Statistics and Computing manuscript No.
(will be inserted by the editor)

Bayesian ODE Solvers: The Maximum A Posteriori Estimate

Filip Tronarp · Simo Särkkä · Philipp Hennig

Received: date / Accepted: date

Abstract There is a growing interest in probabilistic numerical solutions to ordinary differential equations. In this paper, the *maximum a posteriori estimate* is studied under the class of ν times differentiable linear time invariant Gauss–Markov priors, which can be computed with an iterated extended Kalman smoother. The maximum a posteriori estimate corresponds to an optimal interpolant in the reproducing kernel Hilbert space associated with the prior, which in the present case is equivalent to a Sobolev space of smoothness $\nu + 1$. Subject to mild conditions on the vector field, convergence rates of the maximum a posteriori estmate are then obtained via methods from nonlinear analysis and scattered data approximation. These results closely resemble classical convergence results in the sense that a ν times differentiable prior process obtains a global order of ν, which is demonstrated in numerical examples.

Keywords Probabilistic numerical methods, Maximum a posteriori estimation, Kernel methods.

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):

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

Filip Tronarp
University of Tübingen
E-mail: filip.tronarp@uni-tueingen.de

Simo Särkkä
Aalto University
E-mail: simo.sarkka@aalto.fi

Philipp Hennig
University of Tübingen and MPI for Intelligent Systems
E-mail: philipp.hennig@uni-tuebingen.de

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 on it (Butcher, 2008, Deuflhard and Bornemann, 2002, Hairer and Wanner, 1996, Hairer et al., 1987).

Classically, the error of a numerical solver is 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., Kersting et al. 2020, Matsuda and Miyatake 2019), 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., 2016, 2018), 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). 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 satisfy the ODE 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 Markovian (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). Due to 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). Convergence analysis for the deterministic solvers has been initiated, but the theory is as of yet not complete (Kersting et al., 2018).

The formal notion of Bayesian solvers was defined by Cockayne et al. (2019). Under particular conditions on the vector field, the solvers of Kersting and Hennig (2016), Magnani et al. (2017), Schober et al. (2019), 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. Due to the Gauss–Markov prior, the MAP estimate can be computed efficiently by the iterated 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 an optimal interpolant in this space. This enables the use of results from scattered data approximation (Arcangéli et al., 2007) 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. In Section 3, the associated MAP problem is stated and the iterated extended Kalman smoother for computing it is presented (Bell, 1994). In Section 4, the connection between MAP estimation and optimisation in a certain reproducing kernel Hilbert space is reviewed. In Section 5, 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)

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

and 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.

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, the solution of (1) is then modelled by a ν-times differentiable stochastic process prior $Y(t)$ with a state-space representation. That is, the stochastic process $X(t)$ defined by

$$X^{\mathsf{T}}(t) = \Big(Y^T(t)\ DY^{\mathsf{T}}(t)\ \dots\ D^\nu Y^{\mathsf{T}}(t)\Big)$$

solves a certain stochastic differential equation. Furthermore, let $\{\mathrm{e}_m\}_{m=0}^{\nu}$ be the canonical basis on $\mathbb{R}^{\nu+1}$ and I_d is the identity matrix in $\mathbb{R}^{d\times d}$, it is then convenient to define the matrices $\mathrm{E}_m = \mathrm{e}_m \otimes \mathrm{I}_d,\ 0 \le m \le \nu$. That is, the mth subvector of X is given by

$$X^m(t) = \mathrm{E}_m^{\mathsf{T}} X(t) = D^m Y(t), \quad 0 \le m \le \nu.$$

Now let $F_m \in \mathbb{R}^{d\times d}$, $0 \le m \le \nu$ and, $\Gamma \in \mathbb{R}^{d\times d}$ a positive definite matrix, and define the following differential operator

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

and the matrix $F \in \mathbb{R}^{d(\nu+1)\times d(\nu+1)}$ whose non-zero $d\times d$ blocks are given by

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

The class of priors considered herein is then given by

$$Y(t) = \mathrm{E}_0^{\mathsf{T}} \exp(Ft) X(0) + \int_0^T G_Y(t,\tau)\,\mathrm{d}W(\tau), \tag{3}$$

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 $\mathcal{A}$ on $\mathbb{T}$ with initial condition $D^m y(t_0) = 0$, $m = 0,\dots,\nu$. The Green's function is given by

$$G_Y(t,\tau) = \mathrm{E}_0^{\mathsf{T}} G_X(t,\tau), \tag{4a}$$

$$G_X(t,\tau) = \theta(t-\tau)\exp(F(t-\tau))\mathrm{E}_\nu \Gamma^{1/2}, \tag{4b}$$

where θ is Heaviside's step function. By construction, (3) has a state-space representation, which is given by the following stochastic differential equation (Øksendal, 2003)

$$\mathrm{d}X(t) = FX(t)\,\mathrm{d}t + \mathrm{E}_\nu \Gamma^{1/2}\,\mathrm{d}W(t),\ X(0) \sim \mathcal{N}(0, \Sigma(t_0^-)), \tag{5}$$

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 $0 \le m \le \nu$. The transition densities for X are given by (Särkkä and Solin, 2019)

$$X(t+h) \mid X(t) \sim \mathcal{N}(A(h)X(t), Q(h)), \tag{6}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes the Normal distribution with mean and covariance μ and Σ, respectively, and

$$A(h) = \exp(Fh), \tag{7a}$$

$$Q(h) = \int_0^T G_X(h,\tau) G_X^{\mathsf{T}}(h,\tau)\,\mathrm{d}\tau. \tag{7b}$$

Note that the integrand in (7b) has limited support, that is, the effective interval of integration is $[0, h]$. These parameters can practically be computed via the matrix fraction decomposition method (Särkkä and Solin, 2019). Details are given in Appendix A.

2.1.1 The Selection of Prior

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 5 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 $\alpha \ge 2$ 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-classes of (3) are listed below.

- (Released ν times integrated Wiener process onto $\mathbb{R}^d$). The process Y is a ν times integrated Wiener process if $F_m = 0,\ m = 1,\dots,\nu$. 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 Appendix B). 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 $F_m = 0,\ m = 1,\dots,\nu-1$. 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)

  $$F_m = -\binom{\nu+1}{m}\lambda^{\nu+1-m}, \quad m = 0,\dots,\nu,$$
  $$\Gamma = 2\sigma^2\lambda^{2\nu+1},$$

  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., 2018). 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.

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

$$\mathcal{Z} = D - \mathcal{S}_f, \tag{9}$$

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

$$\mathcal{S}_f[y](t) = f(t, y(t)). \tag{10}$$

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

$$z(t, x) \coloneqq x^1 - f(t, x^0),$$

then $\mathcal{Z}[Y](t) = \mathcal{S}_z[X](t) = z(t, X(t))$. 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 $X^1(0) = f(0, y_0)$.

Remark 2 The properties of the Nemytsky operator are entirely determined by the vector field f. For instance, if $f \in C^\alpha(\mathbb{T} \times \mathbb{R}^d, \mathbb{R}^d)$, $\alpha \geq 0$, then $\mathcal{S}_f$ maps $C^\nu(\mathbb{T}, \mathbb{R}^d)$ to $C^{\min(\nu,\alpha)}(\mathbb{T}, \mathbb{R})$, which is fine for present purposes. However, in the subsequent convergence analysis it is more appropriate to view $\mathcal{S}_f$ (and $\mathcal{Z}$) as a mapping between different Sobolev spaces, which is possible if α is sufficiently large (Valent, 2013).

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

3 Maximum A Posteriori Estimation

The MAP estimate for Y, or equivalently for X, is in view of (6) the solution to the optimisation problem

$$\min_{x(t_{0:N})} \quad \mathcal{V}(x(t_{0:N})) \tag{11a}$$

$$\text{subject to} \quad \mathrm{E}_0^\mathsf{T} x(t_0) - y_0 = 0, \tag{11b}$$

$$\mathrm{E}_1^\mathsf{T} x(t_0) - f(t_0, y_0) = 0, \tag{11c}$$

$$z(t_n, x(t_n)) = 0, \ n = 1, \ldots, N, \tag{11d}$$

where $h_n = t_n - t_{n-1}$ is the step size sequence and $\mathcal{V}$ is up to a constant, the negative log-density

$$\begin{aligned}\mathcal{V}(x(t_{0:N})) = \frac{1}{2}\sum_{n=1}^{N} \big\| x(t_n) - A(h_n) x(t_{n-1}) \big\|^2_{Q(h_n)} \\ + \frac{1}{2}\big\| x(t_0) \big\|^2_{\Sigma(t_0^-)}.\end{aligned} \tag{12}$$

If the vector field is affine in y, then the MAP estimate and the full posterior can be computed exactly via Gaussian filtering and smoothing (Särkkä, 2013). However, when this is not the case then, for instance, a Gauss–Newton method can be used, which can be efficiently implemented by Gaussian filtering and smoothing as well. This method for MAP estimation is known as the *iterated extended Kalman smoother* (Bell, 1994).

3.1 Inference with Affine Vector Fields

If the vector field is 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). Define the information sets

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

In Gaussian filtering and smoothing, only the mean and covariance matrix of $X(t)$ are tracked. The 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. The mean and covariance conditioned on $\mathscr{Z}(T)$ at time t are denoted by $\mu_S(t)$ and $\Sigma_S(t)$, respectively.

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

$$\mathrm{E}_0^\mathsf{T} X(0) = y_0, \quad \mathrm{E}_1^\mathsf{T} X(0) = f(t_0, y_0).$$

This is can be done by a Kalman update

$$C^\mathsf{T}(t_0) = \begin{pmatrix} \mathrm{E}_0 & \mathrm{E}_1 \end{pmatrix}, \tag{14a}$$
$$S(t_0) = C(t_0)\Sigma(t_0^-)C^\mathsf{T}(t_0), \tag{14b}$$
$$K(t_0) = \Sigma(t_0^-)C^\mathsf{T}(t_0)S^{-1}(t_0), \tag{14c}$$
$$\mu_F(t_0) = K(t_0)\begin{pmatrix} y_0 \\ f(t_0, y_0) \end{pmatrix}, \tag{14d}$$
$$\Sigma_F(t_0) = \Sigma(t_0^-) - K(t_0)S(t_0)K^\mathsf{T}(t_0). \tag{14e}$$

The filtering mean and covariance on the mesh evolve as

$$\mu_F(t_n^-) = A(h_n)\mu_F(t_{n-1}), \tag{15a}$$
$$\Sigma_F(t_n^-) = A(h_n)\Sigma_F(t_{n-1})A^\mathsf{T}(h_n) + Q(h_n). \tag{15b}$$

The prediction moments at $t \in \mathbb{T}_N$ are then corrected according to the Kalman update

$$C(t_n) = \mathrm{E}_1^\mathsf{T} - \Lambda(t_n)\mathrm{E}_0^\mathsf{T}, \tag{16a}$$
$$S(t_n) = C(t_n)\Sigma_F(t_n^-)C^\mathsf{T}(t_n), \tag{16b}$$
$$K(t_n) = \Sigma_F(t_n^-)C^\mathsf{T}(t_n)S^{-1}(t_n), \tag{16c}$$
$$\mu_F(t_n) = \mu_F(t_n^-) + K(t_n)\big(\zeta(t_n) - C(t_n)\mu_F(t_n^-)\big), \tag{16d}$$
$$\Sigma_F(t_n) = \Sigma_F(t_n^-) - K(t_n)S(t_n)K^\mathsf{T}(t_n). \tag{16e}$$

On the mesh $\mathbb{T}_N$, the smoothing moments are given by

$$G(t_n) = \Sigma_F(t_n)A^\mathsf{T}(h_{n+1})\Sigma_F^{-1}(t_{n+1}^-), \tag{17a}$$
$$\mu_S(t_n) = \mu_F(t_n) + G(t_n)(\mu_S(t_{n+1}) - \mu_F(t_{n+1}^-)), \tag{17b}$$
$$\Sigma_S(t_n) = G(t_n)\big(\Sigma_S(t_{n+1}) - \Sigma_F(t_{n+1}^-)\big)G^\mathsf{T}(t_n) + \Sigma_F(t_n), \tag{17c}$$

with terminal conditions $\mu_S(t_N) = \mu_F(t_N)$, and $\Sigma_S(t_N) = \Sigma_F(t_N)$. The MAP estimate and its derivatives, on the mesh, are then given by

$$D^m\hat{y}(t) = \mathrm{E}_m^\mathsf{T}\mu_S(t), \quad t \in \mathbb{T}_N,\ m = 0, \ldots, \nu.$$

Remark 3 The filtering covariance can be written as

$$\Sigma_F(t_n) = \Sigma_F^{1/2}(t_n^-)\Big(\mathrm{I} - \mathrm{Proj}\left(\Sigma_F^{1/2}(t_n^-)C^\mathsf{T}(t_n)\right)\Big) \times \Sigma_F^{1/2}(t_n^-),$$

where $\mathrm{Proj}(A) = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ is the projection matrix onto the column space of A. By (16a) and $\Sigma_F(t_n^-) \succ 0$, the dimension of the column space of $\Sigma_F^{1/2}(t_n^-)C^\mathsf{T}(t_n)$ is readily seen to be d. That is, the null-space of $\Sigma_F(t_n)$ is of dimension d. By (17a) and (17c), it is also seen that $\Sigma_F(t_n)$ and $\Sigma_S(t_n)$ share null-space. This rank deficiency is not a problem in principle since the addition of $Q(h_n)$ in (15b) ensures $\Sigma_F(t_n^-)$ is of full rank. However, in practice $Q(h_n)$ may become numerically singular for very small step sizes.

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 replaced by an affine approximation (Magnani et al., 2017, Schober et al., 2019, Tronarp et al., 2019b). The iterated extended Kalman smoother approach for doing so is discussed in the following.

3.2 The Iterated Extended Kalman Smoother

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 (16). The iterated extended Kalman smoother linearises f around the smoothing mean in an iterative fashion. That is,

$$\hat{\Lambda}^l(t_n) = J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_S^l(t_n)), \tag{18a}$$
$$\hat{\zeta}^l(t_n) = f(t_n, \mathrm{E}_0^\mathsf{T}\mu_S^l(t_n)) - J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_S^l(t_n))\mathrm{E}_0^\mathsf{T}\mu_S^l(t_n). \tag{18b}$$

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 (18).

As mentioned, this is just the Gauss–Newton algorithm for the maximum a posteriori trajectory (Bell, 1994), and it can be shown that, under some conditions on the Jacobian of the vector field, the fixed-point is at least a local optimum to the MAP problem (11) (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.

3.2.1 Initialisation

In order to implement the IEKS, a method of initialisation needs to be devised. Fortunately, there exists non-iterative Gaussian solvers for this purpose (Schober et al., 2019, Tronarp et al., 2019b). These methods also employ Taylor series expansions to construct an affine approximation of the vector field. These methods select an expansion point at the prediction estimates $\mathrm{E}_0^\mathsf{T}\mu_F(t_n^-)$, and consequently the affine approximation can be constructed on the fly within the filter recursion. The affine approximation due to a zeroth order expansion gives the parameters (Schober et al., 2019)

$$\hat{\Lambda}(t_n) = 0, \tag{19a}$$

$$\hat{\zeta}(t_n) = f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F(t_n)), \tag{19b}$$

and will be referred to as the zeroth order extended Kalman smoother (EKS0). The affine expansion due to a first order expansion (Tronarp et al., 2019b) gives the parameters

$$\hat{\Lambda}(t_n) = J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F(t_n)), \tag{20a}$$

$$\hat{\zeta}(t_n) = f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F(t_n)) - J_f(t_n, \mathrm{E}_0^\mathsf{T}\mu_F(t_n))\mathrm{E}_0^\mathsf{T}\mu_F(t_n), \tag{20b}$$

and will be referred to as the first order extended Kalman smoother (EKS1). Note that EKS0 computes the exact MAP estimate in the event that the vector field f is constant in y, while EKS1 computes the exact MAP estimate in the more general case when f is affine in y. Consequently, as EKS1 makes a more accurate approximation of the vector field than EKS0, it is expected to perform better.

Furthermore, as Jacobians of the vector field will be computed in the IEKS iteration anyway, the preferred method of initialisation is EKS1, which is the method used in the subsequent experiments.

3.2.2 Computational Complexity

The computational complexity of a Gaussian filtering and smoothing method for approximating the solution of (1) can be separated into two parts (i) the cost of linearisation and (ii) the cost of inference. The cost of inference here refers to the computatational cost associated with the filtering and smoothing recursion, which for affine systems is $\mathcal{O}(Nd^3\nu^3)$. Since EKS0 and EKS1 perform the filtering and smoothing recursion once, their cost of inference is the same, $\mathcal{O}(Nd^3\nu^3)$. Furthermore, the linearisation cost of EKS0 amounts to $N+1$ evaluations of f and no evaluations of J_f, while EKS1 evaluates f $N+1$ times and J_f N times, respectively. Assuming IEKS is initialised by EKS1 using L iterations, including the initialisation, then the cost of inference is $\mathcal{O}(LNd^3\nu^3)$, f is evaluated $LN+1$ times, and J_f is evaluated LN times. A summary of the computational costs is given in Table 1.

Table 1 Comparison of the computational cost between EKS0, EKS1, and IEKS, where L denotes the total number of iterations for IEKS and it is assumed that IEKS is initialised by EKS1.

	EKS0	EKS1	IEKS
Inference cost	$\mathcal{O}(Nd^3\nu^3)$	$\mathcal{O}(Nd^3\nu^3)$	$\mathcal{O}(LNd^3\nu^3)$
# Evals of f	$N+1$	$N+1$	$LN+1$
# Evals of J_f	0	N	LN

4 Interpolation in Reproducing Kernel Hilbert Space

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 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 MAP estimator 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 5.

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 as defined by (3), then $\mathbb{Y}^{\nu+1}$ is given by the image of the operator (van der Vaart and van Zanten, 2008, lemmas 7.1, 8.1, and 9.1)

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

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

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

and inner product is given by

$$\begin{aligned}\langle y, y' \rangle_{\mathbb{Y}^{\nu+1}} &= \mathbf{y}_0^\mathsf{T} \Sigma^{-1}(t_0^-)\mathbf{y}_0' + \langle \mathcal{A}y, \mathcal{A}y' \rangle_{\mathcal{L}_2} \\ &= \mathbf{y}_0^\mathsf{T} \Sigma^{-1}(t_0^-)\mathbf{y}_0' + \langle \dot{w}_y, \dot{w}_{y'} \rangle_{\mathcal{L}_2}.\end{aligned}$$

Remark 4 For an element $y \in \mathbb{Y}^{\nu+1}$, the vector $\mathbf{y}_0$ contains the initial values for $D^m y(t)$, $m = 0, \ldots, \nu$, in similarity with the vector $X(0)$ in the definition of the prior process Y in (3). That is, $\mathbf{y}_0$ should not be confused with the initial value of (1).

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

$$D^m y \in \mathrm{AC}(\mathbb{T}, \mathbb{R}^d),\ m = 0, \ldots, \nu, \tag{21a}$$

$$D^{\nu+1} y \in \mathcal{L}_2(\mathbb{T}, \mathbb{R}^d). \tag{21b}$$

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)

$$\begin{aligned}R(t,s) = \mathrm{E}_0^\mathsf{T} \exp(Ft)\Sigma(t_0^-) \exp(F^\mathsf{T} s)\mathrm{E}_0 \\ + \int_0^T G_Y(t,\tau) G_Y^\mathsf{T}(s,\tau)\,\mathrm{d}\tau,\end{aligned}$$

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

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

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

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

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

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

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

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

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

$$\begin{aligned}P(t,s) = \exp(Ft)\Sigma(t_0^-)\exp(F^\mathsf{T} s) \\ + \int_0^T G_X(t,\tau) G_X^\mathsf{T}(s,\tau)\,\mathrm{d}\tau,\end{aligned} \tag{24}$$

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,

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

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

Consider the interpolation problem

$$\hat{y} = \operatorname*{arg\,min}_{y \in \mathcal{I}_N} \frac{1}{2}\|y\|_{\mathbb{Y}}^2, \tag{25}$$

where the feasible set is given by

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

Define the following subspaces of $\mathbb{Y}$

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

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* (25) *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

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

for all $t \in \mathbb{T}_N$. □

By Proposition 2 the optimal point of (25) 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, $\mathcal{R}_N(\nu) \supset \mathcal{R}_N(1)$

$$b(t_n) = \begin{pmatrix} b_0^{\mathsf{T}}(t_n) & \ldots & b_\nu^{\mathsf{T}}(t_n) \end{pmatrix}^{\mathsf{T}}, \tag{27a}$$
$$y = \sum_{n=0}^{N} \begin{pmatrix} \eta_{t_n}^0 & \ldots & \eta_{t_n}^\nu \end{pmatrix} b(t_n), \tag{27b}$$
$$x = \sum_{n=0}^{N} \psi_{t_n} b(t_n), \tag{27c}$$

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

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

or more compactly

$$\|x\|_{\mathbb{X}}^2 = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{P}^{-1} \boldsymbol{x}, \tag{28}$$

where

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

Here $\boldsymbol{P}$ is the kernel matrix associated with function value observations of X at $\mathbb{T}_N$. That is, (28) 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* (25) *is equivalent to the MAP problem* (11).

5 Convergence Analysis

In this section, convergence rates of the kernel interpolant $\hat{y}$ as defined by (25), and by Proposition 3 the MAP estimate are obtained. These rates will be in terms of the fill-distance of the mesh $\mathbb{T}_N$, which is[2]

$$\delta = \sup_{t \in \mathbb{T}} \min_{n=0,\ldots,N} |t - t_n|. \tag{29}$$

In the following results from the scattered data approximation literature (Arcangéli et al., 2007) 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

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

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 *Vector field $f \in C^{\alpha+1}(\tilde{\mathbb{T}} \times \mathbb{R}^d, \mathbb{R}^d)$ with $\alpha \geq \nu$ and some set $\tilde{\mathbb{T}}$ with $\mathbb{T} \subset \tilde{\mathbb{T}} \subset \mathbb{R}$.*

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

5.1 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 *There exists $T^* > 0$ such that Equation* (1) *admits a unique solution $y^* \in C^{\alpha+1}([0, T^*), \mathbb{R}^d)$.*

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

Theorem 1 makes apparent the necessity of the next standing assumption.

Assumption 2 *$T < T^*$. That is, $\mathbb{T} \subset [0, T^*)$.*

The model is thus correctly specified in the following sense.

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

Proof Firstly, $y^* \in C^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$ due to Assumption 1, Theorem 1, and Assumption 2. 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 (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, that is $\|\hat{y}\|_{\mathbb{Y}} \leq \|y^*\|_{\mathbb{Y}}$.

Remark 5 It is in general difficult to determine T^* for a given vector field f and initial condition y_0, which makes Assumption 2 hard to verify in general. However, additional conditions can be imposed which assures $T^* = \infty$. An example of such a condition is that the vector field is *uniformly Lipschitz* as mapping of $\mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ (Kelley and Peterson, 2010, Theorem 8.13). That is, for any $y, y' \in \mathbb{R}^d$ it holds that

$$\sup_{t \in \mathbb{R}_+} \big\| f(t, y) - f(t, y') \big\| \leq \mathrm{Lip}(f) \big\| y - y' \big\| ,$$

where $\mathrm{Lip}(f) < \infty$ is a positive constant.

5.2 Properties of the Information Operator

By Proposition 1, $\mathbb{Y}$ correspond to the Sobolev space $H_2^{\nu+1}(\mathbb{T}, \mathbb{R}^d)$, hence it is crucial to understand how the Nemytsky operator $\mathcal{S}_f$, and consequently $\mathcal{Z}$, act on Sobolev spaces. 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

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

Proof The first claim is just an application of Theorem 4.1 of (Valent, 2013, page 32) and the second claim follows from (ii) in the proof of Theorem 4.5 in (Valent, 2013, page 37). □

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.

Proposition 4 *In the same setting as Theorem 2. The ith 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 $i = 1, \ldots, d$. 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*

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

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

$$\begin{aligned} \|\mathcal{Z}_i[y] - \mathcal{Z}_i[y']\|_{H_2^\nu} \leq{}& \|Dy_i - Dy_i'\|_{H_2^\nu} \\ &+ \|\mathcal{S}_{f_i}[y] - \mathcal{S}_{f_i}[y']\|_{H_2^\nu} , \end{aligned}$$

and clearly

$$\|Dy_i - Dy_i'\|_{H_2^\nu} \leq \|y - y'\|_{H_2^{\nu+1}} .$$

Consequently, by Theorem 2 the statement holds by selecting

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

□

5.3 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 4. 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 $m = 0, \ldots, \nu - 1$*

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

where

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

Proof Firstly, Cauchy–Schwartz inequality yields

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

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 4 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

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

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

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

and by Proposition 4 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 $y^*, \hat{y}' \in B(0,\|\rho\|_{\mathbb{Y}})$
(ii) A bound proportional to δ^γ, $\gamma > 0$, of $\big\|\mathcal{Z}_i[\hat{y}'] \mid \mathbb{T}_N\big\|_\infty$ exists.

Neither (i) nor (ii) appear trivial to establish for Gaussian estimators in general (e.g., the methods of Schober et al. 2019 and Tronarp et al. 2019b). 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 is a constant $c_4(y^*,\nu,f_i,r)$ such that for $\delta < \delta_{0,\nu}$ the following holds:*

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

where $m = 1,\ldots,\nu$.

Proof Firstly, note that $\|\hat{y}\|_{\mathbb{Y}} \leq \|y^*\|_{\mathbb{Y}}$ and $\big|\mathcal{E}_i[\hat{y}]\big|_{H_q^m} = \big|\mathcal{Z}_i[\hat{y}]\big|_{H_q^{m-1}}$. By definition

$$\big\|\mathcal{Z}_i[\hat{y}] \mid \mathbb{T}_N\big\|_\infty = 0,$$

hence $\hat{y} \in B(0,\|\rho\|_{\mathbb{Y}})$, and Lemma 1 gives for $m = 1,\ldots,\nu$

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

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

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

and hence

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

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 $|f_i|_{\nu+1,B(0,r)}$ and 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

$$\dot{y}(t) = \lambda y(t),\ y(0) = y_0. \tag{31}$$

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

$$\|y^*\|_{\mathbb{Y}}^2 = y_0^2\Big(\sum_{m=0}^{\nu}\lambda^{2m} + \frac{\lambda^{2\nu+1}}{2}\big(\exp(2\lambda T) - 1\big)\Big). \tag{32}$$

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)).

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

$$y(t) = y_0 + \int_0^t f(\tau, y(\tau))\,\mathrm{d}\tau, \tag{33}$$

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, $D^m\hat{\dot{y}}$ converges uniformly to $D^m\dot{y}^*$ at a rate of $\delta^{\nu-m-1/2},\ m = 0, \ldots, \nu-1$, thus for $\hat{\dot{y}}$ 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.

6 Selecting the Hyperparameters

In order to calibrate the credible intervals, the parameters $\Sigma(t_0^-)$ and Γ need to be appropriately scaled to the problem being solved. It is practical to work with the parametrisation $\Sigma(t_0^-) = \sigma^2\breve{\Sigma}(t_0^-)$ and $\Gamma = \sigma^2\breve{\Gamma}$ for fixed $\Sigma(t_0^-)$ and $\breve{\Gamma}$. In this case, the quasi maximum likelihood estimate of σ^2 can be computed cheaply, see Appendix B.

In principle, the parameters F_m $(0 \le m \le \nu)$ can be estimated via quasi maximum likelihood as well but this would require iterative optimisation. For a given computational budget this may not be advantageous since the convergence rate obtained in Theorem 3 holds for any selection of these parameters. Thus it is not clear that spending a portion of a computational budget on estimating F_m $(0 \le m \le \nu)$ will yield a smaller solution error than solving the MAP problem on a denser grid (smaller δ) for a fixed parameters, with the same total computational budget. The IWP$(\sigma^2\breve{\Gamma}, \nu)$ class of priors thus seem like a good default choice ($F_m = 0$, $0 \le m \le \nu$).

Nevertheless, the parameters could in principle be selected to optimise the constant appearing in Theorem 3. That is, solving the following optimisation problem

$$\min_{F_0,\ldots,F_\nu} c_4(y^*, \nu, f_i, r)\,\|y^*\|^2, \tag{34}$$

which unfortunately appears to be intractable in general. However, it might be a good idea to use the the second factor, $\|y^*\|^2$ as a proxy. For instance, consider solving the ODE in (31) again, but this time with the prior set to IOUP$(\lambda, 1, \nu)$. In this case, $\mathcal{A} = D^{\nu+1} - \lambda D^\nu$, and the RKHS norm becomes

$$\|y^*\|_{\mathbb{Y}}^2 = y_0^2 \sum_{m=0}^{\nu} \lambda^{2m}, \tag{35}$$

which is strictly smaller than the RKHS norm obtained by IWP(I, ν) in (32).

7 Numerical Examples

In this section, the MAP estimate as implemented by the iterated extended Kalmans smoother (IEKS) is compared to the methods of Schober et al. (2019) (EKS0), and Tronarp et al. (2019b) (EKS1). In particular the convergence rates of the MAP estimator from Section 5 are verified, which appear to generalise to the other methods as well.

In Sections 7.1, 7.2, and 7.3 the logistic equation, Riccati equation, and the Fitz–Hugh–Nagumo model are investigated, respectively. The vector field is a polynomial in these cases, which means it is infinitely many times differentiable and Assumption 1 is satisfied for any $\nu \ge 1$. Lastly, in Section 7.4, a case where the vector field is only continuous is given, which means that Assumption 1 is violated for any $\nu \ge 1$.

7.1 The Logistic Equation

Consider the logistic equation

$$\dot{y}(t) = 10y(t)(1 - y(t)), \quad y(0) = y_0 = 15/100,$$

which has the following solution.

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

The 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 IWP(I, ν), $\nu = 1, \ldots, 4$. 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 EKS1/IEKS tapers off for $\nu = 4$ and small δ. However, it can be verified that this is due

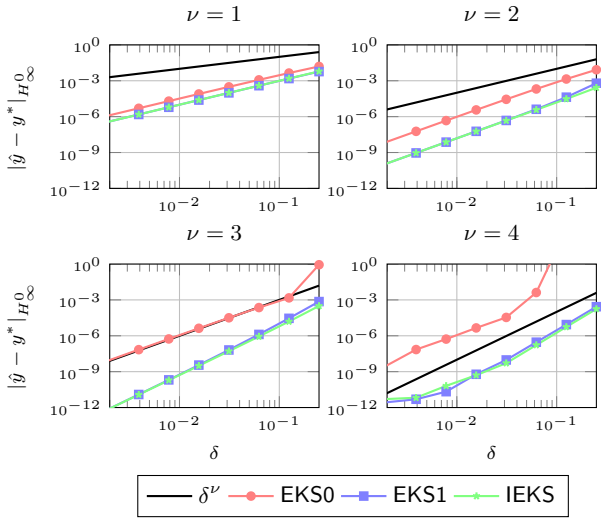

Fig. 1 $\mathcal{L}_\infty$ error of the solution estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

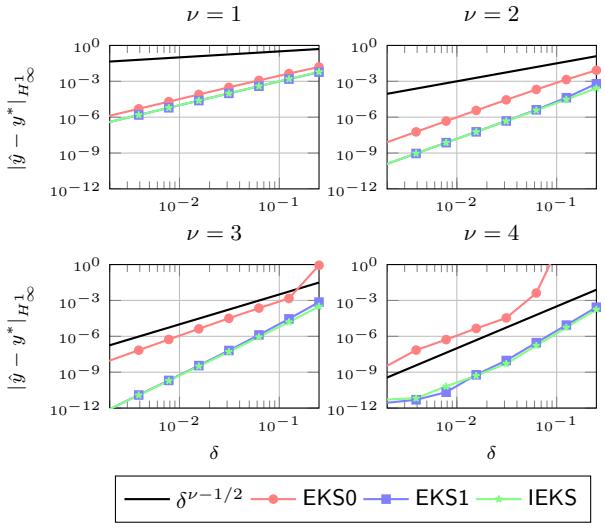

Fig. 2 $\mathcal{L}_\infty$ error of the derivative estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate $\delta^{\nu-1/2}$ (black), versus fill-distance.

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 (17a)). 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). The credible intervals are calibrated via the quasi maximum likelihood method, see B. 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

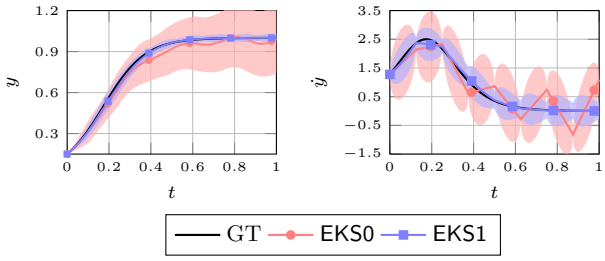

Fig. 3 Reconstruction of the logistic map (left) and its derivative (right) with two standard deviation credible bands for EKS0 (red) and EKS1 (blue).

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

$$\dot{y}(t) = -c\frac{y^3(t)}{2}, \quad y(0) = y_0 = 1,$$

which has the following solution

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

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.

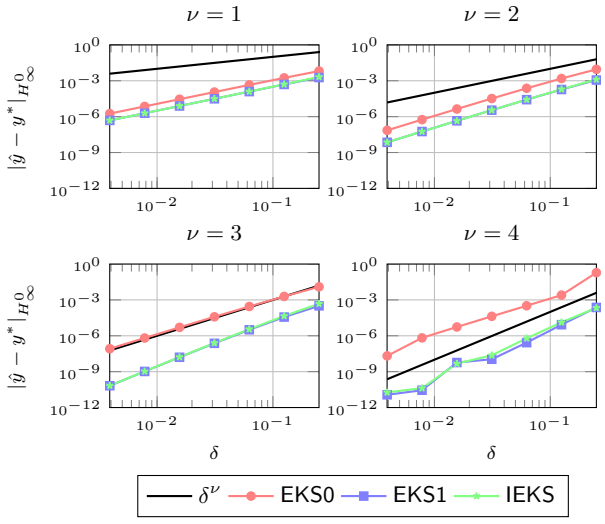

Fig. 4 $\mathcal{L}_\infty$ error of the solution estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

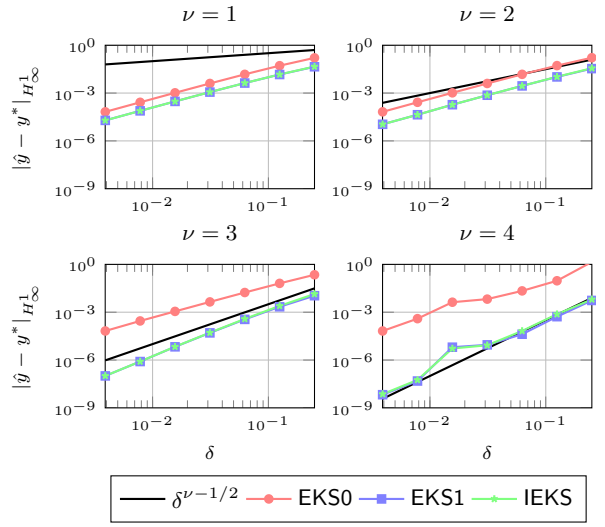

Fig. 5 $\mathcal{L}_\infty$ error of the derivative estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate $\delta^{\nu-1/2}$ (black), versus fill-distance.

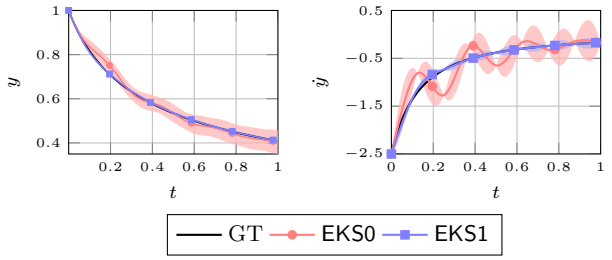

Fig. 6 Reconstruction of the Riccati map (left) and its derivative (right) with two standard deviation credible bands for EKS0 (red) and EKS1 (blue), using a step size of $h = 0.125$.

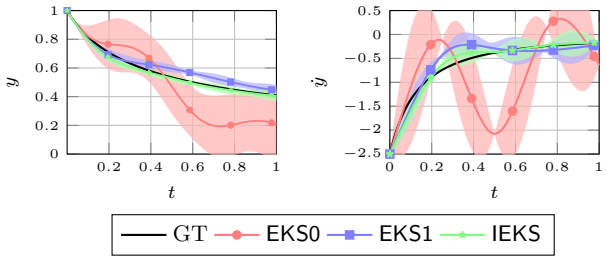

Fig. 7 Reconstruction of the Riccati map (left) and its derivative (right) with two standard deviation credible bands for EKS0 (red), EKS1 (blue), and IEKS (green), using a step size of $h = 0.25$.

7.3 The Fitz–Hugh–Nagumo Model

Consider the Fitz–Hugh–Nagumo model, which is given by

$$D\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} c(y_1(t) - y_1^3(t)/3 + y_2(t)) \\ -\frac{1}{c}(y_1(t) - a + by_2(t)) \end{pmatrix}. \tag{36}$$

The initial conditions and parameters are set to $y_2(0) = -y_1(0) = 1$, and $(a, b, c) = (0.2, 0.2, 2)$, respectively. The solution is estimated by EKS0, EKS1, and IEKS with an IWP(I, ν) prior ($1 \leq \nu \leq 4$) on a uniform grid with $2^{12} + 1$ points on the interval $[0, 2.5]$, using the same decimation scheme as previously. As this ODE does not have a closed form solution, it is approximated with `ode45`[3] in MATLAB, which is called with the parameters `RelTol` $= 10^{-14}$, and `AbsTol` $= 10^{-14}$. The approximate $\mathcal{L}_2$ error of the zeroth and first order derivative estimates of y_1^* are shown in Figures 8 and 9, respectively. The results appear to be consistent with the findings from the previous experiments.

Examples of the solver output of EKS1 and IEKS for $\nu = 2$ and $h = 0.4375$ is in Figures 10 and 11 for the first and second coordinates of y, respectively. The estimate and uncertainty quantification of the IEKS can be seen to be quite good, except for a slight undershoot in the estimate of $\dot{y}_1$ at $t = 1$. The performance of EKS1 is poorer, it overshoots quite a bit in its estimate of y_1 at around $t = 1.5$, which is not appropriately reflected in its credible interval.

7.4 A Non-smooth Example

Let the vector field f be given by

$$f(y) = \begin{cases} \kappa, & y \leq b, \\ \kappa + \lambda(y - b), & y > b, \end{cases} \tag{37}$$

[3] This is an adaptive embedded Runge–Kutta 4/5 method.

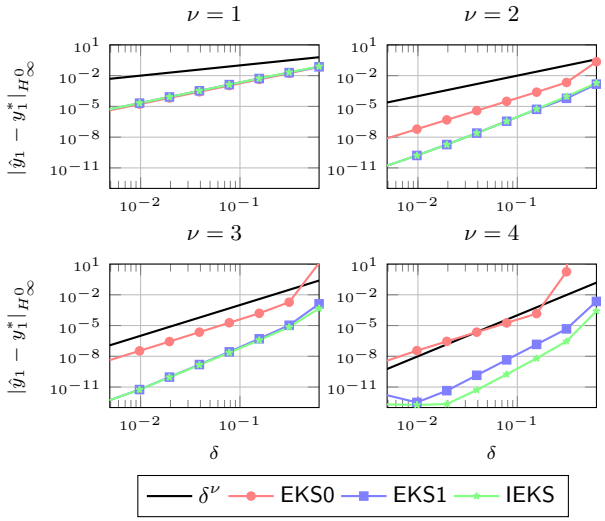

Fig. 8 $\mathcal{L}_\infty$ error of the solution estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

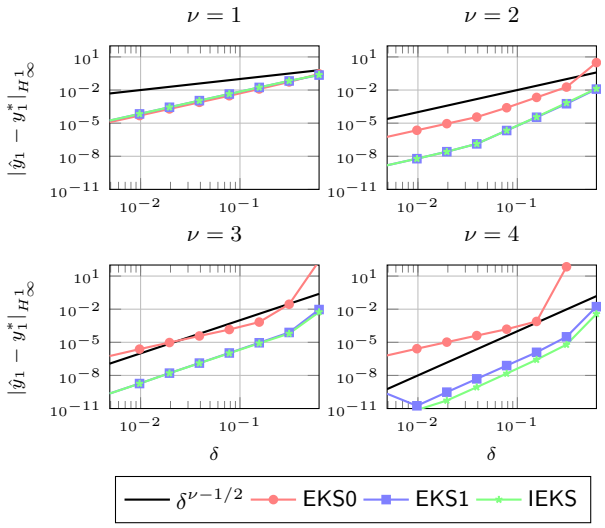

Fig. 9 $\mathcal{L}_\infty$ error of the derivative estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate $\delta^{\nu-1/2}$ (black), versus fill-distance.

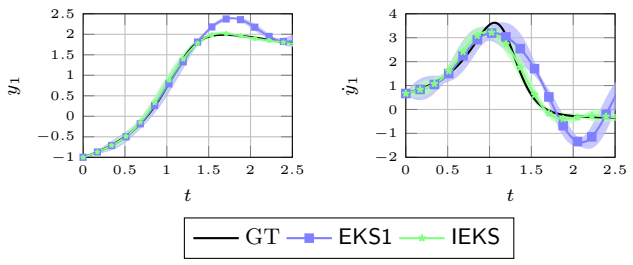

Fig. 10 Reconstruction of the first coordinate, y_1, in the Fitz–Hugh–Nagumo model (left) and its derivative (right) with two standard deviation credible bands for EKS1 (blue) and IEKS (green), using a step size of $h = 0.25$.

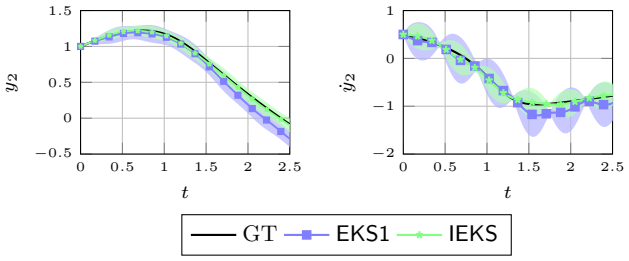

Fig. 11 Reconstruction of the first coordinate, y_2, in the Fitz–Hugh–Nagumo model (left) and its derivative (right) with two standard deviation credible bands for EKS1 (blue) and IEKS (green), using a step size of $h = 0.25$.

and consider the following ODE:

$$\dot{y}(t) = f(y(t)), \quad y(0) = y_0 \leq b. \tag{38}$$

If $\kappa > 0$, the solution is given by

$$y^*(t) = \begin{cases} y_0 + \kappa t, & t \leq \tau^*, \\ b + \frac{1}{\lambda}(\exp(\lambda(t - \tau^*)) - 1)\kappa, & t > \tau^*, \end{cases} \tag{39}$$

where $\tau^* = (b - y_0)/\kappa$. While f is continuous, it has a discontinuity in its derivative at $y = b$ and therefore Assumption 1 is violated for all $\nu \geq 1$. Nonetheless the solution is approximated by EKS0, EKS1, and IEKS using an IWP prior of smoothness $0 \leq \nu \leq 4$, and the parameters are set to $y_0 = 0$, $b = 1$, $\kappa = 2(b - y_0)$, and $\lambda = -5$. The $\mathcal{L}_\infty$ errors of the zeroth and first derivative of the approximate solutions are shown in Figures 12 and 13, respectively. Additionally, a comparison of the solver outputs of EKS1 and IEKS is shown in Figure 14 for $\nu = 2$ and $h = 0.25$.

The estimates still appear to converge as seen in Figures 12 and 13. However, while the rate predicted by Theorem 3 appears to still be obtained for $\nu = 1$, a rate reduction is observed for $\nu > 1$ (in comparison to the rate of Theorem 3). As Assumption 1 is violated, these results cannot be explained by the present theory.

However, note that Theorem 3 was obtained using $y^* \in \mathbb{Y}$ (Corollary 1) and $\mathcal{S}_f$ is locally Lipschitz (Theorem 2), together with the sampling inequalities of Arcangéli et al. (2007). These properties of f and y^* may be obtainable by other means than invoking Assumption 1. This could explain the results for $\nu = 1$.

On the other hand, in the setting of numerical integration, reduction in convergence rates when the RKHS is smoother than the integrand has been investigated by Kanagawa et al. 2020. If these results can be extended to the setting of solving ODEs, it could explain the results for $\nu > 1$.

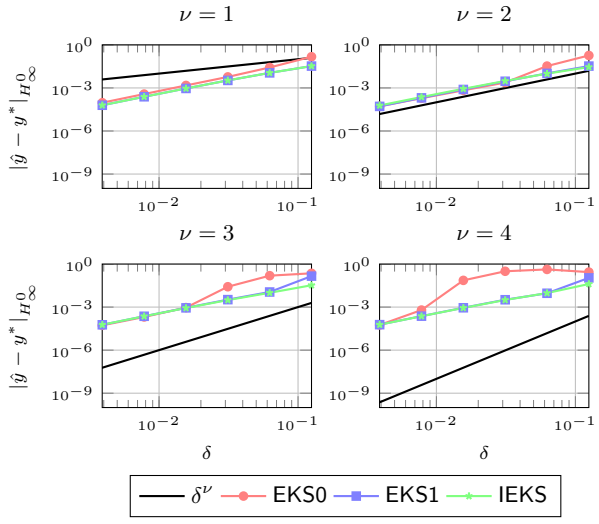

Fig. 12 $\mathcal{L}_\infty$ error of the solution estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

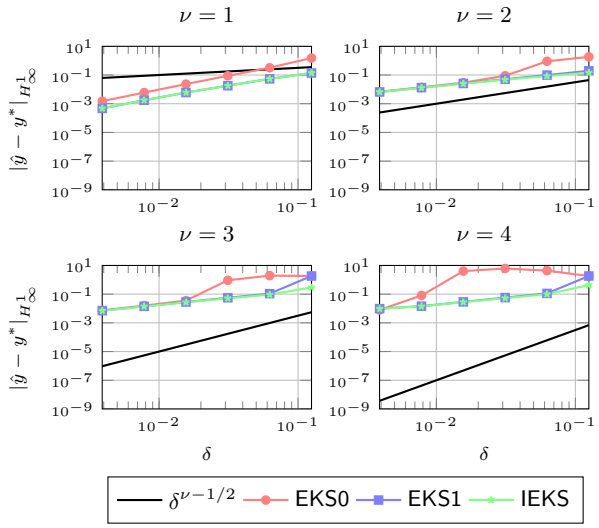

Fig. 13 $\mathcal{L}_\infty$ error of the derivative estimate as produced by EKS0 (red), EKS1 (blue), IEKS (green), and the predicted MAP rate δ^ν (black), versus fill-distance.

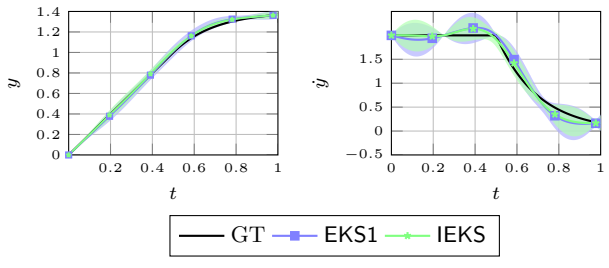

Fig. 14 Reconstruction of the solution to the non-smooth ODE (left) and its derivative (right) with two standard deviation credible bands for EKS1 (blue) and IEKS (green), using a step size of $h = 0.25$.

8 Conclusion

In this paper, the maximum a posteriori estimate associated with the Bayesian solution of initial value problems (Cockayne et al., 2019) was examined and it was shown to enjoy fast convergence rates to the true solution.

In the present setting the MAP estimate is just taken as a given, in the sense that IEKS is not guaranteed to produced the global optimum of the MAP problem. It would therefore be fruitful to study the MAP problem more carefully. In particular, establishing conditions on the vector field and the fill-distance under which the MAP problem admits a unique local optimum would be a point for future research. On the algorithmic side, other MAP estimators can be considered, such as Levenberg–Marquardt (Särkkä and Svensson, 2020) or alternate direction method of multipliers (Boyd et al., 2011, Gao et al., 2019).

Furthermore, the empirical findings of Section 7 suggests, although not being MAP estimators, EKS0 and EKS1 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 methodology developed here can be extended for local convergence analysis as well by considering the filter update as an interpolation problem in some RKHS on each interval $[t_{n-1}, t_n]$.

Acknowledgements The authors have had productive discussions with Toni Karvonen and Hans Kersting.

A Computing Transition Densities

An effective method for computing the parameters of the transition density in (7) is the *matrix fraction decomposition* (Axelsson and Gustafsson, 2014, Särkkä and Solin, 2019, Van Loan, 1978). Define the matrix valued function

$$\Xi(h) = \exp\left(\begin{pmatrix} F & \mathrm{E}_\nu \Gamma \mathrm{E}_\nu^\mathsf{T} \\ 0 & -F^\mathsf{T} \end{pmatrix} h\right).$$

It can then be shown that Ξ has the following structure

$$\Xi(h) = \begin{pmatrix} \Xi_{11}(h) & \Xi_{12}(h) \\ 0 & \Xi_{22}(h) \end{pmatrix},$$

and (Axelsson and Gustafsson, 2014)

$$A(h) = \Xi_{11}(h), \tag{40a}$$

$$Q(h) = \Xi_{12}(h)\Xi_{11}^\mathsf{T}(h). \tag{40b}$$

Furthermore, the Green's functions can be evaluated by the same means by noting that (see (4))

$$G_X(t,\tau) = \theta(t-\tau)A(t-\tau)\mathrm{E}_\nu \Gamma^{1/2}.$$

B Calibration

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 (5)), which the present discussion is just concerned with.

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)

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

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)[4]

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

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 5 *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*

$$\hat{\sigma}^2_N = \frac{1}{d(N+2)}\begin{pmatrix} y_0 \\ f(0,y_0)\end{pmatrix}^{\mathsf{T}} \breve{S}^{-1}(t_0)\begin{pmatrix} y_0 \\ f(0,y_0)\end{pmatrix} + \frac{1}{d(N+2)}\sum_{n=1}^{N}\|\zeta(t_n) - C(t_n)\mu_F(t_n^-)\|^2_{\breve{S}(t_n)}\,. \tag{41}$$

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.

References

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

Adams RA, Fournier JJ (2003) Sobolev spaces, vol 140. Elsevier

Arcangéli R, de Silanes MCL, Torrens JJ (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 VI (1992) Ordinary Differential Equations. Springer-Verlag Berlin Heidelberg

Axelsson P, Gustafsson F (2014) Discrete-time solutions to the continuous-time differential Lyapunov equation with applications to Kalman filtering. IEEE Transactions on Automatic Control 60(3):632–643

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

Boyd S, Parikh N, Chu E, Peleato B, 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

Butcher JC (2008) Numerical Methods for Ordinary Differential Equations, 2nd edn. John Wiley & Sons, Inc.

Chkrebtii OA, Campbell DA, Calderhead B, Girolami MA (2016) Bayesian solution uncertainty quantification for differential equations. Bayesian Analysis 11(4):1239–1267

Cockayne J, Oates CJ, Sullivan TJ, Girolami M (2019) Bayesian probabilistic numerical methods. SIAM Review 61(4):756–789

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

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

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

Gao R, Tronarp F, 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

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

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

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

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

Hartikainen J, 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, IEEE, pp 379–384

Hennig P, 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), JMLR, W&CP, vol 33

Hennig P, Osborne MA, 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, Schober M (2019) GOODE: A Gaussian off-the-shelf ordinary differential equation solver. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 3152–3162

Kalman R, Bucy R (1961) New results in linear filtering and prediction theory. Transactions of the ASME, Journal of Basic

[4] 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.

Engineering 83:95–108
Kalman RE (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(1):35–45
Kanagawa M, Hennig P, Sejdinovic D, Sriperumbudur BK (2018) Gaussian processes and kernel methods: A review on connections and equivalences. arXiv preprint arXiv:180702582
Kanagawa M, Sriperumbudur BK, Fukumizu K (2020) Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. Foundations of Computational Mathematics 20:155–194
Karvonen T, 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 CJ, Särkkä S (2020) Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. arXiv preprint arXiv:200110965
Kelley WG, Peterson AC (2010) The Theory of Differential Equations: Classical and Qualitative. Springer Science & Business Media
Kersting H, Hennig P (2016) Active uncertainty calibration in Bayesian ODE solvers. In: Uncertainty in Artificial Intelligence (UAI) 2016, AUAI, New York City, NY, USA
Kersting H, Sullivan TJ, Hennig P (2018) Convergence rates of Gaussian ODE filters. arXiv preprint arXiv:180709737
Kersting H, Krämer N, Schiegg M, Daniel C, Tiemann M, Hennig P (2020) Differentiable likelihoods for fast inversion of 'likelihood-free' dynamical systems. arXiv preprint arXiv:200209301
Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. Journal of mathematical analysis and applications 33(1):82–95
Kimeldorf GS, 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
Lie HC, Stuart AM, Sullivan TJ (2019) Strong convergence rates of probabilistic integrators for ordinary differential equations. Statistics and Computing 29(6):1265–1283
Magnani E, Kersting H, Schober M, Hennig P (2017) Bayesian Filtering for ODEs with Bounded Derivatives. arXiv:170908471 [csNA]
Marcus M, Mizel VJ (1973) Nemytsky operators on Sobolev spaces. Arch Rational Mech Anal 51:347–370
Matsuda T, Miyatake Y (2019) Estimation of ordinary differential equation models with discretization error quantification. arXiv preprint arXiv:190710565
Nielson OA (1997) An Introduction to Integration and Measure Theory. John Wiley & Sons, Inc., New York.
Oates CJ, Sullivan TJ (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 CE, Williams CKI (2006) Gaussian Processes for Machine learning. MIT Press
Rauch HE, Tung F, Striebel CT (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, Solin A (2019) Applied Stochastic Differential Equations. Cambridge University Press
Särkkä S, Svensson L (2020) Levenberg–Marquardt and line-search extended Kalman smoothers. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Virtual location
Särkkä S, Solin A, 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
Schober M, Duvenaud DK, Hennig P (2014) Probabilistic ODE solvers with Runge-Kutta means. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in Neural Information Processing Systems 27, Curran Associates, Inc., Montréal, Canada, pp 739–747
Schober M, Särkkä S, Hennig P (2019) A probabilistic model for the numerical solution of initial value problems. Statistics and Computing 29(1):99–122
Schultz MH (1970) Error bounds for polynomial spline interpolation. Mathematics of Computation 24(111):507–515
Schumaker LL (1982) Optimal spline solutions of systems of ordinary differential equations. In: Differential Equations, Springer, pp 272–283
Sidhu GS, Weinert HL (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, Springer, pp 23–37
Solin A, 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, Zygalakis K, Calderhead B (2016) Probabilistic linear multistep methods. In: Advances in Neural Information Processing Systems (NIPS)
Teymur O, Lie HC, Sullivan T, Calderhead B (2018) Implicit probabilistic integrators for ODEs. In: Advances in Neural Information Processing Systems (NIPS)
Tronarp F, Karvonen T, Särkkä S (2018) 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, 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, Hennig P (2019b) Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. Statistics and Computing 29(6):1297–1315
van der Vaart AW, van Zanten JH (2008) Reproducing kernel Hilbert spaces of Gaussian priors. In: Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh, Institute of Mathematical Statistics, pp 200–222
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, vol 31. Springer Science & Business Media
Van Loan C (1978) Computing integrals involving the matrix exponential. IEEE Transactions on Automatic Control 23(3):395–404
Wahba G (1973) A class of approximate solutions to linear operator equations. Journal of Approximation Theory 9(1):61–

77
Wang J, Cockayne J, Oates CJ (2018) A role for symmetry in the Bayesian solution of differential equations. Bayesian Analysis
Weinert HL, Kailath T (1974) Stochastic interpretations and recursive algorithms for spline functions. The Annals of Statistics 2(4):787–794