Uncertainty-Aware Numerical Solutions of ODEs by Bayesian Filtering

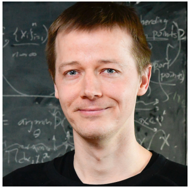
Dr. Hans Kersting

MaxEnt2021, TU Graz
06 July 2021

ENS | PSL
Département Informatique

Contents

Main Collaborators (of presented materials)



Philipp Hennig

Filip Tronarp

Nicholas Krämer

Tim J. Sullivan

Introduction

Numerical methods perform inference

Probabilistic numerics recasts numerics as inferential statistics [Hennig et al., 2015, Oates and Sullivan, 2019]

Both the evaluation of functions and the collection of data are gathering of (Shannon) information.

- ▷ Statistics estimates **parameters** in statistical **models** from **collected data**
- ▷ Numerics approximates **solutions** of numerical **problems** from **function evaluations**

By providing a statistical model that links func. eval. to the solution of a numerical problem, probabilistic numerics solves numerical problems by statistics (or machine learning).

$$\begin{tikzcd} \mathcal{X} \arrow[r, "\text{information } Z"] \arrow[dr, "\text{QoI operator } Q"'] & \mathcal{Z} \arrow[d, "\text{numerical method } B"] \\ & \mathcal{Q} \end{tikzcd}$$

E.g. Integrals: compute QoI $Q(f) = \int f(x)\mathrm{d}\mu(x) \in \mathcal{Q}$:

1. Collect evaluations $f(x_i) \in \mathcal{Z}$ given $f \in \mathcal{X}$
2. Compute approximation of $Q(f)$ based on $f \mid f(x_i)$

An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightharpoonup \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s) \, \mathrm{d}s,$$

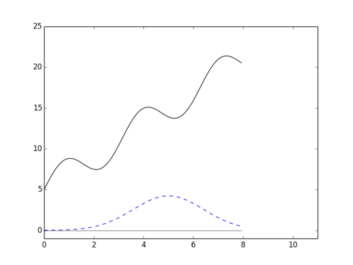
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightarrow \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



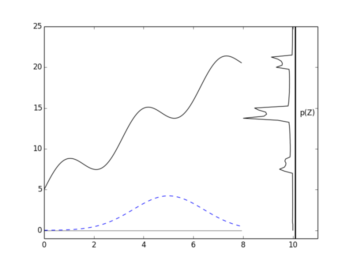
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

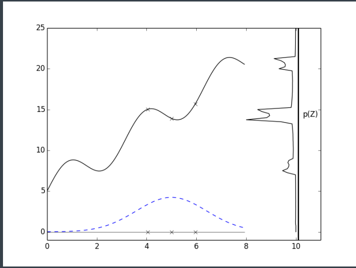
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

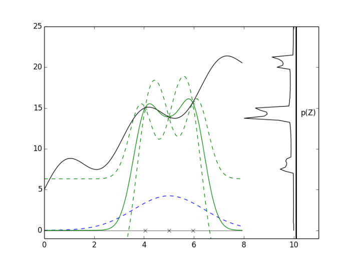
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

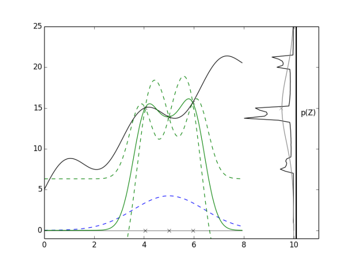
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightarrow \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

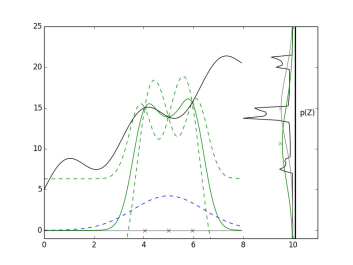
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

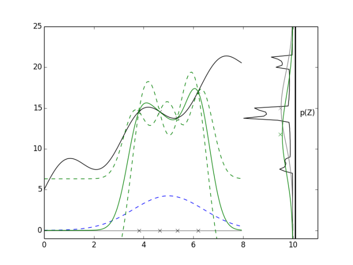
An initial value problem (IVP)

$$\text{ODE}\quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightarrow \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

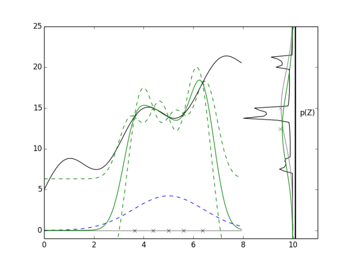
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightarrow \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

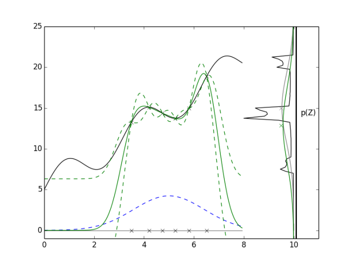
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, \mathrm{d}s,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

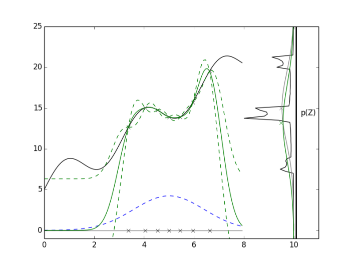
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

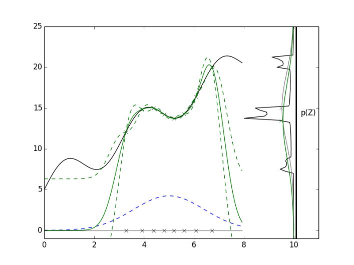
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

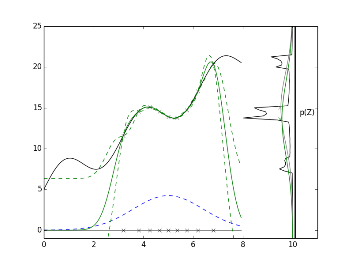
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \to \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

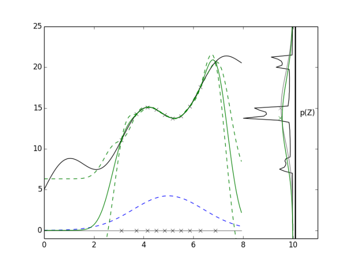
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightharpoonup \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

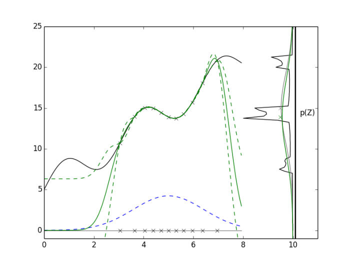
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightarrow \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s) \, \mathrm{d}s,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

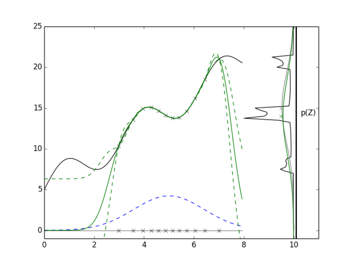
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightharpoonup \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, \mathrm{d}s,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

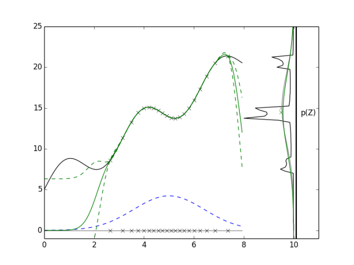
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightharpoonup \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, ds,$$

Bayesian Quadrature (BQ):



Big question: Can we construct a generalization of BQ for ODEs?

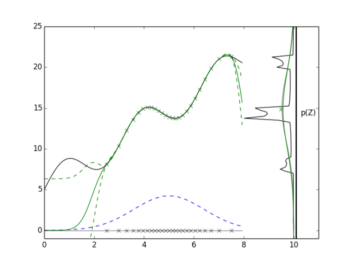
An initial value problem (IVP)

$$\text{ODE} \quad \dot{x}(t) = f(x(t), t) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d$$

generalizes quadrature problems by the fundamental theorem of calculus:

$$x :]0, T] \rightarrow \mathbb{R}^d, \qquad t \mapsto x_0 + \int_0^t f(x(s), s)\, \mathrm{d}s,$$

Bayesian Quadrature (BQ):

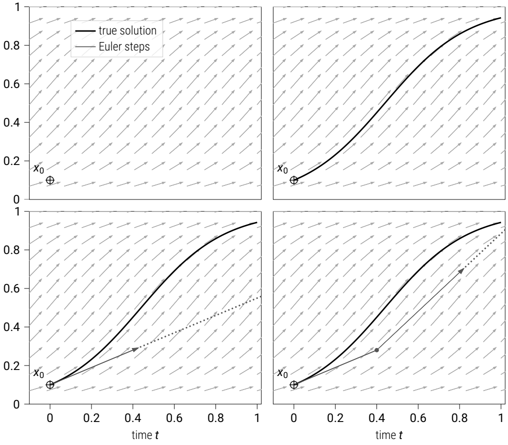


Big question: Can we construct a generalization of BQ for ODEs?

Euler's method

Solving ODEs by iterated first-order Taylor expansions [Euler, 1768]

logistic ODE: $\dot{x}(t) = f(x(t)) = 5x(t)[1 - x(t)]$ on $t \in [0, 1]$, with $x(0) = 0.1 \in \mathbb{R}^d$.



Runge–Kutta (RK) methods

RK methods match higher-order Taylor expansions [Runge, 1895, Kutta, 1901]

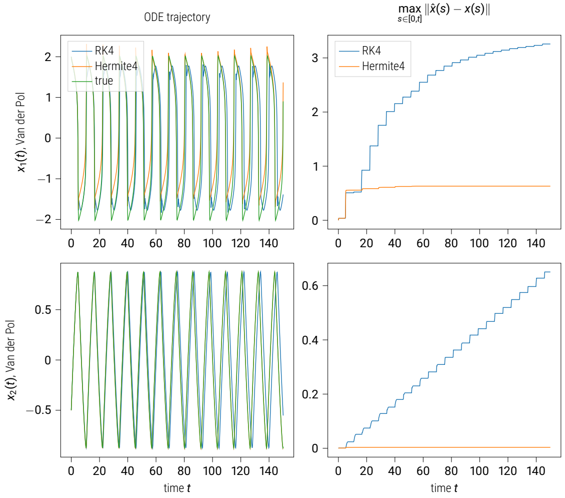
- **Explicit RK methods** (like many other classical solvers) generalize Euler's method. An s-stage RK method **chooses the coefficients** (a, b, c) in

$$\hat{x}(h) = x_0 + h \sum_{i=1}^{s} b_i y_i, \quad \text{with} \quad y_i = f\left(x_0 + h \sum_{j=1}^{i-1} a_{ij} y_j\right),$$

  to match the p-th order **Taylor** polynomial $\sum_{i=1}^{p} \frac{x^{(i)}(0)}{i!} h^i$ for a maximal $p \leq s$.
- E.g.,the **standard RK4 solver** fits a $p = 4$-th order Taylor polynomial with $s = 4$ stages.
- Hence, assuming $x(t) = \hat{x}(t)$, RK **assumes** to perform iterated **Hermite** interpolation with **perfect data** on $x^{(i)}(t)$, $i \in \{1, \dots, 4\}$.

Unaware of Uncertainty: RK falsely assumes perfect data

Since $x(t) \neq \hat{x}(t)$ for $t > 0$, the **uncertainty-unaware** assumptions of RK are **too optimistic**.



Inferring Dynamical Systems...

...from data (statistics) and from mechanistic model (numerics)

In realitas, all (temporal) dynamical systems $x : [0, T] \to \mathbb{R}^d$ follow an ODE:

$$\text{ODE} \quad \dot{x}(t) = f(x(t)) \quad \text{on} \quad t \in [0, T], \qquad \text{under initial condition} \qquad x(0) = x_0 \in \mathbb{R}^d.$$

Statistics: data available, but ODE unknown

Method: Treat the dynamical system like a time series. Data comes from a sensor, e.g. on the derivative

$$p(z_t) = \mathcal{N}(z_t;\ \dot{x}(t), R).$$

Numerics: ODE known, but no data available

Method: Construct iterative extrapolation with information

$$\dot{x}(t) = f(x(t)) \approx f(\hat{x}(t)) =: y_t.$$

where $\hat{x}(t)$ is the estimate of $x(t)$

If we construct a statistical model that treats y_t like z_t, **time-series methods are unlocked for ODEs**!

ODEs as a stochastic filtering problem

This view turns ODEs into a **stochastic filtering problem**:

$$\begin{tikzcd}
\text{hidden:} \arrow[r, dashed] & \mathbf{x}_{k-1} \arrow[r] \arrow[d] & \mathbf{x}_k \arrow[r] \arrow[d] & \mathbf{x}_{k+1} \arrow[r, dashed] \arrow[d] & {} \\
\text{observed:} & \mathbf{y}_{k-1} & \mathbf{y}_k & \mathbf{y}_{k+1} &
\end{tikzcd}$$

modeled by a D-dimensional **stochastic process** $\{X(t);\ t \in [0, T]\}$ from which the **solution**

$$x(t) \sim H_0 X(t), \qquad \text{for some } H_0 \in \mathbb{R}^{d \times D}. \tag{1}$$

and the **derivative**

$$\dot{x}(t) \sim HX(T) \qquad \text{for some } H \in \mathbb{R}^{d \times D}. \tag{2}$$

can be **linearly extracted**.

A Bayesian model for ODEs

1. **Prior:** linear time-invariant SDE

$$\mathrm{d}X(t) = FX(t)\,\mathrm{d}t + L\,\mathrm{d}B(t) \tag{3}$$

   which yield the discretized **dynamic model**

$$p\left(X(t+h) \mid X(t)\right) = \mathcal{N}(A(h)X(t), Q(h)). \tag{4}$$

2. **Likelihood:** The **measurement model** uses the current mismatch between solution and derivative state:

$$p(Z(t) \mid X(t)) = \mathcal{N}\left(f(H_0 X(t)) - HX(t), R\right). \tag{5}$$

3. **Data:** By the ODE this mismatch is observed to be zero

$$Z(t) \equiv 0. \tag{6}$$

This is a complete **probabilistic state space model (SSM)** which unlocks new methods!

Mission complete: a statistical model for ODEs...

...that unlocks all of Bayesian filtering for ODEs

[Tronarp et al., 2019]

Recipe to invent new ODE solvers:

1. Choose a **prior** for x:

$$x(t) \sim \mathrm{d}X(t) = FX(t)\,\mathrm{d}t + L\,\mathrm{d}B(t)$$

2. Construct **SSM** with dynamic model from prior

3. **Copy** a Bayesian filter (or smoother) **from signal processing**

Mission complete: a statistical model for ODEs...

...that unlocks all of Bayesian filtering for ODEs

[Tronarp et al., 2019]

Recipe to invent new ODE solvers:

1. Choose a **prior** for x:

$$x(t) \sim \mathrm{d}X(t) = FX(t)\,\mathrm{d}t + L\,\mathrm{d}B(t)$$

2. Construct **SSM** with dynamic model from prior

3. **Copy** a Bayesian filter (or smoother) **from signal processing**

4. (*optional:* Use **active learning** to choose points $\hat{\xi}$ **for evaluations** $f(\hat{\xi})$, see [Kersting and Hennig, 2016])

Mission complete: a statistical model for ODEs...

...that unlocks all of Bayesian filtering for ODEs [Tronarp et al., 2019]

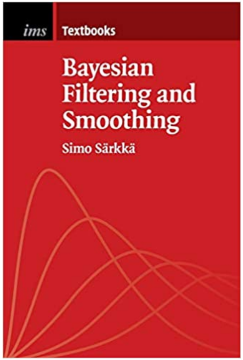
Recipe to invent new ODE solvers:

1. Choose a **prior** for x:

$$x(t) \sim \mathrm{d}X(t) = FX(t)\,\mathrm{d}t + L\,\mathrm{d}B(t)$$

2. Construct **SSM** with dynamic model from prior
3. **Copy** a Bayesian filter (or smoother) **from signal processing**
4. (*optional:* Use **active learning** to choose points $\hat{\xi}$ **for evaluations** $f(\hat{\xi})$, see [Kersting and Hennig, 2016])

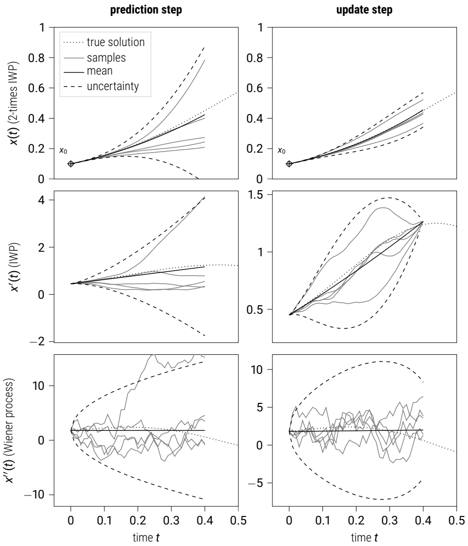
Alternative title: 'Probabilistic Numerical Methods for ODEs'



The resulting **ODE filters and smoothers** inherit the excellent linear-time properties of filters and smoothers!

Illustration of an ODE filtering step

Kalman ODE filtering with IWP prior

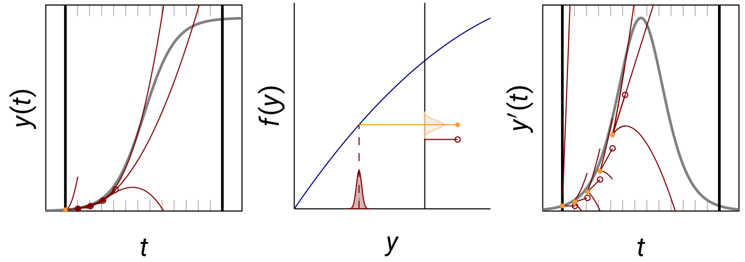


An elementary example: Kalman ODE Filtering

an iterative application of Bayes' rule

Plots from [Schober et al., 2019]

In every step, a Bayes' update is computed...

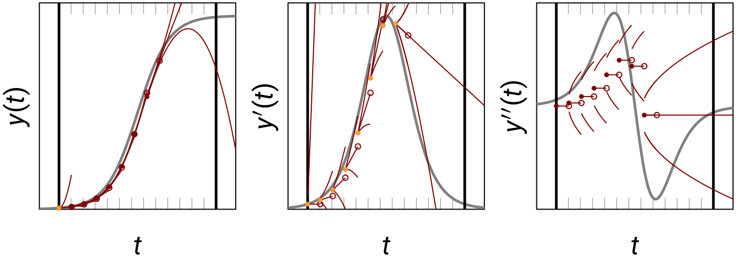


predict (prior) **observe** (likelihood) **update** (on data $z = 0$)

An elementary example: Kalman ODE filter

Plots from [Schober et al., 2019]

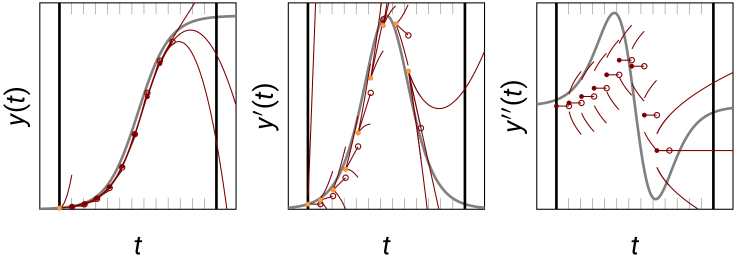
...which when iterated over **[0, T]**...



An elementary example: Kalman ODE filter

Plots from [Schober et al., 2019]

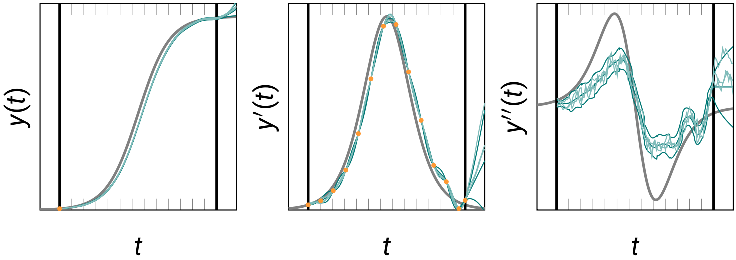
...which when iterated over **[0, *T*]**...



An elementary example: Kalman ODE filter

Plots from [Schober et al., 2019]

...yields a posterior distribution.



Choice of prior

a new way to categorize ODEs

[Kersting and Mahsereci, 2020]

1. **Matérn**: the general Gauss-Markov prior for **highly-specific ODEs**

$$\mathrm{d}\mathbf{X}(t) = \begin{pmatrix} 0 & 1 & 0 \dots & 0 \\ \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 1 \\ c_0 & \dots & \dots & c_q \end{pmatrix} \mathbf{X}(t)\,\mathrm{d}t + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sigma \end{pmatrix} \mathrm{d}B(t),$$

2. **integrated Brownian motion** $(c_0, \dots, c_q) = 0$: for **generic ODEs** (Taylor predictions)
3. **integrated Ornstein–Uhlenbeck process** $(c_0, \dots, c_{q-1}) = 0, c_q = -\theta$: for **ODEs with drift** to equilibrium level explored in [Kersting et al., 2020b]
4. **periodic prior**: Fourier predictions for **oscillators**

$$\mathrm{d}\mathbf{X}(t) = \begin{pmatrix} F_1 & & 0 \\ & \ddots & \\ 0 & & F_J \end{pmatrix} \mathbf{X}(t)\,\mathrm{d}t + 0\,\mathrm{d}B(t), \quad \text{with } F_j = \begin{bmatrix} 0 & -jw_0 \\ jw_0 & 0 \end{bmatrix},$$

which approximates the periodic kernel [Kersting and Mahsereci, 2020].

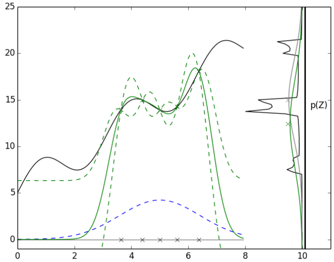
Active learning in ODE filtering

beyond signal-processing literature [Kersting and Hennig, 2016]

- Unlike in signal processing, ODE sovers can **decide** at which point $\hat{\xi} \in \mathbb{R}^d$ to collect inform. $f(\hat{\xi})$
- Given a predictive distribution $p(x(ih)) = \mathcal{N}(x(ih);\ m_i^-, P_i^-)$, the **expected value of $\dot{x}(ih)$** is

$$\mathbb{E}\left[\dot{x}(ih) \mid m_i^-, P_i^-\right] = \int f(\xi)\, \mathrm{d}\mathcal{N}(\xi; m_i^-, P_i^-); \qquad \forall i \in \{1, \dots, N\} \tag{7}$$

- active learning by BQ of eq. (7) for each i gives **better-calibrated uncertainty**, as shown in [Kersting and Hennig, 2016]
- **joint active learning** for all $i = 1, \dots, N$ seems **promising future research** to reduce # func. evals.



Theory

BQ (with GM prior) $\subset$ ODE filtering

ODE filtering generalizes BQ whenever applicable

Proposition

*Consider the **integral***

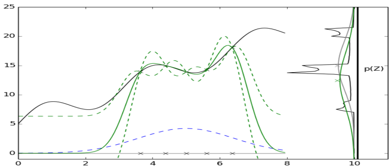
$$\int_0^T g(t)\,\mathrm{d}t \tag{8}$$

and let $\mathbf{g}$ *be **a priori** modeled by a Gauss-Markov* $\{X_t;\ t \in [0,T]\}$ *process. Then, the **Kalman ODE filter** computes the BQ posterior to the IVP*

$$\textit{ODE}\quad \dot{x}(t) = g(t) \quad \textit{on} \quad t \in [0,T], \qquad \textit{with initial condition} \qquad x(0) = 0 \in \mathbb{R}^d \tag{9}$$

in the sense that the distribution $p(x(T) \mid z_{1:T})$ *coincides with the BQ output for Eq.(8) with design points* $\{ih;\ i = 1, \dots, \frac{T}{h}\}$.

Proof: See Appendix A of [Tronarp et al., 2019].



[Kersting et al., 2020b, Theorem 8]

Theorem (local convergence rates)

Let f be sufficiently regular, and $R > 0$ an arbitrary noise model. If

- $q \in \mathbb{N}$, *and*
- *the prior X is a q-times integrated Ornstein–Uhlenbeck process or integrated Brownian motion*

then we locally have:

- ***optimal polynomial convergence**: $\|m(h) - x(h)\| \leq Kh^{q+1}$, and*
- ***asymptotically well-calibrated uncertainties**: $\sqrt{P(h)} \leq Kh^{q+1/2}$.*

Proof idea:

(i) note that the predictive mean deviates from a q^{th} Taylor expansion by $\mathcal{O}(h^{q+1})$,
(ii) apply Taylor's theorem to the predictive mean, and
(iii) use multiple triangle and Lipschitz inequalities. □

Why is a global convergence proof more difficult?

Because we cannot assume the steady-state!

[Kersting et al., 2020b, Proposition 10]

In every **Kalman ODE filtering** step $nh \to (n+1)h$

$$m_{n+1} = m_n + K_n \left[f(m_{n+1}^{(0)-}) - m_{n+1}^{(1)-} \right],$$

the **Kalman gain** $K_n = P_n^- H_n^\intercal \left[H P_n^- H^\intercal + R \right]^{-1}$ **is adaptive** to num. uncertainty P_n^- and eval. noise R.

Proposition (global bounds on Kalman gains)

*For all constant $R \geq 0$, then the **limit steady-state** is*

$$\lim_{n\to\infty} K_n^{(0)} = \frac{\sqrt{4R\sigma^2 h + \sigma^4 h^2}}{\sigma^2 h + \sqrt{4\sigma^2 Rh + \sigma^4 h^2}} h, \qquad \lim_{n\to\infty} K_n^{(1)} = \frac{\sigma^2 h + \sqrt{4\sigma^2 Rh + \sigma^4 h^2}}{\sigma^2 h + \sqrt{4\sigma^2 Rh + \sigma^4 h^2} + 2R}.$$

If moreover the initial covariance P_0 is small enough and $R \equiv Kh^p$ with $p \in [0, \infty]$, then

(global bounds on gains) $$\max_{n\in\{1,\dots,N\}} \left\| K_n^{(0)} \right\| \leq Kh, \qquad \max_{n\in\{1,\dots,N\}} \left\| 1 - K_n^{(1)} \right\| \leq Kh^{(p-1)\vee 0}.$$

Theorem (global convergence rates)

Let f be sufficiently regular, and $R \equiv Kh^p$ for some $p \geq 1$. If

- $q = 1$, *and*
- *prior X is a q-times integrated Brownian motion.*

then we globally have:

- ***optimal polynomial convergence***: $\left\|m(T) - x(T)\right\| \leq K(T)h^q$
- ***asymptotically well-calibrated uncertainties***: $\sqrt{P(T)} \leq K(T)h^q$.

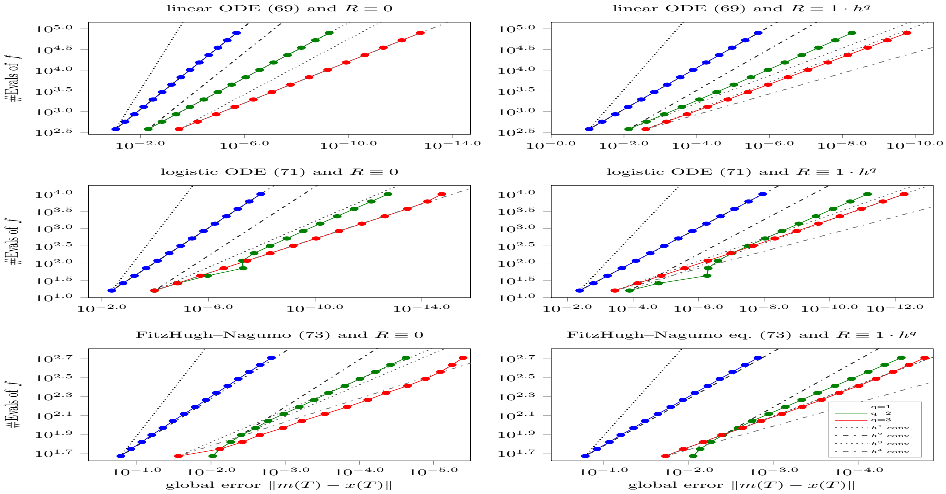
Proof idea: Define $\varepsilon(nh) := \left\|m(nh) - x(nh)\right\|$, find limit of Kalman gains, and *prove that* (**difficult part**)

$$\varepsilon((n+1)h) - \varepsilon(nh) \leq Kh^{q+1} + Kh^q \sum_{l=0}^{n-1} \left[\varepsilon((l+1)h) - \varepsilon(lh)\right]. \qquad (10)$$

and apply a special version of the **discrete Grönwall inequality** from [Clark, 1987] to (10). □

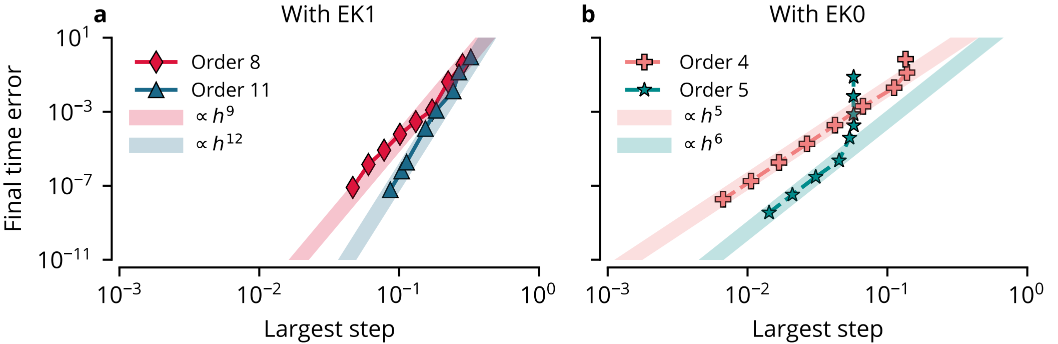
Experiments: $\mathcal{O}(h^q)$ rates of Kalman ODE Filters

are confirmed and seem to extend to $q \in \{2, 3, \ldots\}$! [Kersting et al., 2020b, Section 8]



New Experiments by [Krämer and Hennig, 2020]...

...extend the $\mathcal{O}(h^q)$ rates to up to q = 11!!!   source: N. Krämer, P. Hennig "Stable Implementation of Probabilistic ODE Solvers", 2020



Even **rates of $\mathcal{O}(h^{q+1})$** are observed here (and in previous experiments).

Particle ODE filter captures true posterior

detects Bifurcation and other non-Gaussian phenomena at higher cost

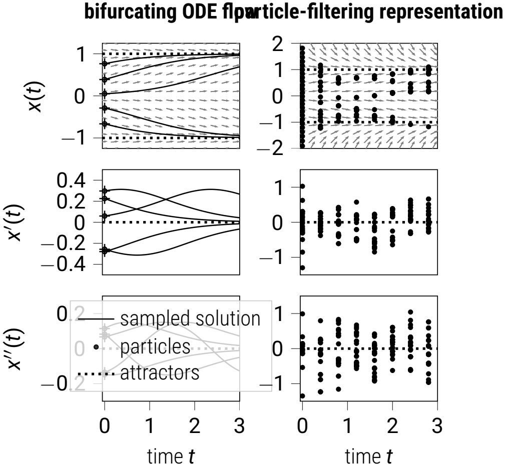
[Tronarp et al., 2019, Theorem 1]

- The **true posterior** is **non-Gaussian**.
- **Particle ODE filtering** approximates the **true posterior** weakly.
- **Standard MCMC rate**
  of $\mathcal{O}(\sqrt{\#\ \textbf{particles}})$

Particle ODE filter captures true posterior

detects Bifurcation and other non-Gaussian phenomena at higher cost

[Tronarp et al., 2019, Theorem 1]



Inverse Problems

ODE Inverse Problems

A mix of numerics and statistics

[Kersting et al., 2020a]

Forward Map (likelihood): $\theta \mapsto F(\theta)$

$$\begin{array}{ccc} \boxed{\begin{array}{c}\text{Parameter}\\ \theta \in \Theta\end{array}} & \begin{array}{c}\longrightarrow\\ \longleftarrow\end{array} & \boxed{\begin{array}{c}\text{Simulation}\\ F(\theta) \in \mathbb{R}^d\end{array}} \end{array}$$

Inverse Problem: $F(\theta_{\text{true}})$ + 'noise' $\mapsto \theta_{\text{true}}$

The **ODE** $\dot{x}(t) = f(x(t), \theta)$ on $t \in [0, T]$, with $x(0) = x_0 \in \mathbb{R}^d$,

has **forward map** $F : \theta \mapsto \left[t \mapsto x_0 + \int_0^t f(x(s), \theta)\, \mathrm{d}s \right]$.

- The forward problem is **well-posed**. (numerical analysis)
- The inverse problem is **ill-posed**. (statistics, machine learning)
- The mix of numerical and statistical estimation invites a treatment by **probabilistic numerics**.

ODE filter inserts uncertainty-aware likelihood...

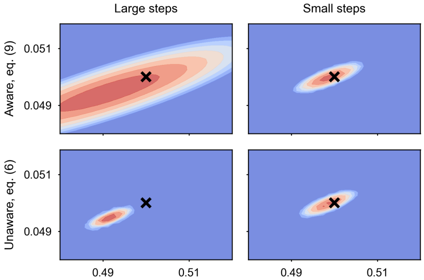
...into 'likelihood-free' ODE inverse problems [Kersting et al., 2020a]

- Inverse problems are called **likelihood-free** if F is **too expensive** to approximate exactly, as is the case for **ODEs**
- Suppose we observe **noisy data** $\mathbf{z} = z(t_{1:M})$ of the true $\mathbf{x} = x(t_{1:M})$, then the lik.-free case uses an

***uncertainty-unaware* likelihood** $$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z};\ \mathbf{x}, \sigma^2 I_M\right) \overset{\text{(lik.-free)}}{=} \mathcal{N}\left(\mathbf{z};\ \hat{\mathbf{x}}, \sigma^2 I_M\right)$$

- If we however use the **ODE-filtering distribution** $\mathcal{N}(\mathbf{x}_\theta; \mathbf{m}_\theta, \mathbf{P})$, we obtain an

***uncertainty-aware* likelihood** $$p(\mathbf{z} \mid \theta) = \mathcal{N}(\mathbf{z};\ x_0 + J\theta,\ \underbrace{\mathbf{P} + \sigma^2 I_M)}_{\text{num. + stat. var.}}\ )$$

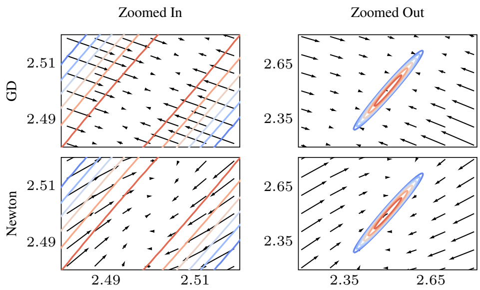


Gradients and Hessians are now available...

...because the filtering mean is twice differentiable [Kersting et al., 2020a]

The ***uncertainty-aware* likelihood** $p(\mathbf{z} \mid \theta) = \mathcal{N}(\mathbf{z};\ \underbrace{x_0 + J\theta}_{\in C^2(\Theta,\mathbb{R}^d)}, \mathbf{P} + \sigma^2 I_M))$ is **twice differentiable**,

using the **Jacobian estimator** $J = J(\hat{\theta})$ of $\theta \mapsto \mathbf{x}_\theta$ which is **freely-available** from ODE-filtering output.

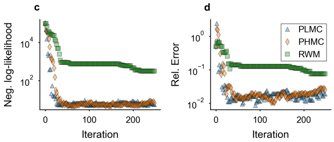


Gradient-based outperforms likelihood-free/gradient-free

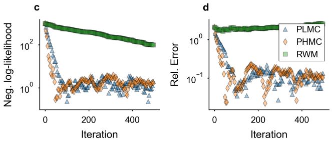
the modeling of uncertainty by PN increases sample-efficiency [Kersting et al., 2020a]

Sampling: Gradient-based methods **more quickly** find and cover **regions of high probability**.

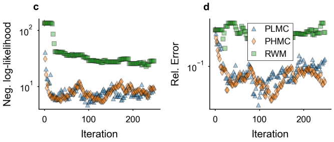
Lotka Volterra



Protein Transduction

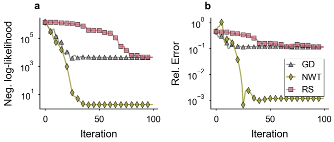


Glucose Uptake in Yeast

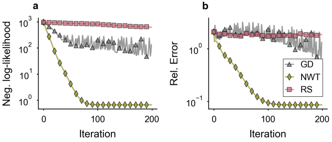


Optimization: Gradient-based optimizers find local maxima **with less samples**.

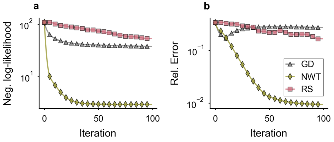
Lotka Volterra



Protein Transduction



Glucose Uptake in Yeast



Summary

We demonstrated how ODE filters (and smoothers) advance the three promises of PN:

1. invention of new statistically-optimal algorithms,
2. more flexible classification of problems by statistical model selection, and
3. comprehensive uncertainty quantification.

Summary

We demonstrated how ODE filters (and smoothers) advance the three promises of PN:

1. invention of new statistically-optimal algorithms,
2. more flexible classification of problems by statistical model selection, and
3. comprehensive uncertainty quantification.

Further material:

- Check out the Python code in the probnum-package: https://probnum.readthedocs.io
- Check out recent publications by F. Tronarp, N. Bosch and N. Krämer (Tübingen)
- Stay tuned for the PN book

Thank you!

Bibliography

- Clark, D. S. (1987).
  Short proof of a discrete Gronwall inequality.
  Discrete Appl. Math., 16(3):279–281.
- Euler, L. (1768).
  Institutionum Calculi Integralis, volume I.
- Hennig, P., Osborne, M. A., and Girolami, M. (2015).
  Probabilistic numerics and uncertainty in computations.
  Proc. Roy. Soc. London A, 471(2179).
- Kersting, H. and Hennig, P. (2016).
  Active uncertainty calibration in Bayesian ODE solvers.
  Uncertainty in Artificial Intelligence (UAI).
- Kersting, H., Krämer, N., Schiegg, M., Daniel, C., Tiemann, M., and Hennig, P. (2020a).
  Differentiable likelihoods for fast inversion of 'likelihood-free' dynamical systems.
  In *International Conference on Machine Learning (ICML)*.
- Kersting, H. and Mahsereci, M. (2020).
  A Fourier state space model for Bayesian ODE filters.
  In *Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models, ICML*.
- Kersting, H., Sullivan, T. J., and Hennig, P. (2020b).
  Convergence rates of Gaussian ODE filters.
  Stat. Comput., 30(6):1791–1816.
- Krämer, N. and Hennig, P. (2020).
  Stable implementation of probabilistic ODE solvers.
  ArXiv e-prints, stat.ML 2012.10106.
- Kutta, W. (1901).
  Beitrag zur näherungsweisen Integration totaler Differentialgleichungen.
  Zeitschrift für Mathematik und Physik, 46:435–453.
- Oates, C. J. and Sullivan, T. J. (2019).
  A modern retrospective on probabilistic numerics.
  Stat. Comput., 29(6):1335–1351.
- Runge, C. (1895).
  Über die numerische Auflösung von Differentialgleichungen.
  Mathematische Annalen, 46:167–178.
- Schober, M., Duvenaud, D., and Hennig, P. (2014).
  Probabilistic ODE solvers with Runge–Kutta means.
  In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Schober, M., Särkkä, S., and Hennig, P. (2019).
  A probabilistic model for the numerical solution of initial value problems.
  Stat. Comput., 29(1):99–122.
- Tronarp, F., Kersting, H., Särkkä, S., and Hennig, P. (2019).
  Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective.
  Stat. Comput., 29(6):1297–1315.

Backup

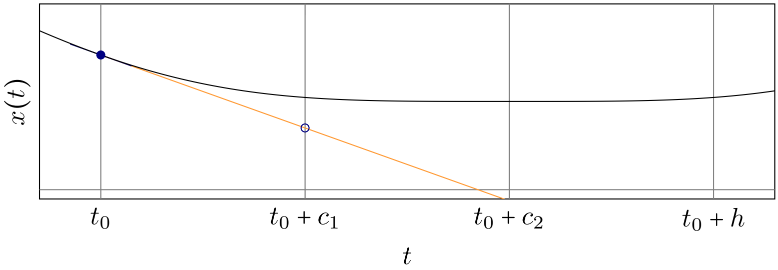
Classical solutions of IVPs

plots: Runge-Kutta of order 3

...we need to extrapolate ($x_k \rightarrow x_{k-1}$), how do classical solvers do that?

- Estimate $\dot{x}(t_i)$, $t_0 \leq t_1 \leq \cdots \leq t_n \leq t_0 + h$ by evaluating $y_i \approx f(\hat{x}(t_i))$, where $\hat{x}(t)$ is itself an estimate for $x(t)$
- Use this data $y_i := \dot{x}(t_i)$ to estimate $x(t_0 + h)$, i.e.

$$\hat{x}(t_0 + h) \approx x(t_0) + h \sum_{i=1}^{b} w_i y_i.$$



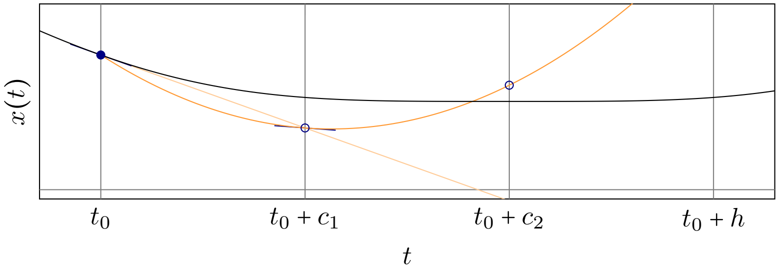
Classical solutions of IVPs

plots: Runge-Kutta of order 3

...we need to extrapolate ($x_k \to x_{k-1}$), how do classical solvers do that?

- Estimate $\dot{x}(t_i)$, $t_0 \leq t_1 \leq \cdots \leq t_n \leq t_0 + h$ by evaluating $y_i \approx f(\hat{x}(t_i))$, where $\hat{x}(t)$ is itself an estimate for $x(t)$
- Use this data $y_i := \dot{x}(t_i)$ to estimate $x(t_0 + h)$, i.e.

$$\hat{x}(t_0 + h) \approx x(t_0) + h \sum_{i=1}^{b} w_i y_i.$$



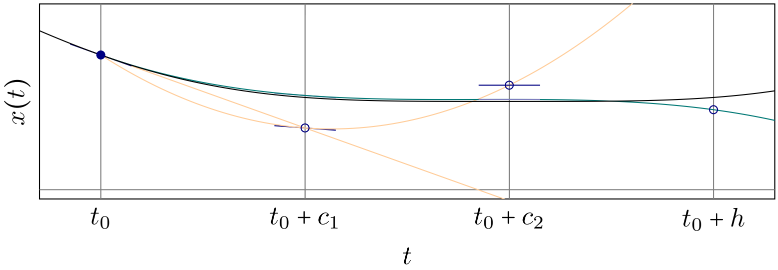
Classical solutions of IVPs

plots: Runge-Kutta of order 3

...we need to extrapolate ($x_k \rightarrow x_{k-1}$), how do classical solvers do that?

- Estimate $\dot{x}(t_i)$, $t_0 \leq t_1 \leq \cdots \leq t_n \leq t_0 + h$ by evaluating $y_i \approx f(\hat{x}(t_i))$, where $\hat{x}(t)$ is itself an estimate for $x(t)$
- Use this data $y_i := \dot{x}(t_i)$ to estimate $x(t_0 + h)$, i.e.

$$\hat{x}(t_0 + h) \approx x(t_0) + h \sum_{i=1}^{b} w_i y_i.$$



Classical solutions of IVPs

plots: Runge-Kutta of order 3

...we need to extrapolate ($x_k \rightharpoonup x_{k-1}$), how do classical solvers do that?

- Estimate $\dot{x}(t_i)$, $t_0 \leq t_1 \leq \cdots \leq t_n \leq t_0 + h$ by evaluating $y_i \approx f(\hat{x}(t_i))$, where $\hat{x}(t)$ is itself an estimate for $x(t)$
- Use this data $y_i := \dot{x}(t_i)$ to estimate $x(t_0 + h)$, i.e.

$$\hat{x}(t_0 + h) \approx x(t_0) + h \sum_{i=1}^{b} w_i y_i.$$

Information in these calculations:

$$\dot{x}(t) = f(x(t)) \approx f(\hat{x}(t))$$

For information, f is evaluated at (or around) the current numerical estimate $\hat{x}$ of x!

Equivalences with Classical Methods

Since the **prediction step** is a Taylor expansion

$$m^{-}(t+h) = \sum_{i=1}^{n} \frac{h^i}{i!} m^{(i)}(t+th) \tag{11}$$

and since generic numerical methods match **Taylor expansions** (as they are worst-case optimal), the Kalman ODE filter with **Integrated Brownian motion prior** has been found to be

- **equivalent to Runge–Kutta** (by some unnatural choices) [Schober et al., 2019],
- **equivalent to Nordsieck methods** (when covariance matrices in stationary state) [Schober et al., 2014].

Choosing the uncertainty scale σ^2

tuning the constant before the asymptotically well-calibrated rates

[Tronarp et al., 2019, Proposition 4]

Recall that the **prior** on $x : [0, T]$ is defined by an **SDE**

$$\mathrm{d}\mathbf{X}(t) = F\mathbf{X}(t)\,\mathrm{d}t + \sigma L\,\mathrm{d}B(t),$$

driven by a **Brownian motion** $B(t)$ with **variance** $\sigma^2 > 0$ which linearly scales the **posterior variance**.

Theorem

The ***maximum-likelihood estimate*** $\hat{\sigma}^2$ *for* σ^2 *is given by*

$$\hat{\sigma}^2 = \frac{1}{Nd} \sum_{n=1}^{N} \left[f(H_0 m^-(nh)) - Hm^-(nh)\right]^\intercal \left[HP_n^- H^\intercal + R\right]^{-1} \left[f(H_0 m^-(nh)) - Hm^-(nh)\right].$$

In particular, for **q***-times* ***IBM prior***, $d = 1$, *and* $R = 0$:

$$\hat{\sigma}^2 = \frac{(2q-1)\cdot(q-1)!^2}{Nh^{2q-1}} \sum_{n=1}^{N} \left(f(H_0 m^-(nh)) - Hm^-(nh)\right)^2.$$

End of Backup