ODE Filters – Forward and Backward

Hans Kersting

Dagstuhl
26 October 2021

You already heard a few talk on ODE filters (smoothers). The goal of this talk is to fill in a few thoughts on the following topics:

- **the underlying SSM,**
- **forward convergence rates of ODE filters, and**
- **their use in inverse problems.**

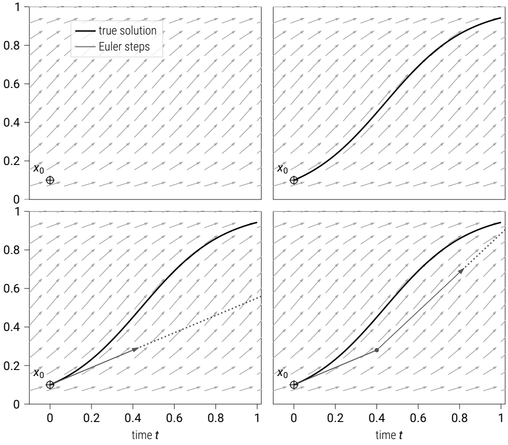
Introduction

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

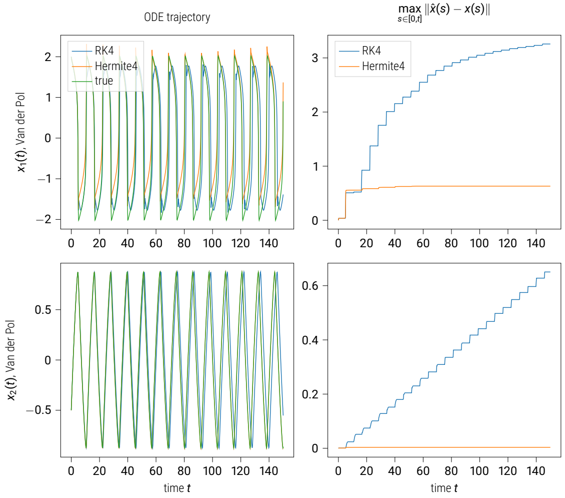
$$\hat{x}(h) = x_0 + h \sum_{i=1}^{s} b_i y_i, \quad \text{with} \quad y_i = f\left(x_0 + h \sum_{j=1}^{i-1} a_{ij} y_j\right),$$

  to match the p-th order **Taylor** polynomial $\quad \sum_{i=1}^{p} \frac{x^{(i)}(0)}{i!} h^i \quad$ 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 $x^{(i)}(t)$, $i \in \{1, \ldots, 4\}$.

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**.



Part I: State Space Models and ODE Filters

The development of the different formulations

Abbreviated History of (state-space) models for ODEs

1st phase: plain GP models

- **1991**: Skilling first proposes to model ODE solutions with **plain GPs**
- **2014–2016**: Hennig and Hauberg, Chkrebtii et al. and Schober et al. refine Skilling's proposition

2nd phase: Probabilistic SSM conditioning on the derivative

- **2016–2018**: Schober et al. and Kersting and Hennig introduce **path-dependent SSMs** conditioning on $y_t = f(\hat{x})$

3rd phase: Probabilistic SSM conditioning on the ODE:

- **2019–2021**: Tronarp et al. introduce **path-independent** SSMs conditioning on the ODE itself, leading to new algorithms

1st phase: GP models

1991–2016

Skilling (1991); Hennig and Hauberg (2014); Chkrebtii et al. (2016); Schober et al. (2014)

ODE: $$\dot{x}(t) = f(x(t))$$

---

prior:

$$x(t) \sim \mathcal{GP}(m(t), k(t, t')) \tag{1}$$

***n*-th data-likelihood-pair (after *n* steps):**

$$y_n := f(m(t_n)|_{y_{1:n-1}}) \tag{2}$$

$$p(y_n \mid x_n) = \mathcal{N}(y_n; \dot{x}_n, R_n). \tag{3}$$

This SSM is **“path-dependent”** because likelihood and data depend via $m_{y_{1:n-1}}(t_n)$ on $y_{1:n-1}$.

2nd phase: path-dependent stochastic filtering

2016–2018 Kersting and Hennig (2016); Schober et al. (2018)

ODE: $$\dot{x}(t) = f(x(t))$$

---

prior:

$$x(t) \sim d\mathbf{X}(t) = F\mathbf{X}(t)\,dt + L\,d\omega_t, \tag{4}$$

n-th data-likelihood-pair (after n steps):

$$y_n := f(m(t_n)|_{y_{1:n-1}}) \tag{5}$$

$$p(y_n \mid x_n) = \mathcal{N}(y_n; \dot{x}_n, R_n). \tag{6}$$

This SSM is **“path-dependent”** because likelihood and data depend via $m_{y_{1:n-1}}(t_n)$ on $y_{1:n-1}$.

3rd phase: path-independent stochastic filtering

2016–2018 Tronarp et al. (2019a)

ODE: $\dot{x}(t) = f(x(t))$

---

prior:

$$x(t) \sim d\mathbf{X}(t) = F\mathbf{X}(t)\,dt + L\,d\omega_t, \tag{7}$$

likelihood:

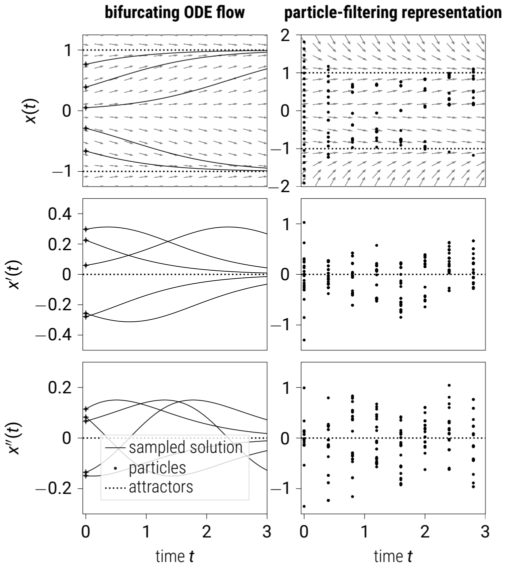
$$p(z_n \mid x(t_n)) = \delta(\dot{x}(t_n) - f(x(t_n))) \tag{8}$$

data:

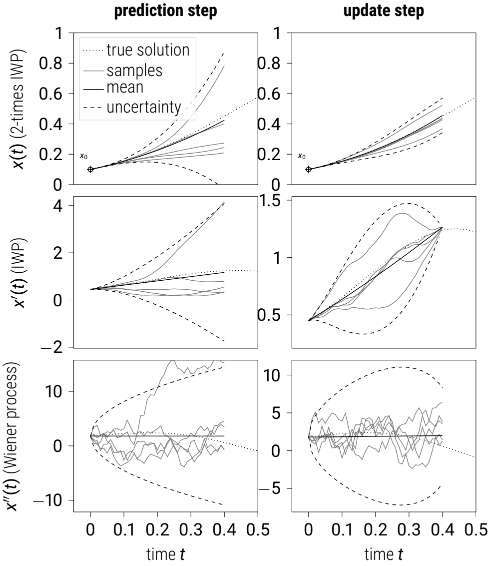
$$z_n := 0, \tag{9}$$

This SSM is **"path-independent"** because likelihood and data are seperate and independent of previous computations.

The path-dependent SSM allows for the use of all Bayesian filters...



...while in the path-dependent SSM only Gaussian filters are possible.



The relation

Locally the path-dependent SSM is a Gaussian approximation of path-independent SSM

Let's consider a local case, i.e. we have arrived at

$$x(t_n) \sim \mathcal{N}(m_n^-, P_n^-). \tag{10}$$

Old logic:

$$y_n = \int f(\xi)\, d\mathcal{N}(\xi, m_n^-, P_n^-). \tag{11}$$

Use data $y_n = f(m^-)$ to update the following likelihood

- $p(y_n \mid x_n) = \mathcal{N}(y_n; \dot{x}_n, 0)$ (**"Kalman ODE filter"**)
- $p(y_n \mid x_n) = \mathcal{N}(y_n; Hx_n + J_f(m_n^-)[y_n - x_n], R_n + J_f(m_n^-)P_n^- J_f(m_n^-)^\top)$ (**"extended Kalman ODE filter"**).

The relation

Locally the path-dependent SSM is a Gaussian approximation of path-independent SSM

Let's consider a local case, i.e. we have arrived at

$$x(t_n) \sim \mathcal{N}(m_n^-, P_n^-). \tag{10}$$

New logic: Replace f by its Taylor expansion $\hat{f}$ around m_n^- and update in the approximated model

$$p(z_n \mid x(t_n)) = \delta(\dot{x}(t_n) - \hat{f}(x(t_n))) \tag{11}$$

For an m-th order Taylor expansion the resulting ODE filters are called **"EKFm"**.

The relation

Locally the path-dependent SSM is a Gaussian approximation of path-independent SSM

Let's consider a local case, i.e. we have arrived at

$$x(t_n) \sim \mathcal{N}(m_n^-, P_n^-). \tag{10}$$

	previous name	new name
$n = 0$	Kalman filter	EKF0
$n = 1$	extended Kalman filter	EKF1

Part II: Convergence Rates

An exposition of three different sets of convergence results

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., 2020, 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**: $\|m(T) - x(T)\| \leq K(T)h^q$*
- ***asymptotically well-calibrated uncertainties**: $\sqrt{P(T)} \leq K(T)h^q$.*

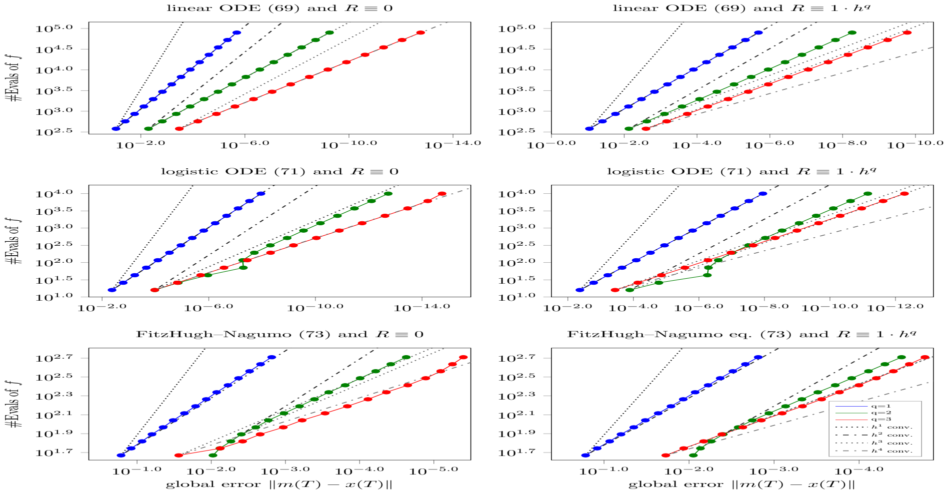
Proof idea: Define $\varepsilon(nh) := \|m(nh) - x(nh)\|$, find limit of Kalman gains, and *prove that* (**difficult part**)

$$\varepsilon((n+1)h) - \varepsilon(nh) \leq Kh^{q+1} + Kh^q \sum_{l=0}^{n-1} \left[\varepsilon((l+1)h) - \varepsilon(lh)\right]. \tag{11}$$

and apply a special version of the **discrete Grönwall inequality** from Clark (1987) to (11). □

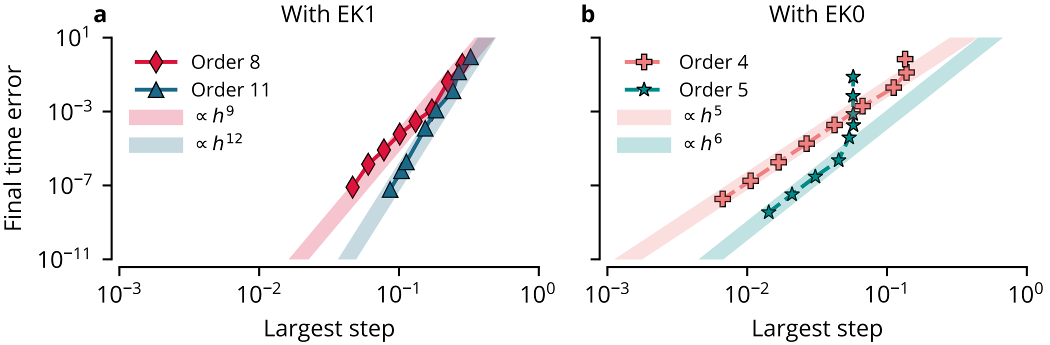
Experiments: $\mathcal{O}(h^q)$ rates of Kalman ODE Filters

are confirmed and seem to extend to $q \in \{2, 3, \dots\}$! (Kersting et al., 2020, 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).

MAP estimate convergence rates

according to “Bayesian ODE Solvers: The Maximum A Posteriori Estimate” by Tronarp et al. (2021)

Assumption

Let $f \in C^{q+1}(\mathbb{R}^d; \mathbb{R}^d)$.

Unlike the above convergence rates, Tronarp et al. (2021) considered the MAP

$$\vec{\mathbf{x}}^*(t_{0:N}) := \underset{\vec{\mathbf{x}}(t_{0:N})}{\arg\min} \left[-\log\left(p\left(\vec{\mathbf{x}}(t_{0:N}) \mid z_{1:N} = 0\right)\right)\right]. \tag{12}$$

Theorem

Under this Assumption and for any prior $\mathbf{X}(t)$ *of smoothness* q, *there exists a constant* $C(T) > 0$ *such that*

$$\max_{n=0,\dots,N} \left\| x^*(t_n) - x(t_n) \right\| \leq C(T) h^q, \tag{13}$$

where $x^*(t_n) = H_0 \vec{\mathbf{x}}^*(t_n)$ *is the MAP estimate of* $x(t_n)$ *given a discretisation* $0 = t_0 \leq t_1 \leq \cdots \leq t_N = T$.

Convergence Rates via Nordsieck Equivalences

according to "A probabilistic model for the numerical solution of initial value problems" by Schober et al. (2018)

Schober et al. (2018) observed the equivalence of the filtering mean with Nordsieck methods

$$\widehat{\vec{\mathbf{x}}}(t+h) = \left[I - l\bar{H}\right] \bar{A}\,\widehat{\vec{\mathbf{x}}}(t) + hlf(H_0\bar{A}\,\widehat{\vec{\mathbf{x}}}(t)), \tag{14}$$

Theorem

A ***Kalman ODE filter (EKF0)*** *with* 2*-times integrated Wiener process prior coincides with the* ***trapezoidal rule***

$$\hat{x}_{n+1} = \hat{x}_n + \frac{h}{2}\left(f(\hat{x}_n) + f(\tilde{x}_{n+1})\right), \tag{15}$$

with $\tilde{x}_{n+1} := \hat{x}_n + hf(\hat{x}_n)$. In particular, if initialized in this steady state, we have

$$\left\|m(T) - x(T)\right\| \leq C(T)h^3 = C(T)h^{q+1}. \tag{16}$$

Overview convergence rates

So, what is known about (extended) Kalman ODE filters with q-times integrated Wiener process prior?

Classical analysis

- local rates: $\mathcal{O}(h^{q+1})$, for all q
- global rates: $\mathcal{O}(h^{q})$, for $q = 1$
- main restriction: $q = 1$
- way forward: Repeat the proof for higher q.

MAP estimate

- local rates: –
- global rates: $\mathcal{O}(h^{q})$ of the MAP
- main restriction: No known method converges to the MAP.
- way forward: Show that some method approximates the MAP.

Nordsieck equivalence

- local rates: $\mathcal{O}(h^{q+2})$, for $q = 1, 2$
- global rates: $\mathcal{O}(h^{q+1})$, for $q = 1, 2$
- main restriction: $q = 1, 2$ and only in the steady-state
- way forward: Find more Nordsieck equivalences.

Part III: Inverse Problems

A first uncertainty-aware computational chain

Inverse Problems...

...are defined by their forward map F

Cranmer et al. (2020)

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

$$\begin{array}{c} \text{Parameter} \\ \theta \in \Theta \end{array} \quad \rightleftarrows \quad \begin{array}{c} \text{Simulation} \\ F(\theta) \in \mathbb{R}^d \end{array}$$

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

- 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**.

Inverse problems are called **likelihood-free** if their **forward map** is **too expensive** to approximate exactly.

ODE Inverse Problems...

...are only likelihood-free because they have a numerical forward map

Cranmer et al. (2020)

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

Parameter $\theta \in \Theta$

Simulation $F(\theta) \in \mathbb{R}^d$

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

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

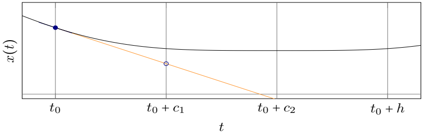
$\forall \theta \in \Theta$, **ODEs** have a **well-defined solution**

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

and hence an **high-fidelity** forward map

$$F : \Theta \to C^1([0, T]; \mathbb{R}^d), \qquad \theta \mapsto x_\theta.$$

- x_θ has to be estimated with **non-zero step size** $h > 0$, i.e. with **low fidelity**!
- With **numerical error**, e.g. **Runge–Kutta**:



ODE Inverse Problems...

...are only likelihood-free because they have a numerical forward map — Cranmer et al. (2020)

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

Parameter $\theta \in \Theta$ → Simulation $F(\theta) \in \mathbb{R}^d$

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

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

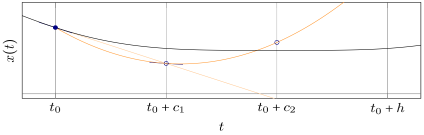
$\forall \theta \in \Theta$, **ODEs** have a **well-defined solution**

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

and hence an **high-fidelity** forward map

$$F : \Theta \to C^1([0, T]; \mathbb{R}^d), \qquad \theta \mapsto x_\theta.$$

- x_θ has to be estimated with **non-zero step size** $h > 0$, i.e. with **low fidelity**!
- With **numerical error**, e.g. **Runge–Kutta**:



ODE Inverse Problems...

...are only likelihood-free because they have a numerical forward map

Cranmer et al. (2020)

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

Parameter
$\theta \in \Theta$

Simulation
$F(\theta) \in \mathbb{R}^d$

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

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

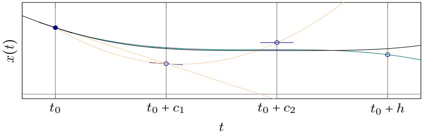
$\forall \theta \in \Theta$, **ODEs** have a **well-defined solution**

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

and hence an **high-fidelity** forward map

$$F : \Theta \to C^1([0, T]; \mathbb{R}^d), \qquad \theta \mapsto x_\theta.$$

- x_θ has to be estimated with **non-zero step size $h > 0$**, i.e. with **low fidelity**!
- With **numerical error**, e.g. **Runge–Kutta**:



ODE Inverse Problems...

...are only likelihood-free because they have a numerical forward map

Cranmer et al. (2020)

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

$$\begin{array}{c} \text{Parameter} \\ \theta \in \Theta \end{array} \rightleftarrows \begin{array}{c} \text{Simulation} \\ F(\theta) \in \mathbb{R}^d \end{array}$$

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

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

$\forall \theta \in \Theta$, **ODEs** have a **well-defined solution**

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

and hence an **high-fidelity** forward map

$$F : \Theta \to C^1([0, T]; \mathbb{R}^d), \qquad \theta \mapsto x_\theta.$$

- x_θ has to be estimated with **non-zero step size** $h > 0$, i.e. with **low fidelity**!
- With **numerical error**, e.g. **Runge–Kutta**:

In classical numerics, ODE inverse problems are likelihood-free!

Probabilistic numerics inserts a likelihood...

...into the 'likelihood-free' ODE inverse problem

Hennig et al. (2015)

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

$$\begin{tikzcd} \boxed{\begin{array}{c}\text{Parameter}\\ \theta \in \Theta\end{array}} \arrow[r, bend left] & \boxed{\begin{array}{c}\text{Simulation}\\ F(\theta) \in \mathbb{R}^d\end{array}} \arrow[l, bend left] \end{tikzcd}$$

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

- Inverse problems are called **likelihood-free** if F is **too expensive** to approximate exactly.
- ODE inverse problems are **likelihood-free** if **numerical error** is **unaccounted**.

Probabilistic numerics inserts a likelihood...

...into the 'likelihood-free' ODE inverse problem

Hennig et al. (2015)

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

Parameter
$\theta \in \Theta$

Simulation
$F(\theta) \in \mathbb{R}^d$

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

- Inverse problems are called **likelihood-free** if F is **too expensive** to approximate exactly.
- ODE inverse problems are **likelihood-free** if **numerical error** is **unaccounted**.

Likelihood-free → **Probabilistic Numerics** captures numerical error → Differentiable Likelihood

Gradient-free methods:

- Density estimation methods
- ABC

Gradient-based methods:

- Gradient descent
- Hamiltonian/Langevin MCMC

We propose the following likelihood.

Schober et al. (2019); Tronarp et al. (2019b); **?**

Assume that we observe **noisy data** $\mathbf{z} = z(t_{1:M})$ of the true $\mathbf{x} = x(t_{1:M})$, i.e:

$$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z};\ \mathbf{x}, \sigma^2 I_M\right). \tag{17}$$

For any θ, **Gaussian ODE Filtering**, a probabilistic numerical method, yields

$$p(\mathbf{z} \mid \theta) = \mathcal{N}(\mathbf{z};\ x_0 + J\theta, \quad \underbrace{\mathbf{P} + \sigma^2 I_M)}_{\text{numerical + statistical var.}} \quad) \tag{18}$$

where J is freely-availabe from the filtering output.

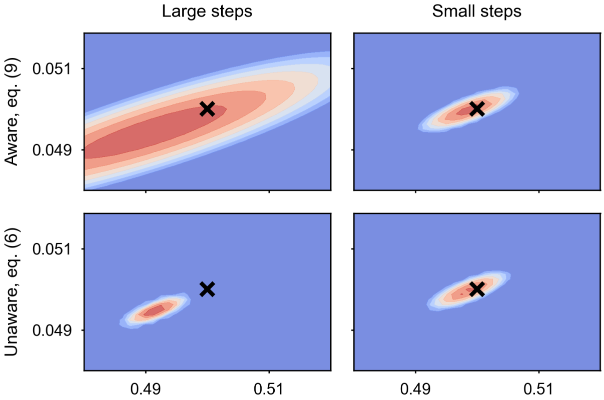
Two advantages:

- $\mathbf{P}$ accounts for then epistemic (numerical) uncertainty for non-zero step size $h > 0$, and
- $J = J(\hat{\theta})$ is an estimate of the Jacobian of $\theta \mapsto \mathbf{x}_\theta$ at some support point $\hat{\theta}$, and implies gradient and Hessian estimators

$$\hat{\nabla}_\theta E(\mathbf{z}) := -J^\intercal \left[\mathbf{P} + \sigma^2 I_M\right]^{-1} [\mathbf{z} - \mathbf{m}_\theta], \quad \text{and} \quad \hat{\nabla}^2_\theta E(\mathbf{z}) := J^\intercal \left[\mathbf{P} + \sigma^2 I_M\right]^{-1} J. \tag{19}$$

The likelihood account for the numerical/epistemic uncertainty!

- The **statistical (aleatoric) variance** $\sigma^2 I_M$ is accounted for in any case.
- The **numerical (epistemic) variance** $\mathbf{P}$ makes the implicit forward model tractable.

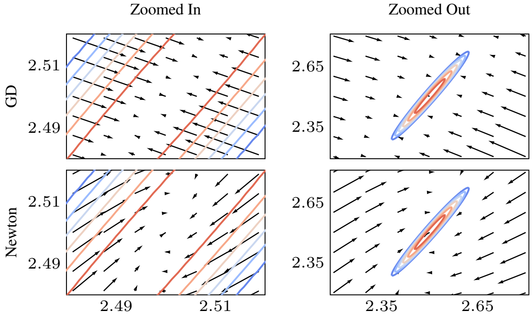


The gradients are accurate enough to point towards modes!

Both the

- **gradient** estimator, and
- the Hessian-precionditioned (**Newton**) gradient estimator

are **useful approximations**.



Insertion of gradients/Hessians unlocks gradient-free methods!

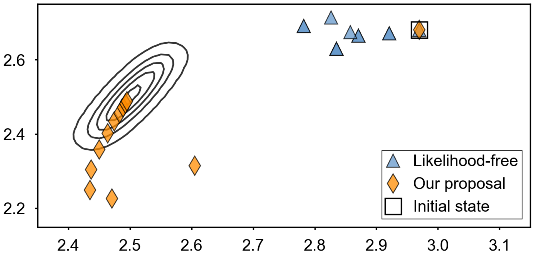
These **gradient-based** methods are more **sample-efficient**.

Sampling:

- Langevin MCMC
- Hamiltonian MCMC

Optimization:

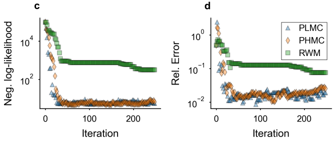
- Gradient descent
- Newton's Method



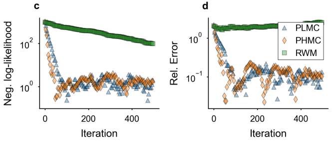
Sampling Experiments

- **Likelihood-free** random-walk **Metropolis** (RWM) **gets lost** in regions of low probability.
- **Gradient-based** sampling quickly finds and covers **regions of high probability**.

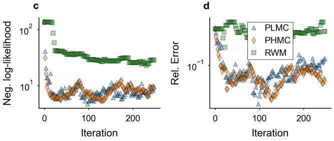
Lotka Volterra



Protein Transduction



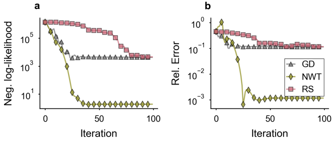
Glucose Uptake in Yeast



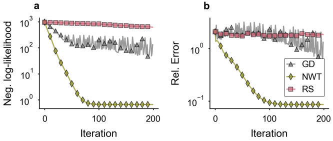
Optimization Experiments

- **Likelihood-free** random-search **hardly learns** at all.
- **Gradient-based** optimization **quickly** finds local maxima.

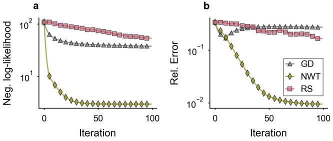
Lotka Volterra



Protein Transduction



Glucose Uptake in Yeast



Thank you for listening!

Bibliography

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

Clark, D. S. (1987). Short proof of a discrete Gronwall inequality. *Discrete Appl. Math.*, 16(3):279–281.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Euler, L. (1768). *Institutionum Calculi Integralis*, volume I.

Hennig, P. and Hauberg, S. (2014). Probabilistic Solutions to Differential Equations and their Application to Riemannian Statistics. In *Proc. of the 17th int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 33. JMLR, W&CP.

Hennig, P., Osborne, M. A., and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *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., Sullivan, T. J., and Hennig, P. (2020). 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.

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. (2018). A probabilistic model for the numerical solution of initial value problems. *Statistics and Computing*.

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.

Skilling, J. (1991). Bayesian solution of ordinary differential equations. *Maximum Entropy and Bayesian Methods, Seattle*.

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

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

Tronarp, F., Särkkä, S., and Hennig, P. (2021). Bayesian ODE solvers: The maximum a posteriori estimate. *Stat. Comput.*, 31(3):1–18.