THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Series Expansion Approximations of Brownian Motion for Non-Linear Kalman Filtering of Diffusion Processes

OPEN ACCESS

Series Expansion Approximations of Brownian Motion for Non-Linear Kalman Filtering of Diffusion Processes

Simon Lyons, Simo Särkkä, *Senior Member, IEEE,* and Amos Storkey

Abstract—In this paper, we describe a novel application of sigma-point methods to continuous-discrete filtering. The nonlinear continuous-discrete filtering problem is often computationally intractable to solve. Assumed density filtering methods attempt to match statistics of the filtering distribution to some set of more tractable probability distributions. Filters such as these are usually decompose the problem into two sub-problems. The first of these is a prediction step, in which one uses the known dynamics of the signal to predict its state at time t_{i+1} given observations up to time t_i. In the second step, one updates the prediction upon arrival of the observation at time t_{i+1}.

The aim of this paper is to describe a novel method that improves the prediction step. We decompose the Brownian motion driving the signal in a generalised Fourier series, which is truncated after a number of terms. This approximation to Brownian motion can be described using a relatively small number of Fourier coefficients, and allows us to compute statistics of the filtering distribution with a single application of a sigma-point method.

Assumed density filters that exist in the literature usually rely on discretisation of the signal dynamics followed by iterated application of a sigma point transform (or a limiting case thereof). Iterating the transform in this manner can lead to loss of information about the filtering distribution in highly non-linear settings. We demonstrate that our method is better equipped to cope with such problems.

I. INTRODUCTION

STOCHASTIC differential equations (SDEs) provide a natural way to describe the evolution of systems that are inherently noisy, or contain unknown phenomena that can be modelled as stochastic processes [1], [2]. Suppose that the evolution of an idealised system could be modelled with the ordinary differential equation (ODE)

$$\frac{d\mathbf{X_t}}{dt} = \mathbf{a}(\mathbf{X}_t), \qquad (1)$$

where $\mathbf{X}_t \in \mathbb{R}^n$ is the state of the system, and $\mathbf{a} : \mathbb{R}^n \to \mathbb{R}^n$. Roughly speaking, to construct an SDE, one adds a 'white' driving noise to the dynamics of an ODE. From the modelling perspective, the purpose of the noise is to capture deviations from the ideal deterministic model. The amplitude of this driving noise may potentially depend on the current state $\mathbf{X}_t$ of the system. The result is a differential equation

S.M.J Lyons and A.J. Storkey are with the department of Informatics, Edinburgh University. E-mail: s.lyons-4@sms.ed.ac.uk, a.storkey@ed.ac.uk

Simo Särkkä is with the Department of Biomedical Engineering and Computational Science, Aalto University. E-mail: simo.sarkka@aalto.fi

$$\frac{d\mathbf{X}_t}{dt} = \mathbf{a}(\mathbf{X}_t) + \mathbf{b}(\mathbf{X}_t)\dot{\mathbf{W}}_t, \qquad (2)$$

where $\dot{\mathbf{W}}_t \in \mathbb{R}^d$ is Gaussian white noise, and $\mathbf{b} : \mathbb{R}^n \to \mathbb{R}^{n \times d}$. Because of the highly irregular nature of continuous-time white noise, one needs to be careful when defining this equation mathematically. In order to do this, the usual approach is re-write (2) as an integral equation and interpret the second term on the right as an Itô stochastic integral [2], [3]:

$$\mathbf{X}_t = \mathbf{X}_0 + \int_0^t \mathbf{a}(\mathbf{X}_u)du + \int_0^t \mathbf{b}(\mathbf{X}_u)d\mathbf{W}_u. \qquad (3)$$

This allows us to interpret the dynamics as an Itô *stochastic differential equation*:

$$d\mathbf{X}_t = \mathbf{a}(\mathbf{X}_t)dt + \mathbf{b}(\mathbf{X}_t)d\mathbf{W}_t, \qquad \mathbf{X}_0 = \mathbf{x}_0. \qquad (4)$$

The solution $\mathbf{X}_t$ will then be an *Itô diffusion process*. Here, the term $d\mathbf{W}_t$ denotes the infinitesimal change in a d-dimensional Brownian motion. We assume $\mathbf{W}$ is a standard Brownian motion, so that its components are independent with variance t at time t (we use the convention that a vector-valued stochastic process is represented with an upper-case bold letter, whereas a stochastic process *evaluated at a given time* also has a subscript). In situations where we need to refer to, say, the k-th component of the vector $\mathbf{X}_t$, we use the notation $\mathbf{X}_k(t)$.

One must make some assumptions on $\mathbf{a}$ and $\mathbf{b}$ to ensure Equation (4) has a unique solution. If both functions are globally Lipschitz and grow at most linearly, one is assured that this will be the case [3].

It is often the case that one cannot observe the process $\mathbf{X}$ directly—instead, one must rely on discrete-time, noisy observations $\{\mathbf{Y}_{t_k} \in \mathbb{R}^s\}_{k \geq 1}$ of the process. In mathematical terms, the model for measurements of this type can often be written as

$$\mathbf{Y}_{t_k} = \mathbf{h}(\mathbf{X}_{t_k}) + \mathbf{V}_{t_k}, \qquad (5)$$

with Gaussian measurement noises $\mathbf{V}_{t_k} \sim \mathcal{N}(0, \mathbf{R}_k)$. One is then often faced with the task of computing the expectation $E[\phi(\mathbf{X}_t)|\mathbf{Y}_{t_1}, \dots, \mathbf{Y}_{t_f}]$, where $t \geq t_f$ for some given function ϕ. This is known as the *continuous-discrete filtering problem*. For simplicity, we assume that the conditional distribution of $\mathbf{X}_t$ has a density with respect to the Lebesgue measure. For filtering problems where this is not the case, such as when part of the system is observed without error, much of our analysis can be applied with only minor modifications. The

estimation problem can be solved for arbitrary ϕ provided that we can compute the filtering density $p_{\mathbf{X}_t}\left(x \mid \{\mathbf{Y}_{t_k} : t_k \le t\}\right)$ for all t. This latter approach is often called the probabilistic or Bayesian approach to the filtering problem [1].

It is only in a small number of special cases that the conditional distribution of $\mathbf{X}_t$ can be described using a finite number of parameters. When the SDE is linear and the function $\mathbf{h}$ in the measurement model is linear, then the Kalman filter can be used to compute the exact solution [4]. Certain other filtering problems also admit closed-form solutions (see, for example, the Beneš filter [5]). However, closed-form filters are rare, and in most cases one must approximate the filtering distribution in some manner. For example, one can discretise the signal and employ a particle filter [6]–[9], which uses Monte Carlo samples to approximate the filtering distribution. Other approaches include variational filtering [10], homotopy filtering [11], and path integral filtering [12].

Another general technique is to take a parametric set of tractable densities (for example a set of densities within the exponential family) and find the density within that set that most closely matches the filtering density. This approach, introduced in [13], is known as *assumed density filtering*.

In this paper, we will attempt to compute statistics of the Gaussian distribution that most closely matches the filtering distribution. This particular special case of assumed density filtering is known as *Gaussian filtering* [14]. There are a number of ways to approach the problem. The *extended Kalman filter* (EKF) [1] uses a Taylor series approximation to the non-linearities in SDE and measurement model. The *unscented Kalman filter* (UKF) [15]–[17] uses a set of sigma-points for computing the mean and covariance of the Gaussian approximation. Quadrature and cubature based filters [14], [18]–[20] use Gaussian numerical integration for computing the mean and covariance. The Gaussian assumption is a natural one when the filtering distribution is known to be unimodal. However, it may lead to significant errors for certain multimodal distributions. It is not advisable to apply a Gaussian filter blindly, without considering the possibility of encountering a multimodal filtering distribution.

The commonly used approaches to filtering in continuous-discrete systems can be divided into two categories: one possibility is that the SDE is first discretised using methods such as Itô–Taylor series or a stochastic Runge–Kutta discretisation [21], [8]. Discrete-time filtering algorithms are the applied to the discretised process. The alternative is that an approximate filter is formed that operates in continuous time, and that filter is discretised. The relative merits of these approaches were recently studied in [22].

In the current paper, we take a different approach. We begin by noting that one can view the Brownian motion $\mathbf{W}$ as a random element of the Hilbert space $L^2[0,T]$. It is an inherently infinite-dimensional object. However, one can construct a finite-dimensional approximation of $\mathbf{W}$ by projecting it onto a finite-dimensional subspace of $L^2[0,T]$ [23]. We use the projection as the driving noise in an approximation of the original signal. The transition map is then approximated as a function that satisfies a certain ordinary differential equation. We refer to the new filter as the *series expansion unscented*

Kalman filter (SE-UKF).

Ideas of this type were first explored by Wong and Zakai [24]. Similar ideas have been explored in [25], [26] in the context of variance reduction for Monte-Carlo simulation, and in [27] in the context of parameter estimation. In this framework, one can interpret the approximation we use as the image of an N-dimensional standard normal distribution under a nonlinear transform. This suggests the possibility of using sigma-point methods such as the unscented transform to construct a Gaussian filter.

Gaussian filters that currently exist in the literature typically rely on discretisation of the signal. The time-t distribution of the discretised signal is repeatedly projected onto the set of Gaussian distributions, for example through moment matching or by minimising some form of generalised metric as in [28]. Our methodology avoids repeated projection onto the space of Gaussian random variables during the prediction phase. For this reason we expect our new prediction step to outperform the prediction steps of existing methods when the inference problem is sufficiently nonlinear.

Our paper is structured as follows. In Section II, we describe our model of the filtering problem and briefly review some methods that are used in the literature at present. In Section III, we describe our method of approximating the time-t marginal distribution of a diffusion process, and we show how the approximation can be exploited to construct a novel Gaussian filter. The accuracy of this approximation is investigated in Section IV, and we show that our filter performs well on a high-dimensional nonlinear problem. In Section V, we review our work and discuss some questions that arise as a result of the study.

II. GAUSSIAN FILTERING

A. Sigma point approximations

One widely-used approach to Gaussian filtering relies on so-called 'sigma point' approximations, perhaps the best known of which is the unscented transform [15], [16]. Given a random variable $\mathbf{U}$ and a function $\mathbf{f}$, we wish to approximate the distribution of $\mathbf{f}(\mathbf{U})$. In order to accomplish this, one chooses a number of points $\{\sigma_i\}$ that represent the distribution of $\mathbf{U}$ in some sense. The number of sigma points used by the unscented transform scales linearly with the dimension of $\mathbf{U}$.

We will restrict our exposition to the case where $\mathbf{U}$ has an n-dimensional multivariate normal distribution, and we wish to fit a multivariate normal distribution to $\mathbf{f}(\mathbf{U})$. Suppose $\mathbf{U}$ has mean $\mathbf{m}$ and covariance $\mathbf{P}$. The unscented transform uses $2n + 1$ sigma points, which are constructed as follows. One chooses two tuning parameters α and κ, then sets $\lambda = \alpha^2(n + \kappa) - n$. The sigma points are then defined by the following expressions:

$$\sigma^0 = \mathbf{m}, \tag{6}$$

$$\sigma^i = \mathbf{m} + (\sqrt{(n+\lambda)\mathbf{P}})_{*i}, \quad 1 \le i \le n, \tag{7}$$

$$\sigma^{n+i} = \mathbf{m} - (\sqrt{(n+\lambda)\mathbf{P}})_{*i}, \quad 1 \le i \le n. \tag{8}$$

The sigma points are determined once one chooses an appropriate matrix square root. Here $(\sqrt{\mathbf{P}})_{*i}$ is the i-th column of the matrix square root of $\mathbf{P}$ defined via $\mathbf{P} = \sqrt{\mathbf{P}}\sqrt{\mathbf{P}}^\top$.

The mean and covariance of $\mathbf{f}(\mathbf{U})$ are approximated by a weighted average of the sigma-point images. Define $\mathcal{Y}_i = \mathbf{f}(\sigma^i)$, and set

$$\mu := \sum_{i=0}^{2n} w_i^{(m)} \mathcal{Y}_i \simeq E[\mathbf{f}(\mathbf{U})]. \tag{9}$$

We can then make the approximations

$$E[(\mathbf{f}(\mathbf{U}) - E[\mathbf{f}(\mathbf{U})])(\mathbf{f}(\mathbf{U}) - E[\mathbf{f}(\mathbf{U})])^\top]$$
$$\simeq \sum_{i=0}^{2n} w_i^{(c)} (\mathcal{Y}_i - \mu)(\mathcal{Y}_i - \mu)^\top, \tag{10}$$
$$E[(\mathbf{U} - m)(\mathbf{f}(\mathbf{U}) - E[\mathbf{f}(\mathbf{U})])^\top]$$
$$\simeq \sum_{i=0}^{2n} w_i^{(c)} (\sigma^i - \mathbf{m})(\mathcal{Y}_i - \mu)^\top. \tag{11}$$

The weights depend on a third tuning parameter β, and are given by

$$w_0^{(m)} = \frac{\lambda}{n + \lambda},$$
$$w_0^{(c)} = \frac{\lambda}{n + \lambda} + (1 - \alpha^2 + \beta),$$
$$w_i^{(m)} = \frac{1}{2(n + \lambda)} \qquad i = 1, \dots, 2n,$$
$$w_i^{(c)} = \frac{1}{2(n + \lambda)} \qquad i = 1, \dots, 2n. \tag{12}$$

It is well known that the unscented transform matches the mean of $\mathbf{f}(\mathbf{U})$ exactly when $\mathbf{f}$ is a polynomial of degree three or less. In general, errors in the estimate of the mean are introduced only by the fourth and higher terms in the Taylor expansion of $\mathbf{f}$ [29].

B. Sigma point Kalman filters for diffusion processes

In the Gaussian filtering paradigm, of which the unscented Kalman filter (UKF) is a special case, the filtering problem is reduced to computation of the conditional mean and covariance of the filtering distribution:

$$\mathbf{m}_t = E[\mathbf{X}_t \mid \{\mathbf{Y}_{t_k} : t_k \le t\}] \tag{13}$$

and

$$\mathbf{P}_t = \mathrm{Cov}[\mathbf{X}_t \mid \{\mathbf{Y}_{t_k} : t_k \le t\}]. \tag{14}$$

This procedure is usually divided up into two steps: the *prediction* step and the *update* step. In the prediction step, we begin with an estimate of the mean and covariance of the filtering distribution at time t_{k-1}. We then use the known dynamics of $\mathbf{X}$ to compute the mean and variance of the filtering distribution the instant before the next observation arrives.

Upon arrival of the observation at time t_k, we proceed to the *update* step. In this step, we update our estimate of the mean and variance of the filtering distribution using information from observation $\mathbf{Y}_{t_k}$.

It is usually necessary to approximate the conditional mean and covariance: for a general nonlinear diffusion, the moments are only known in terms of the solution of a partial differential equation known as the Fokker-Planck equation [1], [2]. In dimensions higher than three, the Fokker-Planck equation is typically numerically intractable.

The simplest application of the UKF to a diffusion relies on discretisation of the process $\mathbf{X}$. Suppose that at time t_{k-1} we have an estimate of $\mathbf{m}_{t_{k-1}}$ and $\mathbf{P}_{t_{k-1}}$. In the prediction step, our aim is to compute an estimate of $\mathbf{m}_t$ and $\mathbf{P}_t$ the instant before the next observation arrives.

We divide the time interval $[t_{k-1}, t_k]$ into a number of sub-intervals of length Δt (for clarity, we will discuss the interval $[0, t_1]$ here). We then approximate the SDE (4) on the grid $\{\mathbf{X}_{\Delta t}, \mathbf{X}_{2\Delta t}, \dots\}$ via the relation

$$\mathbf{X}_{(j+1)\Delta t} = \mathbf{f}(\mathbf{X}_{j\Delta t}, \mathbf{Z}_j), \tag{15}$$

where $\mathbf{Z}_0, \mathbf{Z}_1, \dots$ is a suitable sequence of Gaussian random variables. Here, $\mathbf{f}$ is a transition function that depends on the method of discretisation, and $\mathbf{Z}_k$ is typically draw from a spherical Gaussian distribution of dimension d. For example, in the *Euler-Maruyama* scheme [21],

$$\mathbf{f}(\mathbf{X}_{j\Delta t}, \mathbf{Z}_j) = \mathbf{X}_{j\Delta t} + \mathbf{a}(\mathbf{X}_{j\Delta t})\Delta t + \mathbf{b}(\mathbf{X}_{j\Delta t})\sqrt{\Delta t}\mathbf{Z}_j, \tag{16}$$

where $\mathbf{Z}_j \sim \mathcal{N}(0, \mathbf{I}_d)$.

In this sense, $\mathbf{X}_{(j+1)\Delta t}$ is the image of $(\mathbf{X}_{j\Delta t}, \mathbf{Z}_j)$ under a nonlinear transform $\mathbf{f}$. Given a Gaussian approximation to $\mathbf{X}_{j\Delta t}$, one can apply the unscented transform to $\mathbf{f}$ to find a Gaussian approximation of $\mathbf{X}_{(j+1)\Delta t}$. One proceeds iteratively until t_k, at which point the prediction phase ends and we proceed to the update phase. Instead of the Euler–Maruyama method, one can also use higher order Itô–Taylor expansions, stochastic Runge–Kutta methods or various other methods [21].

Alternatively, instead of iteratively applying the unscented transform at the prediction, one can take a limit as $\Delta t \to 0$, in which case one recovers a system of differential equations for the predictive mean and covariance (see, e.g., [17], [20]):

$$\frac{d\mathbf{m}_t^-}{dt} = E[\mathbf{a}(\mathbf{X}_t^-)]$$
$$\frac{d\mathbf{P}_t^-}{dt} = E[\mathbf{a}(\mathbf{X}_t^-)(\mathbf{X}_t^- - \mathbf{m}_t^-)^\top]$$
$$+ E[(\mathbf{X}_t^- - \mathbf{m}_t^-)\mathbf{a}^\top(\mathbf{X}_t^-)]$$
$$+ E[\mathbf{b}(\mathbf{X}_t^-)\mathbf{b}^\top(\mathbf{X}_t^-)], \tag{17}$$

where the expectations are taken with respect to the Gaussian distribution $\mathbf{X}_t^- \sim \mathcal{N}(\mathbf{m}_t^-, \mathbf{P}_t^-)$.

When $t = t_k$, we will make a new observation. We must then update our predictive distribution with the new information from that observation. Let $\mathbf{m}_{t_k}^-$ and $\mathbf{P}_{t_k}^-$ be the mean and covariance of the predictive distribution immediately before the new observation arrives. We form an approximation $\hat{\mathbf{X}}_{t_k}^-$ of the signal, which is Gaussian with the predictive mean and covariance. The update equations using the linear

minimum mean square error estimator are as follows:

$$\begin{aligned}
\mu_k &= E[h(\hat{\mathbf{X}}_{t_k}^-)] \\
\mathbf{S}_k &= E[(h(\hat{\mathbf{X}}_{t_k}^-) - \mu_k)(h(\hat{\mathbf{X}}_{t_k}^-) - \mu_k)^\top] + \mathbf{R}_k \\
\mathbf{C}_k &= E[(\hat{\mathbf{X}}_{t_k}^- - \mathbf{m}_{t_k}^-)(h(\hat{\mathbf{X}}_{t_k}^-) - \mu_k)^\top] \\
\mathbf{K}_k &= \mathbf{C}_k \mathbf{S}_k^{-1} \\
\mathbf{m}_{t_k} &= \mathbf{m}_{t_k}^- + \mathbf{K}_k(\mathbf{Y}_{t_k} - \mu_i) \\
\mathbf{P}_{t_k} &= \mathbf{P}_{t_k}^- - \mathbf{K}_k \mathbf{S}_k \mathbf{K}_k^\top,
\end{aligned} \tag{18}$$

The updated distribution has mean $\mathbf{m}_{t_k}$ and covariance $\mathbf{P}_{t_k}$. When the observation function $\mathbf{h}$ is nonlinear, one can apply the unscented transform to $\mathbf{h}(\hat{\mathbf{X}}_{t_k}^-)$ to compute an approximation of μ_k, $\mathbf{S}_k$, and $\mathbf{C}_k$ in (18). More complex update rules that have been tuned for numerical stability are also known in the literature [30].

In Section III, we describe a new method of approximating the predictive mean and variance by constructing a function $\mathbf{f}$ so that $\mathbf{X}_t \approx \mathbf{f}(t, \mathbf{X}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_N)$, where the random variables $\{\mathbf{Z}_i\}$ follow a standard normal distribution. We can apply sigma point methods to $\mathbf{f}$ to estimate the mean and variance of $\mathbf{X}_t$. The image of each sigma point is computed by solving an ordinary differential equation. Our method requires *one* application of the unscented transform per observation (though this is generalised in Section IV-C). This is in contrast to the standard UKF, which discretises the system first, then iteratively applies the unscented transform at each timestep.

III. SIGMA POINT FILTERING VIA SMOOTH APPROXIMATIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS

A. Series expansions of Brownian motion

We now describe a method for obtaining a smooth approximation of Brownian motion by decomposing it in a generalised Fourier series. We aim to use the smooth approximation as a driving function in a differential equation. This will enable us to approximate a nonlinear stochastic differential equation with a *randomised ordinary differential equation*, which will prove to be computationally tractable to work with. This approximation was used as the basis of a Markov chain Monte Carlo algorithm for Bayesian parameter estimation of a nonlinear diffusion in [27].

Suppose $\mathbf{W} = (W^{(1)}, \ldots, W^{(d)})$ is a standard d-dimensional Brownian motion, and let $\{\phi_i\}_{i \geq 1}$ be an orthonormal basis of $L^2([0, T], \mathbf{R})$. We use the notation I to denote the indicator function. That is, $I_{\{[0,t]\}}(u) = 1$ when $0 \leq u \leq t$, and $I_{\{[0,t]\}}(u) = 0$ otherwise. One can construct a series expansion of $\mathbf{W}$ in terms of the basis functions $\{\phi_i\}$ as follows [31]:

$$\begin{aligned}
\mathbf{W}_t &= \int_0^T I_{\{[0,t]\}}(u) d\mathbf{W}_u \\
&= \int_0^T \left(\sum_{i=1}^{\infty} \langle I_{\{[0,t]\}}, \phi_i \rangle \phi_i(u) \right) d\mathbf{W}_u \\
&= \sum_{i=1}^{\infty} \left(\int_0^T \phi_i(u) d\mathbf{W}_u \right) \int_0^t \phi_i(u) du. \tag{19}
\end{aligned}$$

We use the standard inner product on $L^2[0, T]$, which is defined as

$$\langle f, g \rangle = \int_0^T f(u) g(u) du. \tag{20}$$

The stochastic integrals are i.i.d d-dimensional standard normal. We can see this by noting that he basis functions are deterministic,

$$E\left[\int_0^T \phi_i(u) d\mathbf{W}_u \right] = 0, \tag{21}$$

and

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{Z}_i, \mathbf{Z}_j) \\
&= E\left[\left(\int_0^T \phi_i(u) d\mathbf{W}_u \right) \left(\int_0^T \phi_j(u) d\mathbf{W}_u \right)^\top \right] \\
&= \left(\int_0^T \phi_i(u) \phi_j(u) du \right) \mathbf{I}_d = \delta_{ij} \mathbf{I}_d. \tag{22}
\end{aligned}$$

Here, $\mathbf{I}_d$ is the $d \times d$ identity matrix.

For ease of notation, we set

$$\mathbf{Z}_i = \int_0^T \phi_i(u) d\mathbf{W}_u. \tag{23}$$

We conclude that

$$\mathbf{W}_t = \sum_{i=1}^{\infty} \mathbf{Z}_i \int_0^t \phi_i(u) du. \tag{24}$$

We can obtain an approximation of a Brownian sample path by drawing i.i.d samples $\mathbf{Z}_i$ from a standard normal distribution and truncating the sum in (24). This allows us to describe a Brownian sample path approximately in terms of a finite number of variates. This representation is crucial for our implementation of sigma-point inference methods.

B. Series Expansion Approximation of SDE

In order to approximate the diffusion $\mathbf{X}$, we truncate the series expansion (24) after N terms, and use the resulting smooth process as an approximation of Brownian motion. We replace the stochastic integral in Equation (3) with the time derivative of the truncated process:

$$\hat{\mathbf{X}}_t = \mathbf{X}_0 + \int_0^t \mathbf{a}(\hat{\mathbf{X}}_u) du + \sum_{i=1}^{N} \int_0^t \mathbf{b}(\hat{\mathbf{X}}_u) \mathbf{Z}_i \phi_i(u) du. \tag{25}$$

Since $\hat{\mathbf{X}}$ is driven by a finite linear combination of basis functions, the resulting process is differentiable. We can therefore interpret $\hat{\mathbf{X}}$ as the solution to an ordinary differential equation with a random driving function.

$$\frac{d\hat{\mathbf{X}}_t}{dt} = \mathbf{a}(\hat{\mathbf{X}}_t) + \sum_{i=1}^{N} \mathbf{b}(\hat{\mathbf{X}}_t) \mathbf{Z}_i \phi_i(t), \qquad \hat{\mathbf{X}}_0 = x_0. \tag{26}$$

Approximations of this type were first investigated by Wong and Zakai [24], who showed that in the one-dimensional case, $\hat{\mathbf{X}}_t$ converges to the Stratonovich solution of the stochastic differential equation [21]. Convergence issues are discussed in a more general multidimensional setting in the appendix.

The approximation (26) has the advantage of re-casting an infinite dimensional problem in finite-dimensional terms. We can view the solution of (26) as a function

$$\hat{\mathbf{X}}_t = \hat{\mathbf{X}}(t, x_0, \mathbf{Z}_1 \dots, \mathbf{Z}_N). \tag{27}$$

In essence, the time-t distribution of the process $\hat{\mathbf{X}}$ can be interpreted as the image of a $d \times N$-dimensional Gaussian distribution under a nonlinear transform. This is precisely the setting for which sigma-point methods were designed.

C. The proposed series expansion filter

Our algorithm proceeds as follows. We assume we have a Gaussian approximation $\mathcal{N}(\mathbf{m}_{t_{k-1}}, \mathbf{P}_{t_{k-1}})$ to the filtering distribution at time t_{k-1}. We wish to compute the filtering distribution at time t. If $t < t_k$, we compute the predictive distribution. If $t = t_k$, we must also update the predictive distribution with the information gained from our observation $\mathbf{Y}_{t_k}$.

We choose a set $\{\sigma_j\}$ of sigma points to represent the joint distribution of the state and the random coefficients $\{\mathbf{Z}_i\}$ in (26). Each sigma point can be thought of as a vector of dimension $n + d \times N$,

$$\sigma^j = (\sigma_x^j, \sigma_z^j). \tag{28}$$

Here, the first n elements σ_x^j of the vector σ^j are the sigma points for the initial condition for the ODE (26), that is, the sigma points that represent $\mathcal{N}(\mathbf{m}_{t_{k-1}}, \mathbf{P}_{t_{k-1}})$. The remaining $d \times N$ elements σ_z^j are the sigma points corresponding to an N-term expansion of a d-dimensional Brownian motion. Together, these data determine an initial value problem. For each sigma point σ^j, we solve the ordinary differential equation (26). The initial condition is $\mathbf{X}_{t_{k-1}} = \sigma_x^j$ and the coefficients representing $\{\mathbf{Z}_i\}_{i \leq N}$ are formed from the appropriate subvectors of σ_z^j (each one having length d). At time T, the solution is an n-dimensional vector

$$\hat{\mathbf{X}}_T^j = \hat{\mathbf{X}}(T, \sigma_x^j, \sigma_z^j). \tag{29}$$

We treat the solution at time T of the initial value problem as the image of the sigma point σ^j. The set of vectors $\{\hat{\mathbf{X}}_t^j\}$ can be thought of as a discrete approximation to the predictive distribution. We can use these vectors to compute an estimate of $\mathbf{m}_t$ and $\mathbf{P}_t$, though the specific computation depends on the choice of sigma-point method. This methodology is in marked contrast to the sigma point Kalman filters of Section II-B. These rely on discretisation of the signal dynamics and sigma point approximation of the Brownian increment $\mathbf{W}_{t+\Delta t} - \mathbf{W}_t$ at each timestep, or a limiting case of this discretisation as $\Delta t \to 0$.

We summarise our algorithm in pseudocode as follows:

for $k = 1 : m$ **do**
  Set $\mathbf{m}_\sigma = (\mathbf{m}_{t_{k-1}}, \mathbf{0}_{1 \times (Nd)})$
  Set $\mathbf{P}_\sigma = \begin{pmatrix} \mathbf{P}_{t_{k-1}} & \mathbf{0}_{n \times (Nd)} \\ \mathbf{0}_{(Nd) \times n} & I_{(Nd) \times (Nd)} \end{pmatrix}$
  Generate $2(n + Nd) + 1$ sigma points, with weighted mean $\mathbf{m}_\sigma$ and weighted covariance $\mathbf{P}_\sigma$
  for Each sigma point $\sigma^{(j)}$ **do**
    Set $x_0 = \sigma_{1:n}^{(j)}$.

Set $\mathbf{Z}_{1:N} = \sigma_{n+1:n+(Nd)}^{(j)}$ (reshaping the right-hand side into a $d \times N$ matrix if appropriate).
Solve numerically Equation (26). Let $\mathbf{X}_T^{(j)}$ be the value of the solution after T units of time.
Set $\mathcal{Y}_j = \mathbf{h}(\mathbf{X}_T^{(j)})$.
end for
Predict the mean and variance of the incoming observation using (9) and (10).
Upon arrival of the observation $\mathbf{y}_{t_k}$, update the mean $\mathbf{m}_{t_k}$ and variance $\mathbf{P}_{t_k}$ of the filtering distribution using (18).
end for

IV. NUMERICAL EXPERIMENTS

A general analysis of the error induced by the series expansion approximation is difficult. One cannot easily exploit the usual tools from the theory of stochastic processes. In general, the truncated driving noise does not possess the Markov property, nor is it a martingale. The truncated driving noise is, however, a Gaussian process, and this structure is exploited in [32] to demonstrate convergence to the true SDE. In the first part of this section we present a numerical investigation into the approximation error.

We then compare the series expansion UKF with the cubature Kalman filter, which was found to be the most accurate and numerically stable amongst standard unscented transform-based filters in this context. There is already a considerable amount of theoretical and empirical evidence in the literature that sigma point methods outperform the extended Kalman filter, especially in tracking models such as the one described below (see, for example, [16] [19] [29]). In addition, one must compute the gradient of the drift function in order to implement the EKF. For some processes, this can be cumbersome. In contrast, our algorithm can be used as a 'black box' filter. We compare our results with the UKF rather than the EKF to provide the most informative experiments. In these experiments, we use a Stratonovich-to-Itô correction term to modify the dynamics of our approximation, so that the solution coincides with the Itô dynamics [21].

A. Approximation error

In the general nonlinear case, analytic solutions for nonlinear multi-dimensional ordinary differential equations are rarely available in closed form. Hence, it is difficult to establish precise bounds on the error induced by the series expansion approximation. In this section we aim to investigate properties of the series expansion approximation numerically. In the example we consider, we will see that one can obtain a good approximation by truncating the series expansion (19) after about ten terms.

We will test our approximation on a model of an aircraft turning in the (x_1, x_3) plane. We model the motion of the aircraft using noisy dynamics that account for imperfections in the control system. The model also accounts for external forces such as wind that might affect the trajectory of the aircraft. We describe the state of the with a seven-dimensional vector $x_{1:7}$. The components (x_1, x_3, x_5) represent the position of the aircraft in rectangular cartesian coordinates, while the

components (x_2, x_4, x_6) describe its velocity. The number x_7 describes the rate at which the aircraft is turning in the (x_1, x_3) plane.

The dynamics of the system are given by (4), with

$$\mathbf{a}(x_{1:7}) = \begin{pmatrix} x_2 \\ -x_7 x_4 \\ x_4 \\ x_7 x_2 \\ x_6 \\ 0 \\ 0 \end{pmatrix} \quad (30)$$

$$\mathbf{b}(x_{1:7}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{\sqrt{1+x_2^2}}{v} & \frac{\sqrt{1+x_4^2}}{v_{xy}} & \frac{\sqrt{(1+x_2^2)(1+x_6^2)}}{vv_{xy}} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{\sqrt{1+x_4^2}}{v} & -\frac{\sqrt{1+x_2^2}}{v_{xy}} & \frac{\sqrt{(1+x_4^2)(1+x_6^2)}}{vv_{xy}} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{\sqrt{1+x_6^2}}{v} & 0 & -\frac{v_{xy}}{v} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (31)$$

Here, $v = \sqrt{1 + x_2^2 + x_4^2 + x_6^2}$ and $v_{xy} = \sqrt{1 + x_2^2 + x_4^2}$. Nonlinearities arise from two sources in this system. Firstly, the state-dependent covariance matrix causes the system to deviate from Gaussianity. Second, the random evolution of the turn rate $\mathbf{X}_7(\cdot)$ causes the aircraft to behave erratically. As the variance of $\mathbf{X}_7(\cdot)$ grows, the system becomes more nonlinear and more non-Gaussian. A similar model was studied in [19], though in that case the diffusion matrix was assumed to be constant. Note that the state dependent covariance matrix makes Itô-Taylor and Runge-Kutta discretisations difficult to implement.

In order to test the series expansion approximation, we simulated paths from $\mathbf{X}$ on the interval $[0, 8]$. We set $\mathbf{X}_0 = (1000, 0, 2650, 150, 200, 0, 6)$, and $\mathrm{Cov}(\mathbf{W}_t) = \mathrm{Diag}(50, 50, 50, 25)t$, resulting in a highly nonlinear process. We took $100,000$ simulations from the Euler-Maruyama scheme as ground truth, having set $\Delta t = .005$. The basis functions were defined by

$$\phi_k(t) = \sqrt{\frac{2}{T}} \sin\left(\frac{(k - \frac{1}{2})\pi t}{T}\right), \quad (32)$$

with $T = 8$. We simulated $100,000$ paths from the series expansion approximation with $N = 1, 4, 6$ and 10. The marginal means and standard deviations are shown in tables 1 and 2. Figure 1 shows a Q-Q plot of the Euler simulation versus the series expansion simulation with $N = 10$, together with a plot of both densities.

B. Filtering Experiments

As the nonlinearity of the system increases, the speed at which the filtering distribution deviates from Gaussianity should also increase. Intuitively, this means the amount of information that the conventional UKF 'throws away' at each timestep grows with nonlinearity of the system. The series expansion method avoids this issue by targeting the predictive density at a given time directly without any intermediate
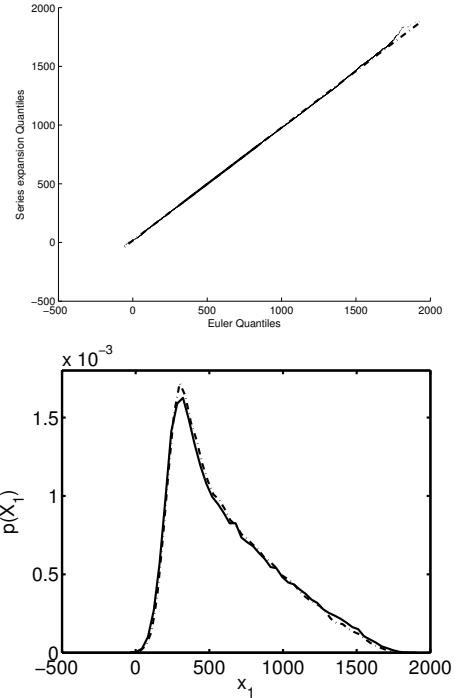

Fig. 1: (Top) Q-Q plot of 100,000 samples from an Euler discretisation of $\mathbf{X}_1(T = 8)$ versus 100,000 samples from the series expansion approximation. Linearity of the plot suggests the distributions are very similar. (Bottom) Density plots of the samples. Draws from the Euler scheme are plotted using the solid line, and draws from the series expansion scheme are represented by the broken line. We used the Fourier sine series as a basis, with $N = 10$.

	Euler	N = 1	N = 4	N = 6	N = 10
$E[\mathbf{X}_1(t)]$	626	549	607	612	619
$E[\mathbf{X}_2(t)]$	-59	-91	-65	-63	-61
$E[\mathbf{X}_3(t)]$	3588	3689	3612	3603	3597
$E[\mathbf{X}_4(t)]$	53	82	58	56	55
$E[\mathbf{X}_5(t)]$	200	200	200	200	200
$E[\mathbf{X}_6(t)]$	0	0	0	0	0
$E[\mathbf{X}_7(t)]$	6	6	5.9	6	5.9

TABLE I: Marginal mean values for $\mathbf{X}_{1:7}(t = 8)$ as computed by the Euler scheme and series expansion approximations

discretisation or projection. As a result, we should expect the series expansion filter to outperform the conventional UKF in systems that are more highly nonlinear.

To test this hypothesis, we set the covariance of the four-dimensional Brownian motion driving the aircraft model to $\mathrm{Cov}(\mathbf{W})(t) = \mathrm{Diag}(10, 0.2, 0.2, Q_W^2)t$. The quantity Q_W determines the variance of the turn rate of the aircraft. We use it as a proxy for the degree of nonlinearity of the system. We chose a number of values for Q_W, ranging between $Q_W = 0.1$ and $Q_W = 1.1$. For each value of the variance, we simulated 1400 trajectories for the aircraft, running both filters on each trajectory. For each trajectory, the initial condition was drawn from a Gaussian distribution with mean $m_0 = (1000, 0, 2650, 150, 200, 0, 6)$. The standard deviation

	Euler	N = 1	N = 4	N = 6	N = 10
Std($\mathbf{X}_1(t)$)	359	151	317	333	346
Std($\mathbf{X}_2(t)$)	90	61	86	88	89
Std($\mathbf{X}_3(t)$)	277	128	250	261	268
Std($\mathbf{X}_4(t)$)	93	66	90	91	92
Std($\mathbf{X}_5(t)$)	29	17	27	28	28
Std($\mathbf{X}_6(t)$)	6.4	5.7	6.2	6.3	6.3
Std($\mathbf{X}_7(t)$)	14.1	12.8	13.7	13.9	14.0

TABLE II: Marginal standard deviations for $\mathbf{X}_{1:7}(t = 8)$ as computed by the Euler scheme and series expansion approximations

of each component was set to 100, with the exception of the standard deviation of $\mathbf{X}_7(0)$ (recall that this notation denotes the seventh component of the vector at time 0, rather than the value at time 7). This was set to 0.1. All components were assumed to be uncorrelated initially.

For each trajectory, we simulated 20 observations, spaced $T = 8$ units of time apart. The observation function $\mathbf{h}$ models radar signals arriving at a dish. For this reason, we assume observations arrive in spherical coordinates, so that $\mathbf{h}$ is given by

$$\mathbf{h}(x_{1:7}) = \begin{pmatrix} \sqrt{x_1^2 + x_3^2 + x_5^2} \\ \tan^{-1}(x_3/x_1) \\ \tan^{-1}(x_5/\sqrt{x_1^2 + x_3^2}). \end{pmatrix} \quad (33)$$

The covariance matrix of the observation noise was set to $\mathbf{R} = \text{diag}(50, 0.1, 0.1)$.

For the standard unscented Kalman filter, an Itô-Taylor scheme such as the one proposed in [19] is impractical to implement as a result of the state-dependent noise. This is due to the presence of iterated stochastic integrals in which the integrand is a function of $\mathbf{X}_t$ (see [21]). Even the simplified order 2.0 Itô-Taylor scheme proposed in [21] is cumbersome to implement. For an n-dimensional process, we need to compute $n^2 + 2n + 1$ terms involving derivatives of the coefficient functions (in our case, $n = 7$ so this means 64 terms). The simplified scheme also involves a number of Bernoulli random variables, and it is not immediately clear how one would incorporate these into an unscented filter.

We chose to use the limiting scheme first proposed in [17]. The system of ODEs (17) was solved by a fourth order Runge-Kutta scheme. The number of Runge-Kutta steps used did not appear to affect the error appreciably. However, with a large step size the predicted covariance can fail to be positive definite, which causes the filter to diverge. We found that a good compromise between computational cost and the divergence issue was to choose a smaller step-size for more highly nonlinear parameter settings. For this reason, we used $200Q_{\mathbf{W}}$ steps per unit time.

The system of ordinary differential equations (26) defining the series expansion method was solved numerically using the Dormand-Prince Runge-Kutta method. This is the default ODE solver implemented in MATLAB.

For the standard Unscented filter, we set $\alpha = 1$, $\kappa = 0$ and $\beta = 0$. This choice of tuning parameters is also known as the cubature Kalman filter [18], [19]. Various other parameter settings produced similar results, though these settings were

most stable and most accurate.

For the series expansion method, we used the orthonormal basis (32) with $T = 8$, and used $N = 8$ basis functions for each component of the Brownian motion.

The series expansion filter takes one large step instead of many small ones. As such, one can expect that the target distribution is less like a Gaussian distribution. As a consequence, the standard heuristics for choosing parameters for the unscented transform may be sub-optimal. We found that 'tweaking' the standard parameters slightly improved performance, though not dramatically. We set $\alpha = 1, \kappa = -32, \beta = 0$ so that $\lambda = 7$. Our motivation for this choice is given in Section V.

For each component (i.e. position, velocity, turn rate), we compute the root mean squared error

$$\epsilon^c = \sqrt{\frac{1}{nl} \sum_{k=1}^{n} \left(\mathbf{X}_c(t_k) - \mathbf{m}_c(t_k)\right)^\top \left(\mathbf{X}_c(t_k) - \mathbf{m}_c(t_k)\right)}, \quad (34)$$

Here, $\mathbf{m}_i$ is the mean of the filtering distribution at time i. The value of c depends on the error component. For position errors, $c = (1, 3, 5)$. For velocity errors, $c = (2, 4, 6)$, and for turn rate errors, $c = 7$. We set $l = 3$ for the position and velocity errors and $l = 1$ for the turn rate error. Mean filter errors and divergences are reported in Table III. A filter was deemed to have diverged if the RMSE position error was greater than 1km. When this occurred, the run was not included in the average.

Both the series expansion filter and unscented filter can diverge and lose track of the signal, in which case the error becomes very large. Even if divergences are discardrd, a few large errors can still dominate the average. For this reason, we report the median absolute error for each component in Figure 2. We also compute the errors $\epsilon_{\text{UKF}}(i)$ and $\epsilon_{\text{SE}}(i)$ for each run i. We report quartiles of the empirical distribution of $\epsilon_{\text{UKF}} - \epsilon_{\text{SE}}$ in Figure 3.

Figure 2 shows the median values of the difference in errors together with the first and third interquartiles. The third interquartile corresponding to $Q_W = 1.1$ is excluded because the plot could not be scaled appropriately. For the position, the value is $77m$. For the velocity, 67m/s, and for the turn rate, 7.8 degrees/s.

Q_W	.1	.3	.5
RMSE UKF (divs)	48.6 (2)	50.6 (4)	55.0 (16)
RMSE SE-UKF (divs)	48.6 (2)	49.7 (7)	55.0 (7)

Q_W	.7	.9	1.1
RMSE UKF (divs)	68.2 (32)	97.3 (97)	131.2 (152)
RMSE SE-UKF (divs)	62.5 (21)	73.1 (36)	81 (69)

TABLE III: Mean position errors and divergences for 1400 runs of the filter. Larger values of Q_W result in more erratic trajectories. The filter was deemed to have diverged if the position error was greater than 1km, or if the filter failed due to the appearance of a non-positive definite covariance matrix. The number of divergences is reported in parentheses

Choice of basis functions made minimal difference in this

Basis	Pos. Error	Vel. Error	Turn Error
Sine	54.6	21.4	292.6
Haar	54.1	21.4	292.5

TABLE IV: Error induced by using a Haar wavelet basis versus error from a sinusoidal basis. Median error from 300 runs of the filter. We used the most highly nonlinear setting, $Q_W = 1.1$.

experiment. We re-ran the experiment using $N = 8$ Haar wavelet functions instead of sinusoidal basis functions. Results for the most nonlinear setting $Q_W = 1.1$ are shown in Table IV. Filtering errors for both sets of basis functions were close to one another. This is because the Gaussian approximation and tuning parameters of the unscented transform have a larger effect on the filter than specifics of the series expansion approximation.

Surprisingly, we found that choosing the symmetric square root of P_t (that is, the matrix that satisfies $\mathbf{S}^2 = \mathbf{P}_t$, implemented in MATLAB as sqrtm()) instead of the Cholesky decomposition improved the accuracy of our algorithm considerably (though this choice did not improve performance of the standard UKF). The choice of matrix square root is known to affect fourth-order and higher terms in the Taylor expansion of the transition function $\mathbf{f}$ [29]. This is in agreement with our intuition: the transition function in the UKF is locally linear, and hence can be approximated with a low-order Taylor series. On the other hand, the series expansion filter uses a more nonlinear transition function and one must consider higher order terms.

C. Series expansion step size

In the prediction step of the standard unscented filter, one discretises the process $\mathbf{X}$, and iteratively applies the unscented transform at each timestep. The aim is to estimate the mean and covariance of $\mathbf{X}_t$ at some time t, given an appropriate initial condition. Repeated applications of the unscented transform at each timestep induce error in this estimate. We will refer to error of this nature as 'projection error'.

On the other hand, the error in the SE-UKF comes from the error induced by the series expansion approximation, coupled with the error induced by a single application of the unscented transform. Empirically, we observe that the accuracy of the series expansion approach improves with the number N of basis functions that we use, and deteriorates with the time T between observations. We will refer to error induced by the series expansion as 'approximation error'.

In one sense, these two approaches represent two extremes of a more general framework. For example, we might use the series expansion approximation to estimate the mean and variance of $\mathbf{X}_{T/2}$. We could then form a Gaussian approximation of its distribution, and use this as the initial condition (starting at time $T/2$) for a second application of the series expansion trick to estimate the mean and covariance of $\mathbf{X}_T$. In effect, we reduce the approximation error at the cost of increasing the projection error.

In order to investigate the effect of trading approximation error for projection error, we ran the filtering experiment of
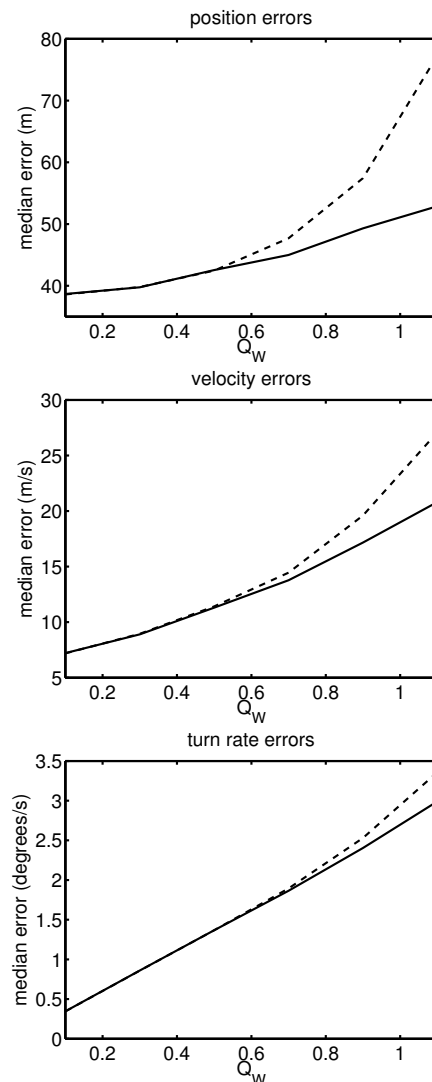

Fig. 2: The x-axis shows the diffusion coefficient Q_W of the Brownian motion driving $\mathbf{X}_7(t)$. We use this as a measure of the nonlinearity of the system. For a range of values of Q_W, we simulated 1400 trajectories of the signal, observed with noise. We plot median values of the error for the unscented Kalman filter (dotted line) and series expansion filter (solid line).

Section IV-B using the most nonlinear setting, $Q_W = 1.1$. Recall that the time interval between observations was $T = 8$ seconds. We divided this interval into K subintervals of length T/K. At the end of each subinterval, we re-initialised the series expansion approximation, using as initial condition the mean and variance computed at the previous sub-interval.

Table V shows that one can reduce the error slightly by repeatedly employing the series expansion approximation over a shorter timescale, thus trading approximation error for projection error. As the number K of projections becomes large, the error grows to match that of the standard UKF.

V. DISCUSSION AND CONCLUSIONS

In this paper, we have presented a Gaussian filter based on the series expansion approximation. The novel contributions
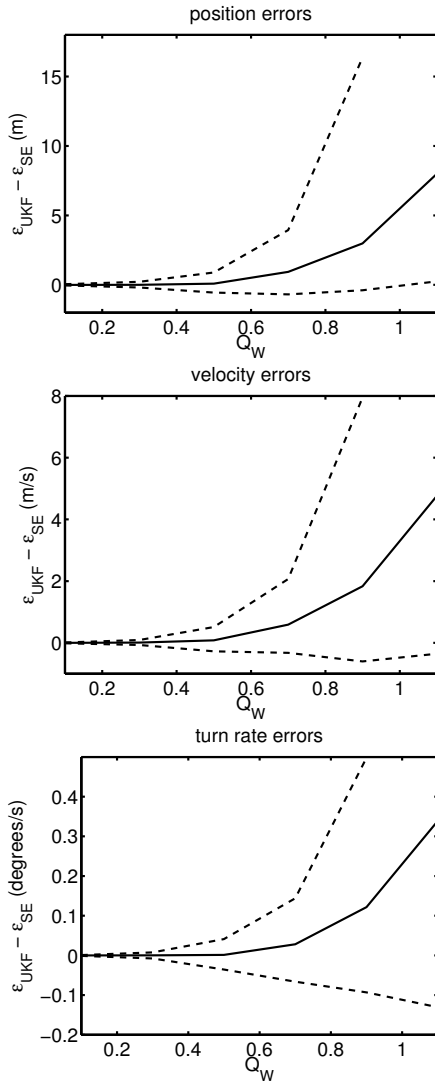
Fig. 3: The x-axis shows the diffusion coefficient Q_W of the Brownian motion driving $\mathbf{X}_7(t)$. We use this as a measure of the nonlinearity of the system. For a range of values of Q_W, we simulated 1400 trajectories of the signal, observed with noise. We plot median values of the difference in error between the unscented Kalman filter and series expansion Kalman filter (solid line), together with the first and third quartiles (dashed lines). Errors were computed seperately for position, velocity and turn rate of the aircraft. The last point in the upper range is omitted because its inclusion would skew the scaling in the image. Values for these points can be found in Section IV-B.

K	Pos. Error	Vel. Error	Turn Error
1	54.4	21.0	296
2	52.3	20.5	291
4	52.7	20.8	293
8	54.6	22.0	299
16	61.2	23.0	307
32	72.3	24.9	324

TABLE V: The effect of trading approximation error for projection error. Median errors over 300 runs of the filter. Rows are indexed by number K of projections per observation. We used the most challenging parametrisation $Q_W = 1.1$ to generate the data. Observe that results for $K = 32$ correspond closely to the errors for the standard UKF in Figure 2.

of this paper focus on improving the predictive distribution, so it is straightforward to construct a smoother using similar methods: for example one can use the unscented smoother [6] or Gaussian smoother [33] directly.

Two questions follow naturally from this work. Firstly, how does one choose parameters for the unscented transform in a sensible way? Secondly, what basis functions should one use in the series expansion? In most cases the optimal solution for either question is likely to be very difficult to compute.

All filters based around the unscented transform must some-

how deal with the first issue. Various heuristics can be found in the literature on how one might choose the tuning parameters: see, for example [34], [35]. In some cases, a poor choice of tuning parameters can cause the covariance matrix in the prediction step to fail to be positive definite. This causes the filter to diverge.

When using a common set of tuning parameters ($\alpha = 1, \kappa = 3-n, \beta = 2$, where n is the dimensionality of the system [36]), we found the matrix degeneracy problem to occur in both the series expansion filter (about 1% of runs) and the standard unscented filter (about 10% of runs). This is a known issue when using these settings in a high-dimensional context [36]. We found that increasing κ slightly to $\kappa = 5-n$ in the series expansion filter removed the divergence issue without affecting performance. On the other hand, the cubature Kalman filter settings ($\alpha = 1, \kappa = 0, \beta = 0$) performed poorly for the series expansion filter. This is because the higher dimensionality of the system causes the sigma points to be spread far out from the mode (that is, $\lambda = \alpha^2(n + \kappa) - n$ is large).

We also compared our algorithm to the third-order Gauss-Hermite Kalman filter (GHKF). This algorithm also exhibited numerical instability, with the predicted covariance matrix often failing to be positive definite. When we discarded test runs on which the GHKF diverged, we found that our algorithm performed comparably to the GHKF. This is despite the fact that the cost of the GHKF scales exponentially with dimension. In the present setting, the GHKF used $3^7 = 2187$ sigma points, and required several days of computation time to perform a comparison for a single value of Q_W.

We now address the issue of the choice of orthonormal basis. We performed the same filtering experiments using a sinusoidal basis, and a basis of Haar wavelets. Results were similar in both cases. Our explanation for this is that we already induce significant error by assuming the filtering distribution is Gaussian. This error is significantly larger than the error induced by the series expansion approximation, so the latter error is difficult to detect.

In any case, we recommend choosing a basis that converges uniformly to Brownian motion, so that one has concrete theoretical guarantees of convergence without the need to invoke rough path theory.

For completeness, we outline one more possible strategy for choosing basis functions for the series expansion. When the

SDE is linear, it is possible to construct a set of basis functions such that the series expansion approximation is error-free. In the univariate case, when b is constant the solution to (26) is given by

$$\hat{X}_T = X_0 \exp(aT) + \sum_{i=1}^{N} Z_i \int_0^T \exp(a(t-u))b\phi_i(u)du. \quad (35)$$

If we set $\phi_1(u) = \exp(-au)/\|\exp(-a\cdot)\|$, then the integral on the right disappears for all $i > 1$. This is because all other basis functions are orthogonal to $\exp(-a\cdot)$ by construction. Thus the approximation is exact at time T (though not necessarily at earlier times $t < T$). The multivariate case is similar, though more involved.

To choose a set of basis functions in a non-linear setting, one can first construct a linear approximation to the non-linear problem. One then computes the optimal basis functions for the linearized dynamics as above. However, in out numerical tests we found that these basis functions were prone to numerical instability, and furthermore they do not come with a guarantee of uniform convergence. We note that this may be a useful strategy in filtering problems that are 'almost' linear.

APPENDIX
CONVERGENCE OF THE APPROXIMATION

We now discuss asymptotic convegence of the approximation (26) to the solution of the true SDE. Wong and Zakai [24] showed that in the univariate case, under mild technical assumptions, the solution of (26) converges to the Stratonovich solution of the SDE that is being approximated.

We use the circle notation to denote Stratonovich integration. Recall that a Stratonovich SDE

$$d\mathbf{X}_t = \mathbf{a}(\mathbf{X}_t)dt + \mathbf{b}(\mathbf{X}_t) \circ d\mathbf{W}_t \quad (36)$$

can be converted to an Itô SDE and vice versa using the relationship

$$\int_0^t \mathbf{b}(\mathbf{X}_t) \circ d\mathbf{W}_t = \int_0^t \mathbf{b}(\mathbf{X}_t)d\mathbf{W}_t + \int_0^t \mathbf{c}(\mathbf{X}_t)dt, \quad (37)$$

where the integral on the left is in the Stratonovich sense, and the i-th component of the vector $\mathbf{c}$ satisfies

$$\mathbf{c}^i(x) = -\frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{d} \mathbf{b}^{j,k}(x) \frac{\partial \mathbf{b}^{i,k}}{\partial x_j}(x). \quad (38)$$

In other words, the Stratonovich solution of an SDE is equivalent to the Itô solution with a modified drift.

The issue of convergence in the multidimensional setting is somewhat more involved than in the univariate case. In general, if $\{\mathbf{W}_n\}$ is a sequence of piecewise smooth processes converging to a Brownian motion, one cannot guarantee $\{\mathbf{W}_n\} \to \mathbf{W}$ implies that the sequence of approximate differential equations converges to the Stratonovich solution of the SDE. One must impose some extra conditions on the so-called 'Levy area' of the Brownian approximations. Let $W_{n,u}^j$ be the j-th component of $\mathbf{W}_n$ at time u, and let W_u^j be the j-th component of $\mathbf{W}$ at time u. Define a set of processes

$$S_{n,t}^{ij} = \int_0^t \left(W_u^j - W_{n,u}^j \right) dW_{n,u}^i - \delta_{ij}t. \quad (39)$$

Many results about the convergence issue are known in the mathematical literature. For example, suppose the following conditions hold with probability 1 for all κ less than some positive number γ:

$$\sup_{u \leq T} \|\mathbf{W}_u - \mathbf{W}_{n,u}\| = O(n^{-\kappa}), \quad (40)$$

$$\sup_{u \leq T} \|S_{n,u}^{ij}\| = O(n^{-\kappa}), \quad (41)$$

$$\int_0^T \left\| \frac{d}{du} S_{n,u}^{ij} \right\| du = O(\log^\delta(n)) \qquad \forall \delta > 0. \quad (42)$$

The thesis of Schmatkov [37] showed that under these assumptions,

$$\sup_{u \leq T} \|\mathbf{X}_u - \hat{\mathbf{X}}_u\| = O(n^{-\kappa}). \quad (43)$$

See [38] for an analogous result about stochastic partial differential equations.

In general, there is no guarantee that the sequence of partial sums in (24) converges uniformly (so (40) is not necessarily satisfied). If one chooses the Haar wavelets as an orthonormal basis in which to expand the driving Brownian motion, then convergence is indeed uniform: in fact this choice corresponds to the Lévy-Ciesielski construction of Brownian motion [39]. For a general choice of basis functions, one can show that $\hat{\mathbf{X}} \to \mathbf{X}$ provided that the processes are interpreted as *rough paths* [32].

ACKNOWLEDGMENT

REFERENCES

[1] A. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Pr, 1970, vol. 63.

[2] B. Oksendal, *Stochastic Differential Equations*. Springer, 2007.

[3] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*. Springer, 1991.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, pp. 35–45, March 1960.

[5] A. Baine and D. Crisan, *Fundamentals of Stochastic Filtering*. Springer, 2009.

[6] S. Särkkä and T. Sottinen, "Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems," *Bayesian Analysis*, vol. 3, no. 3, pp. 555–584, 2008.

[7] P. Fearnhead, O. Papaspiliopoulos, and G. Roberts, "Particle filters for partially observed diffusions," *Journal of the Royal Statistical Society: Series B*, vol. 70, pp. 755–777, 2008. [Online]. Available: http://eprints.lancs.ac.uk/10281/

[8] L. Murray and A. Storkey, "Particle smoothing in continuous time: A fast approach via density estimation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 1017–1026, 2011.

[9] J. H. Kotecha and P. M. Djuric, "Gaussian particle filtering," *Signal Processing, IEEE Transactions on*, vol. 51, no. 10, pp. 2592–2601, 2003.

[10] K. J. Friston, "Variational filtering," *NeuroImage*, vol. 41, no. 3, pp. 747–766, 2008.

[11] F. Daum and J. Huang, "Particle flow for nonlinear filters with log-homotopy," in *SPIE Defense and Security Symposium*. International Society for Optics and Photonics, 2008, pp. 696 918–696 918.

[12] B. Balaji, "Continuous-discrete path integral filtering," *Entropy*, vol. 11, no. 3, pp. 402–430, 2009.

[13] H. Kushner, "Approximations to optimal nonlinear filters," *Automatic Control, IEEE Transactions on*, vol. 12, no. 5, pp. 546–556, 1967.

[14] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *Automatic Control, IEEE Transactions on*, vol. 45, no. 5, pp. 910–927, 2000.

[15] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *American Control Conference, 1995. Proceedings of the*, vol. 3. IEEE, 1995, pp. 1628–1632.

[16] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *Automatic Control, IEEE Transactions on*, vol. 45, no. 3, pp. 477–482, 2000.

[17] S. Särkkä, "On unscented Kalman filtering for state estimation of continuous-time nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 52, pp. 1631–1641, 2007.

[18] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *Automatic Control, IEEE Transactions on*, vol. 54, no. 6, pp. 1254–1269, 2009.

[19] I. Arasaratnam, S. Haykin, and T. Hurd, "Cubature Kalman filtering for continuous-discrete systems: Theory and simulations," *Signal Processing, IEEE Transactions on*, vol. 58, no. 10, pp. 4977–4993, 2010.

[20] S. Särkkä and J. Sarmavuori, "Gaussian filtering and smoothing for continuous-discrete dynamic systems," *Signal Processing*, vol. 93, no. 2, pp. 500 – 510, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168412003210

[21] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*. Springer, 1999.

[22] S. Särkkä and A. Solin, "On continuous-discrete cubature Kalman filtering," *Proc. SYSID 2012*, pp. 1210–1215, to Appear.

[23] D. Stroock, *Probability theory: an analytic view*. Cambridge University Press, 2010.

[24] E. Wong and M. Zakai, "On the convergence of ordinary integrals to stochastic integrals," *The Annals of Mathematical Statistics*, pp. 1560–1564, 1965.

[25] S. Corlay and P. Gilles, "Functional quantization based stratified sampling methods," *Arxiv preprint Arxiv:1008.4441*, 2010.

[26] S. Corlay, "Partial functional quantization and generalized bridges," *Arxiv preprint arXiv:1101.5488*, 2011.

[27] S. Lyons, A. Storkey, and S. Särkkä, "The coloured noise approximation and parameter estimation of stochastic differential equations," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[28] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor, "Variational inference for diffusion processes," in *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2008.

[29] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," in *Proceedings of the IEEE*, 2004, pp. 401–422.

[30] M. Grewal and A. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley-IEEE press, 2011.

[31] W. Luo, "Wiener chaos expansion and numerical solutions of stochastic partial differential equations," Ph.D. dissertation, California Institute of Technology, 2006.

[32] P. Friz and N. Victoir, "Differential equations driven by Gaussian signals," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 46, no. 2. Institut Henri Poincaré, 2010, pp. 369–413.

[33] S. Särkkä and J. Hartikainen, "On gaussian optimal smoothing of nonlinear state space models," *Automatic Control, IEEE Transactions on*, vol. 55, no. 8, pp. 1938–1941, 2010.

[34] R. Turner and C. E. Rasmussen, "Model based learning of sigma points in unscented Kalman filtering," *Neurocomputing*, vol. 80, pp. 47–53, 2012.

[35] F. Gustafsson and G. Hendeby, "Some relations between extended and unscented Kalman filters," *Signal Processing, IEEE Transactions on*, vol. 60, no. 2, pp. 545–555, 2012.

[36] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*. IEEE, 2000, pp. 153–158.

[37] A. Schmatkov, "The rate of convergence of Wong-Zakai approximations for SDEs and SPDEs," Ph.D. dissertation, University of Edinburgh, 2005.

[38] I. Gyongy, A. Shmatkov *et al.*, "Rate of convergence of Wong-Zakai approximations for stochastic partial differential equations," *Applied Mathematics and Optimization*, vol. 54, no. 3, pp. 341–341, 2006.

[39] H. McKean, *Stochastic Integrals*. American Mathematical Society, 1969, vol. 353.