ODE Solvers as Gauss-Markov Regression: An Overview

Filip Tronarp

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Initial value problem:

$$\dot{y}^\star(t) = f(y^\star(t)), \quad y^\star(0), = y_0, \quad t \in [0, T] \tag{1}$$

Problem

- Grid: $0 = t_0 < t_1 < ... < t_N = T$
- Evaluations: $f(\cdot)$

Find approximation: $\hat{y} \approx y^\star$

A probabilistic formulation

At a glance

Probabilistic formulation:

- Prior: $y \sim \mathcal{GP}$
- Initial data: $y(0) = y^\star(0)$
- Data: $\dot{y}(t) = f(y(t))$ for $t = t_0, t_1, \dots, t_N$
- Bayes' rule

Voilá!

State-space realisable priors

Convenient and canonical priors

Prior:

$$\mathrm{d}y^{(\nu)}(t) = \sum_{m=0}^{\nu} A_m y^{(m)}(t)\,\mathrm{d}t + \sqrt{\kappa}\sigma(t)\,\mathrm{d}w(t) \tag{2}$$

Usually ν-times integrated Wiener process: [1]

$$\mathrm{d}y^{(\nu)}(t) = \sqrt{\kappa}\,\mathrm{d}w(t) \tag{3}$$

Corresponds to Taylor polynomial + perturbation:

$$y(t) = \sum_{m=0}^{\nu} y^{(m)}(0)\frac{t^m}{m!} + \sqrt{\kappa}\int_0^t \frac{(t-\tau)^{\nu}}{\nu!}\,\mathrm{d}w(\tau)$$

[1] A probabilistic model for the numerical solution of initial value problems. M. Schober, S. Särkkä, P. Hennig. Statistics and Computing, 2019.

For instance:

$$x^*(t) = \begin{pmatrix} y^{(\nu)*} & y^{(\nu-1)*} & \dots & y^{(0)*} \end{pmatrix}$$

State-space realisation:

$$dx(t) = Ax(t)\,dt + B\sqrt{\kappa}\sigma(t)\,dw(t), \tag{4a}$$

$$y^{(m)}(t) = E_m x(t). \tag{4b}$$

- $x(t)$ is a Gauss–Markov process
- y and its derivatives are linear transforms of x.

State-space realisable priors

The Gauss–Markov property

x is Markov:

$$p(x(t_{0:N})) = p(x(t_0)) \prod_{n=1}^{N} p(x(t_n) \mid x(t_{n-1})) \quad \text{for} \quad t_0 < t_1 < \ldots < t_N. \tag{5}$$

In our case:

$$p(x(t) \mid x(u)) = \mathcal{N}\Big(x(t); \Phi(t,u)x(u), \kappa Q(t,u)\Big) \tag{6}$$

Parameters:

$$\Phi(t,u) = e^{A(t-u)}, \tag{7a}$$

$$Q(t,u) = \int_u^t \Phi(t,\tau) B \sigma(\tau) \sigma^*(\tau) B^* \Phi^*(t,\tau) \, \mathrm{d}\tau. \tag{7b}$$

ODE solvers as Non-linear Gauss–Markov regression

The inference problem

Non-linear Gauss–Markov regression problem: [2]

$$x(t_n) \mid x(t_{n-1}) \sim \mathcal{N}\Big(\Phi(t_n, t_{n-1})x(t_{n-1}), \kappa Q(t_n, t_{n-1})\Big), \tag{8a}$$

$$0 = z(x(t_n)) = E_1 x(t_n) - f(E_0 x(t_n)) = y^{(1)}(t_n) - f(y(t_n)), \quad n = 1, \dots, N. \tag{8b}$$

- Initial value $\mathbf{x_0}$ set to exact value via auto-diff. [3]
- κ can be used to calibrate the numerical uncertainty. [4] [5] [6]

---

[2] Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. F. Tronarp, H. Kersting, S Särkkä, P Hennig. Statistics and Computing, 2019.

[3] Stable implementation of probabilistic ODE solvers. N Krämer, P. Hennig. arXiv:2012.10106, 2020.

[4] Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. F. Tronarp, H. Kersting, S Särkkä, P Hennig. Statistics and Computing, 2019.

[5] A probabilistic model for the numerical solution of initial value problems. M. Schober, S. Särkkä, P. Hennig. Statistics and Computing, 2019.

[6] Calibrated adaptive probabilistic ODE solvers. N. Bosch, P. Hennig, F. Tronarp. AISTATS, 2021.

Affine vector field:

$$f(y) = L(t)y + b(t). \tag{9}$$

Affine measurements:

$$C(t) = E_1 - L(t)E_0, \tag{10a}$$

$$z(x(t_n)) = E_1 x(t_n) - f(E_0 x(t_n)) = C(t_n)x(t_n) - b(t_n). \tag{10b}$$

Solution: Kalman filtering and Rauch–Tung Striebel smoothing. [7]

[7] Bayesian filtering and smoothing. S. Särkkä. Cambridge University Press, 2013.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Posterior marginal for data up to time t_n: $p(x(t_n) \mid z(x(t_{0:n})) = 0) = \mathcal{N}\left(x(t_n); \mu(t_n), \kappa\Sigma(t_n)\right)$

The Kalman filter

Predict:

$$\mu(t_n^-) = \Phi(t_n, t_{n-1})\mu(t_{n-1}), \tag{11a}$$

$$\Sigma(t_n^-) = \Phi(t_n, t_{n-1})\Sigma(t_{n-1})\Phi^*(t_n, t_{n-1}) + Q(t_n, t_{n-1}). \tag{11b}$$

Update:

$$S(t_n) = C(t_n)\Sigma(t_n^-)C^*(t_n), \tag{12a}$$

$$K(t_n) = \Sigma(t_n^-)C^*(t_n)S^{-1}(t_n), \tag{12b}$$

$$\mu(t_n) = \mu(t_n^-) + K(t_n)\Big(b(t_n) - C(t_n)\mu(t_n^-)\Big), \tag{12c}$$

$$\Sigma(t_n) = \Sigma(t_n^-) - K(t_n)S(t_n)K^*(t_n). \tag{12d}$$

Posterior marginal for all data:

$$p(x(t_n) \mid z(x(t_{0:N})) = 0) = \mathcal{N}\left(x(t_n); \xi(t_n), \kappa\Lambda(t_n)\right)$$

Rauch–Tung–Striebel smoother

Backwards prediction:

$$\xi(t_{n-1}) = G(t_{n-1}, t_n)\left(\xi(t_n) - \mu(t_n^-)\right), \tag{13a}$$

$$\Lambda(t_{n-1}) = G(t_{n-1}, t_n)\Lambda(t_n)G^*(t_{n-1}, t_n) + V(t_{n-1}, t_n), \tag{13b}$$

where

$$G(t_{n-1}, t_n) = \Sigma(t_{n-1})\Phi^*(t_n, t_{n-1})\Sigma^{-1}(t_n^-), \tag{14a}$$

$$V(t_{n-1}, t_n) = \Sigma(t_{n-1}) - G(t_{n-1}, t_n)\Sigma(t_n^-)G^*(t_{n-1}, t_n). \tag{14b}$$

Succesive linearisation:

- Zeroth order method (explicit): [8]

$$f(\mathrm{E}_0 x(t)) \approx f(\mathrm{E}_0 \mu(t_n^-)).$$

- First order method (semi-implicit): [9]

$$f(\mathrm{E}_0 x(t_n)) \approx f(\mathrm{E}_0 \mu(t_n^-)) + J_f(\mathrm{E}_0 \mu(t_n^-))\mathrm{E}_0 \left(x(t_n) - \mu(t_n^-) \right)$$

[8] A probabilistic model for the numerical solution of initial value problems. M. Schober, S. Särkkä, P. Hennig. Statistics and Computing, 2019.

[9] Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. F. Tronarp, H. Kersting, S Särkkä, P Hennig. Statistics and Computing, 2019.

$$\hat{x}(t_{1:N}) = \arg\min_{x(t_{1:N})} \frac{1}{2} \sum_{n=1}^{N} \left\| x(t_n) - \Phi(t_n, t_{n-1}) x(t_{n-1}) \right\|^2_{Q^{-1}(t_n, t_{n-1})}, \tag{15}$$
$$\text{subject to} \quad z(x(t_n)) = 0, \quad n = 1, \dots, N.$$

Equivalent to minimum norm interpolation in RKHS: [10]

$$\hat{y} = \arg\min_{y} \int_0^{t_N} \left| \left(y^{(\nu+1)}(t) - \sum_{m=0}^{\nu} A_m y^{(m)}(t) \right) \right|^2 \sigma^{-2}(t) \, \mathrm{d}t,$$
$$\text{subject to} \quad z(x(t_n)) = 0, \quad n = 1, \dots, N.$$

[10] Bayesian ODE solvers: the maximum a posteriori estimate. F. Tronarp, S. Särkkä, P. Hennig. Statistics and Computing, 2021.

Linear test equation:

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

Definition: A-stability

A method $\hat{y}$ using a constant step-size is A-stable if $\hat{y}(t)$ is asymptotically whenever Λ has eigenvalues strictly in the left-half plane.

- Classical approach: analyse roots of discrete time process.
- Probabilistic approach: exploit systems theory results relating to stabilising control.

Stability

A-Stability

- Constant measurement matrix (semi-implicit):

$$C = \mathrm{E}_1 - \Lambda \mathrm{E}_0.$$

- Let $\sigma(t) = \texttt{const}$, implies model matrices Φ, Q, and C are all constant for constant step-size.

Generative form

$$x(t_n) = \Phi x(t_{n-1}) + Q^{1/2} w(t_n), \tag{16a}$$

$$0 = C x(t_n). \tag{16b}$$

Definition (Absolute stabilisability).

The pair $[\Phi, Q^{1/2}]$ is completely stabilisable if $w^* Q^{1/2} = 0$ and $w^* \Phi = \eta w^*$ for some constant η implies either $|\eta| < 1$ or $w = 0$.

Definition (Absolute detectability).

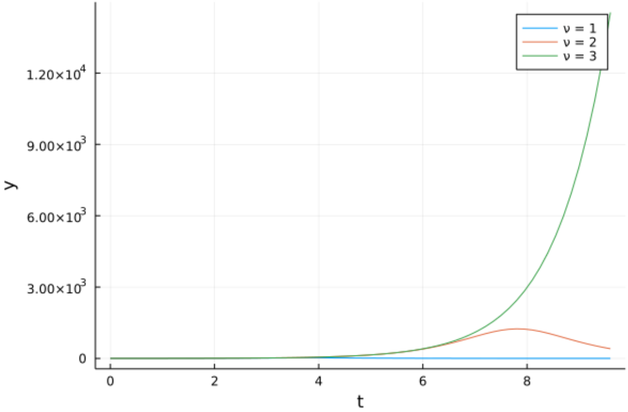
The pair $[\Phi, C]$ is completely detectable if $[\Phi^*, C^*]$ is completely stabilisable.

Theorem

The semi-implicit solver is exponentially (and therefore A-stable) if and only if the pair $[\Phi, Q^{1/2}]$ and $[\Phi, C]$ are complete stabilisable and detectable, respectively.

- Complete detectability of $[\Phi, C]$ is not a function of the real part of the eigenvalues of Λ!
- Let us use a ν-times integrated Wiener process to solve:

$$\dot{y}^\star(t) = y^\star(t), \quad y^\star(0) = 1.$$



Results for explicit methods:

- Matching methods associated with some priors to classical methods. [11] [12]
- Local and global rates for a limited set of priors using “classical” convergence analysis. [13]

Results for semi-implicit methods:

- Only empirical so far. [14] [15]

Results for MAP estimate:

- Quite nice result under mild assumptions using methods from scatterd data approximation. [16]

Contraction rates of the actual posterior has not been investigated at all?

---

[11] A probabilistic model for the numerical solution of initial value problems. M. Schober, S. Särkkä, P .Hennig. Statistics and Computing, 2019.

[12] Probabilistic ODE solvers with Runge-Kutta means. M. Schober, D. K. Duvenaud, P. Hennig. Neurips, 2014.

[13] Convergence rates of Gaussian ODE filters. H. Kersting, T. J. Sullivan, P. Hennig. Statistics and computing, 2020.

[14] Calibrated adaptive probabilistic ODE solvers. N. Bosch, P. Hennig, F. Tronarp. AISTATS, 2021.

[15] Stable implementation of probabilistic ODE solvers. N. Krämer, P. Hennig. arXiv:2012.10106, 2020.

[16] Bayesian ODE solvers: the maximum a posteriori estimate. F. Tronarp, S. Särkkä, P. Hennig. Statistics and Computing, 2021.

Suppose:

- The prior is of the form:

$$\mathrm{d}y^{(\nu)}(t) = \sum_{m=0}^{\nu} A_m y^{(m)}(t)\,\mathrm{d}t + \sqrt{\kappa}\,\mathrm{d}w(t). \tag{17}$$

- The vector field is smooth: $f \in C^{\nu+1}$.
- A unique solution $y^\star(t)$ exists up until $T^\star > t_N$.

Then:

- RKHS is equivalent to $H_2^{\nu+1}$.
- The solution $y^\star(t)$ is in RKHS.
- The operator $S_f[\varphi](t) = f(\varphi(t))$ is locally Lipschitz from $B(0, \|y^\star\|^2_{H_2^{\nu+1}} + \varepsilon) \subset H_2^{\nu+1}$ onto H_2^{ν}. [17]

---

[17] Boundary Value Problems of Finite Elasticity: Local Theorems on Existence, Uniqueness, and Analytic Dependence on Data. T. Valent. Springer, 2013.

Integral form of estimate:

$$\hat{y}(t) = y(0) + \int_0^t \dot{\hat{y}}(\tau)\,\mathrm{d}\tau = y(0) + \int_0^t f(\hat{y}(\tau))\,\mathrm{d}\tau + \int_0^t \dot{R}[\hat{y}; f](\tau)\,\mathrm{d}\tau$$

Derivative of residual:

$$\dot{R}[\hat{y}; f](\tau) = \dot{\hat{y}}(\tau) - f(\hat{y}(\tau)) \tag{18}$$

Sobolev functions with many zeros are small: [18]

$$\left|\dot{R}_i[\hat{y}; f]\right|_{H_q^m} \leq c_2 h^{\nu - m - (1/2 - 1/q)_+} \left|\dot{R}_i[\hat{y}; f]\right|_{H_2^\nu}, \quad m \leq \nu - 1 \tag{19}$$

Lipschitz property and $\hat{y}$ is smaller than $y^\star$:

$$\left|\dot{R}_i[\hat{y}; f]\right|_{H_2^\nu} \leq \left\|\dot{R}_i[\hat{y}; f] - \dot{R}_i[y^\star; f]\right\|_{H_2^\nu} \leq c_3^\star(f)\left\|\hat{y} - y^\star\right\|_{H_2^\nu} \leq 2c_3^\star(f)\|y^\star\|_{H_2^\nu} \tag{20}$$

[18] An extension of a bound for functions in Sobolev spaces, with applications to (m,s)-spline interpolation and smoothing. Arcangéli, R., de Silanes, M.C.L., Torrens, J.J. Numer. Math, 2007.

Conclusions

The MAP estimate converges to the solution quickly in the sense that:

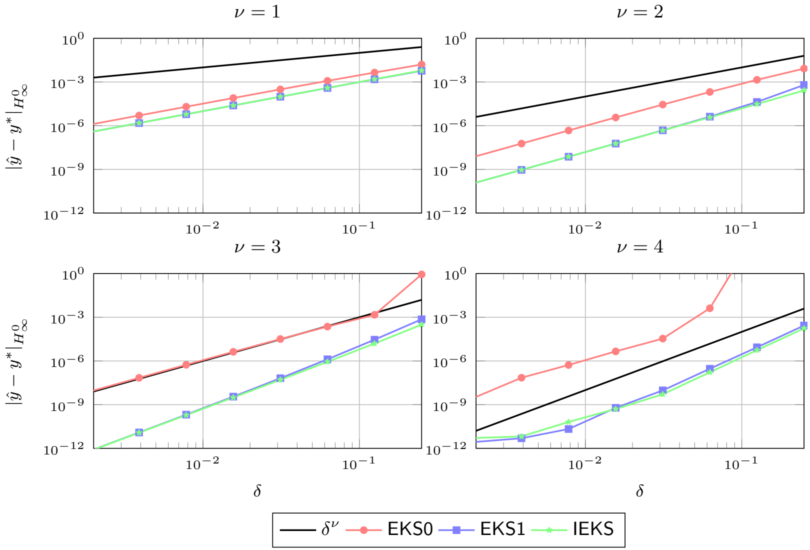
$$\left|\hat{y}(t) - y(0) - \int_0^t f(\hat{y}(\tau))\,\mathrm{d}\tau\right| \sim h^{\nu}. \tag{21}$$

Error estimates may be obtained with Gronwall's inequality:

$$\left|\hat{y}(t) - y^{\star}(t)\right| \sim h^{\nu}. \tag{22}$$

Convergence

line goes down



Some notes:

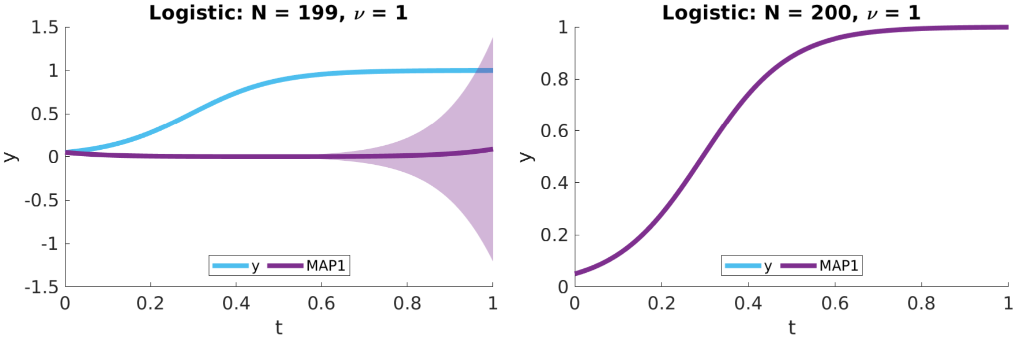
- The MAP estimate is an idealised object in general (non-convex problem).
- The rates only hold “eventually”.
- The minimum norm property has some funky effects:

$$\dot{y}^\star(t) = r y^\star(t)\left(1 - y^\star(t)\right), \quad y^\star(0) = \varepsilon. \tag{23}$$

There is a function close to zero, which interpolates well.

Convergence

The MAP estimate: Eventually can take a while and then happen suddenly



On Pecularities

- Estimates can asymptotically die, even if the solution exploded.
- Small functions are premiered, perhaps too much?

RKHS for large time horizons:

$$\|y\|^2_{\mathrm{RKHS}} = \int_0^\infty \left| \left(y^{(\nu+1)}(t) - \sum_{m=0}^{\nu} A_m y^{(m)}(t) \right) \right|^2 \sigma^{-2}(t)\,\mathrm{d}t \tag{24}$$

Use σ to make the RKHS norm of the solution small somehow?

Compositon of numeric and measurement uncertainty

Problem setting

Parametric ODE:

$$\dot{y}^\star_\theta(t) = f_\theta(y^\star_\theta(t)), \quad y^\star(0) = y_0(\theta).$$

Data:

$$u(t_n) = Hy(t_n) + v(t_n), \quad v(t_n) \sim \mathcal{N}(0, R_\theta). \tag{25}$$

Likelihood functional:

$$L_\mathrm{D}(\theta, \varphi) = \prod_n \mathcal{N}(u(t_n); H\varphi(t_n), R_\theta). \tag{26}$$

Marginal likelihood function:

$$M(\theta) = \int L_\mathrm{D}(\theta, \varphi)\delta(\varphi - y^\star_\theta)\,\mathrm{d}\varphi. \tag{27}$$

Output of probabilistic numerics:

$$\widehat{\delta}_N(\varphi;\theta,\kappa) \approx \delta(\varphi - y^\star_\theta) \tag{28}$$

Marginal likelihood approximation: [19]

$$\widehat{M}(\theta,\kappa) = \int L_{\mathrm{D}}(\theta,\varphi)\widehat{\delta}_N(\varphi;\theta,\kappa)\,\mathrm{d}\varphi. \tag{29}$$

[19] Differentiable likelihoods for fast inversion of'likelihood-free'dynamical systems. H. Kersting, N. Krämer, M. Schiegg, C. Daniel, M. Tiemann, P. Hennig. ICML, 2020.

Gauss–Markov representation of $\widehat{\delta}_N$: [20]

$$\widehat{\gamma}(x(t_{1:N}); \theta, \kappa) = \mathcal{N}(x(t_N); \xi_\theta(t_N), \kappa\Lambda(t_N)) \prod_{n=N-1}^{1} \mathcal{N}(x(t_n); G_\theta(t_n, t_{n+1})x(t_{n+1}) + \zeta_\theta(t_n), \kappa V_\theta(t_n)) \quad (30a)$$

$$\zeta_\theta(t_n) = \mu(t_n) - G_\theta(t_n, t_{n+1})\mu(t_{n+1}^-). \quad (30b)$$

Gauss–Markov regression but backwards:

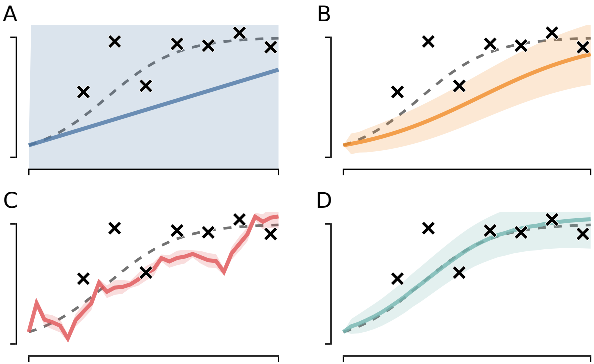
$$x(t_n) \mid x(t_{n+1}) \sim \mathcal{N}(x(t_n); G_\theta(t_n, t_{n+1})x(t_{n+1}) + \zeta_\theta(t_n), \kappa V_\theta(t_n)), \quad (31a)$$

$$u(t_n) \mid x(t_n) \sim \mathcal{N}(Hx(t_n), R_\theta) \quad (31b)$$

[20] Fenrir: Physics-Enhanced Regression for Initial Value Problems F. Tronarp, N. Bosch, P. Hennig. ICML, 2022.

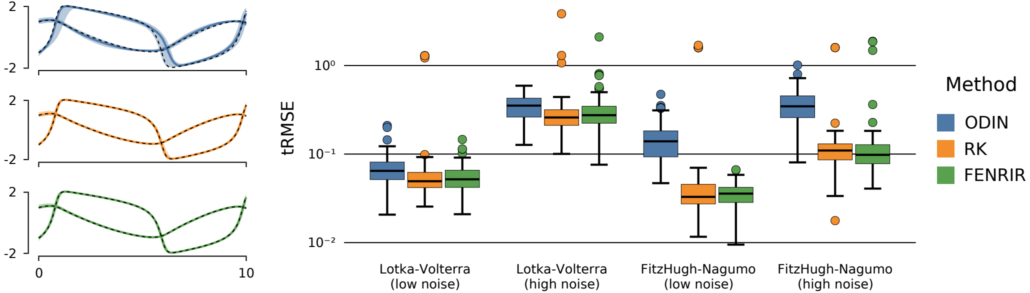
Composition of numeric and measurement uncertainty

Pictoral numerics



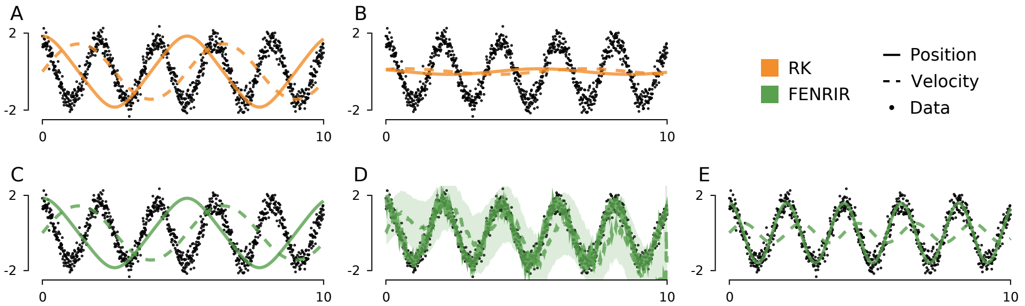
Composition of numeric and measurement uncertainty

Some benchmarking



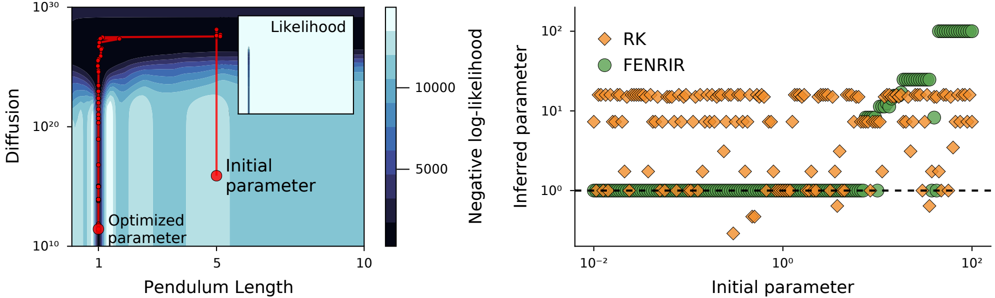
Composition of numeric and measurement uncertainty

The benefits of likelihood smoothing and uncertainty quantification?



Composition of numeric and measurement uncertainty

The benefits of likelihood smoothing and uncertainty quantification?



Some concluding thoughts and marketing

Bells and whitles:

- Numerically stable implementation of probabilistic solvers. [21]
- Solvers for boundary value problems. [22]
- Augmenting measurement model to handle known constraints (e.g. energy conservation). [23]

Software if you care to try:

- Python (ProbNum): `https://probnum.readthedocs.io/en/latest/` [24]
- Julia (ProbNumDiffEq.jl): `https://github.com/nathanaelbosch/ProbNumDiffEq.jl`

Probabilistic ODE solvers are becoming mature in terms of theory, algorithms, and software - BUT!

---

[21] Stable implementation of probabilistic ODE solvers. N Krämer, P. Hennig. arXiv:2012.10106, 2020.

[22] Linear-Time Probabilistic Solutions of Boundary Value Problems. N. Krämer, P. Hennig. Neurips, 2021.

[23] Pick-and-mix information operators for probabilistic ODE solvers. N. Bosch, F. Tronarp, P. Hennig. AISTATS, 2022.

[24] J. Wenger, N. Krämer, M. Pförtner, J. Schmidt, N. Bosch, N. Effenberger, J. Zenn, A. Gessner, T. Karvonen, F.-X. Briol, M. Mahsereci, P. Hennig. ProbNum: Probabilistic Numerics in Python. arXiv:2112.02100, 2021.