Tronarp, Filip; Kersting, Hans; Särkkä, Simo; Hennig, Philipp

Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering

Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective

Filip Tronarp[1] · Hans Kersting[2] · Simo Särkkä[1] · Philipp Hennig[2]

Abstract

We formulate probabilistic numerical approximations to solutions of ordinary differential equations (ODEs) as problems in Gaussian process (GP) regression with nonlinear measurement functions. This is achieved by defining the measurement sequence to consist of the observations of the difference between the derivative of the GP and the vector field evaluated at the GP—which are all identically zero at the solution of the ODE. When the GP has a state-space representation, the problem can be reduced to a nonlinear Bayesian filtering problem and all widely used approximations to the Bayesian filtering and smoothing problems become applicable. Furthermore, all previous GP-based ODE solvers that are formulated in terms of generating synthetic measurements of the gradient field come out as specific approximations. Based on the nonlinear Bayesian filtering problem posed in this paper, we develop novel Gaussian solvers for which we establish favourable stability properties. Additionally, non-Gaussian approximations to the filtering problem are derived by the particle filter approach. The resulting solvers are compared with other probabilistic solvers in illustrative experiments.

Keywords Probabilistic numerics · Initial value problems · Nonlinear Bayesian filtering

1 Introduction

We consider an initial value problem (IVP), that is, an ordinary differential equation (ODE)

$$\dot{y}(t) = f\left(y(t), t\right), \; \forall t \in [0, T], \quad y(0) = y_0 \in \mathbb{R}^d, \qquad (1)$$

with initial value y_0 and vector field $f : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$. Numerical solvers for IVPs approximate $y : [0, T] \to \mathbb{R}^d$ and are of paramount importance in almost all areas of science and engineering. Extensive knowledge about this topic

has been accumulated in numerical analysis literature, for example, in Hairer et al. (1987), Deuflhard and Bornemann (2002) and Butcher (2008). However, until recently, a probabilistic quantification of the inevitable uncertainty—for all but the most trivial ODEs—from the numerical error over their outputs has been omitted.

Moreover, ODEs are often part of a pipeline surrounded by preceding and subsequent computations, which are themselves corrupted by uncertainty from model misspecification, measurement noise, approximate inference or, again, numerical inaccuracy (Kennedy and O'Hagan 2002). In particular, ODEs are often integrated using estimates of its parameters rather than the correct ones. See Zhang et al. (2018) and Chen et al. (2018) for recent examples of such computational chains involving ODEs. The field of *probabilistic numerics* (PN) (Hennig et al. 2015) seeks to overcome this ignorance of numerical uncertainty and the resulting overconfidence by providing *probabilistic numerical methods*. These solvers quantify numerical errors probabilistically and add them to uncertainty from other sources. Thereby, they can take decisions in a more uncertainty-aware and uncertainty-robust manner (Paul et al. 2018).

In the case of ODEs, one family of probabilistic solvers (Skilling 1992; Hennig and Hauberg 2014; Schober et al.

✉ Filip Tronarp
filip.tronarp@aalto.fi

Hans Kersting
hkersting@tuebingen.mpg.de

Simo Särkkä
simo.sarkka@aalto.fi

Philipp Hennig
ph@tue.mpg.de

[1] Department of Electrical Engineering and Automation, Aalto University, 02150 Espoo, Finland

[2] University of Tübingen and Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

2014) first treated IVPs as Gaussian process (GP) regression (Rasmussen and Williams 2006, Chapter 2). Then, Kersting and Hennig (2016) and Schober et al. (2019) sped up these methods by regarding them as stochastic filtering problems (Øksendal 2003). These completely deterministic filtering methods converge to the true solution with high polynomial rates (Kersting et al. 2018). In their methods, data for the 'Bayesian update' is constructed by evaluating the vector field f under the GP predictive mean of $y(t)$ and linked to the model with a Gaussian likelihood (Schober et al. 2019, Section 2.3). See also Wang et al. (2018, Section 1.2) for alternative likelihood models. This conception of data implies that it is the output of the adopted inference procedure. More specifically, one can show that with everything else being equal, two different priors may end up operating on different measurement sequences. Such a coupling between prior and measurements is not standard in statistical problem formulations, as acknowledged in Schober et al. (2019, Section 2.2). It makes the model and the subsequent inference difficult to interpret. For example, it is not clear how to do Bayesian model comparisons (Cockayne et al. 2019, Section 2.4) when two different priors necessarily operate on two different data sets for the same inference task.

Instead of formulating the solution of Eq. (1) as a Bayesian GP regression problem, another line of work on probabilistic solvers for ODEs comprising the methods from Chkrebtii et al. (2016), Conrad et al. (2017), Teymur et al. (2016), Lie et al. (2019), Abdulle and Garegnani (2018) and Teymur et al. (2018) aims to represent the uncertainty arising from the discretisation error by a set of samples. While multiplying the computational cost of classical solvers with the amount of samples, these methods can capture arbitrary (non-Gaussian) distributions over the solutions and can reduce overconfidence in inverse problems for ODEs—as demonstrated in Conrad et al. (2017, Section 3.2.), Abdulle and Garegnani (2018, Section 7) and Teymur et al. (2018). These solvers can be considered as more expensive, but statistically more expressive. This paper contributes a particle filter as a sampling-based filtering method at the intersection of both lines of work, providing a previously missing link.

The contributions of this paper are the following: Firstly, we circumvent the issue of generating synthetic data, by recasting solutions of ODEs in terms of nonlinear Bayesian filtering problems in a well defined state-space model. For any fixed-time discretisation, the measurement sequence and likelihood are also fixed. That is, we avoid the coupling of prior and measurement sequence, that is for example present in Schober et al. (2019). This enables application of all Bayesian filtering and smoothing techniques to ODEs as described, for example, in Särkkä (2013). Secondly, we show how the application of certain inference techniques recovers the previous filtering-based methods. Thirdly, we

discuss novel algorithms giving rise to both Gaussian and non-Gaussian solvers.

Fourthly, we establish a stability result for the novel Gaussian solvers. Fifthly, we discuss practical methods for uncertainty calibration, and in the case of Gaussian solvers, we give explicit expressions. Finally, we present some illustrative experiments demonstrating that these methods are practically useful both for fast inference of the unique solution of an ODE as well as for representing multi-modal distributions of trajectories.

2 Bayesian inference for initial value problems

Formulating an approximation of the solution to Eq. (1) at a discrete set of points $\{t_n\}_{n=0}^N$ as a problem of Bayesian inference requires, as always, three things: a prior measure, data, and a likelihood, which define a posterior measure through Bayes' rule.

We start with examining a continuous-time formulation in Sect. 2.1, where Bayesian conditioning should, in the ideal case, give a Dirac measure at the true solution of Eq. (1) as the posterior. This has two issues: (1) conditioning on the entire gradient field is not feasible on a computer in finite time and (2) the conditioning operation itself is intractable. Issue (1) is present in classical Bayesian quadrature (Briol et al. 2019) as well. Limited computational resources imply that only a finite number of evaluations of the integrand can be used. Issue (2) turns, what is linear GP regression in Bayesian quadrature, into nonlinear GP regression. While this is unfortunate, it appears reasonable that something should be lost as the inference problem is more complex.

With this in mind, a discrete-time nonlinear Bayesian filtering problem is posed in Sect. 2.2, which targets the solution of Eq. (1) at a discrete set of points.

2.1 A continuous-time model

Like previous works mentioned in Sect. 1, we consider priors given by a GP

$$X(t) \sim \mathrm{GP}\left(\bar{x}, k\right),$$

where $\bar{x}(t)$ is the mean function and $k(t, t')$ is the covariance function. The vector $X(t)$ is given by

$$X(t) = \left[\left(X^{(1)}(t)\right)^\mathsf{T}, \dots, \left(X^{(q+1)}(t)\right)^\mathsf{T}\right]^\mathsf{T}, \tag{2}$$

where $X^{(1)}(t)$ and $X^{(2)}(t)$ model $y(t)$ and $\dot{y}(t)$, respectively. The remaining $q - 1$ sub-vectors in $X(t)$ can be used to model higher-order derivatives of $y(t)$ as done by Schober

et al. (2019) and Kersting and Hennig (2016). We define such priors by a stochastic differential equation (Øksendal 2003), that is,

$$X(0) \sim \mathcal{N}\big(\mu^-(0), \Sigma^-(0)\big), \tag{3a}$$

$$\mathrm{d}X(t) = \big[FX(t) + u\big]\mathrm{d}t + L\mathrm{d}B(t), \tag{3b}$$

where F is a state transition matrix, u is a forcing term, L is a diffusion matrix, and $B(t)$ is a vector of standard Wiener processes.

Note that for $X^{(2)}(t)$ to be the derivative of $X^{(1)}$, F, u, and L are such that

$$\mathrm{d}X^{(1)}(t) = X^{(2)}(t)\mathrm{d}t. \tag{4}$$

The use of an SDE—instead of a generic GP prior—is computationally advantageous because it restricts the priors to Markov processes due to Øsendal (2003, Theorem 7.1.2). This allows for inference with linear time-complexity in N, while the time-complexity is N^3 for GP priors in general (Hartikainen and Särkkä 2010).

Inference requires data and an associated likelihood. Previous authors, such as Schober et al. (2019) and Chkrebtii et al. (2016), put forth the view of the prior measure defining an *inference agent*, which cycles through extrapolating, generating measurements of the vector field, and updating. Here we argue that there is no need for generating measurements, since re-writing Eq. (1) yields the requirement

$$\dot{y}(t) - f(y(t), t) = 0. \tag{5}$$

This suggests that a measurement relating the prior defined by Eq. (3) to the solution of Eq. (1) ought to be defined as

$$Z(t) = X^{(2)}(t) - f(X^{(1)}(t), t). \tag{6}$$

While conditioning the process $X(t)$ on the event $Z(t) = 0$ for all $t \in [0, T]$ can be formalised using the concept of *disintegration* (Cockayne et al. 2019), it is intractable in general and thus impractical for computer implementation. Therefore, we formulate a discrete-time inference problem in the sequel.

2.2 A discrete-time model

In order to make the inference problem tractable, we only attempt to condition the process $X(t)$ on $Z(t) = z(t) \triangleq 0$ at a set of discrete time points, $\{t_n\}_{n=0}^N$. We consider a uniform grid, $t_{n+1} = t_n + h$, though extending the present methods to non-uniform grids can be done as described in Schober et al. (2019). In the sequel, we will denote a function evaluated at t_n by subscript n, for example $z_n = z(t_n)$. From Eq. (3), an equivalent discrete-time system can be obtained (Grewal

and Andrews 2001, Chapter 3.7.3).[1] The inference problem becomes

$$X_0 \sim \mathcal{N}(\mu_0^F, \Sigma_0^F), \tag{7a}$$

$$X_{n+1} \mid X_n \sim \mathcal{N}\big(A(h)X_n + \xi(h), Q(h)\big), \tag{7b}$$

$$Z_n \mid X_n \sim \mathcal{N}\big(\dot{C}X_n - f(CX_n, t_n), R\big), \tag{7c}$$

$$z_n \triangleq 0, \quad n = 1, \dots, N, \tag{7d}$$

where z_n is the realisation of Z_n. The parameters $A(h), \xi(h)$, and $Q(h)$ are given by

$$A(h) = \exp(Fh), \tag{8a}$$

$$\xi(h) = \int_0^h \exp(F(h - \tau))u\mathrm{d}\tau, \tag{8b}$$

$$Q(h) = \int_0^h \exp(F(h - \tau))LL^\mathsf{T} \exp(F^\mathsf{T}(h - \tau))\mathrm{d}\tau. \tag{8c}$$

Furthermore, $C = [\mathrm{I}\ 0\ \dots\ 0]$ and $\dot{C} = [0\ \mathrm{I}\ 0\ \dots\ 0]$. That is, $CX_n = X_n^{(1)}$ and $\dot{C}X_n = X_n^{(2)}$. A measurement variance, R, has been added to $Z(t_n)$ for greater generality, which simplifies the construction of particle filter algorithms. The likelihood model in Eq. (7c) has previously been used in the gradient matching approach to inverse problems to avoid explicit numerical integration of the ODE (see, e.g. Calderhead et al. 2009).

The inference problem posed in Eq. (7) is a standard problem in nonlinear GP regression (Rasmussen and Williams 2006), also known as Bayesian filtering and smoothing in stochastic signal processing (Särkkä 2013). Furthermore, it reduces to Bayesian quadrature when the vector field does not depend on y. This is Proposition 1.

Proposition 1 *Let $X_0^{(1)} = 0$, $f(y(t), t) = g(t)$, $y(0) = 0$, and $R = 0$. Then the posteriors of $\{X_n^{(1)}\}_{n=1}^N$ are Bayesian quadrature approximations for*

$$\int_0^{nh} g(\tau)\mathrm{d}\tau, \quad n = 1, \dots, N. \tag{9}$$

A proof of Proposition 1 is given in "Appendix A".

Remark 1 The Bayesian quadrature method described in Proposition 1 conditions on function evaluations outside the domain of integration for $n < N$. This corresponds to the smoothing equations associated with Eq. (7). If the integral on the domain $[0, nh]$ is only conditioned on evaluations of g inside the domain, then the filtering estimates associated with Eq. (7) are obtained.

---

[1] Here 'equivalent' is used in the sense that the probability distribution of the continuous-time process evaluated on the grid coincides with the probability distribution of the discrete-time process (Särkkä 2006, P. 17).

2.3 Gaussian filtering

The inference problem posed in Eq. (7) is a standard problem in statistical signal processing and machine learning, and the solution is often approximated by Gaussian filters and smoothers (Särkkä 2013). Let us define $z_{1:n} = \{z_l\}_{l=1}^n$ and the following conditional moments

$$\mu_n^F \triangleq \mathbb{E}[X_n \mid z_{1:n}], \tag{10a}$$

$$\Sigma_n^F \triangleq \mathbb{V}[X_n \mid z_{1:n}], \tag{10b}$$

$$\mu_n^P \triangleq \mathbb{E}[X_n \mid z_{1:n-1}], \tag{10c}$$

$$\Sigma_n^P \triangleq \mathbb{V}[X_n \mid z_{1:n-1}], \tag{10d}$$

where $\mathbb{E}[\cdot \mid z_{1:n}]$ and $\mathbb{V}[\cdot \mid z_{1:n}]$ are the conditional mean and covariance operators given the measurements $Z_{1:n} = z_{1:n}$. Additionally, $\mathbb{E}[\cdot \mid z_{1:0}] = \mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot \mid z_{1:0}] = \mathbb{V}[\cdot]$ by convention. Furthermore, μ_n^F and Σ_n^F are referred to as the filtering mean and covariance, respectively. Similarly, μ_n^P and Σ_n^P are referred to as the predictive mean and covariance, respectively. In Gaussian filtering, the following relationships hold between μ_n^F and Σ_n^F, and μ_{n+1}^P and Σ_{n+1}^P:

$$\mu_{n+1}^P = A(h)\mu_n^F + \xi(h), \tag{11a}$$

$$\Sigma_{n+1}^P = A(h)\Sigma_n^F A^\mathsf{T}(h) + Q(h), \tag{11b}$$

which are the prediction equations (Särkkä 2013, Eq. 6.6). The update equations, relating the predictive moments μ_n^P and Σ_n^P with the filter estimate, μ_n^F, and its covariance Σ_n^F, are given by (Särkkä 2013, Eq. 6.7)

$$S_n = \mathbb{V}\big[\dot{C}X_n - f(CX_n, t_n) \mid z_{1:n-1}\big] + R, \tag{12a}$$

$$K_n = \mathbb{C}\big[X_n, \dot{C}X_n - f(CX_n, t_n) \mid z_{1:n-1}\big]S_n^{-1}, \tag{12b}$$

$$\hat{z}_n = \mathbb{E}\big[\dot{C}X_n - f(CX_n, t_n) \mid z_{1:n-1}\big], \tag{12c}$$

$$\mu_n^F = \mu_n^P + K_n(z_n - \hat{z}_n), \tag{12d}$$

$$\Sigma_n^F = \Sigma_n^P - K_n S_n K_n^\mathsf{T}, \tag{12e}$$

where the expectation ($\mathbb{E}$), covariance ($\mathbb{V}$) and cross-covariance ($\mathbb{C}$) operators are with respect to $X_n \sim \mathcal{N}(\mu_n^P, \Sigma_n^P)$. Evaluating these moments is intractable in general, though various approximation schemes exist in literature. Some standard approximation methods shall be examined below. In particular, the methods of Schober et al. (2019) and Kersting and Hennig (2016) come out as particular approximations to Eq. (12).

2.4 Taylor series methods

A classical method in filtering literature to deal with non-linear measurements of the form in Eq. (7) is to make a first-order Taylor series expansion, thus turning the problem

into a standard update in linear filtering. However, before going through the details of this it is instructive to interpret the method of Schober et al. (2019) as an even simpler Taylor series method. This is Proposition 2.

Proposition 2 *Let $R = 0$ and approximate $f(CX_n, t_n)$ by its zeroth order Taylor expansion in X_n around the point μ_n^P*

$$f(CX_n, t_n) \approx f(C\mu_n^P, t_n). \tag{13}$$

Then, the approximate posterior moments are given by

$$S_n \approx \dot{C}\Sigma_n^P \dot{C}^\mathsf{T} + R, \tag{14a}$$

$$K_n \approx \Sigma_n^P \dot{C}^\mathsf{T} S_n^{-1}, \tag{14b}$$

$$\hat{z}_n \approx \dot{C}\mu_n^P - f(C\mu_n^P, t_n), \tag{14c}$$

$$\mu_n^F \approx \mu_n^P + K_n(z_n - \hat{z}_n), \tag{14d}$$

$$\Sigma_n^F \approx \Sigma_n^P - K_n S_n K_n^\mathsf{T}, \tag{14e}$$

which is precisely the update by Schober et al. (2019).

A first-order approximation The approximation in Eq. (14) can be refined by using a first-order approximation, which is known as the extended Kalman filter (EKF) in signal processing literature (Särkkä 2013, Algorithm 5.4). That is,

$$\begin{aligned} f(CX_n, t_n) &\approx f(C\mu_n^P, t_n) \\ &\quad + J_f(C\mu_n^P, t_n)C(X_n - \mu_n^P), \end{aligned} \tag{15}$$

where J_f is the Jacobian of $y \to f(y, t)$. The filter update is then

$$\tilde{C}_n = \dot{C} - J_f(C\mu_n^P, t_n)C, \tag{16a}$$

$$S_n \approx \tilde{C}_n \Sigma_n^P \tilde{C}_n^\mathsf{T} + R, \tag{16b}$$

$$K_n \approx \Sigma_n^P \tilde{C}_n^\mathsf{T} S_n^{-1}, \tag{16c}$$

$$\hat{z}_n \approx \dot{C}\mu_n^P - f(C\mu_n^P, t_n), \tag{16d}$$

$$\mu_n^F \approx \mu_n^P + K_n(z_n - \hat{z}_n), \tag{16e}$$

$$\Sigma_n^F \approx \Sigma_n^P - K_n S_n K_n^\mathsf{T}. \tag{16f}$$

Hence the extended Kalman filter computes the residual, $z_n - \hat{z}_n$, in the same manner as Schober et al. (2019). However, as the filter gain, K_n, now depends on evaluations of the Jacobian, the resulting probabilistic ODE solver is different in general.

While Jacobians of the vector field are seldom exploited in ODE solvers, they play a central role in Rosenbrock methods, (Rosenbrock 1963; Hochbruck et al. 2009). The Jacobian of the vector field was also recently used by Teymur et al. (2018) for developing a probabilistic solver.

Although the extended Kalman filter goes as far back as the 1960s (Jazwinski 1970), the update in Eq. (16) results in

a probabilistic method for estimating the solution of (1) that appears to be novel. Indeed, to the best of the authors' knowledge, the only Gaussian filtering-based solvers that have appeared so far are those by Kersting and Hennig (2016), Magnani et al. (2017) and Schober et al. (2019).

2.5 Numerical quadrature

Another method to approximate the quantities in Eq. (12) is by quadrature, which consists of a set of nodes $\{\mathcal{X}_{n,j}\}_{j=1}^J$ with weights $\{w_{n,j}\}_{j=1}^J$ that are associated to the distribution $\mathcal{N}(\mu_n^P, \Sigma_n^P)$. These nodes and weights can either be constructed to integrate polynomials up to some order exactly (see, e.g. McNamee and Stenger 1967; Golub and Welsch 1969) or by Bayesian quadrature (Briol et al. 2019). In either case, the expectation of a function $\psi(X_n)$ is approximated by

$$\mathbb{E}[\psi(X_n)] \approx \sum_{j=1}^J w_{n,j} \psi(\mathcal{X}_{n,j}). \qquad (17)$$

Therefore, by appropriate choices of ψ the quantities in Eq. (12) can be approximated. We shall refer to filters using a third degree fully symmetric rule (McNamee and Stenger 1967) as unscented Kalman filters (UKF), which is the name that was adopted when it was first introduced to the signal processing community (Julier et al. 2000). For a suitable cross-covariance assumption and a particular choice of quadrature, the method of Kersting and Hennig (2016) is retrieved. This is Proposition 3.

Proposition 3 *Let $\{\mathcal{X}_{n,j}\}_{j=1}^J$ and $\{w_{n,j}\}_{j=1}^J$ be the nodes and weights, corresponding to a Bayesian quadrature rule with respect to $\mathcal{N}(\mu_n^P, \Sigma_n^P)$. Furthermore, assume $R = 0$ and that the cross-covariance between $\dot{C}X_n$ and $f(CX_n, t_n)$ is approximated as zero,*

$$\mathbb{C}[\dot{C}X_n, f(CX_n, t_n) \mid z_{1:n-1}] \approx 0. \qquad (18)$$

Then the probabilistic solver proposed in Kersting and Hennig (2016) *is a Bayesian quadrature approximation to Eq. (12).*

A proof of Proposition 3 is given in "Appendix B".

While a cross-covariance assumption of Proposition 3 reproduces the method of Kersting and Hennig (2016), Bayesian quadrature approximations have previously been used for Gaussian filtering in signal processing applications by Prüher and Šimandl (2015), which in this context gives a new solver.

2.6 Affine vector fields

It is instructive to examine the particular case when the vector field in Eq. (1) is affine. That is,

$$f(y(t), t) = \Lambda(t)y(t) + \zeta(t). \qquad (19)$$

In such a case, Eq. (7) becomes a linear Gaussian system, which is solved exactly by a Kalman filter. The equations for implementing this Kalman filter are precisely Eq. (11) and Eq. (12), although the latter set of equations can be simplified. Define $H_n = \dot{C} - \Lambda(t_n)C$, then the update equations become

$$S_n = H_n \Sigma_n^P H_n^\mathsf{T} + R, \qquad (20a)$$
$$K_n = \Sigma_n^P H_n^\mathsf{T} S_n^{-1}, \qquad (20b)$$
$$\mu_n^F = \mu_n^P + K_n(\zeta(t_n) - H_n\mu_n^P) \qquad (20c)$$
$$\Sigma_n^F = \Sigma_n^P - K_n S_n K_n^\mathsf{T}. \qquad (20d)$$

Lemma 1 *Consider the inference problem in Eq. (7) with an affine vector field as given in Eq. (19). Then the EKF reduces to the exact Kalman filter, which uses the update in Eq. (20). Furthermore, the same holds for Gaussian filters using a quadrature approximation to Eq. (12), provided that it integrates polynomials correctly up to second order with respect to the distribution $\mathcal{N}(\mu_n^P, \Sigma_n^P)$.*

Proof Since the Kalman filter, the EKF, and the quadrature approach all use Eq. (11) for prediction, it is sufficient to make sure that the EKF and the quadrature approximation compute Eq. (12) exactly, just as the Kalman filter. Now the EKF approximates the vector field by an affine function for which it computes the moments in Eq. (12) exactly. Since this affine approximation is formed by a truncated Taylor series, it is exact for affine functions and the statement pertaining to the EKF holds. Furthermore, the Gaussian integrals in Eq. (12) are polynomials of degree at most two for affine vector fields and are therefore computed exactly by the quadrature rule by assumption. □

2.7 Particle filtering

The Gaussian filtering methods from Sect. 2.3 may often suffice. However, there are cases where more sophisticated inference methods may be preferable, for instance, when the posterior becomes multi-modal due to chaotic behaviour or 'numerical bifurcations'. That is, when it is numerically unknown whether the true solution is above or below a certain threshold that determines the limit behaviour of its trajectory. While sampling-based probabilistic solvers such as those of Chkrebtii et al. (2016), Conrad et al. (2017), Teymur et al. (2016), Lie et al. (2019), Abdulle and Garegnani (2018)

and Teymur et al. (2018) can pick up such phenomena, the Gaussian filtering-based ODE solvers discussed in Sect. 2.3 cannot. However, this limitation may be overcome by approximating the filtering distribution of the inference problem in Eq. (7) with particle filters that are based on a sequential formulation of importance sampling (Doucet et al. 2001).

A particle filter operates on a set of particles, $\{X_{n,j}\}_{j=1}^{J}$, a set of positive weights $\{w_{n,j}\}_{j=1}^{J}$ associated to the particles that sum to one and an importance density, $g(x_{n+1} \mid x_n, z_n)$. The particle filter then cycles through three steps (1) propagation, (2) re-weighting, and (3) re-sampling (Särkkä 2013, Chapter 7.4).

The propagation step involves sampling particles at time $n + 1$ from the importance density:

$$X_{n+1,j} \sim g(x_{n+1} \mid X_{n,j}, z_n).\tag{21}$$

The re-weighting of the particles is done by a likelihood ratio with the product of the measurement density and the transition density of Eq. (7), and the importance density. That is, the updated weights are given by

$$\rho(x_{n+1}, x_n) = \frac{p(z_{n+1} \mid x_{n+1}) p(x_{n+1} \mid x_n)}{g(x_{n+1} \mid x_n, z_{n+1})},\tag{22a}$$

$$w_{n+1,j} \propto \rho\big(X_{n+1,j}, X_{n,j}\big) w_{n,j},\tag{22b}$$

where the proportionality sign indicates that the weights need to be normalised to sum to one after they have been updated according to Eq. (22). The weight update is then followed by an optional re-sampling step (Särkkä 2013, Chapter 7.4). While not re-sampling in principle yields a valid algorithm, it becomes necessary in order to avoid the degeneracy problem for long time series (Doucet et al. 2001, Chapter 1.3). The efficiency of particle filters depends on the choice of importance density. In terms of variance, the locally optimal importance density is given by (Doucet et al. 2001)

$$g(x_n \mid x_{n-1}, z_n) \propto p(z_n \mid x_n) p(x_n \mid x_{n-1}).\tag{23}$$

While Eq. (23) is almost as intractable as the full filtering distribution, the Gaussian filtering methods from Sect. 2.3 can be used to make a good approximation. For instance, the approximation to the optimal importance density using Eq. (14) is given by

$$S_n = \dot{C} Q(h) \dot{C}^{\mathsf{T}} + R,\tag{24a}$$

$$K_n = Q(h) \dot{C}^{\mathsf{T}} S_n^{-1},\tag{24b}$$

$$\hat{z}_n = \dot{C} A(h) x_{n-1} - f\big(C A(h) x_{n-1}, t_n\big),\tag{24c}$$

$$\mu_n = A(h) x_{n-1} + K_n(z_n - \hat{z}_n),\tag{24d}$$

$$\Sigma_n = Q(h) - K_n S_n K_n^{\mathsf{T}},\tag{24e}$$

$$g(x_n \mid x_{n-1}, z_n) = \mathcal{N}(x_n; \mu_n, \Sigma_n).\tag{24f}$$

An importance density can be similarly constructed from Eq. (16), resulting in:

$$\tilde{C}_n = \dot{C} - J_f\big(C A(h) x_{n-1}, t_n\big) C,\tag{25a}$$

$$S_n = \tilde{C}_n Q(h) \tilde{C}_n^{\mathsf{T}} + R,\tag{25b}$$

$$K_n = Q(h) \tilde{C}_n^{\mathsf{T}} S_n^{-1},\tag{25c}$$

$$\hat{z}_n = \dot{C} A(h) x_{n-1} - f\big(C A(h) x_{n-1}, t_n\big),\tag{25d}$$

$$\mu_n = A(h) x_{n-1} + K_n(z_n - \hat{z}_n),\tag{25e}$$

$$\Sigma_n = Q(h) - K_n S_n K_n^{\mathsf{T}},\tag{25f}$$

$$g(x_n \mid x_{n-1}, z_n) = \mathcal{N}(x_n; \mu_n, \Sigma_n).\tag{25g}$$

Note that we have assumed $\xi(h) = 0$ in Eqs. (24) and (25), which can be extended to $\xi(h) \neq 0$ by replacing $A(h) x_{n-1}$ with $A(h) x_{n-1} + \xi(h)$. We refer the reader to Doucet et al. (2000, Section II.D.2) for a more thorough discussion on the use of local linearisation methods to construct importance densities.

We conclude this section with a brief discussion on the convergence of particle filters. The following theorem is given by Crisan and Doucet (2002).

Theorem 1 *Let $\rho(x_{n+1}, x_n)$ in Eq. (22a) be bounded from above and denote the true filtering measure associated with Eq. (7) at time n by p_n^R, and let $\hat{p}_n^{R,J}$ be its particle approximation using J particles with importance density $g(x_{n+1} \mid x_n, z_{n+1})$. Then, for all $n \in \mathbb{N}_0$, there exists a constant c_n independent of J such that for any bounded Borel function $\phi : \mathbb{R}^{d(q+1)} \to \mathbb{R}$ the following bound holds*

$$\mathbb{E}_{\text{MC}}[(\langle \hat{p}_n^{R,J}, \phi \rangle - \langle p_n^R, \phi \rangle)^2]^{1/2} \leq c_n J^{-1/2} \|\phi\|,\tag{26}$$

where $\langle p, \phi \rangle$ denotes ϕ integrated with respect to p and $\mathbb{E}_{\text{MC}}$ denotes the expectation over realisations of the particle method, and $\|\cdot\|$ is the supremum norm.

Theorem 1 shows that we can decrease the distance (in the weak sense) between $\hat{p}_n^{R,J}$ and p_n^R by increasing J. However, the object we want to approximate is p_n^0 (the exact filtering measure associated with Eq. (7) for $R = 0$) but setting $R = 0$ makes the likelihood ratio in Eq. (22a) ill-defined for the proposal distributions in Eqs. (24) and (25). This is because, when $R = 0$, then $p(z_{n+1} \mid x_{n+1}) p(x_{n+1} \mid x_n)$ has its support on the surface $\dot{C} x_{n+1} = f(C x_{n+1}, t_{n+1})$ while Eqs. (24) or (25) imply that the variance of $\dot{C} X_{n+1}$ or $\tilde{C}_{n+1} X_{n+1}$ will be zero with respect to $g(x_{n+1} \mid x_n, z_{n+1})$, respectively. That is, $g(x_{n+1} \mid x_n, z_{n+1})$ is supported on a hyperplane. It follows that the null-sets of $g(x_{n+1} \mid x_n, z_{n+1})$ are not necessarily null-sets of $p(z_{n+1} \mid x_{n+1}) p(x_{n+1} \mid x_n)$

and the likelihood ratio in Eq. (22a) can therefore be undefined. However, a straightforward application of the triangle inequality together with Theorem 1 gives

$$
\begin{aligned}
\mathbb{E}_{\mathrm{MC}} & [(\langle \hat{p}_n^{R,J}, \phi \rangle - \langle p_n^0, \phi \rangle)^2]^{1/2} \\
& \leq \mathbb{E}_{\mathrm{MC}}[(\langle \hat{p}_n^{R,J}, \phi \rangle - \langle p_n^R, \phi \rangle)^2]^{1/2} \\
& \quad + \mathbb{E}_{\mathrm{MC}}[(\langle p_n^R, \phi \rangle - \langle p_n^0, \phi \rangle)^2]^{1/2} \\
& = \mathbb{E}_{\mathrm{MC}}[(\langle \hat{p}_n^{R,J}, \phi \rangle - \langle p_n^R, \phi \rangle)^2]^{1/2} \\
& \quad + |\langle p_n^R, \phi \rangle - \langle p_n^0, \phi \rangle| \\
& \leq c_n J^{-1/2} \|\phi\| + |\langle p_n^R, \phi \rangle - \langle p_n^0, \phi \rangle|.
\end{aligned}
\tag{27}
$$

The last term vanishes as $R \to 0$. That is, the error can be controlled by increasing the number of particles J and decreasing R. Though a word of caution is appropriate, as particle filters can become ill-behaved in practice if the likelihoods are too narrow (too small R). However, this also depends on the quality of the proposal distribution.

Lastly, while Theorem 1 is only valid if $\rho(x_{n+1}, x_n)$ is bounded, this can be ensured by either inflating the covariance of the proposal distribution or replacing the Gaussian proposal with a Student's t proposal (Cappé et al. 2005, Chapter 9).

3 A stability result for Gaussian filters

ODE solvers are often characterised by the properties of their solution to the linear test equation

$$
\dot{y}(t) = \lambda y(t), \; y(0) = 1,
\tag{28}
$$

where λ is some complex number. A numerical solver is said to be *A-stable* if the approximate solution tends to zero for any fixed step size h whenever the real part of λ resides in the left half-plane (Dahlquist 1963). Recall that if $y_0 \in \mathbb{R}^d$ and $\Lambda \in \mathbb{R}^{d \times d}$ then the ODE $\dot{y}(t) = \Lambda y(t), \; y(0) = y_0$ is said to be asymptotically stable if $\lim_{t \to \infty} y(t) = 0$, which is precisely when the real part of eigenvalues of Λ are in the left half-plane. That is, A-stability is the notion that a numerical solver preserves asymptotic stability of linear time-invariant ODEs.

While the present solvers are not designed to solve complex valued ODEs, a real system equivalent to Eq. (28) is given by

$$
\dot{y}(t) = \Lambda_{\mathrm{test}} y(t), \quad y^{\mathsf{T}}(0) = [1 \; 0],
\tag{29}
$$

where $\lambda = \lambda_1 + i\lambda_2$ and

$$
\Lambda_{\mathrm{test}} = \begin{bmatrix} \lambda_1 & -\lambda_2 \\ \lambda_2 & \lambda_1 \end{bmatrix}.
\tag{30}
$$

However, to leverage classical stability results from the theory of Kalman filtering we investigate a slightly different test equation, namely

$$
\dot{y}(t) = \Lambda y(t), \; y(0) = y_0,
\tag{31}
$$

where $\Lambda \in \mathbb{R}^{d \times d}$ is of full rank. In this case, Eqs. (11) and (20) give the following recursion for μ_n^P

$$
\mu_{n+1}^P = (A(h) - A(h)K_n H)\mu_n^P,
\tag{32a}
$$
$$
\mu_n^F = (\mathrm{I} - K_n H)\mu_n^P,
\tag{32b}
$$

where we recall that $H = \dot{C} - C\Lambda$ and $z_n = 0$. If there exists a limit gain $\lim_{n \to \infty} K_n = K_\infty$ then asymptotic stability of the filter holds provided that the eigenvalues of $(A(h) - A(h)K_\infty H)$ are strictly within the unit circle (Anderson and Moore 1979, Appendix C, p. 341). That is, $\lim_{n \to \infty} \mu_n^P = 0$ and as a direct consequence $\lim_{n \to \infty} \mu_n^F = 0$.

We shall see that the Kalman filter using an IWP(q) prior is asymptotically stable. For the IWP(q) process on $\mathbb{R}^d$ we have $u = 0$, $L = \mathrm{e}_{q+1} \otimes \Gamma^{1/2}$, and $F = (\sum_{i=1}^q \mathrm{e}_i \mathrm{e}_{i+1}^{\mathsf{T}}) \otimes \mathrm{I}$, where $\mathrm{e}_i \in \mathbb{R}^d$ is the ith canonical eigenvector, $\Gamma^{1/2}$ is the symmetric square root of some positive semi-definite matrix $\Gamma \in \mathbb{R}^{d \times d}$, $\mathrm{I} \in \mathbb{R}^{d \times d}$ is the identity matrix, and $\otimes$ is Kronecker's product. By using Eq. (8), the properties of Kronecker products and the definition of the matrix exponential the equivalent discrete-time system are given by

$$
A(h) = A^{(1)}(h) \otimes \mathrm{I},
\tag{33a}
$$
$$
\xi(h) = 0,
\tag{33b}
$$
$$
Q(h) = Q^{(1)}(h) \otimes \Gamma,
\tag{33c}
$$

where $A^{(1)}(h) \in \mathbb{R}^{(q+1) \times (q+1)}$ and $Q^{(1)}(h) \in \mathbb{R}^{(q+1) \times (q+1)}$ are given by (Kersting et al. 2018, Appendix A)[2]

$$
A_{ij}^{(1)}(h) = \mathbb{I}_{i \leq j} \frac{h^{j-i}}{(j-i)!},
\tag{34a}
$$
$$
Q_{ij}^{(1)}(h) = \frac{h^{2q+3-i-j}}{(2q+3-i-j)(q+1-i)!(q+1-j)!},
\tag{34b}
$$

and $\mathbb{I}_{i \leq j}$ is an indicator function. Before proceeding, we need to introduce the notions of stabilisability and detectability from Kalman filtering theory. These notions can be found in Anderson and Moore (1979, Appendix C).

Definition 1 (*Complete stabilisability*) The pair $[A, G]$ is completely stabilisable if $w^{\mathsf{T}} G = 0$ and $w^{\mathsf{T}} A = \eta w^{\mathsf{T}}$ for some constant η implies $|\eta| < 1$ or $w = 0$.

---

[2] Note that Kersting et al. (2018) uses indexing $i, j = 0, \ldots, q$ while we here use $i, j = 1, \ldots, q+1$.

Definition 2 (*Complete detectability*) [3] $[A, H]$ is completely detectable if $[A^\mathsf{T}, H^\mathsf{T}]$ is completely stabilisable.

Before we state the stability result of this section, the following two lemmas are useful.

Lemma 2 *Consider the discretised IWP(q) prior on $\mathbb{R}^d$ as given by Eq. (33). Let $h > 0$ and Γ be positive definite. Then, the $d \times d$ blocks of $Q(h)$, denoted by $Q_{i,j}(h)$, $i, j = 1, 2, \ldots, q + 1$ are of full rank.*

Proof From Eq. (33c), we have $Q_{i,j}(h) = Q_{i,j}^{(1)}(h)\Gamma$. From Eq. (34b) and $h > 0$, we have $Q_{i,j}^{(1)}(h) > 0$, and since Γ is positive definite it is of full rank. It then follows that $Q_{i,j}(h)$ is of full rank as well.   □

Lemma 3 *Let $A(h)$ be the transition matrix of an IWP(q) prior as given by Eq. (33a) and $h > 0$, then $A(h)$ has a single eigenvalue given by $\eta = 1$. Furthermore, the right-eigenspace is given by*

$$\mathrm{span}[\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_d],$$

where $\mathrm{e}_i \in \mathbb{R}^{(q+1)d}$ are canonical basis vectors, and the left-eigenspace is given by

$$\mathrm{span}[\mathrm{e}_{qd+1}, \mathrm{e}_{qd+2}, \ldots, \mathrm{e}_{(q+1)d}].$$

Proof Firstly, from Eqs. (33a) and (34a) it follows that $A(h)$ is block upper-triangular with identity matrices on the block diagonal, hence the characteristic equation is given by

$$\det(A(h) - \eta\mathrm{I}) = (1 - \eta)^{(q+1)d} = 0, \tag{35}$$

we conclude that the only eigenvalue is $\eta = 1$. To find the right-eigenspace let $w^\mathsf{T} = [w_1^\mathsf{T}, w_2^\mathsf{T}, \ldots, w_{q+1}^\mathsf{T}]$, $w_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, q + 1$ and solve $A(h)w = w$, which by using Eqs. (33a) and (34a) can be written as

$$(A(h)w)_l = \sum_{r=0}^{q+1-l} \frac{h^r}{r!} w_{r+l}, \quad l = 1, 2, \ldots, q + 1, \tag{36}$$

where $(\cdot)_l$ is the lth sub-vector of dimension d. Starting with $l = q+1$, we trivially have $w_{q+1} = w_{q+1}$. For $l = q$ we have $w_q + w_{q+1}h = w_q$ but $h > 0$, hence $w_{q+1} = 0$. Similarly for $l = q - 1$ we have $w_{q-1} = w_{q-1} + w_q h + w_{q+1}h^2/2 = w_{q-1} + w_q h + 0 \cdot h^2/2$. Again since $h > 0$ we have $w_q = 0$. By repeating this argument we have $w_1 = w_1$ and $w_i = 0$, $i = 2, 3, \ldots, q + 1$. Therefore all eigenvectors w are

of the form $w^\mathsf{T} = [w_1^\mathsf{T}, 0^\mathsf{T}, \ldots, 0^\mathsf{T}] \in \mathrm{span}[\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_d]$. Similarly, for the left eigenspace we have

$$(w^\mathsf{T} A(h))_l = \sum_{r=0}^{l-1} \frac{h^r}{r!} w_{l-r}^\mathsf{T}, \quad l = 1, 2, \ldots, q + 1. \tag{37}$$

Starting with $l = 1$ we have trivially that $w_1^\mathsf{T} = w_1^\mathsf{T}$. For $l = 2$ we have $w_2^\mathsf{T} + w_1^\mathsf{T} h = w_2^\mathsf{T}$ but $h > 0$, hence $w_1 = 0$. For $l = 3$ we have $w_3^\mathsf{T} = w_3^\mathsf{T} + w_2^\mathsf{T} h + w_1^\mathsf{T} h^2/2 = w_3^\mathsf{T} + w_2^\mathsf{T} h + 0^\mathsf{T} \cdot h^2/2$ but $h > 0$ hence $w_2 = 0$. By repeating this argument, we have $w_i = 0$, $i = 1, \ldots, q$ and $w_{q+1} = w_{q+1}$. Therefore, all left eigenvectors are of the form $w^\mathsf{T} = [0^\mathsf{T}, \ldots, 0^\mathsf{T}, w_{q+1}^\mathsf{T}] \in \mathrm{span}[\mathrm{e}_{qd+1}, \mathrm{e}_{qd+2}, \ldots, \mathrm{e}_{(q+1)d}]$.   □

We are now ready to state the main result of this section. Namely, that the Kalman filter that produces exact inference in Eq. (7) for linear vector fields is asymptotically stable if the linear vector field is of full rank.

Theorem 2 *Let $\Lambda \in \mathbb{R}^{d \times d}$ be a matrix with full rank and consider the linear ODE*

$$\dot{y}(t) = \Lambda y(t). \tag{38}$$

Consider estimating the solution of Eq. (38) using an IWP(q) prior with the same conditions on Γ as in Lemma 2. Then the Kalman filter estimate of the solution to Eq. (38) is asymptotically stable.

Proof From Eq. (7), we have that the Kalman filter operates on the following system

$$X_{n+1} = A(h)X_n + Q^{1/2}(h)W_{n+1}, \tag{39a}$$
$$Z_n = HX_n, \tag{39b}$$

where $H = [-\Lambda, \mathrm{I}, 0, \ldots, 0]$ and W_n are i.i.d. standard Gaussian vectors. It is sufficient to show that $[A(h), H]$ is completely detectable and $[A(h), Q^{1/2}(h)]$ is completely stabilisable (Anderson and Moore 1979, Chapter 4, p. 77). We start by showing complete detectability. If we let $w^\mathsf{T} = [w_1^\mathsf{T}, \ldots, w_{q+1}^\mathsf{T}]$, $w_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, q + 1$, then by Lemma 3 we have that $w^\mathsf{T} A^\mathsf{T}(h) = \eta w^\mathsf{T}$ for some η implies that either $w = 0$ or $w^\mathsf{T} = [w_1^\mathsf{T}, 0^\mathsf{T}, \ldots, 0^\mathsf{T}]$ for some $w_1 \in \mathbb{R}^d$ and $\eta = 1$. Furthermore, $w^\mathsf{T} H^\mathsf{T} = -w_1^\mathsf{T} \Lambda^\mathsf{T} + w_2^\mathsf{T} = 0$ implies that $w_2 = \Lambda w_1$. However, by the previous argument, we have $w_2 = 0$; therefore, $0 = \Lambda w_1$ but Λ is full rank by assumption so $w_1 = 0$. Therefore, $[A^\mathsf{T}(h), H^\mathsf{T}]$ is completely detectable. As for complete stabilisability, again by Lemma 3, we have $w^\mathsf{T} A(h) = \eta w^\mathsf{T}$ for some η, which implies either $w = 0$ or $w^\mathsf{T} = [0^\mathsf{T}, \ldots, 0^\mathsf{T}, w_{q+1}^\mathsf{T}]$ and $\eta = 1$. Furthermore, since the nullspace of $Q^{1/2}(h)$ is the same as

[3] Anderson and Moore (1979) denotes the measurement matrix by H^T while we denote it by H. With this in mind, our notion of complete detectability does not differ from Anderson and Moore (1979).

the nullspace of $Q(h)$, we have that $w^\mathsf{T} Q^{1/2}(h) = 0$ is equivalent to $w^\mathsf{T} Q(h) = 0$, which is given by

$$w^\mathsf{T} Q(h) = \left[w_{q+1}^\mathsf{T} Q_{q+1,1}(h) \dots w_{q+1}^\mathsf{T} Q_{q+1,q+1}(h) \right] = 0,$$

but by Lemma 2 the blocks $Q_{i,j}(h)$ have full rank so $w_{q+1} = 0$ and thus $w = 0$. To conclude, we have that $[A(h), Q^{1/2}(h)]$ is completely stabilisable and $[A(h), H]$ is completely detectable and therefore the Kalman filter is asymptotically stable. □

Corollary 1 *In the same setting as Theorem 2, the EKF and UKF are asymptotically stable.*

Proof Since the vector field is linear and therefore affine Lemma 1 implies that EKF and UKF reduce to the exact Kalman filter, which is asymptotically stable by Theorem 2.□

It is worthwhile to note that Λ_{test} is of full rank for all $[\lambda_1 \, \lambda_2]^\mathsf{T} \in \mathbb{R}^2 \setminus \{0\}$, and consequently Theorem 2 and Corollary 1 guarantee A-stability for the EKF and UKF in the sense of Dahlquist (1963).[4] Lastly, a peculiar fact about Theorem 2 is that it makes no reference to the eigenvalues of Λ (i.e. the stability properties of the ODE). That is, the Kalman filter will be asymptotically stable even if the underlying ODE is not, provided that, Λ of full rank. This may seem awkward but it is rarely the case that the ODE that we want to integrate is unstable, and even in such a case most solvers will produce an error that grows without a bound as well. Though all of the aforementioned properties are at least partly consequences of using IWP(q) as a prior and they may thus be altered by changing the prior.

4 Uncertainty calibration

In practice the model parameters, (F, u, L), might depend on some parameters that need to be estimated for the probabilistic solver to report appropriate uncertainty in the estimated solution to Eq. (1). The diffusion matrix L is of particular importance as it determines the gain of the Wiener process entering the system in Eq. (3) and thus determines how 'diffuse' the prior is. Herein we shall only concern ourselves with estimating L, though, one might anticipate future interest in estimating F and u as well. However, let us start with a few words on the monitoring of errors in numerical solvers in general.

4.1 Monitoring of errors in numerical solvers

An important aspect of numerical analysis is to monitor the error of a method. While the goal of probabilistic solvers is to do so by calibration of a probabilistic model, the approach of classical numerical analysis is to examine the local and global errors. The global error can be bounded but is typically impractical for monitoring error (Hairer et al. 1987, Chapter II.3). A more practical approach is to monitor (and control) the accumulation of local errors. This can be done by using two step sizes together with Richardson extrapolation (Hairer et al. 1987, Theorem 4.1). Though, perhaps more commonly this is done via embedded Runge–Kutta methods (Hairer et al. 1987, Chapter II.4) or the Milne device Byrne and Hindmarsh (1975).

In the context of filters, the relevant object in this regard is the scaled residual $S_n^{-1/2}(z_n - \hat{z}_n)$. Due to its role in the prediction-error decomposition, which is defined below, it directly monitors the calibration of the predictive distribution. Schober et al. (2019) showed how to use this quantity to effectively control step sizes in practice. It was also recently shown in (Kersting et al. 2018, Section 7), that in the case of $q = 1$, fixed σ^2 (amplitude of the Wiener process) and Integrated Wiener Process prior, the posterior standard deviation computed by the solver of Schober et al. (2019) contracts at the same rate as the worst-case error as the step size goes to zero—thereby preventing both under- and overconfidence.

In the following, we discuss effective strategies for calibrating L when it is given by $L = \sigma \check{L}$ for fixed $\check{L}$, thus providing a probabilistic quantification of the error in the proposed solvers.

4.2 Uncertainty calibration for affine vector fields

As noted in Sect. 2.6, the Kalman filter produces the exact solution to the inference problem in Eq. (7) when the vector field is affine. Furthermore, the marginal likelihood $p(z_{1:N})$ can be computed during the execution of the Kalman filter by the prediction error decomposition (Schweppe 1965), which is given by:

$$p(z_{1:N}) = p(z_1) \prod_{n=2}^{N} p(z_n \mid z_{1:n-1})$$
$$= \prod_{n=1}^{N} \mathcal{N}(z_n; \hat{z}_n, S_n). \tag{40}$$

While the marginal likelihood in Eq. (40) is certainly straightforward to compute without adding much computational cost, maximising it is a different story in general. In the particular case when the diffusion matrix L and the initial covariance Σ_0 are given by re-scaling fixed matrices $L = \sigma \check{L}$

---

[4] Some authors require stability on the line $\lambda_1 = 0$ as well (Hairer and Wanner 1996). Due to the exclusion of origin EKF and UKF cannot be said to be A-stable in this sense.

and $\Sigma_0 = \sigma^2 \check{\Sigma}_0$ for some scalar $\sigma > 0$, then uncertainty calibration can be done by a simple post-processing step after running the Kalman filter, as is shown in Proposition 4.

Proposition 4 *Let $f(y, t) = \Lambda(t)y + \zeta(t)$, $\Sigma_0 = \sigma^2 \check{\Sigma}_0$, $L = \sigma \check{L}$, $R = 0$ and denote the equivalent discrete-time process noise covariance for the prior model $(F, u, \check{L})$ by $\check{Q}(h)$. Then the Kalman filter estimate to the solution of*

$$\dot{y}(t) = f(y(t), t)$$

that uses the parameters $(\mu_0^F, \Sigma_0, A(h), \xi(h), Q(h))$ is equal to the Kalman filter estimate that uses the parameters $(\mu_0^F, \check{\Sigma}_0, A(h), \xi(h), \check{Q}(h))$. More specifically, if we denote the filter mean and covariance at time n using the former parameters by (μ_n^F, Σ_n^F) and the corresponding filter mean and covariance using the latter parameters by $(\check{\mu}_n^F, \check{\Sigma}_n^F)$, then $(\mu_n^F, \Sigma_n^F) = (\check{\mu}_n^F, \sigma^2 \check{\Sigma}_n^F)$. Additionally, denote the predicted mean and covariance of the measurement Z_n by $\check{z}_n$ and $\check{S}_n$, respectively, when using the parameters $(\mu_0^F, \check{\Sigma}_0, A(h), \xi(h), \check{Q}(h))$. Then the maximum likelihood estimate of σ^2, denoted by $\widehat{\sigma_N^2}$, is given by

$$\widehat{\sigma_N^2} = \frac{1}{Nd} \sum_{n=1}^{N} (z_n - \check{z}_n)^\mathsf{T} \check{S}_n^{-1} (z_n - \check{z}_n). \tag{41}$$

Proposition 4 is just an amalgamation of statements from Tronarp et al. (2019). Nevertheless, we provide an accessible proof in "Appendix C".

4.3 Uncertainty calibration for non-affine vector fields

For non-affine vector fields, the issue of parameter estimation becomes more complicated. The Bayesian filtering problem is not solved exactly and consequently any marginal likelihood will be approximate as well. Nonetheless, a common approach in the Gaussian filtering framework is to approximate the marginal likelihood in the same manner as the filtering solution is approximated (Särkkä 2013, Chapter 12.3.3), that is:

$$p(z_{1:N}) \approx \prod_{n=1}^{N} \mathcal{N}(z_n; \hat{z}_n, S_n), \tag{42}$$

where $\hat{z}_n$ and S_n are the quantities in Eq. (12) approximated by some method (e.g. EKF). Maximising Eq. (42) is a common approach in signal processing (Särkkä 2013) and referred to as *quasi maximum likelihood* in time series literature (Lindström et al. 2015). Both Eqs. (14) and (16) can be thought of as Kalman updates for the case where the vector field is approximated by a piece-wise affine function, without

modifying Σ_0, $Q(h)$, and R. For instance the affine approximation of the vector field due to the EKF on the discretisation interval $[t_n, t_{n+1})$ is given by

$$\hat{\zeta}_n(t) = f(C\mu_n^P, t_n) - J_f(C\mu_n^P, t_n)C\mu_n^P, \tag{43a}$$

$$\hat{\Lambda}_n(t) = J_f(C\mu_n^P, t_n), \tag{43b}$$

$$\hat{f}_n(y, t) = \hat{\Lambda}_n(t)y + \hat{\zeta}_n(t). \tag{43c}$$

While the vector field is approximated by a piece-wise affine function, the discrete-time filtering problem Eq. (7) is still simply an affine problem, without modifications of Σ_0, $Q(h)$, and R. Therefore, the results of Proposition 4 still apply and the σ^2 maximising the approximate marginal likelihood in Eq. (42) can be computed in the same manner as in Eq. (41).

On the other hand, it is clear that dependence on σ^2 in Eq. (12) is non-trivial in general, which is also true for the quadrature approaches of Sect. 2.5. Therefore, maximising Eq. (42) for the quadrature approaches is not as straightforward. However, by Taylor series expanding the vector field in Eq. (12) one can see that the numerical integration approaches are roughly equal to the Taylor series approaches provided that $\check{\Sigma}_n^P$ is small. Therefore, we opt for plugging in the corresponding quantities from the quadrature approximations into Eq. (41) in order to achieve computationally cheap calibration of these approaches.

Remark 2 A local calibration method for σ^2 is given by [Schober et al. 2019, Eq. (45)], which in fact corresponds to an h-dependent prior, with the diffusion matrix in Eq. (3) $L = L(t)$ being piece-wise constant over integration steps. Moreover, Schober et al. (2019) had to neglect the dependence of Σ_n^P on the likelihood. Here we prefer the estimator given in Eq. (41) since it is attempting to maximise the likelihood from the globally defined probability model in Eq. (7), and it succeeds for affine vector fields.

More advanced methods for calibrating the parameters of the prior can be developed by combining the Gaussian smoothing equations (Särkkä 2013, Chapter 10) with the expectation maximisation method (Kokkala et al. 2014) or variational Bayes (Taniguchi et al. 2017).

4.4 Uncertainty calibration of particle filters

If calibration of Gaussian filters was complicated by having a non-affine vector field, the situation for particle filters is even more challenging. There is, to the authors' knowledge, no simple estimator of the scale of the Wiener process (such as Proposition 4) even for the case of affine vector fields. However, the literature on parameter estimation using particle methods is vast so we proceed to point the reader towards some alternatives. In the class of off-line methods,

Schön et al. (2011) uses a particle smoother to implement an expectation maximisation algorithm, while Lindsten (2013) uses a particle Markov chain Monte Carlo methods to implement a stochastic approximation expectation maximisation algorithm. One can also use the iterated filtering method of Ionides et al. (2011) to get a maximum likelihood estimator, or particle Markov chain Monte Carlo (Andrieu et al. 2010).

On the other hand, if online calibration is required then the gradient based recursive maximum likelihood estimator by Doucet and Tadić (2003) can be used, or the online version of iterated filtering by Lindström et al. (2012). Furthermore, Storvik (2002) provides an alternative for online calibration when sufficient statics of the parameters are finite dimensional and can be computed recursively in n. An overview on parameter estimation using particle filters was also given by Kantas et al. (2009).

5 Experimental results

In this section, we evaluate the different solvers presented in this paper in different scenarios. Though before we proceed to the experiments we define some summary metrics with which assessments of accuracy and uncertainty quantification can be made. The root-mean-square error (RMSE) is often used to assess accuracy of filtering algorithms and is defined by

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left\| y(nh) - C \mu_n^F \right\|^2}.$$

In fact $y(nh) - C\mu_n^F$ is precisely the *global error* at time t_n (Hairer et al. 1987, Eq. (3.16)). As for assessing the uncertainty quantification, the χ^2-statistics is commonly used (Bar-Shalom et al. 2001). That is, in a linear Gaussian model the following quantities

$$\left(y(nh) - C\mu_n^F\right)^{\mathsf{T}} [C \Sigma_n^F C^{\mathsf{T}}]^{-1} \left(y(nh) - C\mu_n^F\right), \ n = 1, \ldots, N,$$

are i.i.d. $\chi^2(d)$. For a trajectory summary we define the average χ^2-statistics as

$$\bar{\chi}^2 = \frac{1}{N} \sum_{n=1}^{N} \left(y(nh) - C\mu_n^F\right)^{\mathsf{T}} [C \Sigma_n^F C^{\mathsf{T}}]^{-1} \left(y(nh) - C\mu_n^F\right).$$

For an accurate and well-calibrated model, the RMSE is small and $\bar{\chi}^2 \approx d$. In the succeeding discussion, we shall refer to a method producing $\bar{\chi}^2 < d$ or $\bar{\chi}^2 > d$ as *underconfident* or *overconfident*, respectively.

5.1 Linear systems

In this experiment, we consider a linear system given by

$$\Lambda = \begin{bmatrix} \lambda_1 & -\lambda_2 \\ \lambda_2 & \lambda_1 \end{bmatrix}, \tag{44a}$$

$$\dot{y}(t) = \Lambda y(t), \quad y(0) = e_1. \tag{44b}$$

This makes for a good test model as the inference problem in Eq. (7) can be solved exactly, and consequently its adequacy can be assessed. We compare exact inference by the Kalman filter (KF)[5] (see Sect. 2.6) with the approximation due to Schober et al. (2019) (SCH) (see Proposition 2) and the covariance approximation due to Kersting and Hennig (2016) (KER) (see Proposition 3). The integration interval is set to [0, 10], and all methods use an IWP(q) prior for $q = 1, 2, \ldots, 6$, and the initial mean is set to $\mathbb{E}[X^{(j)}(0)] = \Lambda^{j-1} y(0)$ for $j = 1, \ldots, q + 1$, with variance set to zero (exact initialisation). The uncertainty of the methods is calibrated by the maximum likelihood method (see Proposition 4), and the methods are examined for 10 step sizes uniformly placed on the interval $[10^{-3}, 10^{-1}]$.

We examine the parameters $\lambda_1 = 0$ and $\lambda_2 = \pi$ (half a revolution per unit of time with no damping). The RMSE is plotted against step size in Fig. 1. It can be seen that SCH is a slightly better than KF and KER for $q = 1$ and small step sizes, and KF becomes slightly better than SCH for large step size while KER becomes significantly worse than both KF and SCH. For $q > 1$, it can be seen that the RMSE is significantly lower for KF than for SCH/KER in general with performance differing between one and two orders of magnitude. Particularly, the superior stability properties of KF are demonstrated (see Theorem 2) for $q > 3$ where both SCH and KER produce massive errors for larger step sizes.

Furthermore, the average χ^2-statistic is shown in Fig. 2. All methods appear to be overconfident for $q = 1$ with SCH performing best, followed by KER. On the other hand, for $1 < q < 5$, SCH and KER remain overconfident for the most part, while KF is underconfident. Our experiments also show that unsurprisingly all methods perform better for smaller $|\lambda_2|$ (frequency of the oscillation). However, we omit visualising this here.

Finally, a demonstration of the error trajectory for the first component of y and the reported uncertainty of the solvers is shown in Fig. 3 for $h = 10^{-2}$ and $q = 2$. Here it can be seen that all methods produce similar errors bars, though SCH and KER produce errors that oscillate far outside their reported uncertainties.

---

[5] Again note that the EKF and appropriate numerical quadrature methods are equivalent to this estimator here (see Lemma 1).
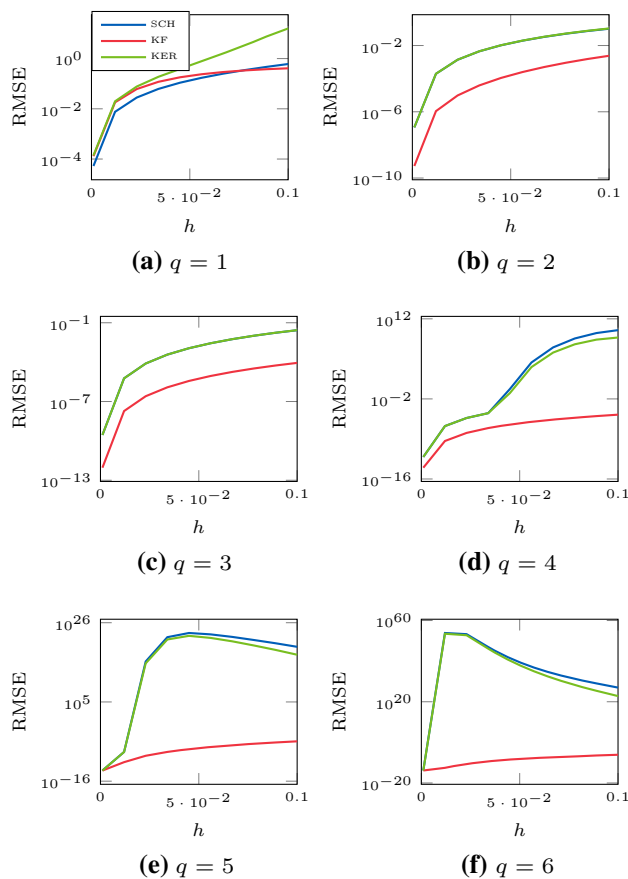
Fig. 1 RMSE of KF, SCH, and KER on the undamped oscillator using IWP(q) priors for $q = 1, \ldots, 6$ plotted against step size
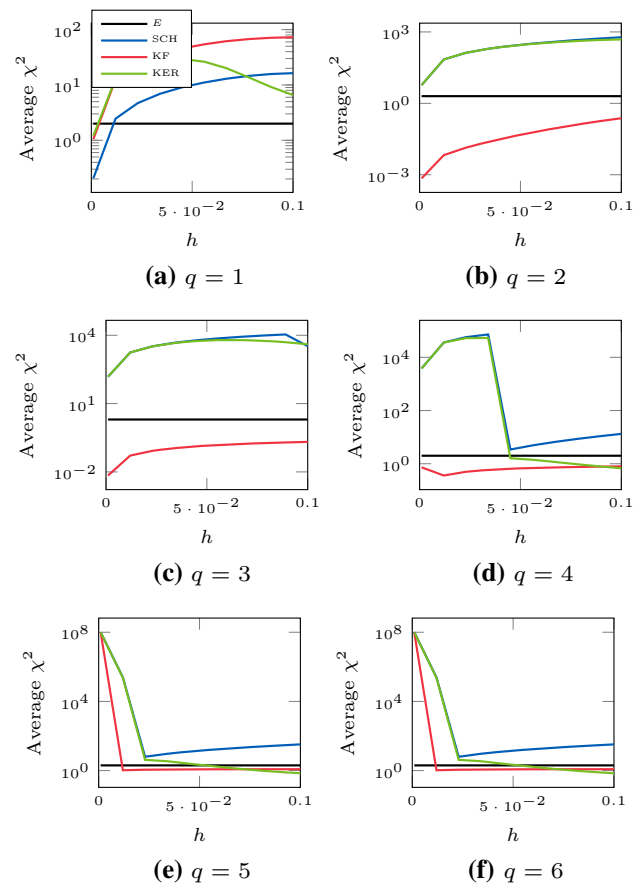

Fig. 2 Average χ^2-statistic of KF, SCH, and KER on the undamped oscillator using IWP(q) priors for $q = 1, \ldots, 6$ plotted against step size. The expected χ^2-statistic is shown in black (E)
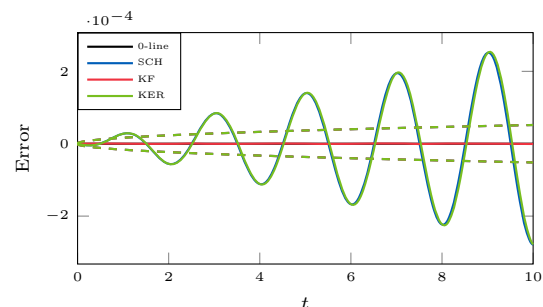
5.2 The logistic equation

In this experiment, the logistic equation is considered:

$$\dot{y}(t) = r y(t)(1 - y(t)), \quad y(0) = 1 \cdot 10^{-1}, \tag{45}$$

which has the solution:

$$y(t) = \frac{\exp(rt)}{1/y_0 - 1 + \exp(rt)}. \tag{46}$$

In the experiments, r is set to $r = 3$. We compare the zeroth order solver (Proposition 2) (Schober et al. 2019) (SCH), the first-order solver in Eq. (16) (EKF), a numerical integration solver based on the covariance approximation in Proposition 3 (Kersting and Hennig 2016) (KER), and a numerical integration solver based on approximating Eq. (12) (UKF). Both numerical integration approaches use a third degree fully symmetric rule (see McNamee and Stenger 1967). The integration interval is set to [0, 2.5], and all methods use an IWP(q) prior for $q = 1, 2, \ldots, 4$, and the initial mean of $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ are set to $y(0)$, $f(y(0))$, and $J_f(y(0))f(y(0))$, respectively (correct values), with zero

Fig. 3 The errors (solid lines) and ± 2 standard deviation bands (dashed) for KF, SCH, and KER on the undamped oscillator with $q = 2$ and $h = 10^{-2}$. A line at 0 is plotted in solid black

covariance. The remaining state components $X^{(j)}$, $j > 3$ are set to zero mean with unit variance. The uncertainty of the methods is calibrated by the quasi maximum likelihood method as explained in Sect. 4.3, and the methods are examined for 10 step sizes uniformly placed on the interval $[10^{-3}, 10^{-1}]$.

The RMSE is plotted against step size in Fig. 4. It can be seen that EKF and UKF tend to produce smaller
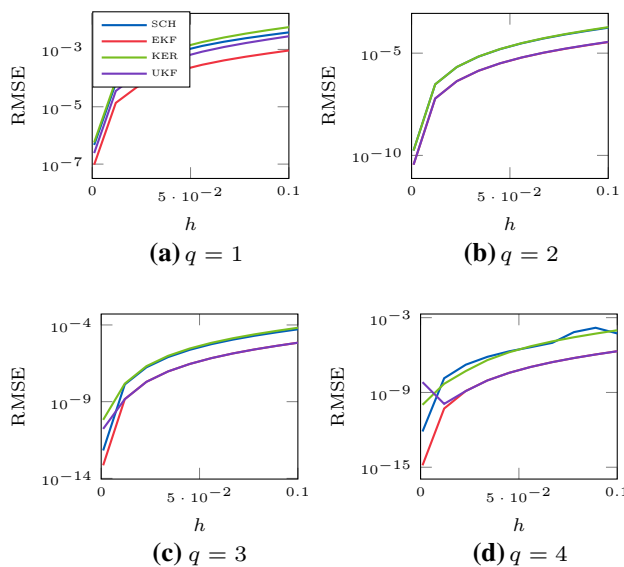
Fig. 4 RMSE of SCH, EKF, KER, and UKF on the logistic equation using IWP(q) priors for $q = 1, \ldots, 4$ plotted against step size
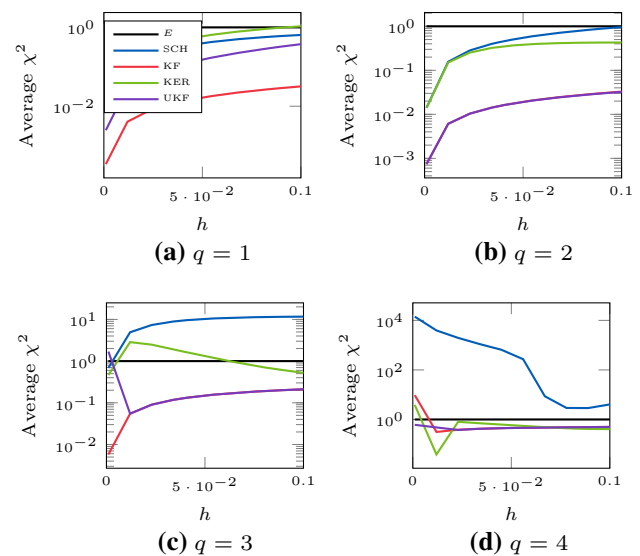

Fig. 5 Average χ^2-statistic of SCH, EKF, KER, and UKF on the logistic equation using IWP(q) priors for $q = 1, \ldots, 4$ plotted against step size. The expected χ^2-statistic is shown in black (E)
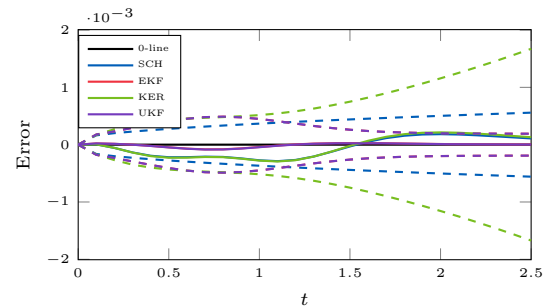
errors by more than an order of magnitude than SCH and KER in general, with the notable exception of the UKF behaving badly for small step sizes and $q = 4$. This is probably due to numerical issues for generating the integration nodes, which requires the computation of matrix square roots (Julier et al. 2000) that can become inaccurate for ill-conditioned matrices. Additionally, the average χ^2-statistic is plotted against step size in Fig. 5. Here it appears that all methods tend to be underconfident for $q = 1, 2$, while SCH becomes overconfident for $q = 3, 4$.

A demonstration of the error trajectory and the reported uncertainty of the solvers is shown in Fig. 3 for $h = 10^{-1}$ and $q = 2$. SCH and KER produce similar errors, and they are hard to discern in the figure. The same goes for EKF and UKF. Additionally, it can be seen that the solvers produce qualitatively different uncertainty estimates. While the uncertainty of EKF and UKF first grows to then shrink as the solution approaches the fixed point at $y(t) = 1$, the uncertainty of SCH grows over the entire interval with the uncertainty of KER growing even faster (Fig. 6).

5.3 The FitzHugh–Nagumo model

The FitzHugh–Nagumo model is given by:

$$\begin{bmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \end{bmatrix} = \begin{bmatrix} c\left(y_1(t) - \frac{y_1^3(t)}{3} + y_2(t)\right) \\ -\frac{1}{c}\left(y_1(t) - a + by_2(t)\right) \end{bmatrix}, \tag{47}$$

Fig. 6 The errors (solid lines) and ± 2 standard deviation bands (dashed) for EKF, SCH, KER, and UKF on the logistic with $q = 2$ and $h = 10^{-1}$. A line at 0 is plotted in solid black

where we set $(a, b, c) = (.2, .2, 3)$ and $y(0) = [-1\ 1]^\mathsf{T}$. As previous experiments showed that the behaviour of KER and UKF are similar to SCH and EKF, respectively, we opt for only comparing the latter to increase readability of the presented results. As previously, the moments of $X^{(1)}(0)$, $X^{(2)}(0)$, and $X^{(3)}(0)$ are initialised to their exact values and the remaining derivatives are initialised with zero mean and unit variance. The integration interval is set to [0, 20], and all methods use an IWP(q) prior for $q = 1, \ldots, 4$ and the uncertainty is calibrated as explained in Sect. 4.3. A baseline solution is computed using MATLAB's ode45 function with an absolute tolerance of 10^{-15} and relative tolerance of 10^{-12}, all errors are computed under the assumption that ode45 provides the exact solution. The methods are examined for 10 step sizes uniformly placed on the interval $[10^{-3}, 10^{-1}]$.

The RMSE is shown in Fig. 7. For $q = 1$, EKF produces an error orders of magnitude larger than SCH and for $q = 2$ both methods produce similar errors until the step size grows too large, causing SCH to start producing orders of magnitude
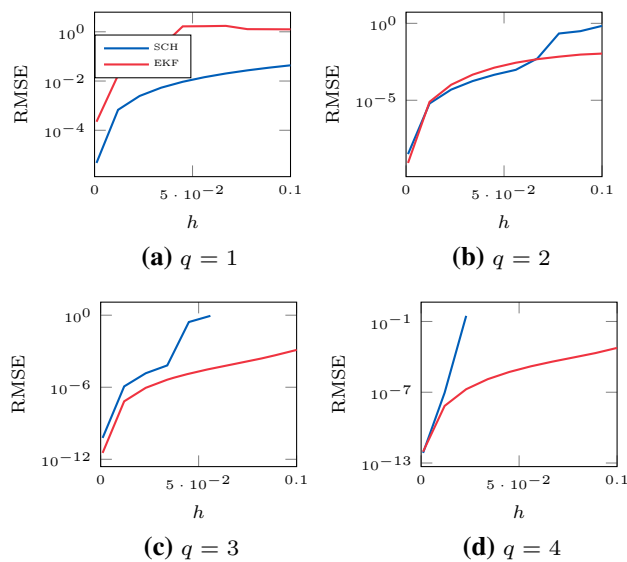
Fig. 7 RMSE of SCH and EKF on the FitzHugh–Nagumo model using IWP(q) priors for $q = 1, \ldots, 4$ plotted against step size

Fig. 8 Average χ^2-statistic of SCH and EKF on the FitzHugh–Nagumo model using IWP(q) priors for $q = 1, \ldots, 4$ plotted against step size

larger errors than EKF. For $q = 3, 4$ EKF is superior in producing lower errors and additionally SCH can be seen to become unstable for larger step sizes (at $h \approx 5 \cdot 10^{-2}$ for $q = 3$ and at $h \approx 2 \cdot 10^{-2}$ for $q = 4$). Furthermore, the averaged χ^2-statistic is shown in Fig. 8. It can be seen that EKF is overconfident for $q = 1$ while SCH is underconfident. For $q = 2$, both methods are underconfident while EKF remains underconfident for $q = 3, 4$, but SCH becomes overconfident for almost all step sizes.

The error trajectory for the first component of y and the reported uncertainty of the solvers is shown in Fig. 9 for $h = 5 \cdot 10^{-2}$ and $q = 2$. It can be seen that both methods have periodically occurring spikes in their errors with EKF being larger in magnitude but also briefer. However, the uncertainty estimate of the EKF is also spiking at the same time giving an adequate assessments of its error. On the other hand, the uncertainty estimate of SCH grows slowly and monotonically over the integration interval, with the error estimate going outside the two standard deviation region at the first spike (slightly hard to see in the figure).

5.4 A Bernoulli equation

In this following experiment, we consider a transformation of Eq. (45), $\eta(t) = \sqrt{y(t)}$, for $r = 2$. The resulting ODE for $\eta(t)$ now has two stable equilibrium points $\eta(t) = \pm 1$ and an unstable equilibrium point at $\eta(t) = 0$. This makes it a simple test domain for different sampling-based ODE solvers, because different types of posteriors ought to arise. We compare the proposed particle filter using both the proposal Eq. (24) [PF(1)] and EKF proposals [Eq. (25)] [PF(2)]
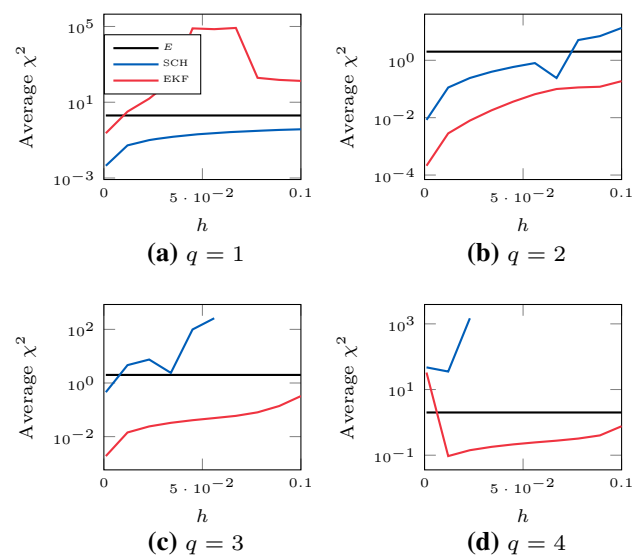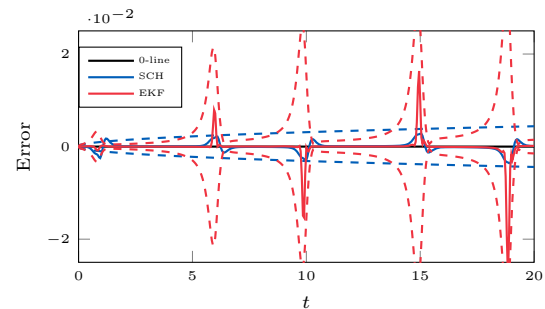

Fig. 9 The errors (solid lines) and ± 2 standard deviation bands (dashed) for EKF and SCH on the FitzHugh–Nagumo model with $q = 2$ and $h = 5 \cdot 10^{-2}$. A line at 0 is plotted in solid black

with the method by Chkrebtii et al. (2016) (CHK) and the one by Conrad et al. (2017) (CON) for estimating $\eta(t)$ on the interval $t \in [0, 5]$ with initial condition set to $\eta_0 = 0$. Both PF and CHK use and IWP(q) prior and set $R = \kappa h^{2q+1}$. CON uses a Runge–Kutta method of order q with perturbation variance $h^{2q+1}/[2q(q!)^2]$ as to roughly match the incremental variance of the noise entering PF(1), PF(2), and CHK, which is determined by $Q(h)$ and not R.

First we attempt to estimate $y(5) = 0$ for 10 step sizes uniformly placed on the interval $[10^{-3}, 10^{-1}]$ with $\kappa = 1$ and $\kappa = 10^{-10}$. All methods use 1000 samples/particles, and they estimate $y(5)$ by taking the mean over samples/empirical measures. The estimate of $y(5)$ is plotted against the step size in Fig. 10. In general, the error increases with the step size for all methods, though most easily discerned in Fig. 10b, d. All in all, it appears that CHK, PF(1), and PF(2) behave similarly with regards to the estimation, while CON appears to produce a bit larger errors. Furthermore, the effect of κ appears to be the greatest on PF(1) and PF(2) as best illustrated in Fig. 10c.
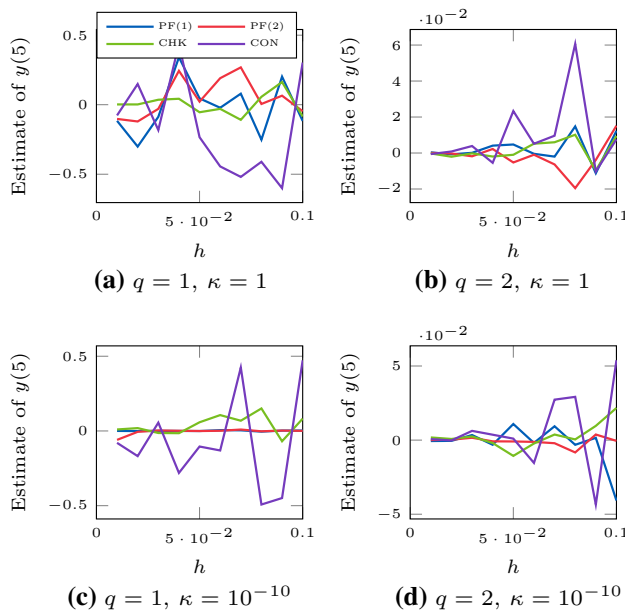
Fig. 10 Sample mean estimate of the solution at $T = 5$

Additionally, kernel density estimates for the different methods are made for time points $t = 1, 3, 5$ for $\kappa = 1$, $q = 1, 2$ and $h = 10^{-1}, 5 \cdot 10^{-2}$. In Fig. 11 kernel density estimates for $h = 10^{-1}$ are shown. At $t = 1$, all methods produce fairly concentrated unimodal densities that then disperse as time goes on, with CON being a least concentrated and dispersing quicker followed by PF(1)/PF(2) and then last CHK. Furthermore, CON goes bimodal as time goes on, which is best seen in for $q = 1$ in Fig. 11e. On the other hand, the alternatives vary between unimodal (CHK in 11f, also to some degree PF(1) and PF(2)), bimodal (PF(1) and CHK in Fig. 11e), and even mildly trimodal (PF(2) in Fig. 11e).

Similar behaviour of the methods is observed for $h = 5 \cdot 10^{-2}$ in Fig. 11, though here all methods are generally more concentrated (Fig. 12).

6 Conclusion and discussion

In this paper, we have presented a novel formulation of probabilistic numerical solution of ODEs as a standard problem in GP regression with a nonlinear measurement function, and with measurements that are identically zero. The new model formulation enables the use of standard methods in signal processing to derive new solvers, such as EKF, UKF, and PF. We can also recover many of the previously proposed sequential probabilistic ODE solvers as special cases.

Additionally, we have demonstrated excellent stability properties of the EKF and UKF on linear test equations, that is, A-stability has been established. The notion of A-stability is closely connected with the solution of stiff equations, which is typically achieved with *implicit* or *semi-implicit*
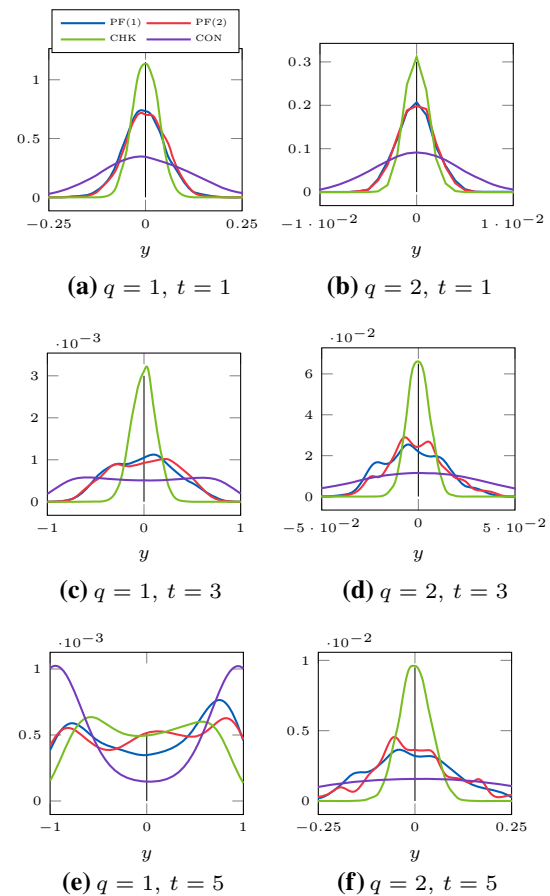

Fig. 11 Kernel density estimates of the solution of the Bernoulli equation for $h = 10^{-1}$ and $\kappa = 1$. Mind the different scale of the axes

methods (Hairer and Wanner 1996). In this respect, our methods (EKF and UKF) most closely fit into the class of semi-implicit methods such as the methods of Rosenbrock type (Hairer and Wanner 1996, Chapter IV.7). Though it does seem feasible, the proposed methods can be nudged towards the class of implicit methods by means of iterative Gaussian filtering (Bell and Cathey 1993; Garcia-Fernandez et al. 2015; Tronarp et al. 2018).

While the notion of A-stability has been fairly successful in discerning between methods with good and bad stability properties, it is not the whole story (Alexander 1977, Section 3). This has lead to other notions of stability such as *L-stability* and *B-stability* (Hairer and Wanner 1996, Chapter IV.3 and IV.12). It is certainly an interesting question whether the present framework allows for the development of methods satisfying these more strict notions of stability.

An advantage of our model formulation is the decoupling of the prior from the likelihood. Thus future work would involve investigating how well the exact posterior to our inference problem approximates the ODE and then analysing how well different approximate inference strategies behave. However, for $h \to 0$, we expect that the novel Gaussian
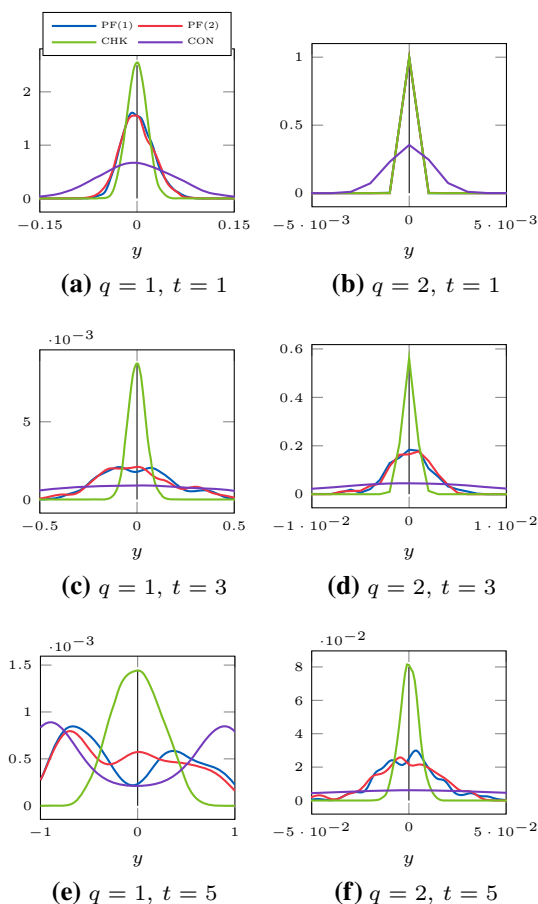
Fig. 12 Kernel density estimates of the solution of the Bernoulli equation for $h = 5 \cdot 10^{-2}$ and $\kappa = 1$. Mind the different scale of the axes

filters (EKF,UKF) will exhibit polynomial worst-case convergence rates of the mean and its credible intervals, that is, its Bayesian uncertainty estimates, as has already been proved in Kersting et al. (2018) for 0th order Taylor series filters with arbitrary constant measurement variance R (see Sect. 2.4).

Our Bayesian recast of ODE solvers might also pave the way towards an average-case analysis of these methods, which has already been executed in Ritter (2000) for the special case of Bayesian quadrature. For the PF, a thorough convergence analysis similar to Chkrebtii et al. (2016), Conrad et al. (2017), Abdulle and Garegnani (2018) and Del Moral (2004) appears feasible. However, the results on spline approximations for ODEs (see, e.g. Loscalzo and Talbot 1967) might also apply to the present methodology via the correspondence between GP regression and spline function approximations (Kimeldorf and Wahba 1970).

A Proof of Proposition 1

In this section, we prove Proposition 1. First note that, by Eq. (4), we have

$$\frac{d\mathbb{C}\big[X^{(1)}(t), X^{(2)}(s)\big]}{dt} = \mathbb{C}\big[X^{(2)}(t), X^{(2)}(s)\big], \qquad (48)$$

where $\mathbb{C}$ is the cross-covariance operator. That is the cross-covariance matrix between $X^{(1)}(t)$ and $X^{(2)}(t)$ is just the integral of the covariance matrix function of $X^{(2)}$. Now define

$$\big(\mathbf{X}^{(i)}\big)^{\mathsf{T}} = \Big[\big(X_1^{(i)}\big)^{\mathsf{T}} \ldots \big(X_N^{(i)}\big)^{\mathsf{T}}\Big], \ i = 1, \ldots, q+1, \quad (49a)$$

$$\mathbf{g}^{\mathsf{T}} = \big[g^{\mathsf{T}}(h) \ldots g^{\mathsf{T}}(Nh)\big], \qquad (49b)$$

$$\mathbf{z}^{\mathsf{T}} = \big[z_1^{\mathsf{T}} \ldots z_N^{\mathsf{T}}\big]. \qquad (49c)$$

Since Equation (3) defines a Gaussian process, we have that $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are jointly Gaussian distributed and from Eq. (48) the blocks of $\mathbb{C}[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}]$ are given by

$$\mathbb{C}\big[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\big]_{n,m} = \int_0^{nh} \mathbb{C}\big[X^{(2)}(t), X^{(2)}(mh)\big]dt$$

which is precisely the kernel mean, with respect to the Lebesgue measure on $[0, nh]$, evaluated at mh, see (Briol et al. 2019, Section 2.2). Furthermore,

$$\mathbb{V}\big[\mathbf{X}^{(2)}\big]_{n,m} = \mathbb{C}\big[X^{(2)}(nh), X^{(2)}(mh)\big],$$

that is, the covariance matrix function (referred to as kernel matrix in Bayesian quadrature literature (Briol et al. 2019)) evaluated at all pairs in $\{h, \ldots, Nh\}$. From Gaussian conditioning rules, we have for the conditional means

and covariance matrices given $\mathbf{X}^{(2)} - \mathbf{g} = 0$, denoted by $\mathbb{E}_{\mathcal{D}}[X^{(1)}(nh)]$ and $\mathbb{V}_{\mathcal{D}}[X^{(1)}(nh)]$, respectively, that

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\big[X^{(1)}(nh)\big] &= \mathbb{E}\big[X^{(1)}(nh)\big] + \mathbf{w}_n\big(\mathbf{z} + \mathbf{g} - \mathbb{E}[\mathbf{X}^{(2)}]\big) \\
&= \mathbb{E}\big[X^{(1)}(nh)\big] + \mathbf{w}_n\big(\mathbf{g} - \mathbb{E}[\mathbf{X}^{(2)}]\big), \\
\mathbb{V}_{\mathcal{D}}\big[X^{(1)}(nh)\big] &= \mathbb{V}\big[X^{(1)}(nh)\big] - \mathbf{w}_n\mathbb{V}[\mathbf{X}^{(2)}]\mathbf{w}_n^\mathsf{T},
\end{aligned}
$$

where we used the fact that $\mathbf{z} = 0$ by definition and $\mathbf{w}_n$ are the Bayesian quadrature weights associated to the integral of g over the domain $[0, nh]$, given by (see Briol et al. 2019, Proposition 1)

$$
\mathbf{w}_n^\mathsf{T} = \mathbb{V}[\mathbf{X}^{(2)}]^{-1}
\begin{bmatrix}
\mathbb{C}\big[X^{(1)}(nh), X^{(2)}(h)\big]^\mathsf{T} \\
\vdots \\
\mathbb{C}\big[X^{(1)}(nh), X^{(2)}(Nh)\big]^\mathsf{T}
\end{bmatrix}.
$$

$\square$

B Proof of Proposition 3

To prove Proposition 3, expand the expressions for S_n and K_n as given by Eq. (12):

$$
\begin{aligned}
S_n &= \dot{C}\Sigma_n^P\dot{C}^\mathsf{T} + \mathbb{V}\big[f(CX_n, t_n) \mid z_{1:n-1}\big] \\
&\quad - \dot{C}\mathbb{C}\big[X_n, f(CX_n, t_n) \mid z_{1:n-1}\big] \\
&\quad - \mathbb{C}\big[X_n, f(CX_n, t_n) \mid z_{1:n-1}\big]^\mathsf{T}\dot{C}^\mathsf{T} \\
&\approx \dot{C}\Sigma_n^P\dot{C}^\mathsf{T} + \mathbb{V}\big[f(CX_n, t_n) \mid z_{1:n-1}\big] \\
K_n &= \big(\Sigma_n^P\dot{C}^\mathsf{T} - \mathbb{C}\big[X_n, f(CX_n, t_n) \mid z_{1:n-1}\big]\big)S_n^{-1} \\
&\approx \Sigma_n^P\dot{C}^\mathsf{T}S_n^{-1},
\end{aligned}
$$

where in the second steps the approximation $\mathbb{C}[X_n, f(CX_n, t_n) \mid z_{1:n-1}] \approx 0$ was used. Lastly, recall that $z_n \triangleq 0$; hence, the update equations become

$$
\begin{aligned}
S_n &\approx \dot{C}\Sigma_n^P\dot{C}^\mathsf{T} + \mathbb{V}\big[f(CX_n, t_n) \mid z_{1:n-1}\big], &\text{(50a)} \\
K_n &\approx \Sigma_n^P\dot{C}^\mathsf{T}S_n^{-1}, &\text{(50b)} \\
\mu_n^F &\approx \mu_n^P + K_n\big(\mathbb{E}\big[f(CX_n, t_n) \mid z_{1:n-1}\big] - \dot{C}\mu_n^P\big), &\text{(50c)} \\
\Sigma_n^F &\approx \Sigma_n^P - K_nS_nK_n^\mathsf{T}. &\text{(50d)}
\end{aligned}
$$

When $\mathbb{E}[f(CX_n, t_n) \mid z_{1:n-1}]$ and $\mathbb{V}[f(CX_n, t_n) \mid z_{1:n-1}]$ are approximated by Bayesian quadrature using a squared exponential kernel and a uniform set of nodes translated and scaled by μ_n^P and Σ_n^P, respectively, the method of Kersting and Hennig (2016) is obtained. $\square$

C Proof of Proposition 4

Note that $(\breve{\mu}_n^F, \breve{\Sigma}_n^F)$ is the output of a misspecified Kalman filter (Tronarp et al. 2019, Algorithm 1). We indicate that a quantity from Eqs. (11) and (12) is computed by the misspecified Kalman filter by $\breve{\ }$. For example $\breve{\mu}_n^P$ is the predictive mean of the misspecified Kalman filter. If $\Sigma_n^F = \sigma^2\breve{\Sigma}_n^F$ and $\breve{\mu}_n^F = \mu_n^F$ holds then for the prediction step we have

$$
\begin{aligned}
\mu_{n+1}^P &= A(h)\mu_n^F + \xi(h) = A(h)\breve{\mu}_n^F + \xi(h) = \breve{\mu}_{n+1}^P, \\
\Sigma_{n+1}^P &= A(h)\Sigma_n^F A^\mathsf{T}(h) + Q(h), \\
&= \sigma^2\big(A(h)\breve{\Sigma}_n^F A^\mathsf{T}(h) + \breve{Q}(h)\big), \\
&= \sigma^2\breve{\Sigma}_{n+1}^P,
\end{aligned}
$$

where we used the fact that $Q(h) = \sigma^2\breve{Q}(h)$, which follows from $L = \sigma\breve{L}$ and Eq. (8). Furthermore, recall that $H_{n+1} = \dot{C} - \Lambda(t_{n+1})C$, which for the update gives

$$
\begin{aligned}
S_{n+1} &= H_{n+1}\Sigma_{n+1}^P H_{n+1}^\mathsf{T} \\
&= \sigma^2 H_{n+1}\breve{\Sigma}_{n+1}^P H_{n+1}^\mathsf{T} \\
&= \sigma^2\breve{S}_{n+1}. \\
K_{n+1} &= \Sigma_{n+1}^P H_{n+1}^\mathsf{T} S_{n+1}^{-1} \\
&= \sigma^2\breve{\Sigma}_{n+1}^P H_{n+1}^\mathsf{T}[\sigma^2\breve{S}_{n+1}]^{-1} \\
&= \breve{\Sigma}_{n+1}^P H_{n+1}^\mathsf{T}\breve{S}_{n+1}^{-1} \\
&= \breve{K}_{n+1}. \\
\hat{z}_{n+1} &= H_{n+1}\mu_{n+1}^P - \zeta(t_n) \\
&= H_{n+1}\breve{\mu}_{n+1}^P - \zeta(t_n) \\
&= \breve{z}_{n+1}, \\
\mu_{n+1}^F &= \mu_{n+1}^P + K_{n+1}(z_{n+1} - \hat{z}_{n+1}) \\
&= \breve{\mu}_{n+1}^P + \breve{K}_{n+1}(z_{n+1} - \breve{z}_{n+1}) \\
&= \breve{\mu}_{n+1}^F. \\
\Sigma_{n+1}^F &= \Sigma_{n+1}^P - K_{n+1}S_{n+1}K_{n+1}^\mathsf{T} \\
&= \sigma^2\big(\breve{\Sigma}_{n+1}^P - \breve{K}_{n+1}\breve{S}_{n+1}\breve{K}_{n+1}^\mathsf{T}\big) \\
&= \sigma^2\breve{\Sigma}_{n+1}^F.
\end{aligned}
$$

It thus follows by induction that $\mu_n^F = \breve{\mu}_n^F$, $\Sigma_n^F = \sigma^2\breve{\Sigma}_n^F$, $\hat{z}_n = \breve{z}_n$, and $S_n = \sigma^2\breve{S}_n$ for $n \geq 0$. From Eq. (40), we have that the log-likelihood is given by

$$
\begin{aligned}
\log p(z_{1:N}) &= \log\prod_{n=1}^N \mathcal{N}(z_n; \hat{z}_n, S_n) \\
&= \log\prod_{n=1}^N \mathcal{N}(z_n; \breve{z}_n, \sigma^2\breve{S}_n)
\end{aligned}
$$

$$= -\frac{Nd}{2}\log\sigma^2$$
$$-\sum_{n=1}^{N}\frac{(z_n - \check{z}_n)^\top \check{S}_n^{-1}(z_n - \check{z}_n)}{2\sigma^2}.$$

Taking the derivative of log-likelihood with respect to σ^2 and setting it to zero gives the following estimating equation

$$0 = -\frac{Nd}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(z_n - \check{z}_n)^\top \check{S}_n^{-1}(z_n - \check{z}_n),$$

which has the following solution

$$\sigma^2 = \frac{1}{Nd}\sum_{n=1}^{N}(z_n - \check{z}_n)^\top \check{S}_n^{-1}(z_n - \check{z}_n).$$

References

Abdulle, A., Garegnani, G.: Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration (2018). arXiv:1703.03680 [mathNA]

Alexander, R.: Diagonally implicit Runge–Kutta methods for stiff ODEs. SIAM J. Numer. Anal. **14**(6), 1006–1021 (1977)

Anderson, B., Moore, J.: Optimal Filtering. Prentice-Hall, Englewood Cliffs (1979)

Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**(3), 269–342 (2010)

Bar-Shalom, Y., Li, X.R., Kirubarajan, T.: Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software. Wiley, New York (2001)

Bell, B.M., Cathey, F.W.: The iterated Kalman filter update as a Gauss–Newton method. IEEE Trans. Autom. Control **38**(2), 294–297 (1993)

Briol, F.X., Oates, C.J., Girolami, M., Osborne, M.A., Sejdinovic, D.: Probabilistic integration: a role for statisticians in numerical analysis? (with discussion and rejoinder). Stat. Sci. **34**(1), 1–22 (2019). **(Rejoinder on pp 38–42)**

Butcher, J.C.: Numerical Methods for Ordinary Differential Equations, 2nd edn. Wiley, New York (2008)

Byrne, G.D., Hindmarsh, A.C.: A polyalgorithm for the numerical solution of ordinary differential equations. ACM Trans. Math. Softw. **1**(1), 71–96 (1975)

Calderhead, B., Girolami, M., Lawrence, N.D.: Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21 (NIPS), pp. 217–224. Curran Associates, Inc. (2009)

Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer, Berlin (2005)

Chen, T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. In: Bengio, S., Wallach, H., Larochelle, H.,Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31 (NIPS), pp. 6571–6583. Curran Associates, Inc. (2018)

Chkrebtii, O.A., Campbell, D.A., Calderhead, B., Girolami, M.A.: Bayesian solution uncertainty quantification for differential equations. Bayesian Anal. **11**(4), 1239–1267 (2016)

Cockayne, J., Oates, C., Sullivan, T., Girolami, M.: Bayesian probabilistic numerical methods. Siam Rev. (2019). **(to appear)**

Conrad, P.R., Girolami, M., Särkkä, S., Stuart, A., Zygalakis, K.: Statistical analysis of differential equations: introducing probability measures on numerical solutions. Stat. Comput. **27**(4), 1065–1082 (2017)

Crisan, D., Doucet, A.: A survey of convergence results on particle filtering methods for practitioners. IEEE Trans. Signal Process. **50**(3), 736–746 (2002)

Dahlquist, G.G.: A special stability problem for linear multistep methods. BIT Numer. Math. **3**(1), 27–43 (1963)

Del Moral, P.: Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Springer, Berlin (2004)

Deuflhard, P., Bornemann, F.: Scientific Computing with Ordinary Differential Equations. Springer, Berlin (2002)

Doucet, A., Tadić, V.B.: Parameter estimation in general state-space models using particle methods. Ann. Inst. Stat. Math. **55**(2), 409–422 (2003)

Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. **10**(3), 197–208 (2000)

Doucet, A., De Freitas, N., Gordon, N.: An introduction to sequential Monte Carlo methods. In: Sequential Monte Carlo methods in practice, pp 3–14. Springer, New York (2001)

Garcia-Fernandez, A.F., Svensson, L., Morelande, M.R., Särkkä, S.: Posterior linearization filter: principles and implementation using sigma points. IEEE Trans. Signal Process. **63**(20), 5561–5573 (2015)

Golub, G.H., Welsch, J.H.: Calculation of Gauss quadrature rules. Math. Comput. **23**(106), 221–230 (1969)

Grewal, M.S., Andrews, A.P.: Kalman Filtering: Theory and Practice Using MATLAB. Wiley, New York (2001)

Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. Springer Series in Computational Mathematics, vol. 14. Springer, Berlin (1996)

Hairer, E., Nørsett, S., Wanner, G.: Solving Ordinary Differential Equations I—Nonstiff Problems. Springer, Berlin (1987)

Hartikainen, J., Särkkä, S.: Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp 379–384 (2010)

Hennig, P., Hauberg, S.: Probabilistic solutions to differential equations and their application to Riemannian statistics. In: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR, W&CP, vol. 33 (2014)

Hennig, P., Osborne, M., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. **471**, 2179 (2015)

Hochbruck, M., Ostermann, A., Schweitzer, J.: Exponential Rosenbrock-type methods. SIAM J. Numer. Anal. **47**(1), 786–803 (2009)

Ionides, E.L., Bhadra, A., Atchadé, Y., King, A., et al.: Iterated filtering. Ann. Stat. **39**(3), 1776–1802 (2011)

Jazwinski, A.: Stochastic Processes and Filtering Theory. Academic Press, London (1970)

Julier, S., Uhlmann, J., Durrant-Whyte, H.F.: A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans. Autom. Control **45**(3), 477–482 (2000)

Kantas, N., Doucet, A., Singh, S.S., Maciejowski, J.M.: An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. IFAC Proc. Vol. **42**(10), 774–785 (2009)

Kennedy, M., O'Hagan, A.: Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B **63**(3), 425–464 (2002)

Kersting, H., Hennig, P.: Active uncertainty calibration in Bayesian ODE solvers. In: 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), pp. 309–318. Curran Associates, Inc. (2016)

Kersting, H., Sullivan, T., Hennig, P.: Convergence rates of Gaussian ODE filters (2018). arXiv:1807.09737 [mathNA]

Kimeldorf, G., Wahba, G.: A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. **41**(2), 495–502 (1970)

Kokkala, J., Solin, A., Särkkä, S.: Expectation maximization based parameter estimation by sigma-point and particle smoothing. In: 2014 17th International Conference on Information Fusion (FUSION), pp 1–8. IEEE (2014)

Lie, H., Stuart, A., Sullivan, T.: Strong convergence rates of probabilistic integrators for ordinary differential equations. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09898-6

Lindsten, F.: An efficient stochastic approximation EM algorithm using conditional particle filters. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp 6274–6278. IEEE (2013)

Lindström, E., Ionides, E., Frydendall, J., Madsen, H.: Efficient iterated filtering. IFAC Proc. Vol. **45**(16), 1785–1790 (2012)

Lindström, E., Madsen, H., Nielsen, J.N.: Statistics for Finance. Chapman and Hall, London (2015)

Loscalzo, F., Talbot, T.: Spline function approximations for solutions of ordinary differential equations. SIAM J. Numer. Anal. **4**, 3 (1967)

Magnani, E., Kersting, H., Schober, M., Hennig, P.: Bayesian filtering for ODEs with bounded derivatives (2017). arXiv:1709.08471 [csNA]

McNamee, J., Stenger, F.: Construction of fully symmetric numerical integration formulas. Numer. Math. **10**(4), 327–344 (1967)

Øksendal, B.: Stochastic Differential Equations: An Introduction with Applications, 5th edn. Springer, Berlin (2003)

Paul, S., Chatzilygeroudis, K., Ciosek, K., Mouret, J.B., Osborne, M.A., Whiteson, S.: Alternating optimisation and quadrature for robust control. In: AAAI Conference on Artificial Intelligence (2018)

Prüher, J., Šimandl, M.: Bayesian quadrature in nonlinear filtering. In: 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), vol. 01, pp. 380–387 (2015)

Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)

Ritter, K.: Average-Case Analysis of Numerical Problems. Springer, Berlin (2000)

Rosenbrock, H.H.: Some general implicit processes for the numerical solution of differential equations. Comput. J. **5**(4), 329–330 (1963)

Särkkä, S.: Recursive Bayesian inference on stochastic differential equations. Ph.D. thesis, Helsinki University of Technology (2006)

Särkkä, S.: Bayesian Filtering and Smoothing. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge (2013)

Schober, M., Duvenaud, D., David, K., Hennig, P.: Probabilistic ODE solvers with Runge–Kutta means. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27 (NIPS), pp. 739–747. Curran Associates, Inc. (2014)

Schober, M., Särkkä, S., Hennig, P.: A probabilistic model for the numerical solution of initial value problems. Stat. Comput. **29**(1), 99–122 (2019)

Schön, T.B., Wills, A., Ninness, B.: System identification of nonlinear state-space models. Automatica **47**(1), 39–49 (2011)

Schweppe, F.: Evaluation of likelihood functions for Gaussian signals. IEEE Trans. Inf. Theory **11**(1), 61–70 (1965)

Skilling, J.: Bayesian solution of ordinary differential equations. In: Smith, C.R., Erickson, G.J., Neudorfer, P.O. (eds.) Maximum Entropy and Bayesian Methods, pp. 23–37. Springer, Dordrecht (1992)

Storvik, G.: Particle filters for state-space models with the presence of unknown static parameters. IEEE Trans. Signal Process. **50**(2), 281–289 (2002)

Taniguchi, A., Fujimoto, K, Nishida, Y.: On variational Bayes for identification of nonlinear state-space models with linearly dependent unknown parameters. In: 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), pp. 572–576. IEEE (2017)

Teymur, O., Zygalakis, K., Calderhead, B.: Probabilistic linear multistep methods. In: Lee, D.D., Sugiyama, M. Luxburg, U.V. Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29 (NIPS), pp. 4321–4328. Curran Associates, Inc. (2016)

Teymur, O., Lie, HC., Sullivan, T., Calderhead, B.: Implicit probabilistic integrators for ODEs. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31 (NIPS), pp. 7244–7253. Curran Associates, Inc. (2018)

Tronarp, F., Garcia-Fernandez, A.F., Särkkä, S.: Iterative filtering and smoothing in non-linear and non-Gaussian systems using conditional moments. IEEE Signal Process. Lett. **25**(3), 408–412 (2018). https://doi.org/10.1109/LSP.2018.2794767

Tronarp, F., Karvonen, T., Särkkä, S.: Student's t-filters for noise scale estimation. IEEE Signal Process. Lett. **26**(2), 352–356 (2019)

Wang, J., Cockayne, J., Oates, C.: On the Bayesian solution of differential equations. In: Proceedings of the 38th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (2018)

Zhang, J., Mokhtari, A., Sra, S., Jadbabaie, A.: Direct Runge–Kutta discretization achieves acceleration. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Prcessing Systems 31 (NIPS), pp. 3900–3909. Curran Associates, Inc. (2018)