Differentiable Likelihoods for Fast Inversion of ‘Likelihood-Free’ Dynamical Systems

Hans Kersting*, Nicholas Krämer*, Martin Schiegg, Christian Daniel, Michael Tiemann, Philipp Hennig

ICML, 2020

EBERHARD KARLS UNIVERSITY TÜBINGEN

Max Planck Institute for Intelligent Systems

BOSCH
**TL;DR Summary**

**ODE Forward Problem:** Given $\theta$, estimate $x : [0, T] \rightarrow \mathbb{R}^d$ which satisfies the

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

**ODE Inverse Problems:**

Given data $z(t_{1:M}) = x_\theta(t_{1:M}) + \varepsilon \in \mathbb{R}^d$, $\varepsilon \sim \mathcal{N}(0, \Sigma)$, estimate $\theta$.

**Question:** Are ODE inverse problems really likelihood-free inference?

**Answer:** No! If we use probabilistic numerics to account for the numerical forward error, there is a differentiable likelihood!

**Practical Benefit:** New gradient-based methods are now available.
Inverse Problems... are defined by their forward map $F$.

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

Parameter $\theta \in \Theta$ \hspace{1cm} Simulation $F(\theta) \in \mathbb{R}^d$

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

- 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

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

Parameter $\theta \in \Theta$

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

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

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 $x_\theta : [0, T] \rightarrow \mathbb{R}^d$, $t \mapsto x_0 + \int_0^t f(x(s), \theta) \, ds$, and hence an high-fidelity forward map $F : \Theta \rightarrow C^1([0, T]; \mathbb{R}^d)$, $\theta \mapsto x_\theta$.

+ $x_\theta$ 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

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

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

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

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

\( \forall \theta \in \Theta, \text{ODEs have a well-defined solution} \)

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

and hence an high-fidelity forward map

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

\( x_\theta \) 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]

\[
\text{Parameter } \theta \in \Theta \\
\text{Simulation } F(\theta) \in \mathbb{R}^d \\
\]

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

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

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

\[\forall \theta \in \Theta, \text{ODEs have a well-defined solution } x_\theta : [0, T] \rightarrow \mathbb{R}^d, \quad t \mapsto x_0 + \int_0^t f(x(s), \theta) \, ds, \]

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

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

- \( x_\theta \) has to be estimated with **non-zero** step size \( h > 0 \), i.e. with **low fidelity**!

- With **numerical error**, e.g. Runge–Kutta:

\[
\begin{align*}
\dot{x}(t) &= \frac{x(t) - x(t_{0+c})}{h} \\
\end{align*}
\]
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**: $F(\theta_{\text{true}}) + \text{‘noise’} \mapsto \theta_{\text{true}}$

**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**

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

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

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

- $x_\theta$ 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

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

Parameter $\theta \in \Theta$

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

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

- 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

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

Parameter \( \theta \in \Theta \) \[ \rightarrow \] Simulation \( F(\theta) \in \mathbb{R}^d \)

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

- 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 \[ \rightarrow \] Probabilistic Numerics captures numerical error \[ \rightarrow \] Differentiable Likelihood

Gradient-free methods:
- Density estimation methods
- ABC

Gradient-based methods:
- Gradient descent
- Hamiltonian/Langevin MCMC
We propose the following likelihood.

**Uncertainty-Aware Likelihood by Gaussian ODE Filtering**

[Schober et al., 2019, Tronarp et al., 2019, Kersting et al., 2019]

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

\[
p(z \mid x) = \mathcal{N}(z; x, \sigma^2 I_M).
\]

(1)

For any \( \theta \), **Gaussian ODE Filtering**, a probabilistic numerical method, yields

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

(2)

where \( J \) is freely-available from the filtering output.

**Two advantages:**

- \( P \) 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 \( \theta \mapsto x_\theta \) at some support point \( \hat{\theta} \), and implies gradient and Hessian estimators

\[
\hat{\nabla}_\theta E(z) := -J^T \left[ P + \sigma^2 I_M \right]^{-1} [z - m_\theta], \quad \text{and} \quad \hat{\nabla}_\theta^2 E(z) := J^T \left[ P + \sigma^2 I_M \right]^{-1} J.
\]

(3)
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** $P$ makes the implicit forward model tractable.
The gradients are accurate enough to point towards modes!

Both the

- **gradient** estimator, and
- the Hessian-preconditioned (**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
- **Likelihood-free** random-walk Metropolis (RWM) **gets lost** in regions of low probability.
- **Gradient-based** sampling quickly finds and covers **regions of high probability**.
- **Likelihood-free** random-search **hardly learns** at all.
- **Gradient-based** optimization **quickly** finds local maxima.
Collaborators
University of Tübingen (top row) and Bosch Center for AI (bottom row)

Nicholas Krämer (joint primary author)

Philipp Hennig

Martin Schiegg

Christian Daniel

Michael Tiemann
▶ Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. Proceedings of the National Academy of Sciences.

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

▶ Kersting, H., Sullivan, T. J., and Hennig, P. (2019). Convergence rates of Gaussian ODE filters. arXiv:1807.09737v2 [math.NA].

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

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