta^1)^4 + 4(\theta^2)^2 (\theta^1)^2 + 6(\theta^2)^4 \right) + \epsilon \left( 4(\theta^1)^2 + 6(\theta^2)^2 + 1 \right) \right) dt \right. \\ &\quad \left. + \left( (\theta^2)^2 \left( 3\epsilon \left( (\theta^1)^2 + (\theta^2)^2 \right) + 1 \right) \right) dY_t \right) \\ d\theta^2 &= \left( -\frac{27\epsilon^2 (\theta^2)^8 + (\theta^2)^4 \left( 15\epsilon^2 (\theta^1)^4 + 12\epsilon (\theta^1)^2 + 1 \right) + 9\epsilon (\theta^2)^6 \left( 6\epsilon (\theta^1)^2 + 1 \right) - 1}{2\theta^2} dt \right. \\ &\quad \left. + \left( 3\epsilon \theta^1 (\theta^2)^3 \right) dY_t \right) \end{aligned}$$

### 6.3 Itô-jet projections

Using the formulae from Theorem 5 together with the formulae and techniques of Section 6.2 we can explicitly calculate the Itô-vector projections of the filtering equation in both the  $L^2$  and Hellinger metrics.

To minimize notation, let us concentrate on the 1-dimensional state space filtering problem and project using the  $L^2$  metric.

We can formally write the filtering equation in the form:

$$dp_t = \mu(p_t)dt + \Sigma(p_t)dW_t \quad (29)$$

where  $p_t$  is an  $L^2$  function and

$$\begin{aligned} \mu(p)(x) &:= \frac{1}{2} \frac{d^2(\sigma(x)^2 p(x))}{dx^2} - \frac{d(f(x)p(x))}{dx} \\ &\quad - p(x) \left( b(x) - \int_{\mathbb{R}} p(t)b(t)dt \right) \int_{\mathbb{R}} p(t)b(t)dt, \\ \Sigma(p)(x) &:= p(x) \left( b(x) - \int_{\mathbb{R}} p(t)b(t)dt \right). \end{aligned} \quad (30)$$

We now suppose that  $p_t$  is parameterized as  $p_t(x) = \phi(\theta)(x)$  as in Section 6.2. Using Theorem 5 we can write down the Itô-jet projection which is an SDE for the components of  $\theta$ .

To write down the result it will be useful to define functions  $\pi^i(\theta)$  by:

$$\pi^i(\theta) = h^{ij} \frac{\partial \phi}{\partial \theta^j}(\theta).$$

We will also use angle brackets to denote the  $L^2$  inner product. With this understood, the Itô-jet projection of the filtering equations in the  $L^2$  metric is given by:

$$d\theta_t^i = A^i(\theta)dt + B^i(\theta)dW_t$$

where we have in turn

$$B^i(\theta) = \langle \pi^i(\theta), \Sigma(\phi(\theta)) \rangle dt$$

and

$$\begin{aligned} A^i(\theta) &= \langle \pi^i(\theta), \mu(\phi(\theta)) \rangle \\ &\quad - \frac{1}{2} \left\langle \frac{\partial^2 \phi}{\partial \theta^\alpha \partial \theta^\beta}(\theta), \pi^i(\theta) \right\rangle \langle \Sigma(\theta), \pi^\alpha(\theta) \rangle \langle \Sigma(\theta), \pi^\beta(\theta) \rangle \\ &\quad + \left\langle \frac{\partial^2 \phi}{\partial \theta^\alpha \partial \theta^\beta}(\theta), \Sigma(\phi(\theta)) \right\rangle \langle \pi^\beta(\theta), \Sigma(\phi(\theta)) \rangle h^{i\alpha}(\theta) \\ &\quad - \left\langle \frac{\partial^2 \phi}{\partial \theta^\alpha \partial \theta^\beta}, \pi^\eta(\theta) \right\rangle \langle \pi^\beta(\theta), \Sigma(\phi(\theta)) \rangle \langle \pi^\xi(\theta), \Sigma(\phi(\theta)) \rangle h_{\eta\xi}(\theta) h^{i\alpha}(\theta) \end{aligned}$$

**Example 6.3** (Itô-jet projection filter for cubic sensor in direct metric). For the filtering problem of 6.1 the Itô-jet projection in the  $L^2$  metric is

$$\begin{aligned} d\theta^1 &= \left( -\frac{1}{4}\theta^1(\theta^2)^2 \left( 3\epsilon^2 \left( 4(\theta^1)^4 - 4(\theta^2)^2(\theta^1)^2 - 9(\theta^2)^4 \right) + 16\epsilon(\theta^1)^2 + 4 \right) \right. \\ &\quad \left. + \left( \frac{1}{2}(\theta^2)^2 \left( 3\epsilon \left( 2(\theta^1)^2 + (\theta^2)^2 \right) + 2 \right) \right) dY_t \right) \\ d\theta^2 &= \left( \frac{3\epsilon^2(\theta^2)^8 - 4(\theta^2)^4 \left( 15\epsilon^2(\theta^1)^4 + 12\epsilon(\theta^1)^2 + 1 \right) - 2\epsilon(\theta^2)^6 \left( 15\epsilon(\theta^1)^2 + 2 \right) + 4}{8\theta^2} \right) dt \\ &\quad + \left( 3\epsilon\theta^1(\theta^2)^3 \right) dY_t \end{aligned}$$

The Itô-jet projection of the filtering equation in the Hellinger metric can be computed in the same way. Indeed we can formally write the filtering equation in the form:

$$dq_t = \mu(q_t)dt + \Sigma(q_t)dW_t \quad (31)$$

where  $q_t$  is the square root of the density and the coefficients now satisfy

$$\begin{aligned} \mu(q)(x) &:= \frac{1}{2q(x)} \left( \frac{1}{2} \frac{d^2(\sigma(x)^2 q(x)^2)}{dx^2} - \frac{d(f(x)q(x)^2)}{dx} \right) \\ &\quad - \frac{1}{8} q(x) \left( b(x) - \int_{\mathbb{R}} q(t)^2 b(t) dt \right) \left( b(x) + 3 \int_{\mathbb{R}} q(t)^2 b(t) dt \right), \quad (32) \\ \Sigma(q)(x) &:= \frac{1}{2} q(x) \left( b(x) - \int_{\mathbb{R}} q(t)^2 b(t) dt \right). \end{aligned}$$

Thus we can use the same formulae as above to compute the Hellinger projection except we must use the coefficients from (32) rather than those from (30).

**Example 6.4** (Itô-jet projection filter for cubic sensor: Hellinger metric). For the filtering problem of 6.1, the Itô-jet projection in the Hellinger metric is

$$\begin{aligned} d\theta^1 &= \left( -\theta^1 (\theta^2)^2 \left( 3\epsilon^2 \left( (\theta^1)^4 + 4 (\theta^2)^2 (\theta^1)^2 + 3 (\theta^2)^4 \right) + \epsilon \left( 4 (\theta^1)^2 + 6 (\theta^2)^2 \right) + 1 \right) \right) dt \\ &\quad + \left( (\theta^2)^2 \left( 3\epsilon \left( (\theta^1)^2 + (\theta^2)^2 \right) + 1 \right) \right) dY_t \\ d\theta^2 &= \left( -\frac{18\epsilon^2 (\theta^2)^8 + (\theta^2)^4 (15\epsilon^2 (\theta^1)^4 + 12\epsilon (\theta^1)^2 + 1) + 3\epsilon (\theta^2)^6 (15\epsilon (\theta^1)^2 + 2) - 1}{2\theta^2} \right) dt \\ &\quad + \left( 3\epsilon\theta^1 (\theta^2)^3 \right) dY_t \end{aligned}$$

## 6.4 Other Gaussian Approximate Filters

Many other Gaussian approximate filters have been proposed in the past. We will briefly review a number of different Gaussian approximate filters that can be found in the literature and calculate the relevant stochastic differential equations for our example 6.1. We will then compare the performance of these filters numerically.

### 6.4.1 The Stratonovich projection filter

Instead of using the Itô-vector projection, one can use the Stratonovich projection.

**Example 6.5** (Stratonovich proj. filter for cubic sensor: direct metric). General formulae for performing the Stratonovich  $L^2$  projection are given in [3]. In the specific case of example 6.1 the resulting Itô SDEs are:

$$\begin{aligned}
d\theta^1 &= \left( -\frac{1}{4}\theta^1 (\theta^2)^2 \left( 3\epsilon^2 \left( 4(\theta^1)^4 - 4(\theta^2)^2 (\theta^1)^2 - 3(\theta^2)^4 \right) + 16\epsilon (\theta^1)^2 + 4 \right) \right) dt \\
&\quad + \left( \frac{1}{2} (\theta^2)^2 \left( 3\epsilon \left( 2(\theta^1)^2 + (\theta^2)^2 \right) + 2 \right) \right) dY_t \\
d\theta^2 &= \left( -\frac{47\epsilon^2 (\theta^2)^8 + (\theta^2)^4 \left( 60\epsilon^2 (\theta^1)^4 + 48\epsilon (\theta^1)^2 + 4 \right) + 2\epsilon (\theta^2)^6 \left( 33\epsilon (\theta^1)^2 + 8 \right) - 4}{8\theta^2} \right) dt \\
&\quad + \left( 3\epsilon\theta^1 (\theta^2)^3 \right) dY_t
\end{aligned}$$

**Example 6.6** (Stratonovich proj. filter for cubic sensor: Hellinger metric). General formulae for performing the Stratonovich Hellinger projection are given in [10]. In the specific case of example 6.1 the resulting SDEs are:

$$\begin{aligned}
d\theta^1 &= \left( -\theta^1 (\theta^2)^2 \left( 3\epsilon^2 \left( (\theta^1)^4 + 4(\theta^2)^2 (\theta^1)^2 + 6(\theta^2)^4 \right) + \epsilon \left( 4(\theta^1)^2 + 6(\theta^2)^2 \right) + 1 \right) \right) dt \\
&\quad + \left( (\theta^2)^2 \left( 3\epsilon \left( (\theta^1)^2 + (\theta^2)^2 \right) + 1 \right) \right) dY_t \\
d\theta^2 &= \left( -\frac{36\epsilon^2 (\theta^2)^8 + (\theta^2)^4 \left( 15\epsilon^2 (\theta^1)^4 + 12\epsilon (\theta^1)^2 + 1 \right) + 9\epsilon (\theta^2)^6 \left( 6\epsilon (\theta^1)^2 + 1 \right) - 1}{2\theta^2} \right) dt \\
&\quad + \left( 3\epsilon\theta^1 (\theta^2)^3 \right) dY_t
\end{aligned}$$

#### 6.4.2 The Extended Kalman Filter

The Extended Kalman Filter (EKF) is a heuristically derived method of finding approximate solutions to the filtering problem based on the idea of linearising the problem and then using the solution to the linear problem. In particular one assumes that the solution can be well approximated by a Gaussian distribution. For the EKF see [21, 1]. A definition and heuristic derivation is given in [6] (which is based, in turn, on the derivation given in [27]).

The EKF can be shown to work well on condition that the initial position of the signal is approximated well, the non-linearities of  $f$  are small,  $b$  is injective and the observation noise is small [28]. Moreover, the EKF is widely used in practice, see [6] for references to applications.

**Example 6.7.** For the example problem  $b(x) = x + \epsilon x^3$  the EKF is:

$$\begin{aligned}
d\theta^1 &= \left( (\theta^2)^2 \left( -\left( 3\epsilon (\theta^1)^2 + 1 \right) \right) \left( \theta^1 + \epsilon (\theta^1)^3 \right) \right) dt \\
&\quad + \left( (\theta^2)^2 \left( 3\epsilon (\theta^1)^2 + 1 \right) \right) dY_t \\
d\theta^2 &= \left( \frac{1 - (\theta^2)^4 \left( 3\epsilon (\theta^1)^2 + 1 \right)^2}{2\theta^2} \right) dt \\
&\quad + (0) dY_t
\end{aligned}$$

#### 6.4.3 Assumed density filters

Assumed density filters (ADFs) provide a finite dimensional method of finding approximate solutions to the filtering problem. They have been considered in,



for example, [23], [24] and [10].

The general setup is to consider a statistical family  $\pi(\cdot, \eta)$  of probability measures parameterized by some coordinates  $\eta = (\eta^1, \dots, \eta^m)$ . This parameterization is not arbitrary. It must be chosen in such a way that, for elements of the statistical family, the values of  $\eta$  correspond to the expectations of some twice differentiable scalar functions  $\{c^1, \dots, c^m\}$  defined on  $\mathbb{R}^n$ .

$$\eta^i = E_{\pi(\cdot, \eta)}(c^i) =: E_{\eta_t}(c^i)$$

where for brevity we are using the abbreviation  $E_{\eta_t}$  for  $E_{\pi(\cdot, \eta)}$ .

For example one might take the statistical family of normal distributions parameterized by its first and second moments  $\eta_1$  and  $\eta_2$ , so  $c^1(x) = x$ ,  $c^2(x) = (x)^2$ .

Given a statistical family parameterized in this way, we define the Itô ADF to be:

$$d\eta_t^i = E_{\eta_t}(\mathcal{L}_t c^i) dt + (E_{\eta_t}(b_t c^i) - E_{\eta_t}(b_t) \eta_t^i)^T (dY_t - E_{\eta_t}(b_t) dt).$$

This is motivated by the fact that under the conditions used to derive equation (26), we have that the  $c_i$ -moments of  $\pi_t$ , the true solution to the filtering problem, satisfy the Itô equation:

$$d\pi_t(c_i) = \pi_t(\mathcal{L}_t c^i) dt - \frac{1}{2}(\pi_t(bc^i) - \pi_t(b)\pi_t(c^i))(dY - \pi_t(b) dt).$$

Thus if it were true that the true density was a member of our chosen statistical family then the Itô ADF would certainly be satisfied. One just hopes that the Itô ADF will continue to give a reasonable approximation even though we know that the true density isn't a member of the chosen statistical family.

With a similar motivation we define the Stratonovich ADF to be:

$$\begin{aligned} d\eta_t^i &= E_{\eta_t}(\mathcal{L}_t c^i) dt - \frac{1}{2}(E_{\eta_t}(|b_t|^2 c^i) - E_{\eta_t}(|b_t|^2) \eta_t^i) dt \\ &+ (E_{\eta_t}(b_t c^i) - E_{\eta_t}(b_t) \eta_t^i)^T \circ dY_t. \end{aligned}$$

If it were true that the density was a member of our statistical family then the Itô ADF and the Stratonovich ADF would be equivalent equations. Since we only expect to be able to approximate the true density with our statistical family, we must expect that the Itô ADF and Stratonovich ADF are in fact inequivalent equations. Intuitively, we can say that the local moment matching approximation on which the ADF heuristics are based and the Itô-Stratonovich transformation do not commute.

The justification just given for ADFs is far from convincing. We are relying on little other than hope that these equations will give good approximations. However, it was shown in [10] that in fact for exponential families, the Stratonovich projection filter in the Hellinger metric coincides with the Stratonovich ADF, and in [8] that for the Gaussian case, this filter approaches the optimal filter under small observation noise. [12] show that the equivalence

between projection using the  $L^2$  direct structure and the assumed density approximation holds for the prediction step of the filtering algorithm, namely the Kolmogorov equation, when using mixture families.

**Example 6.8.** If we calculate the Itô assumed density filter corresponding to example 6.1 and the family of normal distributions, and then change coordinates to  $\theta^1$  and  $\theta^2$  as used in the previous examples, we obtain the SDEs:

$$\begin{aligned} d\theta^1 &= \left( -\theta^1 (\theta^2)^2 \left( 3\epsilon^2 \left( (\theta^1)^4 + 4 (\theta^2)^2 (\theta^1)^2 + 3 (\theta^2)^4 \right) + \epsilon \left( 4 (\theta^1)^2 + 6 (\theta^2)^2 + 1 \right) \right) dt \right. \\ &\quad \left. + \left( (\theta^2)^2 \left( 3\epsilon \left( (\theta^1)^2 + (\theta^2)^2 \right) + 1 \right) \right) dY_t \right) \\ d\theta^2 &= \left( -\frac{9\epsilon^2 (\theta^2)^8 + (\theta^2)^4 \left( 15\epsilon^2 (\theta^1)^4 + 12\epsilon (\theta^1)^2 + 1 \right) + 3\epsilon (\theta^2)^6 \left( 15\epsilon (\theta^1)^2 + 2 \right) - 1}{2\theta^2} dt \right. \\ &\quad \left. + \left( 3\epsilon\theta^1 (\theta^2)^3 \right) dY_t \right) \end{aligned}$$

**Example 6.9.** The family of normal distributions is an exponential family, therefore the Stratonovich assumed density filter is equivalent to the Stratonovich projection filter in the Hellinger metric.

## 6.5 Results

Our explicit calculations show that the two Itô projections give rise to new, distinct, Gaussian approximations.

All our calculations of the resulting filters for the cubic sensor  $b(x, t) = x + \epsilon x^3$  are equal when  $\epsilon = 0$ . This provides a basic sanity check that our formulae correspond to the Kalman filter in the case of a linear sensor. In general, if we know that the solution lies in a particular manifold and we project onto that manifold, the three projection methods will all be exact.

We simulated the example problem  $b(x) = x + \epsilon x^3$  for all of the above approximate filters with  $\epsilon = 0.05$ . We also computed an “exact” solution using a finite difference method on a grid of 1000 intervals spaced evenly from  $-10.0$  to  $10.0$  and a time step of  $0.0002$ . We define the  $L^2$  residual to be the  $L^2$  distance between the approximate solution and the “exact” solution. We define the Hellinger residual similarly, as the  $L^2$  distance between the square roots of the solution densities.

In Figure 5 we see the  $L^2$  residuals for the various methods. All the projection methods shown are taken using the  $L^2$  metric in this case. The Itô-vector projection in the  $L^2$  metric results in the lowest residuals over short time horizons. The Stratonovich projection comes a close second. Over medium term time horizons, the Itô-jet projection outperforms the Itô-vector projection. We have not shown longer term behaviour because over long time horizons, all the methods become inaccurate and any comparison becomes meaningless. The projection methods outperformed all other methods. Although our plot shows only a single run, it is reasonably representative of the typical behaviour.

In Figure 6 we have plotted the ratio of the Hellinger residual for each method to the residual of the Itô-jet projection w.r.t. the Hellinger metric.

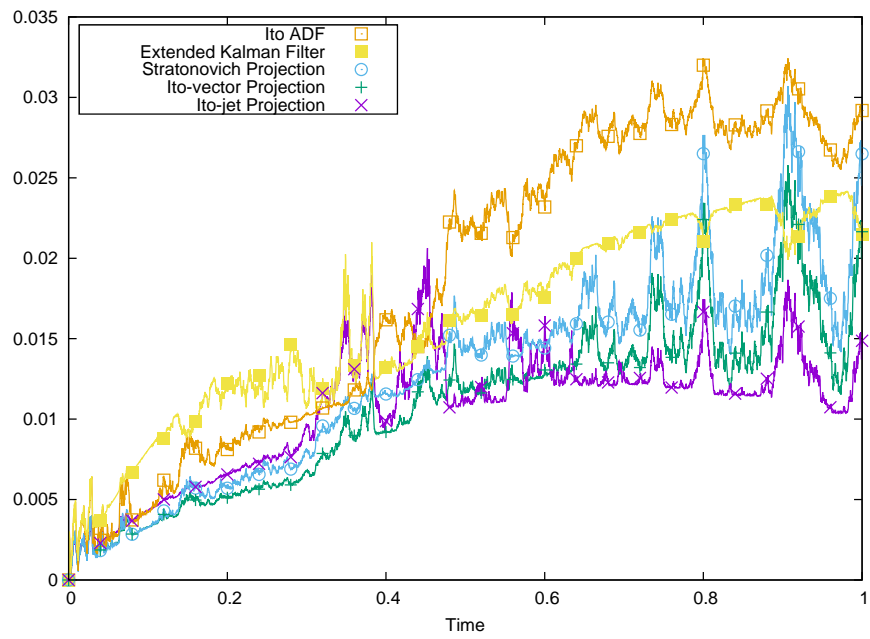


Figure 5:  $L^2$  residuals for each approximation method. All projections are taken relative to the  $L^2$  metric.

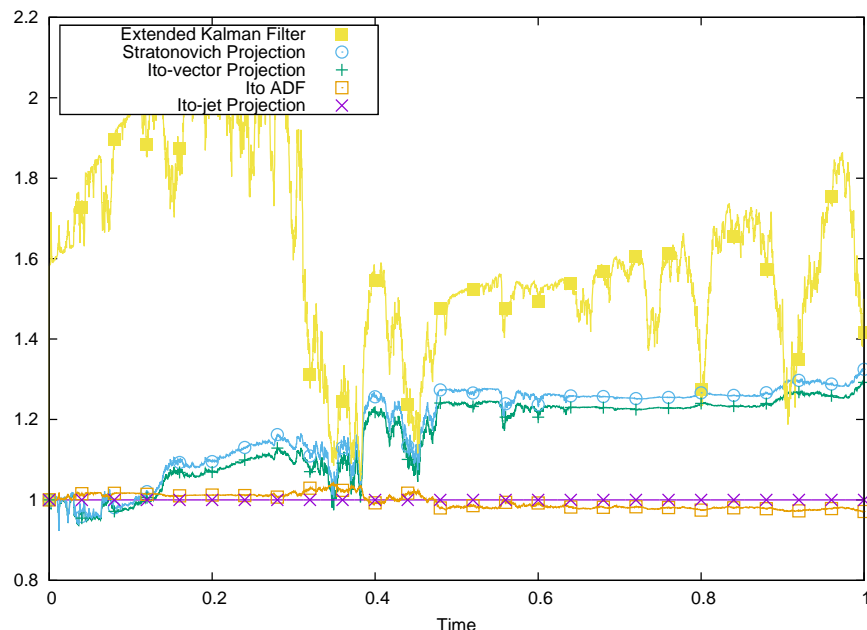


Figure 6: Hellinger residuals for various approximation method divided by residual for the Itô-jet projection. All projections are taken relative to the Hellinger metric.

This is because the residuals themselves are too difficult to distinguish visually. Thus values exceeding 1 show a larger error than the Itô-jet projection and values less than one show a lower error. All the projection methods shown in this plot are taken w.r.t. the Hellinger metric.

This plot indicates that the Itô ADF and the Itô-jet projection are almost indistinguishable in their performance. A look at the explicit formulae reveals that the difference between these two equations is of order  $\epsilon^2$  whereas the difference between the other equations is of order only  $\epsilon$ . Over the short term, the Itô-vector projection gives the best results. Over the medium term, the Itô-jet projection and the Itô ADF give the best results. Again, over the longer term all the filters become highly inaccurate.

## 7 Conclusions

The notion of projecting a vector field onto a manifold is unambiguous. By contrast, there are multiple distinct generalizations of this notion to SDEs, as summarized in Table 1.

The two Itô projections we introduced in this work can both be derived from minimization arguments. However, the Itô-jet projection has some clear

advantages.

- The Itô-jet projection is the best approximation to the metric projection of the true solution and has an error of  $O(t)$ . By contrast the Itô-vector projection only tracks the true solution an accuracy of  $O(t^{\frac{1}{2}})$ .
- The Itô-jet projection gives a more intuitive answer than the Itô-vector projection for the low dimensional example considered in Section 4.
- The Itô-jet projection gives better numerical results in the medium term than the Itô-vector projection in our application to filtering.
- The Itô-jet projection has an elegant definition when written in terms of 2-jets.
- The Itô-jet projection has a pictorial interpretation, shown in Figure 4.

We have also seen that the Stratonovich projection satisfies an ad hoc minimization that is less appealing than the ones of the Itô projections, since it requires a deterministic anchor point. The Itô-jet and Itô-vector projection arguments allow one to derive new Gaussian approximations to non-linear filters. Unlike previous Gaussian approximations to non-linear filters, these approximations are derived by minimization arguments rather than heuristic arguments. Thus the notion of projecting an SDE onto a manifold is able to give new results even for this well-worn topic.

## A Appendix: The cross-diffusion process

We briefly study and give some intuition on the cross-diffusion process (20), whose equations we repeat here:

$$\begin{aligned} dX_t &= \sigma Y_t dW_t, \\ dY_t &= \sigma X_t dW_t, \end{aligned}$$

with deterministic initial condition  $(X_0, Y_0)$ . We call this a cross diffusion, since each state crosses over as diffusion coefficient of the other state. Moreover, as we explain below, depending on the location on the plane of the initial condition, the paths group around the left or right arms of a St Andrew cross. This is another reason for the name of the process.

The process equation can be solved analytically: add and subtract both sides and solve the resulting geometric Brownian motion equations for  $X + Y$  and  $X - Y$ . One obtains:

$$\begin{aligned} X_t &= e^{-\frac{1}{2}\sigma^2 t} (X_0 \cosh(\sigma W_t) + Y_0 \sinh(\sigma W_t)), \\ Y_t &= e^{-\frac{1}{2}\sigma^2 t} (Y_0 \cosh(\sigma W_t) + X_0 \sinh(\sigma W_t)). \end{aligned}$$

One can see that the solution satisfies  $(X_t + Y_t)(X_t - Y_t) = Ke^{-\sigma^2 t}$ , with  $K = (X_0 + Y_0)(X_0 - Y_0)$ . For large times the product will tend to be closer

and closer to zero, so that either  $Y = X$  or  $Y = -X$ , with the solution paths grouping along these two lines while approaching zero. Notice that if  $(X, Y)$  is in the origin, the process does not move. This behaviour of the process can be easily seen when plotting a few paths. In Figure 7 we show a few paths of the cross-diffusion example in coordinates  $(x, y)$  and the process  $\theta_t = \arctan(Y_t/X_t)$  for the angular position. Clearly, when  $X + Y$  is near zero  $\theta$  will tend to be close to  $-\pi/4 = -0.785\dots$ , whereas when  $X - Y$  is near zero  $\theta$  will tend to be close to  $\pi/4 = 0.785\dots$

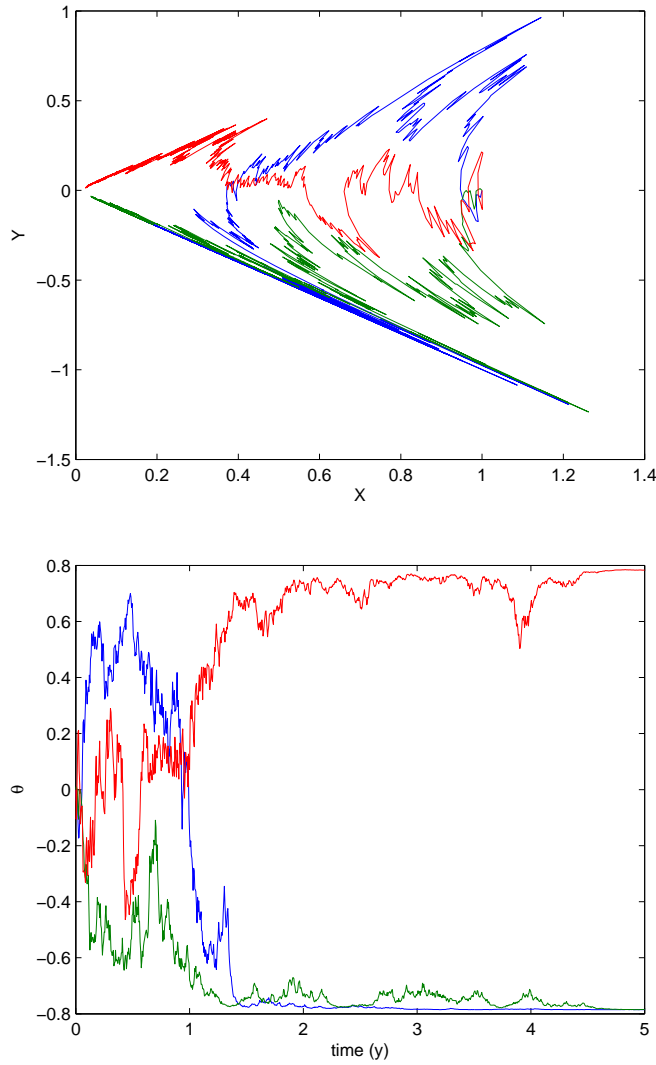


Figure 7: Top: Three paths of the cross-diffusion SDE with  $(X_0, Y_0) = (1, 0)$ ,  $\sigma = 1$ , up to 5 years time. Bottom: The corresponding three paths for  $\theta$  plotted against time.

## References

- [1] N. U. Ahmed. *Linear and Nonlinear Filtering for Scientists and Engineers*. World Scientific, Singapore, 1998.
- [2] J. Armstrong and D. Brigo. Coordinate free stochastic differential equations

- as jets. <http://arxiv.org/abs/1602.03931>, 2016.
- [3] John Armstrong and Damiano Brigo. Stochastic filtering via L2 projection on mixture manifolds with computer algorithms and numerical examples. *arXiv preprint arXiv:1303.6236*, 2013.
  - [4] John Armstrong and Damiano Brigo. Extrinsic projection of Itô SDEs on submanifolds with applications to non-linear filtering. In: *Nielsen, F., Critchley, F., & Dodson, K. (Eds), Computational Information Geometry for Image and Signal Processing, Springer Verlag*, 2016.
  - [5] John Armstrong and Damiano Brigo. Nonlinear filtering via stochastic PDE projection on mixture manifolds in  $L^2$  direct metric. *Mathematics of Control, Signals and Systems*, 28(1):1–33, 2016.
  - [6] Alan Bain and Dan Crisan. *Fundamentals of stochastic filtering*, volume 3. Springer, 2009.
  - [7] L. C. Brewin. Riemann normal coordinates. *Preprint Department of Mathematics, Monash University, Clayton, Victoria*, 3168, 1997.
  - [8] Damiano Brigo. On the nice behaviour of the gaussian projection filter with small observation noise. *Syst. Control Lett.*, 26(5):363–370, 1995.
  - [9] Damiano Brigo, Bernard Hanzon, and François LeGland. A differential geometric approach to nonlinear filtering: the projection filter. *Automatic Control, IEEE Transactions on*, 43(2):247–252, 1998.
  - [10] Damiano Brigo, Bernard Hanzon, and François LeGland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5(3):495–534, 1999.
  - [11] Damiano Brigo and Giovanni Pistone. Dimensionality reduction for measure-valued evolution equations in statistical manifolds. In *Proceedings of the conference on Computational Information Geometry for Image and Signal Processing*. Springer, 2016.
  - [12] Damiano Brigo and Giovanni Pistone. Optimal approximations of the Fokker-Planck-Kolmogorov equation: projection, maximum likelihood eigenfunctions and Galerkin methods. *arXiv preprint arXiv:1603.04348*, 2016.
  - [13] Z. Brzeźniak and K. D. Elworthy. Stochastic differential equations on Banach manifolds. *Methods Funct. Anal. Topology*, 6(1):43–84, 2000.
  - [14] David Elworthy. Geometric aspects of diffusions on manifolds. In *École d’Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, pages 277–425. Springer, 1988.
  - [15] M. Emery. *Stochastic calculus in manifolds*. Springer-Verlag, Heidelberg, 1989.



- [16] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.
- [17] Yuri E. Gliklikh. *Global and Stochastic Analysis with Applications to Mathematical Physics*. Theoretical and Mathematical Physics. Springer, London, 2011.
- [18] Michiel Hazewinkel, SI Marcus, and HJ Sussmann. Nonexistence of finite-dimensional filters for conditional statistics of the cubic sensor problem. *Systems & control letters*, 3(6):331–340, 1983.
- [19] Elton P. Hsu. *Stochastic Analysis on Manifolds*. Contemporary Mathematics. American Mathematical Society, 2002.
- [20] K. Itô. Stochastic differential equations in a differentiable manifold. *Nagoya Math. J.*, (1):35–47, 1950.
- [21] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.
- [22] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*. Applications of mathematics. Springer, Berlin, New York, Third printing, 1999.
- [23] Harold J Kushner. Approximations to optimal nonlinear filters. *Automatic Control, IEEE Transactions on*, 12(5):546–556, 1967.
- [24] Peter S Maybeck. *Stochastic models, estimation, and control*, volume 3. Academic press, 1982.
- [25] Nigel J. Newton. An infinite-dimensional statistical manifold modelled on Hilbert space. *J. Funct. Anal.*, 263(6):1661–1681, 2012.
- [26] Nigel J. Newton. Information geometric nonlinear filtering. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.*, 2(18):1550014, 24, 2015.
- [27] Étienne Pardoux. Filtrage non linéaire et équations aux dérivées partielles stochastiques associées. In *Ecole d’Eté de Probabilités de Saint-Flour XIX 1989*, pages 68–163. Springer, 1991.
- [28] Jean Picard. Efficiency of the extended kalman filter for nonlinear systems with small noise. *SIAM Journal on Applied Mathematics*, 51(3):843–885, 1991.
- [29] Giovanni Pistone and Carlo Sempì. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, 10 1995.