Bayesian ODE Solvers: The Maximum A Posteriori Estimate

Filip Tronarp[1], Simo Särkkä[2], and Philipp Hennig[1,3]

[1]University of Tübingen
[2]Department of Electrical Engineering and Automation, Aalto University
[3]MPI for Intelligent Systems, Tübingen

April 3, 2020

Abstract

It has recently been established that the numerical solution of ordinary differential equations can be posed as a nonlinear Bayesian inference problem, which can be approximately solved via Gaussian filtering and smoothing, whenever a Gauss–Markov prior is used. In this paper the class of ν times differentiable linear time invariant Gauss–Markov priors is considered. A taxonomy of Gaussian estimators is established, with the maximum a posteriori estimate at the top of the hierarchy, which can be computed with the iterated extended Kalman smoother. The remaining three classes are termed explicit, semi-implicit, and implicit, which are in similarity with the classical notions corresponding to conditions on the vector field, under which the filter update produces a *local maximum a posteriori estimate*. The maximum a posteriori estimate corresponds to an optimal interpolant in the reproducing Hilbert space associated with the prior, which in the present case is equivalent to a Sobolev space of smoothness $\nu+1$. Consequently, using methods from scattered data approximation and nonlinear analysis in Sobolev spaces, it is shown that the maximum a posteriori estimate converges to the true solution at a polynomial rate in the fill-distance (maximum step size) subject to mild conditions on the vector field. The methodology developed provides a novel and more natural approach to study the convergence of these estimators than classical methods of convergence analysis. The methods and theoretical results are demonstrated in numerical examples.

1 Introduction

Let $\mathbb{T} = [0, T],\ T < \infty,\ f\colon \mathbb{T} \times \mathbb{R}^d \to \mathbb{R}^d,\ y_0 \in \mathbb{R}^d$ and consider the following ordinary differential equation (ODE):

$$Dy(t) = f(t, y(t)),\ y(0) = y_0, \tag{1}$$

where D denotes the time derivative operator. Approximately solving (1) on a discrete mesh $\mathbb{T}_N = \{t_n\}_{n=0}^N,\ 0 = t_0 < t_1 < \ldots < t_N = T$, involves finding a function $\hat{y}$ such that $\hat{y}(t_n) \approx y(t_n),\ n = 0, 1, \ldots, N$ and a procedure for finding $\hat{y}$ is called a *numerical solver*. This is an important problem in science and engineering, and vast base of knowledge has accumulated, as summarised by, for example, Deuflhard and Bornemann (2002), Hairer et al. (1987), Hairer and Wanner (1996), and Butcher (2008).

Classically, the error of a numerical solver has been quantified in terms of the worst case error. However, in applications where a numerical solution is sought as a component of

a larger statistical inference problem (see, e.g., Matsuda and Miyatake 2019 and Kersting et al. 2020), it is desirable that the error can be quantified with the same semantic, that is to say, *probabilistically* (Hennig et al., 2015, Oates and Sullivan, 2019). Hence the recent endeavour to develop probabilistic ODE solvers.

Probabilistic ODE solvers can roughly be divided into two classes, sampling based solvers and deterministic solvers. The former class includes classical ODE solvers that are stochastically perturbed (Abdulle and Garegnani, 2020, Conrad et al., 2017, Lie et al., 2019, Teymur et al., 2018, 2016), solvers that approximately sample from a Bayesian inference problem (Tronarp et al., 2019b), and solvers that perform Gaussian process regression on stochastically generated data (Chkrebtii et al., 2016). On the other hand, deterministic solvers formulate the problem as a Gaussian process regression problem, either with a data generation mechanism (Hennig and Hauberg, 2014, Kersting and Hennig, 2016, Magnani et al., 2017, Schober et al., 2014, 2019, Skilling, 1992) or by attempting to constrain the estimate to solve the ODE at each point on the mesh (John et al., 2019, Tronarp et al., 2019b). For computational reasons it is fruitful to select the Gaussian process prior to be of Markov type (Kersting and Hennig, 2016, Magnani et al., 2017, Schober et al., 2019, Tronarp et al., 2019b), as this reduces cost of inference from $O(N^3)$ to $O(N)$ (Hartikainen and Särkkä, 2010, Särkkä et al., 2013). Because of the connection between inference with Gauss–Markov processes priors and spline interpolation (Kimeldorf and Wahba, 1970, Sidhu and Weinert, 1979, Weinert and Kailath, 1974), the Gaussian process regression approaches are intimately connected with the spline approach to ODEs (Schumaker, 1982, Wahba, 1973).

The notion of Bayesian solvers was defined by Cockayne et al. (2019), which poses the approximation to the solution of the ODE as a Bayesian inference problem. Under particular conditions on the vector field, the solvers of Kersting and Hennig (2016), Magnani et al. (2017), Schober et al. (2019), and Tronarp et al. (2019b) produce the exact posterior, if in addition a smoothing recursion is implemented, which corresponds to solving the batch problem as posed by John et al. (2019). In some cases, the exact Bayesian solution can also be obtained by exploiting Lie theory (Wang et al., 2018).

In this paper, the Bayesian formalism of Cockayne et al. (2019) is adopted for probabilistic solvers and priors of Gauss–Markov type are considered. However, rather than the exact posterior, the maximum a posteriori (MAP) estimate is studied. Many of the aforementioned Gaussian inference approaches are related to the MAP estimate. They can similarly to classical solvers, be classified as explicit (Schober et al., 2019), semi-implicit (Tronarp et al., 2019b), and implicit, which correspond to cases under which conditions they produce the exact posterior. Due to the Gauss–Markov prior, the method of John et al. (2019) can be implemented efficiently by the extended Kalman smoother (Bell, 1994). Furthermore, the Gauss–Markov prior corresponds to a reproducing kernel Hilbert space (RKHS) of Sobolev type and the MAP estimate is equivalent to the optimum of a minimum norm problem with nonlinear constraints in the RKHS. This enables the use of results from scattered data approximation (Arcangéli et al., 2007, Wendland and Rieger, 2005) to establish, under mild conditions, that the MAP estimate converges to the true solution at a high polynomial rate in terms of the fill-distance (or equivalently, the maximum step size).

The rest of the paper is organised as follows. In Section 2, the solution of the ODE (1) is formulated as a Bayesian inference problem and the associated MAP problem is stated. In Section 3, various methods for inference, which are based on Gaussian filtering and smoothing are presented. In the context of ODE solvers, two new variants are introduced, which are based on the iterated extended Kalman filter (Bell and Cathey, 1993) and the

iterated extended Kalman smoother (Bell, 1994), respectively. In Section 4, the connection between MAP estimation and optimisation in a certain reproducing kernel Hilbert space is reviewed. In Section 6, the error of the MAP estimate is analysed, for which polynomial convergence rates in the fill-distance are obtained. These rates are demonstrated in Section 7, and the paper is finally concluded by a discussion in Section 8.

1.1 Notation

Let $\Omega \subset \mathbb{R}$, then for a (weakly) differentiable function $u\colon \Omega \to \mathbb{R}^d$, its (weak) derivative is denoted by Du, or sometimes $\dot{u}$. The space of m times continuously differentiable functions from Ω to $\mathbb{R}^d$ is denoted by $C^m(\Omega, \mathbb{R}^d)$. The space of absolutely continuous functions is denoted by $\mathrm{AC}(\Omega, \mathbb{R}^d)$. The vector valued Lesbegue spaces are denoted by $\mathcal{L}_p(\Omega, \mathbb{R}^d)$ and the related Sobolev spaces of m times weakly differentiable functions are denoted by $H_p^m(\Omega, \mathbb{R}^d)$, that is, if $u \in H_p^m(\Omega, \mathbb{R}^d)$ then $D^m u \in \mathcal{L}_p(\Omega, \mathbb{R}^d)$. The norm of $y \in \mathcal{L}_p(\Omega, \mathbb{R}^d)$ is given by

$$\|y\|_{\mathcal{L}_p(\Omega,\mathbb{R}^d)} = \sum_{i=1}^{d} \|y_i\|_{\mathcal{L}_p(\Omega,\mathbb{R})} .$$

If $p = 2$, the equivalent norm

$$\|y\|_{\mathcal{L}_p(\Omega,\mathbb{R}^d)} = \sqrt{\sum_{i=1}^{d} \|y_i\|^2_{\mathcal{L}_p(\Omega,\mathbb{R})}}$$

is sometimes used. The Sobolev (semi-)norms are given by (Adams and Fournier, 2003, Valent, 2013)

$$\begin{aligned} |y|_{H_p^\alpha(\Omega,\mathbb{R})} &= \|D^\alpha y\|_{\mathcal{L}_p(\Omega,\mathbb{R})} , \\ \|y\|_{H_p^\alpha(\Omega,\mathbb{R})} &= \Big(\sum_{m=1}^{\alpha} |y|^p_{H_p^m(\Omega,\mathbb{R})} \Big)^{1/p}, \\ \|y\|_{H_p^\alpha(\Omega,\mathbb{R}^d)} &= \sum_{i=1}^{d} \|y_i\|_{H_p^\alpha(\Omega,\mathbb{R})} , \end{aligned}$$

an equivalent norm on $H_p^\alpha(\Omega, \mathbb{R}^d)$ is

$$\|y\|'_{H_p^\alpha(\Omega,\mathbb{R}^d)} = \Big(\sum_{i=1}^{d} \|y_i\|^p_{H_p^\alpha(\Omega,\mathbb{R})} \Big)^{1/p} .$$

Henceforth the domain and codomain of the function spaces will be omitted unless required for clarity. Furthermore, for a function ϕ with domain $\Omega \subset \mathbb{R}$, its left-limit at t is denoted by

$$\phi(t^-) = \lim_{s \uparrow t} \phi(t). \tag{3}$$

For a positive definite matrix Σ, its symmetric square root is denoted by $\Sigma^{1/2}$, and the associated Mahalanobis norm of a vector a is denoted by $\|a\|_\Sigma = a^\mathsf{T} \Sigma^{-1} a$.

2 A Probabilistic State-Space Model

The present approach involves defining a probabilistic state-space model, from which the approximate solution to (1) is inferred. This is essentially the same approach as that of Tronarp et al. (2019b). The class of priors considered is defined in Section 2.1 and the data model is introduced in Section 2.2.

2.1 The Prior

Let ν be a positive integer, $F_m \in \mathbb{R}^{d\times d}$, $0 \leq m \leq \nu$ and, $\Gamma \in \mathbb{R}^{d\times d}$ a positive definite matrix, and define the following differential operator:

$$\mathcal{D} = \Gamma^{-1/2}\Big(\mathrm{I}_d D^{\nu+1} - \sum_{m=0}^{\nu} F_m D^m\Big). \tag{4}$$

The class of priors considered herein is then given by

$$Y(t) = \mathrm{E}_0^\mathsf{T} \exp(Ft)X(0) + \int_0^T G_Y(t,\tau)\,\mathrm{d}W(\tau), \tag{5}$$

where W is a standard Wiener process onto $\mathbb{R}^d$, $X(0) \sim \mathcal{N}(0, \Sigma(t_0^-))$, and G_Y is the Green's function associated with $\mathcal{D}$ on $\mathbb{T}$ with initial condition $D^m y(t_0) = 0$, $m = 0, \ldots, \nu$. The Green's function is given by

$$G_Y(t,\tau) = \mathrm{E}_0^\mathsf{T} G_X(t,\tau), \tag{6a}$$

$$G_X(t,\tau) = \theta(t-\tau)\exp(F(t-\tau))\mathrm{E}_\nu\Gamma^{1/2}, \tag{6b}$$

where $\mathrm{E}_m = \mathrm{e}_m \otimes \mathrm{I}_d$, $m = 0, \ldots, \nu$, $\{\mathrm{e}_m\}_{m=0}^{\nu}$ is the canonical basis on $\mathbb{R}^{\nu+1}$, I_d is the identity matrix in $\mathbb{R}^{d\times d}$, θ is Heaviside's step function, and $F \in \mathbb{R}^{d(\nu+1)\times d(\nu+1)}$ whose $d \times d$ blocks are given by

$$F_{ij} = \begin{cases} \mathrm{I}_d,\ j = i+1,\ 0 \leq i,j < \nu, \\ 0,\ j \neq i+1,\ 0 \leq i,j < \nu, \\ F_j,\ i = \nu,\ 0 \leq j \leq \nu. \end{cases}$$

By construction, (5) has a state-space representation, which is given by the following stochastic differential equation (Øksendal, 2003)

$$\mathrm{d}X(t) = FX(t)\,\mathrm{d}t + \mathrm{E}_\nu\Gamma^{1/2}\,\mathrm{d}W(t),\ X(0) \sim \mathcal{N}(0, \Sigma(t_0^-)), \tag{7}$$

where X takes values in $\mathbb{R}^{d(\nu+1)}$ and the mth sub-vector of X is given by $X^m = D^m Y$ and takes values in $\mathbb{R}^d$ for $0 \leq m \leq \nu$. The transition densities for X are given by (Särkkä and Solin, 2019)

$$X(t+h) \mid X(t) \sim \mathcal{N}(A(h)X(t), Q(h)), \tag{8}$$

where

$$A(h) = \exp(Fh), \tag{9a}$$

$$Q(h) = \int_0^T G_X(h,\tau)G_X^\mathsf{T}(h,\tau)\,\mathrm{d}\tau. \tag{9b}$$

2.1.1 The Selection of Prior

Selecting the prior can be quite an intricate task. While ν determines the smoothness of the prior, the actual estimator will be of smoothness $\nu + 1$ (see Section 4) and the convergence results of Section 6 pertain to the case when the solution is of smoothness $\nu + 1$ as well. Consequently, if it is known that the solution is of smoothness $\alpha \geq 2$ then setting $\nu = \alpha - 1$ ensures the present convergence guarantees are in effect. Though it is likely convergence rates can be obtained for priors that are "too smooth" as well (see Kanagawa et al. 2020 for such results pertaining to numerical integration).

Once the degree of smoothness ν has been selected, the parameters $\Sigma(t_0^-)$, $\{F_m\}_{m=0}^{\nu}$, and Γ need to be selected. Some common sub-families of priors are listed below.

- (Released ν times integrated Wiener process onto $\mathbb{R}^d$). The process Y is a ν times integrated Wiener process if $F_m = 0,\ m = 1, \ldots, \nu$. The parameters $\Sigma(t_0^-)$ and Γ are free. Though it is advisable to set $\Gamma = \sigma^2 \mathrm{I}_d$ for some scalar σ^2. In this case σ^2 can be fit (estimated) to the particular ODE being solved (see Section 5.1). This class of processes is denoted by $Y \sim \mathrm{IWP}(\Gamma, \nu)$.

- (ν times integrated Ornstein–Uhlenbeck process onto $\mathbb{R}^d$). The process Y is a ν times integrated Ornstein–Uhlenbeck process if $F_m = 0,\ m = 1, \ldots, \nu - 1$. The parameters $\Sigma(t_0^-)$, F_ν, and Γ are free. As with $\mathrm{IWP}(\Gamma, \nu)$, it is advisable to set $\Gamma = \sigma^2 \mathrm{I}_d$. These processes are denoted by $Y \sim \mathrm{IOUP}(F_\nu, \Gamma, \nu)$.

- (Mateŕn processes of smoothness ν onto $\mathbb{R}$). If $d = 1$ then Y is a Mateŕn process of smoothness ν if (cf. Hartikainen and Särkkä 2010)

$$\begin{aligned} F_m &= -\binom{\nu+1}{m} \lambda^{\nu+1-m}, \quad m = 0, \ldots, \nu, \\ \Gamma &= 2\sigma^2 \lambda^{2\nu+1}, \end{aligned}$$

  for some $\lambda, \sigma^2 > 0$, and $\Sigma(t_0^-)$ is set to the stationary covariance matrix of the resulting X process. If $d > 1$ then each coordinate of the solution can be modelled by an individual Mateŕn process.

Remark 1. *Many popular choices of Gaussian processes not mentioned here also have state-space representations or can be approximated by a state-space model (Hartikainen and Särkkä, 2010, Karvonen and Sarkkä, 2016, Solin and Särkkä, 2014, Tronarp et al., 2018b). A notable example is Gaussian processes with squared exponential kernel (Hartikainen and Särkkä, 2010). See Chapter 12 of Särkkä and Solin (2019), for a thorough exposition.*

The IWP class of priors corresponds to polynomial splines in the limit $\Sigma(t_0^-) \to \infty$ (Wahba, 1978), and produces methods that are intimately connected with classical Nordsieck methods (Schober et al., 2019). Hence it appears as a natural choice if no further information pertaining to the solution (1) is available. On the other hand, suppose the problem is semi-linear:

$$Dy(t) = \Lambda y(t) + \varepsilon(t, y(t)). \tag{11}$$

It follows that

$$D^{\nu+1} y(t) = \Lambda D^\nu y(t) + D^\nu \varepsilon(t, y(t)),$$

or in differential form:

$$\mathrm{d}D^\nu y(t) = \Lambda D^\nu y(t)\,\mathrm{d}t + D^\nu \varepsilon(t, y(t))\,\mathrm{d}t.$$

Consequently, $\mathrm{IOUP}(\Lambda, \Gamma, \nu)$ corresponds to modelling $D^\nu \varepsilon(t, y(t))\,\mathrm{d}t$ by the increment $\Gamma^{1/2}\,\mathrm{d}W(t)$, which may be a good idea when $D^\nu \varepsilon(t, y(t))$ is expected to be small along solution curves. The IOUP class of priors has previously been investigated for "ODEs with bounded derivatives" by Magnani et al. (2017).

2.2 The Data Model

For the Bayesian formulation of probabilistic numerical methods, the data model is defined in terms of an *information operator* (Cockayne et al., 2019). In this paper, the information operator is given by

$$\mathcal{Z} = D - \mathcal{S}_f, \tag{12}$$

where $\mathcal{S}_f$ is the Nemytsky operator associated with the vector field f (Marcus and Mizel, 1973),[1] that is,

$$\mathcal{S}_f[y](t) = f(t, y(t)). \tag{13}$$

Clearly, $\mathcal{Z}$ maps the solution of (1) to a known quantity, the zero function. Consequently, inferring Y reduces to conditioning on

$$\mathcal{Z}[Y](t) = 0, \quad t \in \mathbb{T}_N.$$

The function $\mathcal{Z}[Y](t)$ can be expressed in simpler terms by use of the process X. That is, define the function

$$z(t, x) \coloneqq x^1 - f(t, x^0), \tag{14}$$

then

$$\mathcal{Z}[Y](t) = \mathcal{Z}[X^0](t) = \mathcal{S}_z[X](t) = z(t, X(t)). \tag{15}$$

Furthermore, it is necessary to account for the initial condition, $X^0(0) = y_0$, and with small additional cost the initial condition of the derivative can also be enforced $X^1(0) = f(0, y_0)$.

2.3 Maximum A Posteriori Estimation

The MAP estimate for Y restricted to $\mathbb{T}_N$ is given as the solution to the optimisation problem

$$\begin{aligned}
\max_{y(t_{0:N}), \dot{y}(t_{0:N}))} \quad & \log p(y(t_{0:N}), \dot{y}(t_{0:N})) && \text{(16a)} \\
\text{subject to} \quad & y(t_0) - y_0 = 0, && \text{(16b)} \\
& \dot{y}(t_0) - f(t_0, y_0) = 0, && \text{(16c)} \\
& \dot{y}(t_n) - f(t_n, y(t_n)) = 0, \ n = 1, \ldots, N, && \text{(16d)}
\end{aligned}$$

where $p(y(t_{0:N}), \dot{y}(t_{0:N}))$ is probability density of $(Y, \dot{Y})$ restricted to $\mathbb{T}_N$. However, since X is a Markov process it is advantageous to pose the equivalent MAP problem for X restricted to $\mathbb{T}_N$, which in view of (8) is given by

$$\begin{aligned}
\min_{x(t_{0:N})} \quad & \frac{1}{2}\left(\|x(t_0)\|^2_{\Sigma(t_0^-)} + \sum_{n=1}^{N} \|x(t_n) - A(h_n)x(t_{n-1})\|^2_{Q(h_n)} \right) && \text{(17a)} \\
\text{subject to} \quad & \mathrm{E}_0^{\mathsf{T}} x(t_0) - y_0 = 0, && \text{(17b)} \\
& \mathrm{E}_1^{\mathsf{T}} x(t_0) - f(t_0, y_0) = 0, && \text{(17c)} \\
& z(t_n, x(t_n)) = 0, \ n = 1, \ldots, N, && \text{(17d)}
\end{aligned}$$

where $h_n = t_n - t_{n-1}$ is the step size sequence and $\|\cdot\|_\Sigma$ is the Mahalanobis norm associated with the positive definite matrix Σ.

3 Gaussian Inference

In the previous section, the following probabilistic state-space model was defined, and in this section methods for inference are developed. All these methods are based on Gaussian filtering and smoothing (Särkkä, 2013, Särkkä and Solin, 2019), where the vector field is linearised in various ways. Some of these methods have already appeared in the literature

[1] Nemytsky operators are also known as composition operators and superposition operators.

(John et al., 2019, Schober et al., 2019, Tronarp et al., 2019b), while the iterative ones, as applied to solving ODEs, are new. Define the information sets

$$\mathscr{Z}(t) = \{z(\tau, X(\tau)) = 0 \colon \tau \in \mathbb{T}_N,\ \tau \leq t\},$$
$$\mathscr{Z}(t^-) = \{z(\tau, X(\tau)) = 0 \colon \tau \in \mathbb{T}_N,\ \tau < t\}.$$

In Gaussian filtering and smoothing, only (approximations of) the mean and covariance matrix of $X(t)$ are tracked. The (approximate) mean and covariance at time t, conditioned on $\mathscr{Z}(t)$ are denoted by $\mu_F(t)$ and $\Sigma_F(t)$, respectively, and $\mu_F(t^-)$ and $\Sigma_F(t^-)$ correspond to conditioning on $\mathscr{Z}(t^-)$, which are limits from the left[2]. The (approximate) mean and covariance conditioned on $\mathscr{Z}(T)$ at time t are denoted by $\mu_S(t)$ and $\Sigma_S(t)$, respectively.

Remark 2. *The mean vectors μ_F and μ_S contain estimates of $D^m y,\ m = 0, \ldots, \nu$ for the solution of* (1)*. The idea of tracking derivatives of the solution goes back to Nordsieck (1962). The connection between Nordsieck methods and Gaussian inference based solvers was discussed by Schober et al. (2019).*

3.1 Inference with Affine Vector Fields

While the information operator is not affine in general, all the methods that are discussed in the sequel are based on replacing it with an affine approximations via affine approximation of the vector field. This results in an affine and Gaussian approximation to the state-space model, for which the Bayesian filtering and smoothing, and consequently MAP problem can be solved exactly (Särkkä, 2013). Hence it is instructive to consider inference for the case of affine vector fields first as it provides the template for the approximate methods. That is, let the vector field be affine:

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

Then the information operator reduces to

$$z(t, x) = x^1 - \Lambda(t)x^0 - \zeta(t),$$

and the inference problem reduces to Gaussian process regression (Rasmussen and Williams, 2006) with a linear combination of function and derivative observations. In the spline literature this is known as (extended) Hermite–Birkhoff data (Sidhu and Weinert, 1979). In this case, the inference problem can be solved exactly with Gaussian filtering and smoothing (Kalman and Bucy, 1961, Kalman, 1960, Rauch et al., 1965, Särkkä, 2013, Särkkä and Solin, 2019), which is reviewed in the following.

Before starting the filtering and smoothing recursions, the process X needs to be conditioned on the initial values

$$\mathrm{E}_0^\mathsf{T} X(0) = y_0, \tag{19a}$$
$$\mathrm{E}_1^\mathsf{T} X(0) = f(t_0, y_0). \tag{19b}$$
$$\tag{19c}$$

[2]Recall that the filtering distribution is right continuous with left limits. More specifically, $\mu_F(t^-) = \mu_F(t)$ and $\Sigma_F(t^-) = \Sigma_F(t)$, unless $t \in \mathbb{T}_N$, where they jump.

This is can be done by a Kalman update

$$C(t_0) = \begin{pmatrix} \mathrm{E}_0^\mathsf{T} \\ \mathrm{E}_1^\mathsf{T} \end{pmatrix}, \tag{20a}$$

$$S(t_0) = C(t_0)\Sigma(t_0^-)C^\mathsf{T}(t_0), \tag{20b}$$

$$K(t_0) = \Sigma(t_0^-)C^\mathsf{T}(t_0)S^{-1}(t_0), \tag{20c}$$

$$\mu_F(t_0) = K(t_0)\begin{pmatrix} y_0 \\ f(t_0, y_0) \end{pmatrix}, \tag{20d}$$

$$\Sigma_F(t_0) = \Sigma(t_0^-) - K(t_0)S(t_0)K^\mathsf{T}(t_0). \tag{20e}$$

The filtering mean and covariance for each interval $[t_{n-1}, t_n)$ and $n = 1, \ldots, N$ is governed by

$$\dot{\mu}_F(t) = F\mu_F(t), \tag{21a}$$

$$\dot{\Sigma}_F(t) = F\Sigma_F(t) + \Sigma_F(t)F^\mathsf{T} + \mathrm{E}_\nu \Gamma \mathrm{E}_\nu^\mathsf{T}, \tag{21b}$$

which on the mesh is solved by

$$\mu_F(t_n^-) = A(h_n)\mu_F(t_{n-1}), \tag{22a}$$

$$\Sigma_F(t_n^-) = A(h_n)\Sigma_F(t_{n-1})A^\mathsf{T}(h_n) + Q(h_n), \tag{22b}$$

where $h_n = t_n - t_{n-1},\ n = 1, \ldots, N$ is the step size sequence. The prediction moments at $t \in \mathbb{T}_N$ are then corrected according to the Kalman update

$$C(t_n) = \mathrm{E}_1^\mathsf{T} - \Lambda(t_n)\mathrm{E}_0^\mathsf{T}, \tag{23a}$$

$$S(t_n) = C(t_n)\Sigma_F(t_n^-)C^\mathsf{T}(t_n), \tag{23b}$$

$$K(t_n) = \Sigma_F(t_n^-)C^\mathsf{T}(t_n)S^{-1}(t_n), \tag{23c}$$

$$\mu_F(t_n) = \mu_F(t_n^-) + K(t_n)[\zeta(t_n) - C(t_n)\mu_F(t_n^-)], \tag{23d}$$

$$\Sigma_F(t_n) = \Sigma_F(t_n^-) - K(t_n)S(t_n)K^\mathsf{T}(t_n). \tag{23e}$$

The smoothing mean and covariance are continuous and evolve according to

$$G_c(t) = \mathrm{E}_\nu \Gamma \mathrm{E}_\nu^\mathsf{T} \Sigma_F^{-1}(t), \tag{24a}$$

$$\dot{\mu}_S(t) = F\mu_S(t) + G_c(t)(\mu_S(t) - \mu_F(t)), \tag{24b}$$

$$\dot{\Sigma}_S(t) = [F + G_c(t)]\Sigma_S(t) + \Sigma_S(t)[F + G_c(t)]^\mathsf{T} - \mathrm{E}_\nu \Gamma \mathrm{E}_\nu^\mathsf{T}, \tag{24c}$$

with terminal conditions $\mu_S(t_N) = \mu_F(t_N)$, and $\Sigma_S(t_N) = \Sigma_F(t_N)$. On the mesh $\mathbb{T}_N$ the smoothing moments are given by

$$G(t_n) = \Sigma_F(t_n)A^\mathsf{T}(h_{n+1})\Sigma_F^{-1}(t_{n+1}^-), \tag{25a}$$

$$\mu_S(t_n) = \mu_F(t_n) + G(t_n)(\mu_S(t_{n+1}) - \mu_F(t_{n+1}^-)), \tag{25b}$$

$$\Sigma_S(t_n) = \Sigma_F(t_n) + G(t_n)\Big[\Sigma_S(t_{n+1}) - \Sigma_F(t_{n+1}^-)\Big]G^\mathsf{T}(t_n). \tag{25c}$$

While Gaussian filtering and smoothing only provides the posterior for affine vector fields, it forms the template for nonlinear problems as well. That is, the vector field is the vector field is replaced by an affine approximation. Approaches for doing this are discussed in the following.

3.2 Approximate Gaussian Inference

For non-affine vector fields, only the update becomes intractable. Approximation methods involve different ways of approximating the vector field with an affine function

$$f(t, y) \approx \hat{\Lambda}(t)y + \hat{\zeta}(t),$$

whereafter approximate filter means and covariances are obtained by plugging $\hat{\Lambda}$ and $\hat{\zeta}$ into (23). In the present setting, only Taylor series methods are considered, some of which have already appeared in the literature (Schober et al., 2019, Tronarp et al., 2019b). Though, there exists other approximation methods based on statistical linear regression (García-Fernández et al., 2015, Lefebvre et al., 2002, Tronarp et al., 2018a), which are based on cubature integration (Kersting and Hennig, 2016, Tronarp et al., 2019b). For classification purposes it is fruitful to study these methods in terms how well they approximate Bayes' rule in terms of the local MAP problem:

$$\min_{x} \quad \|x - \mu_F(t_n^-)\|^2_{\Sigma(t_n^-)} \tag{26a}$$

$$\text{subject to} \quad z(t_n, x) = 0. \tag{26b}$$

The method of classification is then provided by Definition 1.

Definition 1. *A Gaussian solver is said to be one of the following.*

- *Explicit if $\mu_F(t_n)$ is the solution to* (26) *when $f(t, y) = \zeta(t)$ for some ζ.*
- *Semi-implcit if $\mu_F(t_n)$ is the solution to* (26) *when $f(t, y) = \Lambda(t)y + \zeta(t)$ for some Λ and ζ.*
- *Implicit if $\mu_F(t_n)$ is the solution to* (26)*.*

Remark 3. *If $f(t, y) = \Lambda(t)y + \zeta(t)$ then semi-implicit methods solve the global MAP problem* (17) *as well. This also holds for explicit methods when $f(t, y) = \zeta(t)$.*

Remark 4. *Definition 1 is merely an analogue to the classifications of standard numerical analysis (Hairer and Wanner, 1996). Classically, implicit methods have to solve a root-finding problem, which is solved exactly for linear vector fields by semi-implicit methods, and is also typically solved exactly for vector fields that are constant in y by explicit methods. Here the root finding problem is replaced by the constrained minimisation problem, namely the local MAP problem* (26)*.*

The simplest approach is to make a zeroth order approximation of f around $\mu_F(t_n^-)$, which is due to Schober et al. (2019), and the linearisation parameters are given by

$$\hat{\Lambda}(t) = 0, \tag{27a}$$

$$\hat{\zeta}(t) = f(t_n, \mathrm{E}_0^\mathsf{T} \mu_F(t_n^-)). \tag{27b}$$

If the vector field is constant in y, $f(t, y) = \zeta(t)$, then it is clear that the approximation (27) is exact and the local MAP problem (26) is solved exactly. Therefore this is an explicit method.

The next best approach is to make a first order approximation around $\mu_F(t_n^-)$ (Tronarp et al., 2019b), and the linearisation parameters are given by

$$\hat{\Lambda}(t_n) = J_f(t_n, \mathrm{E}_0^\mathsf{T} \mu_F(t_n^-)), \tag{28a}$$

$$\hat{\zeta}(t_n) = f(t_n, \mathrm{E}_0^\mathsf{T} \mu_F(t_n^-)) - J_f(t_n, \mathrm{E}_0^\mathsf{T} \mu_F(t_n^-)) \mathrm{E}_0^\mathsf{T} \mu_F(t_n^-), \tag{28b}$$

where J_f is the Jacobian of f with respect to the second argument. This method is referred to as the extended Kalman filter (EKF) in signal processing literature (Särkkä, 2013), and was used to design probabilistic ODE solvers by Tronarp et al. (2019b). Clearly that the approximation (28) is exact if f is affine in y. Consequently, this method is semi-implicit.

It is not possible to utilise Taylor series expansions to get exact inference for more general classes of vector fields. However, improvements can be made by iteratively re-linearising the vector field at the filter update. Let $\mu_F^0(t_n) = \mu_F(t_n^-)$, $\Sigma_F^0(t_n) = \Sigma_F(t_n^-)$ and

$$\hat{\Lambda}^l(t_n) = J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F^l(t_n)), \tag{29a}$$

$$\hat{\zeta}^l(t_n) = f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F^l(t_n)) - J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F^l(t_n))\mathrm{E}_0^\mathsf{T}\mu_F^l(t_n). \tag{29b}$$

Inserting these parameters into (23) then gives $\mu_F^{l+1}(t_n)$ and $\Sigma_F^{l+1}(t_n)$, which leads to the iterated extended Kalman filter (IEKF) (Bell and Cathey, 1993). It is easy to show that $z(t_n, \mu_F(t_n))$ holds at the fixed point, therefore this method is implicit. However, for the smoothing mean $z(t_n, \mu_S(t_n)) = 0$ will in general only hold when the vector field is affine.

By a slight modification of the IEKF, namely re-linearising around the smoothing mean rather than the filtering mean, the iterated extended Kalman smoother (IEKS) is retrieved

$$\hat{\Lambda}^l(t_n) = J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_S^l(t_n)), \tag{30a}$$

$$\hat{\zeta}^l(t_n) = f(t_n, \mathrm{E}_0^\mathsf{T}\mu_S^l(t_n)) - J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_S^l(t_n))\mathrm{E}_0^\mathsf{T}\mu_S^l(t_n). \tag{30b}$$

The smoothing mean and covariance at iteration $l+1$, $\mu_S^{l+1}(t)$ and $\Sigma_S^{l+1}(t)$, are then obtained by running the filter and smoother with the parameters in (30). Furthermore, the IEKS is just the Gauss–Newton algorithm for the maximum a posteriori trajectory (Bell, 1994), consequently, at the fixed point

$$z(t, \mu_S(t)) = 0, \quad t \in \mathbb{T}_N,$$

and under some conditions on the Jacobian of the vector field it can be shown that the fixed-point is at least a local optimum to the MAP problem (17) (Knoth, 1989). Moreover, the IEKS is just a clever implementation of the method of John et al. (2019) whenever the prior process has a state-space representation.

4 The Reproducing Kernel Hilbert Space Perspective

The correspondence between inference in stochastic processes and optimisation in reproducing kernel Hilbert spaces is well known (Kimeldorf and Wahba, 1970, Sidhu and Weinert, 1979, Weinert and Kailath, 1974). This correspondence is indeed present in the current setting as well, in the sense that MAP estimation as discussed in Section 2.3 is equivalent to optimisation in the reproducing kernel Hilbert space (RKHS) associated with Y and X (see Kanagawa et al. 2018, Proposition 3.6 for standard Gaussian process regression). The purpose of this section is thus to establish that the RKHS associated with Y, which establishes what function space the estimators discussed in Section 3.2 lie in. Furthermore, it is shown that the MAP estimate is equivalent to an interpolation problem in this RKHS, which implies properties on its norm. These results will then be used in the convergence analysis of the MAP estimate in Section 6.

4.1 The Reproducing Kernel Hilbert Space of the Prior

The RKHS of the Wiener process with domain $\mathbb{T}$ and codomain $\mathbb{R}^d$ is the set (cf. van der Vaart and van Zanten 2008, section 10)

$$\mathbb{W}_0 = \{w\colon w \in \mathrm{AC}(\mathbb{T}, \mathbb{R}^d),\ w(0) = 0,\ \dot{w} \in \mathcal{L}_2(\mathbb{T}, \mathbb{R}^d)\},$$

with inner product given by

$$\langle w, w' \rangle_{\mathbb{W}_0} = \int_0^T \dot{w}^\mathsf{T}(\tau)\dot{w}'(\tau)\,\mathrm{d}\tau = \langle \dot{w}, \dot{w}' \rangle_{\mathcal{L}_2}.$$

Let $\mathbb{Y}^{\nu+1}$ denote the reproducing kernel Hilbert space associated with the prior process Y, then $\mathbb{Y}^{\nu+1}$ is given by (van der Vaart and van Zanten, 2008, lemmas 7.1, 8.1, and 9.1) the image of the operator

$$\mathcal{T}(\vec{y}_0, \dot{w}_y)(t) = E_0^\mathsf{T} \exp(Ft)\vec{y}_0 + \int_0^T G_Y(t, \tau)\dot{w}_y(\tau)\,\mathrm{d}\tau, \tag{31}$$

where $\vec{y}_0 \in \mathbb{R}^{d(\nu+1)}$ and $\dot{w}_y \in \mathcal{L}_2(\mathbb{T}, \mathbb{R}^d)$. That is,

$$\mathbb{Y}^{\nu+1} = \{y\colon y = \mathcal{T}(\vec{y}_0, \dot{w}_y),\ \vec{y}_0 \in \mathbb{R}^{d(\nu+1)},\ \dot{w}_y \in \mathcal{L}_2(\mathbb{T}, \mathbb{R}^d)\}, \tag{32}$$

and inner product is given by

$$\langle y, y' \rangle_{\mathbb{Y}^{\nu+1}} = \vec{y}_0^\mathsf{T}\Sigma^{-1}(t_0^-)\vec{y}_0' + \langle \mathcal{D}y, \mathcal{D}y' \rangle_{\mathcal{L}_2} = \vec{y}_0^\mathsf{T}\Sigma^{-1}(t_0^-)\vec{y}_0' + \langle \dot{w}_y, \dot{w}_y' \rangle_{\mathcal{L}_2}. \tag{33}$$

Since G_Y is the Green's function of a differential operator of order $\nu + 1$ with smooth coefficients, $\mathbb{Y}^{\nu+1}$ can be identified as follows. A function $y\colon \mathbb{T} \to \mathbb{R}^d$ is in $\mathbb{Y}^{\nu+1}$ if and only if

$$D^m y \in \mathrm{AC}(\mathbb{T}, \mathbb{R}^d),\ m = 0, \ldots, \nu, \tag{34a}$$

$$D^{\nu+1} y \in \mathcal{L}_2(\mathbb{T}, \mathbb{R}^d). \tag{34b}$$

Hence by similar arguments as for the released ν times integrated Wiener process, Proposition 1 holds (see proposition 2.6.24 and remark 2.6.25 of Giné and Nickl 2016).

Proposition 1. *The reproducing kernel Hilbert space $\mathbb{Y}^{\nu+1}$ as a set is equal to the Sobolev space $H_2^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$ and their norms are equivalent.*

The reproducing kernel of $\mathbb{Y}^{\nu+1}$ is given by (cf. Sidhu and Weinert 1979)

$$R(t, s) = E_0^\mathsf{T} \exp(Ft)\Sigma(t_0^-) \exp(F^\mathsf{T}s)E_0 + \int_0^T G_Y(t, \tau)G_Y^\mathsf{T}(s, \tau)\,\mathrm{d}\tau, \tag{35}$$

which is also the covariance function of Y. The linear functionals

$$y \mapsto v^\mathsf{T} D^m y(s), \quad v \in \mathbb{R}^d,\ t \in \mathbb{T},\ m = 0, \ldots, \nu,$$

are continuous and their representers are given by

$$\eta_s^{m,v} = R^{(0,m)}(t, s)v,$$
$$\langle \eta_s^{m,v}, y \rangle_{\mathbb{Y}^{\nu+1}} = v^\mathsf{T} D^m y(s),$$

where $R^{(m,k)}$ denotes R differentiated m and k times with respect to the first and second arguments, respectively. Furthermore, define the matrix

$$\eta_s^m = \begin{pmatrix} \eta_s^{m,\mathrm{e}_1} & \dots & \eta_s^{m,\mathrm{e}_d} \end{pmatrix},$$

and with notation overloaded in the obvious way, the following identities hold

$$D^m y(t) = \langle \eta_t^m, y \rangle_{\mathbb{Y}^{\nu+1}},$$
$$R^{(m,k)}(t,s) = \langle \eta_t^m, \eta_s^k \rangle_{\mathbb{Y}^{\nu+1}}.$$

Since there is a one-to-one correspondence between the processes Y and X, the RKHS associated with X is isometrically isomorphic to $\mathbb{Y}^{\nu+1}$, and it is given by

$$\mathbb{X}^{\nu+1} = \{x \colon x^0 \in \mathbb{Y}^{\nu+1},\ x^m = D^m x^0,\ m = 1, \dots, \nu+1\},$$

where x^m is the mth sub-vector of x of dimension d. The kernel associated with $\mathbb{X}^{\nu+1}$ is given by

$$P(t,s) = \exp(Ft)\Sigma(t_0^-)\exp(F^{\mathsf{T}}s) + \int_0^T G_X(t,\tau) G_X^{\mathsf{T}}(s,\tau)\,\mathrm{d}\tau, \tag{38}$$

and the $d \times d$ blocks of P are given by

$$P_{m,k}(t,s) = R^{(m,k)}(t,s),$$

and $\psi_s = P(t,s)$ is the representer of evaluation at s,

$$x(s) = \langle \psi_s, x \rangle_{\mathbb{X}^{\nu+1}}. \tag{39}$$

In the following, the short-hands $\mathbb{Y} = \mathbb{Y}^{\nu+1}$ and $\mathbb{X} = \mathbb{X}^{\nu+1}$ are in effect.

4.2 Nonlinear Kernel Interpolation

Now consider the kernel interpolation problem

$$\hat{y} = \operatorname*{arg\,min}_{y \in \mathcal{I}_N} \ \frac{1}{2}\|y\|_{\mathbb{Y}}^2, \tag{40}$$

where the feasible set is given by

$$\mathcal{I}_N = \{y \in \mathbb{Y} \colon y(0) = y_0,\ \dot{y}(0) = f(0, y_0),\ \mathcal{Z}[y](t) = 0,\ t \in \mathbb{T}_N\}. \tag{41}$$

Define the following subspaces $\mathbb{Y}$

$$\mathcal{R}_N(m) = \operatorname{span}\left\{\eta_{t_n}^{l,\mathrm{e}_i}\right\}_{l=0,n=0,i=1}^{m,N,d}, \quad m \leq \nu + 1.$$

Since $\mathcal{R}_N(m)$ is a closed linear sub-space of $\mathbb{Y}$ it follows that any $y \in \mathbb{Y}$ can be written as $y = y_\parallel + y_\perp$ with $y_\parallel \in \mathcal{R}_N(m)$ and $y_\perp \in \mathcal{R}_N^\perp(m)$, where $\mathcal{R}_N^\perp(m)$ is the orthogonal complement to $\mathcal{R}_N(m)$. Similarly to other situations (Cox and O'Sullivan, 1990, Girosi et al., 1995, Kimeldorf and Wahba, 1971) our optimum can be expanded in a finite sub-space spanned by representers, which is the statement of Proposition 2.

Proposition 2. *The solution to* (40) *is contained in* $\mathcal{R}_N(1)$.

Proof. Any $y \in \mathbb{Y}$ has the orthogonal decomposition $y = y_\parallel + y_\perp$, where $y_\parallel \in \mathcal{R}_N(1)$ and $y_\perp \in \mathcal{R}_N^\perp(1)$. However, it must be the case that $\|y_\perp\|_{\mathbb{Y}} = 0$, since

$$\frac{1}{2}\|y\|_{\mathbb{Y}}^2 = \frac{1}{2}\|y_\parallel\|_{\mathbb{Y}}^2 + \frac{1}{2}\|y_\perp\|_{\mathbb{Y}}^2 \geq \frac{1}{2}\|y_\parallel\|_{\mathbb{Y}}^2$$

and

$$\begin{aligned} D^m y(0) &= \langle \eta_0^m, y_\parallel \rangle_{\mathbb{Y}}, \quad m = 0, \ldots, \nu + 1, \\ \mathcal{Z}[y](t) &= Dy(t) - f(t, y(t)) = \langle \eta_t^1, y_\parallel \rangle_{\mathbb{Y}} - f\Big(t, \langle \eta_t^0, y_\parallel \rangle_{\mathbb{Y}}\Big), \end{aligned}$$

for all $t \in \mathbb{T}_N$. □

By Proposition 2 the optimal point of (40) can be written as

$$y = \sum_{n=0}^{N} \begin{pmatrix} \eta_{t_n}^0 & \eta_{t_n}^1 \end{pmatrix} \begin{pmatrix} b_0(t_n) \\ b_1(t_n) \end{pmatrix}.$$

However, it is more convenient to expand the optimal point in the larger subspace, $\mathcal{R}_N(\nu) \supset \mathcal{R}_N(1)$

$$b(t_n) = \begin{pmatrix} b_0^\mathsf{T}(t_n) & \ldots & b_\nu^\mathsf{T}(t_n) \end{pmatrix}^\mathsf{T}, \tag{43a}$$

$$y = \sum_{n=0}^{N} \begin{pmatrix} \eta_{t_n}^0 & \ldots & \eta_{t_n}^\nu \end{pmatrix} b(t_n), \tag{43b}$$

$$x = \sum_{n=0}^{N} \psi_{t_n} b(t_n), \tag{43c}$$

where x is the equivalent element in $\mathbb{X}$ and

$$\|y\|_{\mathbb{Y}}^2 = \|x\|_{\mathbb{X}}^2 = \sum_{n,m=0}^{N} b^\mathsf{T}(t_n) P(t_n, t_m) b(t_m), \tag{44}$$

or more compactly

$$\|x\|_{\mathbb{X}}^2 = \boldsymbol{x}^\mathsf{T} \boldsymbol{P}^{-1} \boldsymbol{x}, \tag{45}$$

where

$$\begin{aligned} \boldsymbol{x} &= \begin{pmatrix} x^\mathsf{T}(t_0) & \ldots & x^\mathsf{T}(t_N) \end{pmatrix}^\mathsf{T}, \\ \boldsymbol{P}_{n,m} &= P(t_n, t_m). \end{aligned}$$

Here $\boldsymbol{P}$ is the kernel matrix associated with function value observations of X at $\mathbb{T}_N$. That is, (45) is up to a constant equal to the negative log-density of X restricted to $\mathbb{T}_N$. Proposition 3 immediately follows.

Proposition 3. *The optimisation problem* (40) *is equivalent to the MAP problem* (17)*.*

Proof. Note that the kernel P associated with $\mathbb{X}$ (see (38)) is also is the covariance function for the X process. Consequently, $\|x\|_{\mathbb{X}}^2$ is up to some scaling and constant equal to the negative log-likelihood of X restricted to $\mathbb{T}_N$. Now X is a Markov process, so in view of (8) and (9)

$$\|x\|_{\mathbb{X}}^2 = \|x(t_0)\|_{\Sigma(t_0^-)}^2 + \sum_{n=1}^{N} \|x(t_n) - A(h_n)x(t_{n-1})\|_{Q(h_n)}^2,$$

and the conclusion follows. □

Remark 5. *The smoothing mean μ_S defined by* (24) *as produced by any of the approximate inference methods discussed in Section 3.2, defines an element in* $\mathbb{X}$.

5 Uncertainty Quantification and Calibration

An important aspect of the probabilistic approach to numerical analysis is that the method produces an error estimate in the language of probability. Any such error estimate is only meaningful if it can be related to the actual error of the method. Denote the true solution of the problem by $y^* \in \mathbb{Y}$, then the optimal interpolant can be written as a linear projection of y^* onto $\mathcal{R}_N(1)$, $\hat{y} = \Pi_N y^*$. Suppose an estimate of $L^\alpha y^*$ is sought for some class of continuous linear functional $\mathscr{L}^{\boldsymbol{\alpha}} = \{L^\alpha \colon \alpha \in \boldsymbol{\alpha}, L^\alpha \colon \mathbb{Y} \to \mathbb{R}\}$ indexed by some compact set $\boldsymbol{\alpha}$. Then by Cauchy–Schwartz inequality, the error can be written as

$$\begin{aligned} |L^\alpha y^* - L^\alpha \hat{y}| &= |\langle \eta^\alpha_{e,N}, y^* \rangle_{\mathbb{Y}}| \\ &\leq \|\eta^\alpha_{e,N}\|_{\mathbb{Y}} \, \|y^*\|_{\mathbb{Y}} \,, \end{aligned} \tag{47}$$

where $\eta^\alpha_{e,N}$ is the representer of $L^\alpha - L^\alpha \circ \Pi_N$. The norm of the error functional coincides with the posterior variance (see e.g., Briol et al. 2019 for the numerical integration case)

$$\mathbb{V}[L^\alpha Y(t) \mid \mathscr{Z}(t_N)] = \|\eta^\alpha_{e,N}\|_{\mathbb{Y}} \,. \tag{48}$$

Consequently, the worst case error for the class $\mathscr{L}^{\boldsymbol{\alpha}}$ in the unit ball of $\mathbb{Y}$ is

$$\sup_{\alpha \in \boldsymbol{\alpha}} \|\eta^\alpha_{e,N}\|_{\mathbb{Y}} = \sup_{\alpha \in \boldsymbol{\alpha}} \mathbb{V}[L^\alpha Y(t) \mid \mathscr{Z}(t_N)]. \tag{49}$$

Unfortunately it appears this line of reasoning generally breaks down in the present case. That is, if the vector field is not affine, then the optimal interpolant can not be expressed as a linear projection of the true solution, and the situation is significantly more complicated. However, in Section 6 bounds are obtained for the *nonlinear* functionals indexed by $t \in \mathbb{T}$ defined by

$$\mathbb{Y} \ni y \mapsto y(t) - y_0 - \int_0^t f(\tau, y(\tau)) \, \mathrm{d}\tau \tag{50}$$

and its derivatives up to order ν, though not in terms of the posterior variance.

In any case, as the error is generally problem dependent, the kernel parameters should be calibrated to the problem. How to do this for the noise scale of the prior is discussed in the sequel.

5.1 Calibrating the Noise Scale

For a full statistical treatment of the inference problem, the parameters F_m $m = 0, \ldots, \nu$, Γ and $\Sigma(t_0^-)$ need to be estimated. Of particular importance in terms of calibrating uncertainty properly are $\Sigma(t_0^-)$ and Γ (see (7)). As discussed in Section 6, the parameters F_m $m = 0, \ldots, \nu$ can have a significant impact on the constants appearing in the convergence rates of the MAP estimator. Nevertheless, the present discussion is just concerned with the calibration of uncertainty.

It can be shown that the logarithm of (quasi-) likelihood as produced by the Gaussian inference methods is, up to an unimportant constant, given by (cf. Tronarp et al. 2019a)

$$\begin{aligned} \ell = &-\frac{1}{2} \log \det S(t_0) - \frac{1}{2} \begin{pmatrix} y_0 \\ f(0, y_0) \end{pmatrix}^\intercal S^{-1}(t_0) \begin{pmatrix} y_0 \\ f(0, y_0) \end{pmatrix} \\ &- \frac{1}{2} \sum_{n=1}^N \log \det S(t_n) - \frac{1}{2} \sum_{n=1}^N \|\zeta(t_n) - C(t_n) \mu_F(t_n^-)\|^2_{S(t_n)} \,. \end{aligned}$$

Additionally, if $\Sigma(t_0^-) = \sigma^2 \breve{\Sigma}(t_0^-)$ and $\Gamma = \sigma^2 \breve{\Gamma}$ for some positive definite matrices $\breve{\Sigma}_F(t_0^-)$ and $\breve{\Gamma}$, then it can be shown that the log-likelihood, up to some unimportant constant, reduces to (see Appendix C of Tronarp et al. 2019b for details)[3]

$$\begin{aligned}\ell(\sigma) = &-\frac{d(N+2)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\begin{pmatrix} y_0 \\ f(0,y_0)\end{pmatrix}^{\mathsf{T}} \breve{S}^{-1}(t_0)\begin{pmatrix} y_0 \\ f(0,y_0)\end{pmatrix} \\ &- \frac{1}{2\sigma^2}\sum_{n=1}^{N} \|\zeta(t_n) - C(t_n)\mu_F(t_n^-)\|^2_{\breve{S}(t_n)},\end{aligned}$$

where $\breve{\cdot}$ denotes the output of the filter using the parameters $(\breve{\Sigma}(t_0^-), \breve{\Gamma})$ rather than $(\Sigma(t_0^-), \Gamma)$. This yields the following proposition, which is proven in Appendix C of Tronarp et al. (2019b), *mutatis mutandis*.

Proposition 4. *Let $\Sigma(t_0^-) = \sigma^2\breve{\Sigma}(t_0^-)$ and $\Gamma = \sigma^2\breve{\Gamma}$ for some positive definite matrices $\breve{\Sigma}(t_0^-)$ and $\breve{\Gamma}$, then the (quasi-) maximum likelihood estimate of σ^2 is given by*

$$\hat{\sigma}_N^2 = \frac{1}{d(N+2)}\left(\begin{pmatrix} y_0 \\ f(0,y_0)\end{pmatrix}^{\mathsf{T}} \breve{S}^{-1}(t_0)\begin{pmatrix} y_0 \\ f(0,y_0)\end{pmatrix} + \sum_{n=1}^{N}\|\zeta(t_n) - C(t_n)\mu_F(t_n^-)\|^2_{\breve{S}(t_n)}\right). \tag{51}$$

Bounds for worst case overconfidence and underconfidence under maximum likelihood estimation of σ^2 has recently been obtained by Karvonen et al. (2020). These results appear to carry over to the present setting for affine vector fields. However, it is not immediately clear how to generalise this to a larger class of vector fields.

6 Convergence Analysis

In this section, error bounds of the kernel interpolant $\hat{y}$ as defined by (40), and by Proposition 3 the MAP estimate is obtained. These bounds will be in terms of the fill-distance of the mesh $\mathbb{T}_N$, which is given by[4]

$$\delta = \sup_{t\in\mathbb{T}} \max_{n=0,\ldots,N} |t - t_n|. \tag{52}$$

In the following results from the scattered data approximation literature (Arcangéli et al., 2007, Wendland and Rieger, 2005) are employed. More specifically, for any $y \in \mathbb{Y}$, which satisfies the initial condition $y(0) = y_0$, formally has the following representation

$$y(t) = y_0 + \int_0^t f(\tau, y(\tau))\,\mathrm{d}\tau + \mathcal{E}[y](t),$$

where the error operator $\mathcal{E}$ is defined as

$$\mathcal{E}[y](t) = \int_0^t \mathcal{Z}[y](\tau)\,\mathrm{d}\tau.$$

[3] There is a slight difference in the log-likelihood expression from that of Tronarp et al. (2019b). This is because here the initial conditions are inferred while Tronarp et al. (2019b) encodes them directly in the prior.

[4] Classically the error of a numerical integrator is assessed in terms of the maximum step size which is twice the fill-distance.

Of course any reasonable estimator $\hat{y}'$ ought to have the property that $\mathcal{Z}[\hat{y}'](t) \approx 0$ for $t \in \mathbb{T}_N$. The approach is thus to bound $\mathcal{Z}[\hat{y}'](t)$ in some suitable norm, which in turn gives a bound on $\mathcal{E}[\hat{y}'](t)$.

Throughout the discussion $\nu \geq 1$ is some fixed integer, which corresponds to the differentiability of the prior, that is, the kernel interpolant is in $H_2^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$. Furthermore, some regularity of the vector field will be required, namely Assumption 1, given below.

Assumption 1. *The vector field f is in $C^{\nu+1}(\tilde{\mathbb{T}} \times \mathbb{R}^d, \mathbb{R}^d)$ for some set $\tilde{\mathbb{T}}$ with $\mathbb{T} \subset \tilde{\mathbb{T}} \subset \mathbb{R}$.*

Remark 6. *Because $f \in C^{\nu+1}(\mathbb{T} \times \mathbb{R}^d, \mathbb{R}^d)$, the derivatives $D^\alpha f$ are locally Lipschitz continuous for any multi-index α with $|\alpha| \leq \nu$. A convergence analysis of the filter based on the zeroth order linearisation (see Eq. (27)) was carried out by Kersting et al. (2018), where they assumed that f was in C^ν and $D^\alpha f$ Lipschitz continuous and bounded for any $|\alpha| \leq \nu$. That is, for the purposes of proving convergence of the MAP estimate, one extra degree of smoothness is imposed on f, while the rather strong assumptions on its derivatives are relaxed.*

Assumption 1 will, without explicit mention, be in force throughout the discussion of this section. Essentially, it implies that (i) the model is well specified and (ii) the information operator is well behaved. This shall be made precise in the following.

6.1 Model Correctness and Regularity of the Solution

Since $\nu \geq 1$, Assumption 1 implies f is locally Lipschitz, and the classical existence and uniqueness results for the solution of Equation (1) apply. The extra smoothness on f ensures the solution itself is sufficiently smooth for present purposes. These facts are summarised in Theorem 1. For proof(s) refer to (Arnol'd, 1992, chapter 4, paragraph 32).

Theorem 1. *Equation (1) admits a unique solution $y^* \in C^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$ on $\mathbb{T}$.*

It immediately follows that the model is correctly specified in the sense that $y^* \in \mathbb{Y}$, which is Corollary 1.

Corollary 1 (Correct model)**.** *The solution y^* of Equation (1) is in $\mathbb{Y}$.*

Proof. Since $D^{\nu+1}y^*$ is continuous and $\mathbb{T}$ is compact, it follows that $D^{\nu+1}y^*$ is bounded and $D^{\nu+1}y^* \in \mathcal{L}_p(\mathbb{T}, \mathbb{R}^d)$ for any $p \in [1, \infty]$. Therefore by the fundamental theorem of Lebesgue calculus (see e.g., Nielson 1997, Theorem 20.8) $D^m y^* \in \mathrm{AC}(\mathbb{T}, \mathbb{R}^d),\ m = 0, \ldots, \nu$. □

Corollary 1 essentially ensures that there is an *a priori* bound on the norm of the MAP estimate. Namely $\|\hat{y}\|_{\mathbb{Y}} \leq \|y^*\|_{\mathbb{Y}}$, which follows from the definition (see (40)).

6.2 Properties of the Information Operator

By Proposition 1, $\mathbb{Y}$ correspond to the Sobolev space $H_2^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$, consequently it is crucial to understand how the Nemytsky operator $\mathcal{S}_f$, and consequently $\mathcal{Z}$, act on Sobolev spaces. Fortunately, for the Nemytsky operator, the work has already been done (Valent, 1985, 2013), and Theorem 2 is immediate.

Theorem 2. *Let $\mathscr{U}$ be an open subset of $H_2^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$ such that $y(\mathbb{T}) \subset U$ for any $y \in \mathscr{U}$, where U some open subset of $\mathbb{R}^d$. The Nemytsky operator $\mathcal{S}_{f_i}$, associated with the ith coordinate of f is then C^1 mapping from $\mathscr{U}$ onto $H_2^\nu(\mathbb{T}, \mathbb{R})$ for $i = 1, \ldots, d$. If in addition, U is convex and bounded, then for any $y' \in \mathscr{U}$ there is number $c_0(y') > 0$ such that*

$$\|\mathcal{S}_{f_i}[y] - \mathcal{S}_{f_i}[y']\|_{H_2^\nu} \leq c_0(y') |f_i|_{\nu+1,U} \|y - y'\|_{H_2^{\nu+1}},$$

for all $y \in \mathscr{U}$*, where*

$$|f_i|_{\nu+1,U} \coloneqq \sum_{m=0}^{\nu+1} \sup_{(t,a)\in\mathbb{T}\times U} |D^m f_i(t,a)| \,.$$

Proof. A direct application of Theorem 4.1 of (Valent, 2013, page 32) establishes the first claim. Furthermore, since the conditions of Theorem 4.5 by Valent (2013) are satisfied, the second claim follows (see point (ii) in the proof of Theorem 4.5 by Valent 2013, page 37). □

Essentially, Theorem 2 establishes that $\mathcal{S}_{f_i}$ as a mapping of $\mathscr{U}$ onto $H_2^\nu(\mathbb{T},\mathbb{R})$ is locally Lipschitz. This property is inherited by the information operator, which is Proposition 5.

Proposition 5. *In the same setting as Theorem 2. The* i*th coordinate of the information operator,* $\mathcal{Z}_i$*, is a* C^1 *mapping from* $\mathscr{U}$ *onto* $H_2^\nu(\mathbb{T},\mathbb{R})$*, for* $i = 1,\dots,d$*. If in addition,* U *is convex and bounded, then for any* $y' \in \mathscr{U}$ *there is number* $c_1(y',\nu,f_i,U) > 0$ *such that*

$$\|\mathcal{Z}_i[y] - \mathcal{Z}_i[y']\|_{H_2^\nu} \le c_1(y',\nu,f_i,U)\,\|y - y'\|_{H_2^{\nu+1}} \,,$$

for all $y \in \mathscr{U}$*.*

Proof. The differential operator $D\mathrm{e}_i^\mathsf{T}$ is a C^1 mapping of $\mathscr{U}$ onto $H_2^\nu(\mathbb{T},\mathbb{R})$. Consequently, by Theorem 2 the same holds for the operator $D\mathrm{e}_i^\mathsf{T} - \mathcal{S}_{f_i} = \mathcal{Z}_i$. For the second part, the triangle inequality gives

$$\|\mathcal{Z}_i[y] - \mathcal{Z}_i[y']\|_{H_2^\nu} \le \|Dy_i - Dy_i'\|_{H_2^\nu} + \|\mathcal{S}_{f_i}[y] - \mathcal{S}_{f_i}[y']\|_{H_2^\nu} \,, \tag{53}$$

and clearly

$$\|Dy_i - Dy_i'\|_{H_2^\nu} \le \|y - y'\|_{H_2^{\nu+1}} \,. \tag{54}$$

Consequently, by Theorem 2 the statement holds by selecting

$$c_1(y',\nu,f_i,U) = 1 + c_0(y')|f_i|_{\nu+1,U} \,.$$

□

6.3 Convergence of the MAP Estimate

Proceeding with the convergence analysis of the MAP estimate can finally be done in view of the regularity properties of the solution y^* and the information operator $\mathcal{Z}$ established by Corollary 1 and Proposition 5. Combining these results with Theorem 4.1 of Arcangéli et al. (2007) leads to Lemma 1.

Lemma 1. *Let* $\rho \in \mathbb{Y}$ *with* $\|\rho\|_{\mathbb{Y}} > \|y^*\|_{\mathbb{Y}}$ *and* $q \in [1,\infty]$*. Then there are positive constants* c_2*,* $\delta_{0,\nu}$*,* r *(depending on* ρ*), and* $c_3(y^*,\nu,f_i,r)$ *such that for any* $y \in B(0,\|\rho\|_{\mathbb{Y}})$ *the following estimate holds for all* $\delta < \delta_{0,\nu}$ *and* $m = 0,\dots,\nu-1$

$$\big|\mathcal{Z}_i[y]\big|_{H_q^m} \le c_2\Big(\delta^{\nu-m-(1/2-1/q)_+} c_3(y^*,\nu,f_i,r)\,\|y - y^*\|_{H_2^{\nu+1}} + \delta^{-m}\big\|\mathcal{Z}_i[y] \mid \mathbb{T}_N\big\|_\infty\Big),$$

where

$$\big\|\mathcal{Z}_i[y] \mid \mathbb{T}_N\big\|_\infty \coloneqq \max_{t\in\mathbb{T}_N} |\mathcal{Z}_i[y](t)| \,.$$

Proof. Firstly, Cauchy–Schwartz inequality yields

$$|y_i(t)| = |\langle \eta_t^{0,\mathrm{e}_i}, y\rangle_{\mathbb{Y}}| \leq \sqrt{R_{ii}(t,t)}\|y\|_{\mathbb{Y}}\,,$$

hence there is a positive constant $\tilde{c}$ such that

$$\|y_i\|_{\mathcal{L}_\infty} \leq \tilde{c}\|y\|_{\mathbb{Y}}\,.$$

Consequently, there exists a radius r (depending on ρ) such that $y(\mathbb{T}) \subset B(0,r)$ whenever $y \in B(0,\|\rho\|_{\mathbb{Y}})$. The set $B(0,\|\rho\|_{\mathbb{Y}})$ is open in $\mathbb{Y}$ and by Proposition 1 it is an open set in $H_2^{\nu+1}(\mathbb{T},\mathbb{R}^d)$. Therefore, all the conditions of Proposition 5 are met for the sets $B(0,\|\rho\|_{\mathbb{Y}})$ and $B(0,r)$. In particular, $\mathcal{Z}_i[y] \in H_2^\nu(\mathbb{T})$ for all $y \in B(0,\|\rho\|_{\mathbb{Y}})$. Consequently, for appropriate selection of parameters (Arcangéli et al., 2007, Theorem 4.1 page 193) gives

$$|\mathcal{Z}_i[y]|_{H_q^m} \leq c_2\Big(\delta^{\nu-m-(1/2-1/q)_+}|\mathcal{Z}_i[y]|_{H_2^\nu} + \delta^{-m}\|\mathcal{Z}_i[y] \mid \mathbb{T}_N\|_\infty\Big)$$

for all $\delta < \delta_{0,\nu}$ and $m = 0,\ldots,\nu-1$. Since $\mathcal{Z}[y^*] = 0$ it follows that

$$|\mathcal{Z}_i[y]|_{H_2^\nu} = |\mathcal{Z}_i[y] - \mathcal{Z}_i[y^*]|_{H_2^\nu} \leq \|\mathcal{Z}_i[y] - \mathcal{Z}_i[y^*]\|_{H_2^\nu}\,,$$

and by Proposition 5 the Lemma holds by selecting

$$c_3(y^*,\nu,f_i,r) = c_1(y^*,\nu,f_i,B(0,r)),$$

which concludes the proof. □

In view of Lemma 1, for any estimator $\hat{y}' \in \mathbb{Y}$, its convergence rate can be established provided the following is shown:

(i) There is $\rho \in \mathbb{Y}$ independent of $\hat{y}'$ such that $\hat{y}' \in B(0,\|\rho\|_{\mathbb{Y}})$

(ii) A bound proportional to δ^γ, $\gamma > 0$, of $\|\mathcal{Z}_i[\hat{y}'] \mid \mathbb{T}_N\|_\infty$ exists.

Neither (i) nor (ii) appear trivial to establish for any of the estimators discussed Section 3 in general. However, (i) and (ii) hold for the optimal (MAP) estimate $\hat{y}$, which yields Theorem 3.

Theorem 3. *Let $q \in [1,\infty]$, then under the same assumptions as in Lemma 1, there exists a constant $c_4(y^*,\nu,f_i,r)$ such that for $\delta < \delta_{0,\nu}$ the following holds for $i = 1,\ldots,d$:*

$$|\mathcal{E}_i[\hat{y}]|_{H_q^0} \leq \delta^\nu T^{1/q} c_4(y^*,\nu,f_i,r)\|y^*\|_{\mathbb{Y}}\,,$$
$$|\mathcal{E}_i[\hat{y}]|_{H_q^m} \leq \delta^{\nu+1-m-(1/2-1/q)_+} c_4(y^*,\nu,f_i,r)\|y^*\|_{\mathbb{Y}}\,,\ m = 1,\ldots,\nu.$$

Proof. Firstly, note that $\|\hat{y}\|_{\mathbb{Y}} \leq \|y^*\|_{\mathbb{Y}}$, $|\mathcal{E}_i[\hat{y}]|_{H_q^m} = |\mathcal{Z}_i[\hat{y}]|_{H_q^{m-1}}$, and $\|\mathcal{Z}_i[\hat{y}] \mid \mathbb{T}_N\|_\infty = 0$ by definition, hence $\hat{y} \in B(0,\|\rho\|_{\mathbb{Y}})$, and Lemma 1 gives for $m = 1,\ldots,\nu$

$$|\mathcal{Z}_i[\hat{y}]|_{H_q^{m-1}} \leq \delta^{\nu+1-m-(1/2-1/q)_+} c_2 c_3(y^*,\nu,f_i,r)\,\|\hat{y} - y^*\|_{H_2^{\nu+1}}\,.$$

By Proposition 1, the fact that $\|\hat{y}\|_{\mathbb{Y}} \leq \|y^*\|_{\mathbb{Y}}$, and the triangle inequality, there exists a constant c_B (independent of $\hat{y}$ and y^*) such that

$$\|\hat{y} - y^*\|_{H_2^{\nu+1}} \leq c_B\|y^*\|_{\mathbb{Y}}$$

and thus the second bound holds by selecting

$$c_4(y^*, \nu, f_i, r) = c_2 c_B c_3(y^*, \nu, f_i, r).$$

For the first bound, the triangle inequality for integrals gives

$$\left|\mathcal{E}_i[\hat{y}](t)\right| \leq \left|\mathcal{Z}_i[\hat{y}]\right|_{H_1^0},$$

wherefore

$$\left|\mathcal{E}_i[\hat{y}](t)\right|_{H_q^0} \leq T^{1/q} \left|\mathcal{Z}_i[\hat{y}]\right|_{H_1^0},$$

which combined with the second bound gives the first. □

At first glance, it may appear that there is an appalling absence of dependence on T in the constants of the convergence rates provided by Theorem 3. This is not the case, the T dependence have conveniently been hidden in $\|y^*\|_{\mathbb{Y}}$ and possibly $c_4(y^*, \nu, f_i, r)$. Now $c_4(y^*, \nu, f_i, r)$ depends on $c_0(y^*)$ and $|f_i|_{\nu+1,B(0,r)}$, unfortunately an explicit expression for $c_0(y^*)$ is not provided by Valent (2013), which makes the effect of $c_4(y^*, \nu, f_i, r)$ difficult to untangle. Nevertheless, the factor $\|y^*\|_{\mathbb{Y}}$ does indeed depend on the interval length T. For example, let $\lambda, y_0 \in \mathbb{R}$ and consider the following ODE

$$\dot{y}(t) = \lambda y(t), \;\; y(0) = y_0.$$

Setting $\Sigma(t_0^-) = \mathrm{I}$ and selecting the prior $\mathrm{IWP}(\mathrm{I}, \nu)$ gives the following (in this case $\mathcal{D} = D^{\nu+1}$)

$$\|y^*\|_{\mathbb{Y}}^2 = y_0^2 \Big(\sum_{m=0}^{\nu} \lambda^{2m} + \frac{\lambda^{2\nu+1}}{2} \Big(\exp(2\lambda T) - 1 \Big) \Big).$$

Consequently, the global error can be quite bad when $\lambda > 0$ and T is large even when δ is very small, which is the usual situation (cf. Theorem 3.4 of Hairer et al. 1987).

6.3.1 Discussion of Convergence Results

In the present context it is instructive to view the solution of (1) as a family of a quadrature problems

$$y(t) = y_0 + \int_0^t f(\tau, y(\tau)) \, \mathrm{d}\tau, \tag{56}$$

where $\dot{y}(t) = f(t, y(t))$ is modelled by an element of $H_2^\nu(\mathbb{T}, \mathbb{R}^d)$. In view of Theorem 3, $D^m \dot{\hat{y}}$ converges uniformly to $D^m \dot{y}^*$ at a rate of $\delta^{\nu - m - 1/2}$, $m = 0, \ldots, \nu - 1$, thus for $\dot{\hat{y}}$ the same rate as for standard spline interpolation is obtained (Schultz, 1970). Furthermore, the rate obtained for $\hat{y}$ by Theorem 3 matches the rate for integral approximations using Sobolev kernels (Kanagawa et al., 2020, Proposition 1). That is, although dealing with a nonlinear interpolation/integration problem, Assumption 1 ensures the problem is still nice enough for the optimal interpolant to enjoy the classical convergence rates.

Global convergence of the filter associated with the zeroth order linearisation was examined by Kersting et al. (2018) under some different assumption on the vector field (see Remark 6). However, there the discussion of global convergence was limited to the priors in the class $\mathrm{IWP}(\Gamma, 1)$, for which a global convergence rate of δ is demonstrated, which matches the rate for the MAP estimator obtained by Theorem 3 ($\nu = 1$). It is important to stress that the statement of Theorem 3 only pertains to the MAP estimate, and the method examined by Kersting et al. (2018) is in general not the MAP estimate (unless $f(t, y) = \zeta(t)$ for some ζ). Consequently, Theorem 3 is not a generalisation of the

results by Kersting et al. (2018). However, in view of the discussion in Section 6.3 and the empirical findings in Section 7, it appears that Theorem 3 can be generalised to the estimator examined by Kersting et al. (2018), and indeed all the estimators discussed in Section 3.2.

7 Numerical Examples

In this section, the methods discussed in Section 3, the smoother based on the zeroth order method is denoted by EKS0, the smoother based on the first order method is denoted by EKS1, and the iterated extended Kalman smoother is denoted by IEKS. In particular the convergence rates of the MAP estimator from Section 6 are verified, which appear to be generalisable to the other methods as well.

7.1 The Logistic Equation

Consider the logistic equation

$$\dot{y}(t) = 10y(t)(1 - y(t)), \quad y(0) = y_0, \tag{57}$$

which has the following solution.

$$y(t) = \frac{\exp(10t)}{\exp(10t) + 1/y_0 - 1}. \tag{58}$$

The initial condition is set to $y_0 = 15/100$ and approximate solutions are computed by EKS0, EKS1, and IEKS on the interval $[0, 1]$ on a uniform, dense using, grid with interval length 2^{-12} using a prior in the class $\mathrm{IWP}(\mathrm{I}, \nu)$, $\nu = 1, \ldots, 4$. The filter updates only occur on a decimation of this dense grid by a factor of 2^{3+m}, $m = 1, \ldots, 8$, which yields the fill-distances $\delta_m = 2^{m-10}$, $m = 1, \ldots, 8$. The $\mathcal{L}_\infty$ error of the zeroth and first derivative estimates of the methods are computed on the dense grid and compared to δ^ν and $\delta^{\nu-1/2}$ (predicted rates), respectively. The errors of the approximate solutions versus fill-distance are shown in Figure 1 and it appears that EKS0, EKS1, and IEKS all attain at worst the predicted rates once δ is small enough. It appears the rate for IEKS1/IEKS tapers off for $\nu = 4$ and small δ. However, it can be verified that this is due to numerical instability when computing the smoothing gains as the prediction covariances $\Sigma_F(t_n^-)$ become numerically singular for too small h_n (see (25a)). The results are similar for the derivative of the approximate solution, see Figure 2.

Solution estimates by EKS0 and EKS1 are illustrutated in Figure 3 for $\nu = 2$ and $\delta = 2^{-4}$ (IEKS is very similar EKS1 and therefore not shown). While both methods produce credible intervals that cover the true solution, those of EKS1 are much tighter. That is, here the EKS1 estimate is of higher quality than that of EKS0, which is particularly clear when looking at the derivative estimates.

7.2 A Riccati Equation

The convergence rates are examined for a Riccati equation as well. That is, consider the following ODE

$$\dot{y}(t) = -c\frac{y^3(t)}{2}, \quad y(0) = y_0, \tag{59}$$

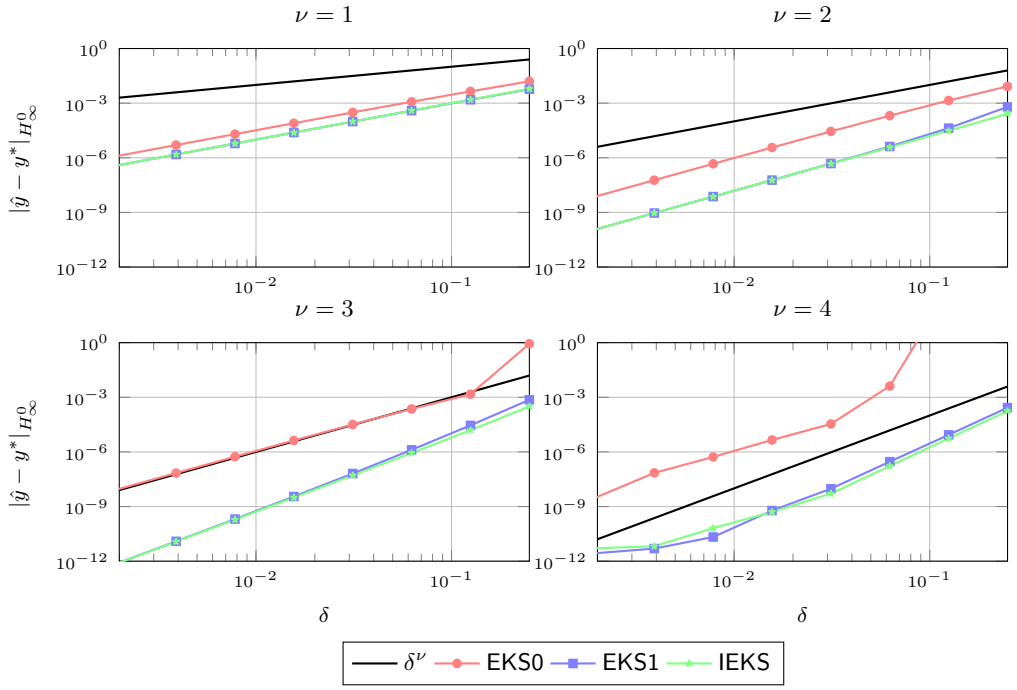

Figure 1: $\mathcal{L}_\infty$ error of the solution estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

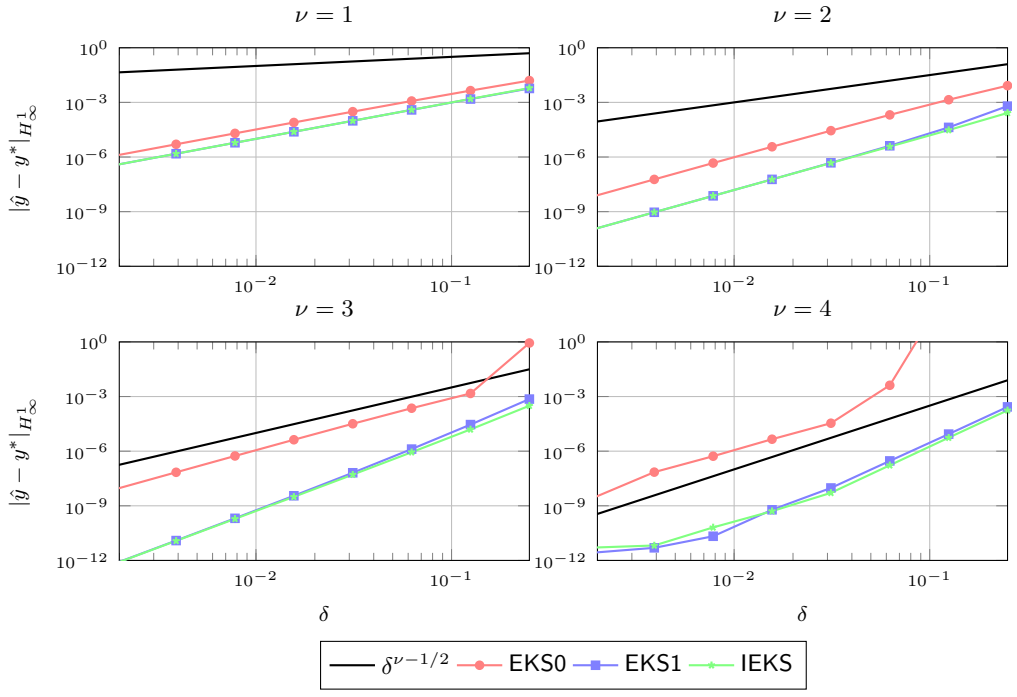

Figure 2: $\mathcal{L}_\infty$ error of the derivative estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate $\delta^{\nu-1/2}$ (black), versus fill-distance.

which has the following solution

$$y(t) = \frac{1}{\sqrt{ct + 1/y_0^2}}. \tag{60}$$

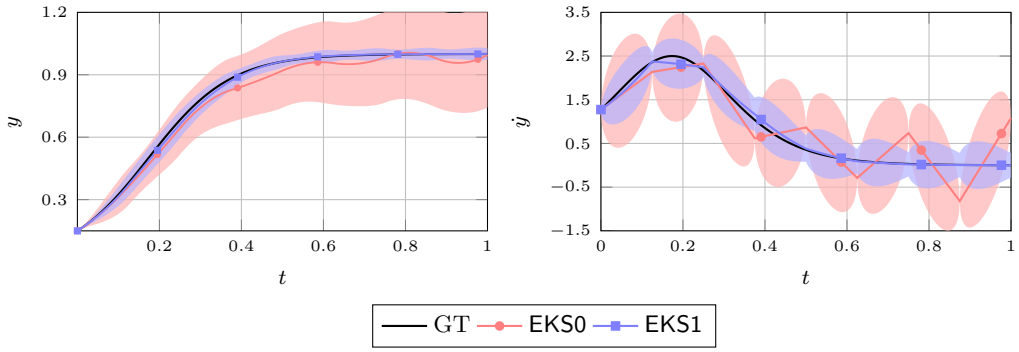

Figure 3: Reconstruction of the logistic map (left) and its derivative (right) with two standard deviation credible bands for EKS0 (red) and EKS1 (blue).

The initial condition is set to $y_0 = 1$. Just as for the logistic map, the solution is approximated by EKS0, EKS1, and IEKS on the interval $[0, 1]$, using a IWP(I, ν), $\nu = 1, \ldots, 4$, for various fill-distances δ. The $\mathcal{L}_\infty$ errors of the zeroth and first derivative estimates are shown in Figures 4 and 5, respectively. The general results are the same as before, EKS1 and IEKS are very similar, and EKS0 is some orders of magnitude worse while still appearing to converge at a similar rate as the former. The numerical instability in the computation of smoothing gains is still present for large ν and small δ.

Additionally, the output of the solvers for $\nu = 2$ is visualised for step-sizes of $h = 0.125$ and $h = 0.25$ in Figures 6 and 7, respectively. It can be seen that already for $h = 0.25$, the solution estimate and uncertainty quantification of the IEKS, while EKS0 and EKS1 leave room for improvement in terms of both accuracy and uncertainty quantification. By halving the step-size EKS1 and IEKS become near identical (wherefore IEKS is not shown in Figure 6), though the error of the EKS0 is still oscillating quite a bit, particularly for the derivative.

8 Conclusion and Discussion

In this paper, the maximum a posteriori estimate associated with the Bayesian solution of initial value problems (Cockayne et al., 2019) was examined. Several Gaussian approximations were reviewed and classified as explicit (Schober et al., 2019), semi-implicit (Tronarp et al., 2019b), and implicit depending on when they solve a local MAP problem, which for explicit and semi-implicit methods means they also solve the global MAP problem in the present setting. Furthermore, it was shown that the MAP estimate corresponds to the optimal interpolant in a Sobolev space, which along with tools from nonlinear analysis Valent (2013) and semi-norm estimates for Sobolev functions with scattered zeroes (Arcangéli et al., 2007) was exploited to obtain convergence rates for the MAP estimate when the vector field is sufficiently smooth.

While the present results are encouraging, there is a lot of open topics for future research. For example, in the present setting the MAP estimate is just taken as a given. Though of course, in practice a reliable method to evaluate it is required. For this end the MAP problems (17) and (26a) need to be analysed more carefully. In particular it would be fruitful to establish which conditions on the vector field and the fill-distance are required to ensure the local optima are global optima for the MAP problem(s), or the convexity of the problem (Boyd and Vandenberghe, 2004). This would of course imply that

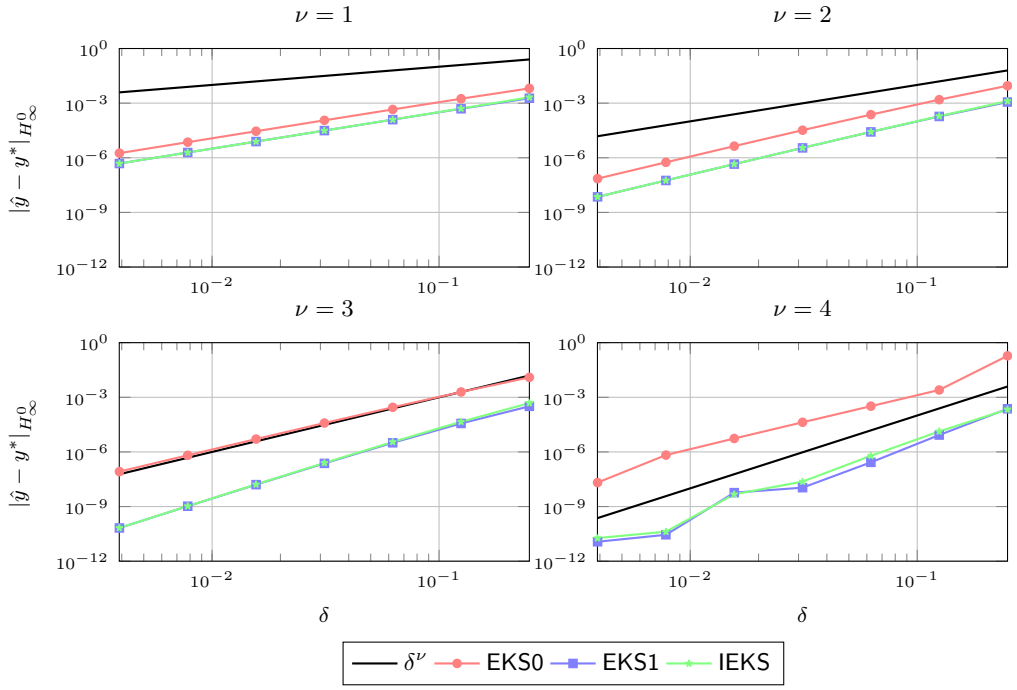

Figure 4: $\mathcal{L}_\infty$ error of the solution estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

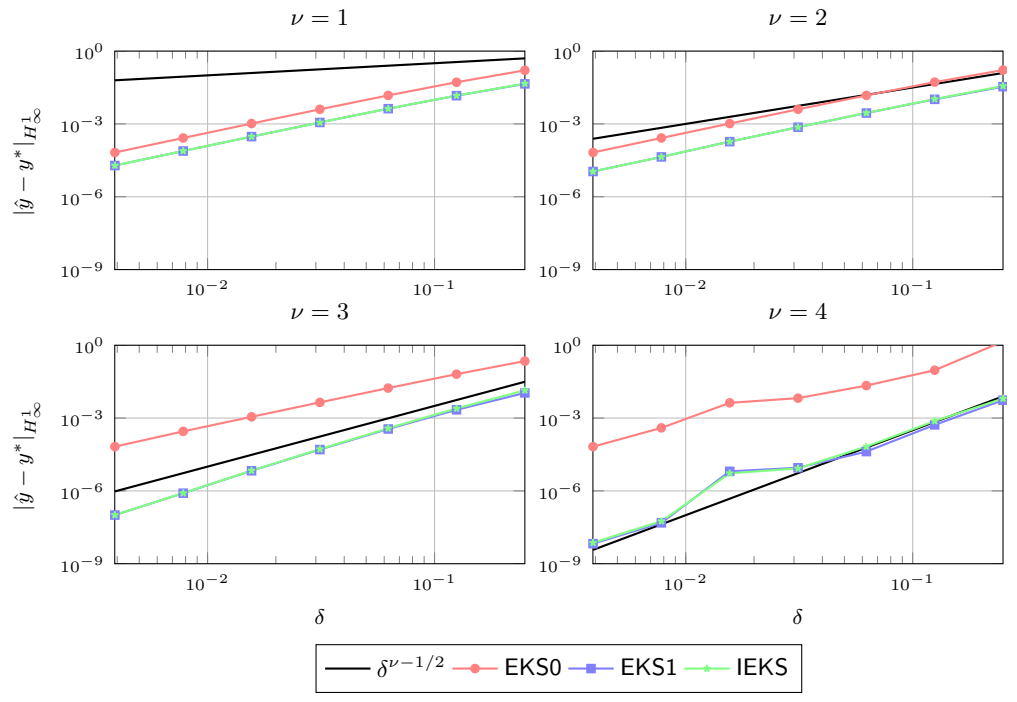

Figure 5: $\mathcal{L}_\infty$ error of the derivative estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate $\delta^{\nu-1/2}$ (black), versus fill-distance.

the IEKS and IEKF produce the global and local MAP estimates, respectively, whenever they converge to a stationary point (under some mild assumptions on f, see e.g., Knoth

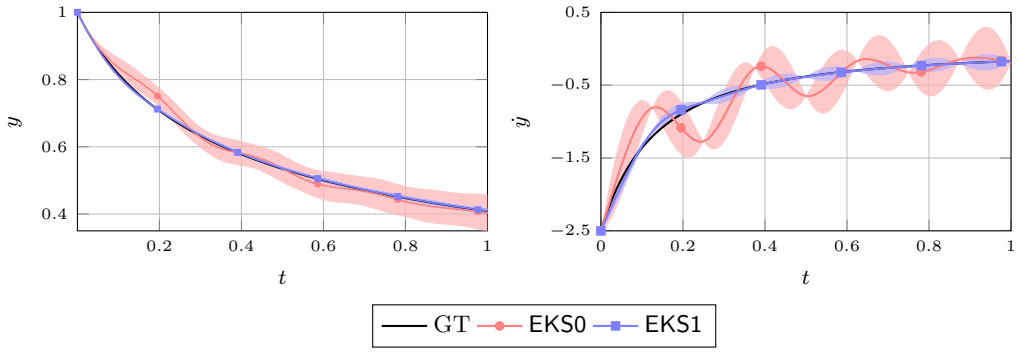

Figure 6: Reconstruction of the Riccati map (left) and its derivative (right) with two standard deviation credible bands for EKS0 (red) and EKS1 (blue), using a step size of $h = 0.125$.

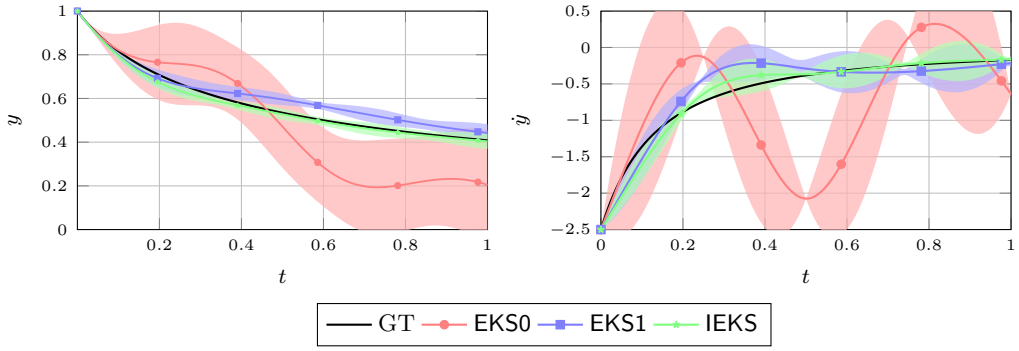

Figure 7: Reconstruction of the Riccati map with (left) and its derivative (right) with two standard deviation credible bands for EKS0 (red), EKS1 (blue), and IEKS (green), using a step size of $h = 0.25$.

1989). If convergence becomes an issue, more advanced MAP estimators can be considered, such as Levenberg–Marquardt (Särkkä and Svensson, 2020) or alternate direction method of multipliers (Aravkin et al., 2017, Boyd et al., 2011, Gao et al., 2019).

Furthermore, the empirical findings of Section 7 suggests, although not being MAP estimators, EKS0, EKS1, and IEKF can likely be given convergence statements similar to Theorem 3. It is not immediately clear what the most effective approach for this purpose is. On one hand, one can attempt to significantly extend the results of Kersting et al. (2018), which is more in line with how convergence rates are obtained for classical solvers. On the other hand, it seems like the local MAP problem (26a) can also be cast as a constrained optimisation problem in an RKHS $\mathbb{Y}_n$ corresponding to the Sobolov space $H_2^{\nu+1}([t_{n-1}, t_n], \mathbb{R}^d)$, in which case a lot of the arguments from Section 7 could be recycled for local convergence analysis. In either case a complicating factor is that the filtering and smoothing defects $z(t_n, \mu_F(t_n))$ and $z(t_n, \mu_S(t_n))$, respectively, need to be controlled. Another issue is the need for a prior bound on the RKHS norm of $\mu_S \in \mathbb{X}$ (recall ρ in Lemma 1).

Another issue is the designing of the mesh, $\mathbb{T}_N$, which has been completely omitted in the present work, classically this is referred to as *step size control* (Hairer et al., 1987). This is in fact one of the more important aspects of designing solvers, which ought to use available computational resources economically while still producing solution estimates of

acceptable accuracy. On the other hand, one of the possible advantages of the probabilistic approach is that in some situations it may be the case that an inaccurate solution estimate is acceptable if the provided uncertainty accurately reflects the error.

In statistical terminology, the mesh design/step size control is an issue of *experimental design* (Cockayne et al., 2019). A heuristic method for designing the mesh on the fly, as the filter is run, was proposed by Schober et al. (2019). This approach monitors the whitened residuals

$$\xi(t_n) = S^{-1/2}(t_n)(\zeta(t_n) - C(t_n)\mu_F(t_n^-)). \tag{61}$$

The statistics of $\xi(t_n)$ can be calibrated on-line as the estimate $\hat{\sigma}_N^2$ can be calculated resursive using the filter output (see Proposition 4). This method is structurally similar to classical methods of step size control (Byrne and Hindmarsh, 1975). In the context of probabilistic ODE solvers, an information theoretic approach to step size control was recently suggested by Chkrebtii and Campbell (2019) for their sampling based solver (Chkrebtii et al., 2016). More generally, a Bayesian experimental design viewpoint for probabilistic numerics was recently explored by Oates et al. (2019).

Another important topic that needs to be considered in practice is the stability properties of the filter and smoother, which is of utmost importance for integrating stiff systems (Hairer and Wanner, 1996). These stability properties depend on the parameters $A(h_n)$, $Q(h_n)$, $C(t_n)$, and $\zeta(t_n)$. The latter two parameters depend on the linearisation method, while the latter two parameters depend on the selection of prior (an issue that was omitted from the discussion in Section 2.1.1) (Anderson and Moore, 1979, 1981). The most basic notion of stability is that of *A-stability* (Dahlquist, 1963), which for a fixed step-size $h_n = h$ considers the following test equation

$$\dot{y}(t) = \Lambda y(t). \tag{62}$$

An ODE solver is said to be A-stable if its estimate of the solution of (62) converges to 0 as $t \to \infty$ whenever the real part of the eigenvalues of Λ are strictly negative. The inference problem can be solved exactly by a Kalman filter (and EKF/IEKF). A peculiar result is that for a prior in the IWP(Γ, ν) class, the convergence to zero of the estimate does not depend on the spectrum of Λ but rather on its rank. That is, the Kalman filter estimate converges to zero as $t \to \infty$ provided Λ is of full rank (Tronarp et al., 2019b, Theorem 2). While this is a solid start, stability analysis for the full class of priors discussed in Section 2.1 and linearisation methods of Section 3.2 ought to be carried out.

Acknowledgements

Filip Tronarp and Philipp Hennig gratefully acknowledge financial support by the German Federal Ministry of Education and Research (BMBF) through Project ADIMEM (FKZ 01IS18052B), and financial support by the European Research Council through ERC StG Action 757275 / PANAMA; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. Simo Särkkä gratefully acknowledges financial support by Academy of Finland.

Additionally, the authors are grateful to Toni Karvonen for his guidance through the scattered data approximation literature and to Hans Kersting for his keen remarks on an early draft version of this paper.

References

Abdulle, A. and Garegnani, G. (2020). Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration. *Statistics and Computing*, pages 1–26.

Adams, R. A. and Fournier, J. J. (2003). *Sobolev spaces*, volume 140. Elsevier.

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Information and System Sciences Series. Prentice-Hall.

Anderson, B. D. O. and Moore, J. B. (1981). Detectability and stabilizability of time-varying discrete-time linear systems. *SIAM Journal on Control and Optimization*, 19(1):20–32.

Aravkin, A., Burke, J. V., Ljung, L., Lozano, A., and Pillonetto, G. (2017). Generalized Kalman smoothing: Modeling and algorithms. *Automatica*, 86:63–86.

Arcangéli, R., de Silanes, M. C. L., and Torrens, J. J. (2007). An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211.

Arnol'd, V. I. (1992). *Ordinary Differential Equations*. Springer-Verlag Berlin Heidelberg.

Bell, B. M. (1994). The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4(3):626–636.

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

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019). Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22.

Butcher, J. C. (2008). *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Inc., 2 edition.

Byrne, G. D. and Hindmarsh, A. C. (1975). A polyalgorithm for the numerical solution of ordinary differential equations. *ACM Transactions on Mathematical Software (TOMS)*, 1(1):71–96.

Chkrebtii, O. A. and Campbell, D. A. (2019). Adaptive step-size selection for state-space probabilistic differential equation solvers. *Statistics and Computing*, 29(6):1285–1295.

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

Cockayne, J., Oates, C. J., Sullivan, T. J., and Girolami, M. (2019). Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789.

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

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, pages 1676–1695.

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

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

Gao, R., Tronarp, F., and Särkkä, S. (2019). Iterated extended Kalman smoother-based variable splitting for L_1-regularized state estimation. *IEEE Transactions on Signal Processing*, 67(19):5078–5092.

García-Fernández, Á. F., Svensson, L., Morelande, M. R., and Särkkä, S. (2015). Posterior linearization filter: Principles and implementation using sigma points. *IEEE Transactions on Signal Processing*, 63(20):5561–5573.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269.

Hairer, E., Nørsett, S., and Wanner, G. (1987). *Solving Ordinary Differential Equations I – Nonstiff Problems*. Springer.

Hairer, E. and Wanner, G. (1996). *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *2010 IEEE international workshop on machine learning for signal processing*, pages 379–384. IEEE.

Hennig, P. and Hauberg, S. (2014). Probabilistic solutions to differential equations and their application to Riemannian statistics. In *Proc. of the 17th int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 33. JMLR, W&CP.

Hennig, P., Osborne, M. A., and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142.

John, D., Heuveline, V., and Schober, M. (2019). GOODE: A Gaussian off-the-shelf ordinary differential equation solver. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3152–3162, Long Beach, California, USA. PMLR.

Kalman, R. and Bucy, R. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME, Journal of Basic Engineering*, 83:95–108.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2020). Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20:155–194.

Karvonen, T. and Sarkkä, S. (2016). Approximate state-space Gaussian processes via spectral transformation. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C. J., and Särkkä, S. (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *arXiv preprint arXiv:2001.10965*.

Kersting, H. and Hennig, P. (2016). Active uncertainty calibration in Bayesian ODE solvers. In *Uncertainty in Artificial Intelligence (UAI) 2016*, New York City, NY, USA. AUAI.

Kersting, H., Krämer, N., Schiegg, M., Daniel, C., Tiemann, M., and Hennig, P. (2020). Differentiable likelihoods for fast inversion of 'likelihood-free' dynamical systems. *arXiv preprint arXiv:2002.09301*.

Kersting, H., Sullivan, T. J., and Hennig, P. (2018). Convergence rates of Gaussian ODE filters. *arXiv preprint arXiv:1807.09737*.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Knoth, O. (1989). A globalization scheme for the generalized Gauss–Newton method. *Numerische Mathematik*, 56(6):591–607.

Lefebvre, T., Bruyninckx, H., and De Schuller, J. (2002). Comment on "A new method for the nonlinear transformation of means and covariances in filters and estimators" [with authors' reply]. *IEEE Transactions on Automatic Control*, 47(8):1406–1409.

Lie, H. C., Stuart, A. M., and Sullivan, T. J. (2019). Strong convergence rates of probabilistic integrators for ordinary differential equations. *Statistics and Computing*, 29(6):1265–1283.

Magnani, E., Kersting, H., Schober, M., and Hennig, P. (2017). Bayesian Filtering for ODEs with Bounded Derivatives. *arXiv:1709.08471 [cs.NA]*.

Marcus, M. and Mizel, V. J. (1973). Nemytsky operators on Sobolev spaces. *Arch. Rational Mech. Anal.*, 51:347–370.

Matsuda, T. and Miyatake, Y. (2019). Estimation of ordinary differential equation models with discretization error quantification. *arXiv preprint arXiv:1907.10565*.

Nielson, O. A. (1997). *An Introduction to Integration and Measure Theory*. John Wiley & Sons, Inc., New York.

Nordsieck, A. (1962). On numerical integration of ordinary differential equations. *Mathematics of Computation*, 16(77):22–49.

Oates, C., Cockayne, J., Prangle, D., Sullivan, T. J., and Girolami, M. (2019). Optimality criteria for probabilistic numerical methods. *arXiv preprint arXiv:1901.04326*.

Oates, C. J. and Sullivan, T. J. (2019). A modern retrospective on probabilistic numerics. *Statistics and Computing*, 29(6):1335–1351.

Øksendal, B. (2003). *Stochastic Differential Equations - An Introduction with Applications*. Springer.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine learning*. MIT Press.

Rauch, H. E., Tung, F., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic system. *AIAA Journal*, 3(8):1445–1450.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Särkkä, S. and Solin, A. (2019). *Applied Stochastic Differential Equations*. Cambridge University Press.

Särkkä, S., Solin, A., and Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61.

Särkkä, S. and Svensson, L. (2020). Levenberg–Marquardt and line-search extended Kalman smoothers. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Virtual location. IEEE.

Schober, M., Duvenaud, D. K., and Hennig, P. (2014). Probabilistic ODE solvers with Runge-Kutta means. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 739–747, Montréal, Canada. Curran Associates, Inc.

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

Schultz, M. H. (1970). Error bounds for polynomial spline interpolation. *Mathematics of Computation*, 24(111):507–515.

Schumaker, L. L. (1982). Optimal spline solutions of systems of ordinary differential equations. In *Differential Equations*, pages 272–283. Springer.

Sidhu, G. S. and Weinert, H. L. (1979). Vector-valued Lg-splines I. interpolating splines. *Journal of Mathematical Analysis and Applications*, 70(2):505–529.

Skilling, J. (1992). Bayesian solution of ordinary differential equations. In *Maximum entropy and Bayesian methods*, pages 23–37. Springer.

Solin, A. and Särkkä, S. (2014). Gaussian quadratures for state space approximation of scale mixtures of squared exponential covariance functions. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

Teymur, O., Lie, H. C., Sullivan, T., and Calderhead, B. (2018). Implicit probabilistic integrators for ODEs. In *Advances in Neural Information Processing Systems (NIPS)*.

Teymur, O., Zygalakis, K., and Calderhead, B. (2016). Probabilistic linear multistep methods. In *Advances in Neural Information Processing Systems (NIPS)*.

Tronarp, F., Garcia-Fernandez, A. F., and Särkkä, S. (2018a). Iterative filtering and smoothing in non-linear and non-Gaussian systems using conditional moments. *IEEE Signal Processing Letters*, 25(3):408–412.

Tronarp, F., Karvonen, T., and Särkkä, S. (2018b). Mixture representation of the Matérn class with applications in state space approximations and Bayesian quadrature. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*.

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

Tronarp, F., Kersting, H., Särkkä, S., and Hennig, P. (2019b). Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. *Statistics and Computing*, 29(6):1297–1315.

Valent, T. (1985). A property of multiplication in Sobolev spaces. Some applications. *Rendiconti del Seminario Matematico della Università di Padova*, 74:63–73.

Valent, T. (2013). *Boundary value problems of finite elasticity: local theorems on existence, uniqueness, and analytic dependence on data*, volume 31. Springer Science & Business Media.

van der Vaart, A. W. and van Zanten, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics.

Wahba, G. (1973). A class of approximate solutions to linear operator equations. *Journal of Approximation Theory*, 9(1):61–77.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):364–372.

Wang, J., Cockayne, J., and Oates, C. J. (2018). A role for symmetry in the Bayesian solution of differential equations. *Bayesian Analysis*.

Weinert, H. L. and Kailath, T. (1974). Stochastic interpretations and recursive algorithms for spline functions. *The Annals of Statistics*, 2(4):787–794.

Wendland, H. and Rieger, C. (2005). Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748.