Probabilistic ODE Solvers with Runge-Kutta Means

Michael Schober*, David Duvenaud‡, Philipp Hennig*

*Research Group Elementary Intelligence
Department of Empirical Inference
Max Planck Institute for Intelligent Systems
Tübingen, Germany

‡Computational and Biological Learning Lab
Department of Engineering
Cambridge University

Can we assign a probability distribution
over the solution to
an ordinary differential equation
(initial value problem)?

$$x(t_0) = x_0 \qquad\qquad x'(t) = f(x(t), t)$$

The Probabilistic View on Computation

computing as the collection of information [Poincaré, 1896, Diaconis, 1988, O'Hagan, 1992]

A numerical method
estimates a function's latent property
given the result of computations.

quadrature	estimates $\int_a^b f(x)\,dx$	given $\{f(x_i)\}$
linear algebra	estimates x s.t. $Ax = b$	given $\{As = y\}$
optimization	estimates x s.t. $\nabla f(x) = 0$	given $\{\nabla f(x_i)\}$
analysis	estimates $x(t)$ s.t. $x' = f(x,t)$,	given $\{f(x_i, t_i)\}$

- computations yield "data" / "observations"
- non-analytic quantities are "latent"
- even deterministic quantities can be uncertain.

Numerical Methods and Statistical Estimators

several classic numerical algorithms identified precisely as maximum a-posteriori estimators

quadrature [Diaconis, 1988, O'Hagan, 1991]

Gaussian quadrature ⟷ Gaussian process regression

linear algebra [Hennig, 2015]

conjugate gradients ⟷ Gaussian conditioning

nonlinear optimization [Hennig & Kiefel, 2013]

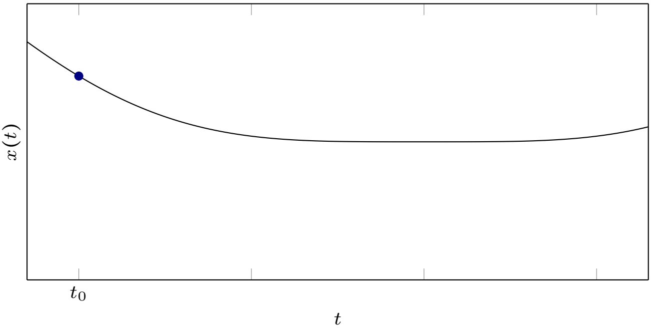
BFGS ⟷ autoregressive filtering

ordinary differential equations [Schober et al., 2014]

Runge-Kutta ⟷ Gauss-Markov extrapolation

Runge-Kutta methods

are linear extrapolators of high convergence order [Hairer et al., 1987]

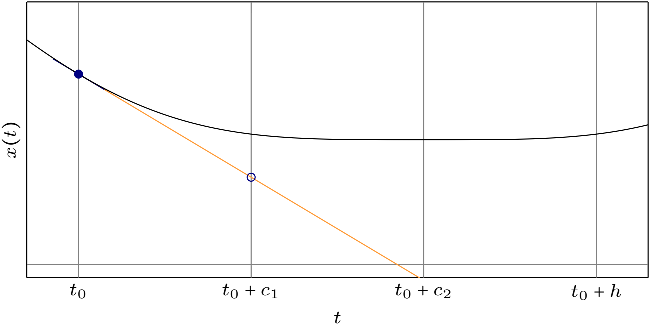


$$
\begin{array}{c|cccc}
0 & 1 & & & \\
c_1 & 1 & w_{11} & & \\
c_2 & 1 & w_{21} & w_{22} & \\
\hline
h & 1 & b_1 & b_2 & b_3
\end{array}
$$

$$
\begin{aligned}
Y_1 &= f(1x_0, t_0 + 0) \\
Y_2 &= f\left(1x_0 + w_{11}Y_1, t_0 + c_1\right) \\
Y_{s+1} &= f\left(1x_0 + \textstyle\sum_i^s w_{si}Y_i, t_0 + c_s\right) \\
\hat{x}(t_0 + h) &= 1x_0 + \textstyle\sum_i b_i Y_i
\end{aligned}
$$

Runge-Kutta methods

are linear extrapolators of high convergence order [Hairer et al., 1987]

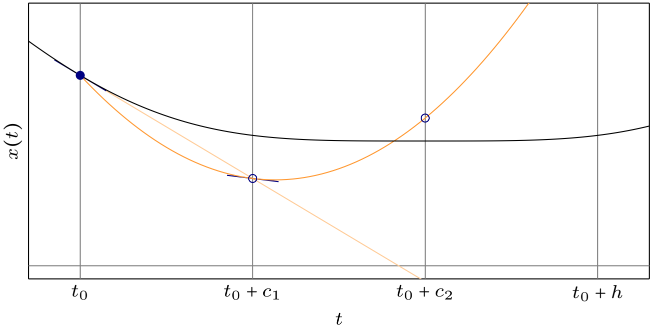


$$
\begin{array}{c|cccc}
0 & 1 & & & \\
c_1 & 1 & w_{11} & & \\
c_2 & 1 & w_{21} & w_{22} & \\
\hline
h & 1 & b_1 & b_2 & b_3
\end{array}
$$

$$
\begin{aligned}
Y_1 &= f(1 x_0, t_0 + 0) \\
Y_2 &= f\left(1 x_0 + w_{11} Y_1, t_0 + c_1\right) \\
Y_{s+1} &= f\left(1 x_0 + \textstyle\sum_i^s w_{si} Y_i, t_0 + c_s\right) \\
\hat{x}(t_0 + h) &= 1 x_0 + \textstyle\sum_i b_i Y_i
\end{aligned}
$$

Runge-Kutta methods

are linear extrapolators of high convergence order [Hairer et al., 1987]



0	1			
c_1	1	w_{11}		
c_2	1	w_{21}	w_{22}	
h	1	b_1	b_2	b_3

$$Y_1 = f(1x_0, t_0 + 0)$$

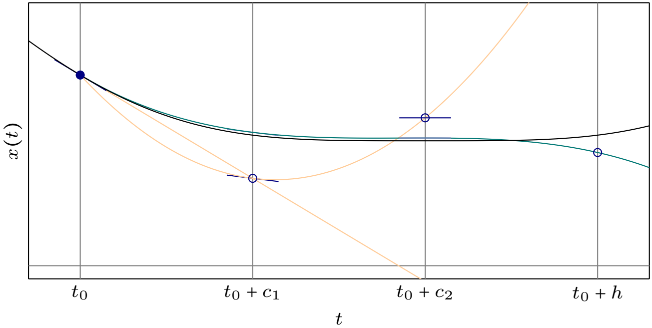
$$Y_2 = f\left(1x_0 + w_{11} Y_1, t_0 + c_1\right)$$

$$Y_{s+1} = f\left(1x_0 + \sum_i^s w_{si} Y_i, t_0 + c_s\right)$$

$$\hat{x}(t_0 + h) = 1x_0 + \sum_i b_i Y_i$$

Runge-Kutta methods

are linear extrapolators of high convergence order [Hairer et al., 1987]



0	1			
c_1	1	w_{11}		
c_2	1	w_{21}	w_{22}	
h	1	b_1	b_2	b_3

$$Y_1 = f(1x_0, t_0 + 0)$$

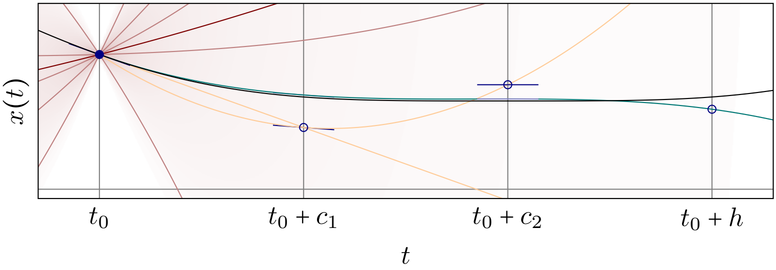
$$Y_2 = f\left(1x_0 + w_{11}Y_1, t_0 + c_1\right)$$

$$Y_{s+1} = f\left(1x_0 + \sum_i^s w_{si}Y_i, t_0 + c_s\right)$$

$$\hat{x}(t_0 + h) = 1x_0 + \sum_i b_i Y_i$$

Gaussian process solvers

are also linear extrapolators



- Linear extrapolation suggests Gaussian process model
- Gaussian process solvers previously studied
  [Skilling (1991), Chrekbtii et al. (2014), Hennig & Hauberg (2014)]

Some properties of Gaussian measures

The only two equations you really need (in this group)

- closure under affine transformations ($\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{y} \in \mathbb{R}^M$)

$$p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{P}), \quad p(\boldsymbol{y}|\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{x} + \boldsymbol{\nu}, \boldsymbol{R})$$

$$\Rightarrow p\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}\right) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{H}\boldsymbol{m} + \boldsymbol{\nu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{P} & \boldsymbol{P}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{P} & \boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^\top + \boldsymbol{R} \end{bmatrix}\right)$$

- inference involves only linear algebra operations

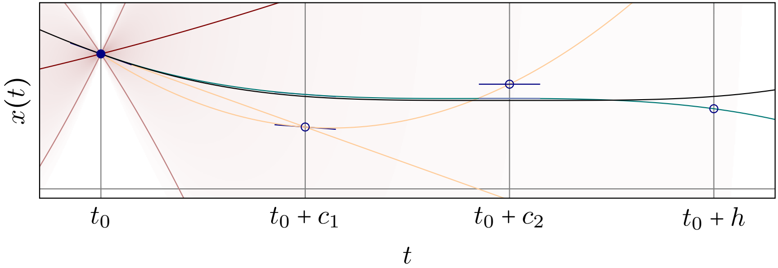
$$p\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}\right) \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{P}_2 \end{bmatrix}\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{m}_1 + \boldsymbol{C}\boldsymbol{P}_2^{-1}(\boldsymbol{y} - \boldsymbol{m}_2), \boldsymbol{P}_1 - \boldsymbol{C}\boldsymbol{P}_2^{-1}\boldsymbol{C}^\top)$$

$\Rightarrow$ sequential Gaussian inference at linear cost (‘filtering’)

Gaussian process solvers

implicitly define a Butcher tableau



0	1			
c_1	1	w_{11}		
c_2	1	w_{21}	w_{22}	
h	1	b_1	b_2	b_3

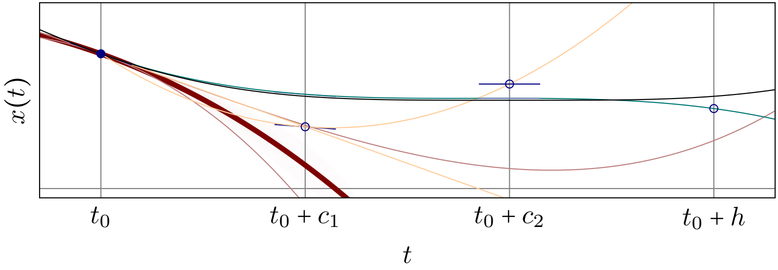
$$y_1 = f\left(\mu_{\mid x_0}(t_0 + 0), t_0 + 0\right)$$

$$\mu_{\mid x_0}(t_0) := \left[k(t_0, t_0)\right]\left[k(t_0, t_0)\right]^{-1}(x_0)$$

$$= 1 x_0$$

Gaussian process solvers

implicitly define a Butcher tableau



0	1			
c_1	1	w_{11}		
c_2	1	w_{21}	w_{22}	
h	1	b_1	b_2	b_3

$$y_1 = f\left(\mu_{|x_0}(t_0 + 0), t_0 + 0\right)$$

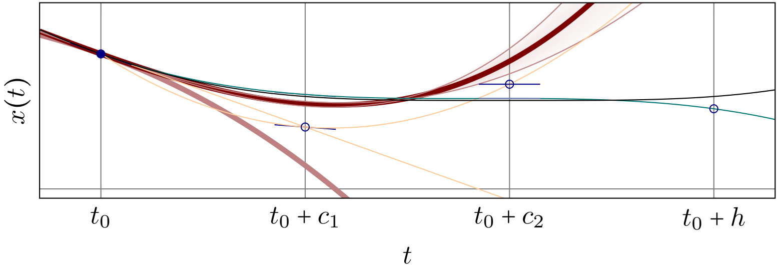
$$y_2 = f\left(\mu_{|x_0,y_1}(t_0 + c_1), t_0 + c_1\right)$$

$$\mu_{|x_0,y_1}(t_0 + c_1) := \left[k(t_0 + c_1, t_0) \quad k^{\partial}(t_0 + c_1, t_0)\right] \begin{bmatrix} k(t_0, t_0) & k^{\partial}(t_0, t_0) \\ {}^{\partial}k(t_0, t_0) & {}^{\partial}k^{\partial}(t_0, t_0) \end{bmatrix}^{-1} \begin{pmatrix} x_0 \\ y_1 \end{pmatrix}$$

$$= w_{10} x_0 + w_{11} y_1$$

Gaussian process solvers

implicitly define a Butcher tableau



0	1			
c_1	1	w_{11}		
c_2	1	w_{21}	w_{22}	
h	1	b_1	b_2	b_3

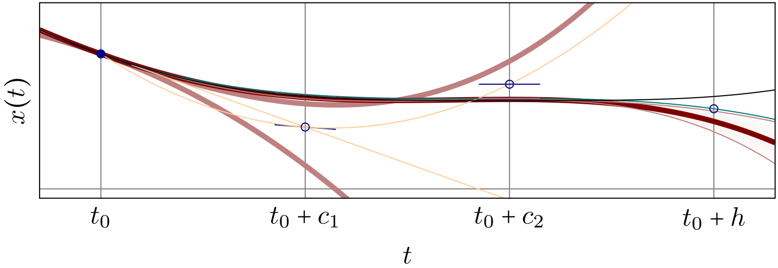
$$
\begin{aligned}
y_1 &= f\left(\mu_{|x_0}(t_0 + 0), t_0 + 0\right) \\
y_2 &= f\left(\mu_{|x_0,y_1}(t_0 + c_1), t_0 + c_1\right) \\
y_{s+1} &= f\left(\mu_{|x_0,y_i}(t_0 + c_s), t_0 + c_s\right)
\end{aligned}
$$

$$
\mu_{|x_0,y_i}(t_0 + c_s) := \left[k(t_0 + c_s, t_0) \quad k^{\partial}(t_0 + c_s, t_0 + c_i)\right] K^{-1} \begin{pmatrix} x_0 \\ y_i \end{pmatrix}
$$

$$
= w_{20} x_0 + \textstyle\sum_{i=1}^{s} w_{2i} y_i
$$

Gaussian process solvers

implicitly define a Butcher tableau



0	1			
c_1	1	w_{11}		
c_2	1	w_{21}	w_{22}	
h	1	b_1	b_2	b_3

$$
\begin{aligned}
y_1 &= f\left(\mu_{|x_0}(t_0 + 0), t_0 + 0\right) \\
y_2 &= f\left(\mu_{|x_0,y_1}(t_0 + c_1), t_0 + c_1\right) \\
y_{s+1} &= f\left(\mu_{|x_0,y_i}(t_0 + c_s), t_0 + c_s\right) \\
\hat{x}(t_0 + h) &= \mu_{|x_0,y_i}(t_0 + h)
\end{aligned}
$$

$$
\mu_{|x_0,y_i}(t_0 + h) := \left[k(t_0 + h, t_0) \quad k^{\partial}(t_0 + h, t_0 + c_i)\right] K^{-1} \begin{pmatrix} x_0 \\ y_i \end{pmatrix}
$$

$$
= b_0 x_0 + \textstyle\sum_{i=1}^{s} b_i y_i
$$

Gauss-Markov-Runge-Kutta methods

a GP solver whose mean matches RK exactly

- RK choose (c, w, b) such that $\|\hat{x}(t_0 + h) - x(t_0 + h)\| = \mathcal{O}(h^p)$
- polynomial form suggests integrated Wiener (polynomial spline) process

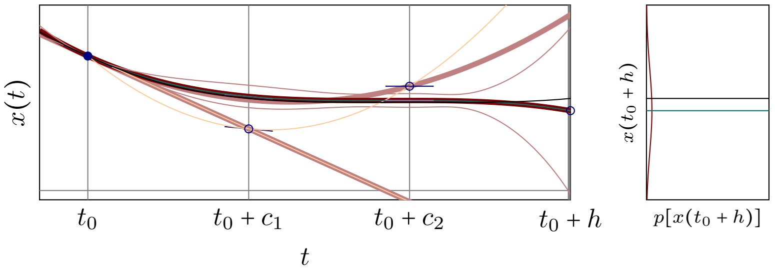
$$p(x(t)) = \mathcal{GP}(x(t); 0, k_s(t, t')) \qquad \text{where}$$

$$k_s(t, t') = \int \cdots \int_\tau^t \int \cdots \int_\tau^{t'} \min(\tilde{t}, \tilde{t}')\, d\tilde{t}\, d\tilde{t}'$$

- $\tau \to -\infty$: improper prior $p(x(t))$, proper posterior after s observations.
- kth-times integrated Wiener process gives k-order RK solver!
- Inherets RK guarantees. Gives closed-form solution for tableau (used to use numerical search!)
- a Markov (state-space) model, so inference is $\mathcal{O}(s)$ (as opposed to usual $\mathcal{O}(s^3)$ cost

Calibrating Uncertainty

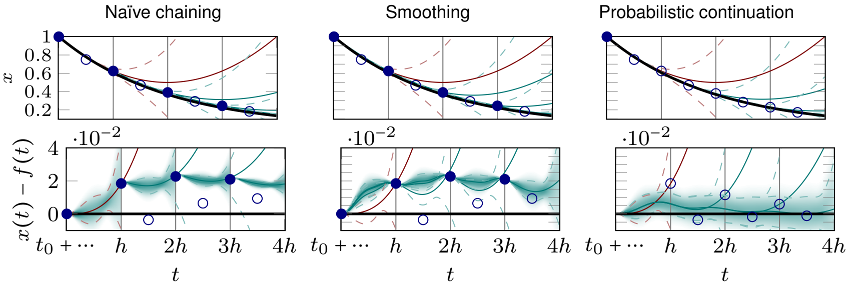
within the parametrized class



- posterior mean $\mu_{|y} = kK^{-1}y$ invariant under $k \to \theta^2 k$
- posterior covariance $k_{|y} = k - kK^{-1}k$ scaled by θ^2
- initial ideas for uncertainty calibration in paper (more to come)

Multi-Step Extension

[A. Nordsieck, 1962]



- probabilistic interpretation questions RK beyond s steps
- 'obvious' solution is to continue filtering process
- result very similar, though not identical, to multi-step methods

Some Conceptual Open Questions

precise interpretation of posterior measure still evolving

How precise can the connection to multi-step methods be?

- order / stability conditions currently not fully understood
- flexibility is also a design criterion
- what about stiff problems?

What, precisely, does the posterior mean?

- width of Gaussian posterior should be inferred from regularity of ‘observed’ gradients. How, precisely, should this be done? (We have one particular solution)
- is the Gaussian family enough? How expensive is it to move beyond Gauss?

What we've done so far:

- Numerical methods can be interpreted as performing statistical inference from noise-free data
- in some cases, e.g. Runge-Kutta, this link can be made precise
- Inherets convergence guarantees, but also get extensibility & uncertainty estimates

What we're working on next:

- understand the connection to multi-step methods
- construct a robust probabilistic IVP solver
- Continue finding model-based interpretations of numerical solvers.

Bibliography

P. Diaconis. Bayesian numerical analysis. *Statistical decision theory and related topics*, IV(1):163–175, 1988.

E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I – Nonstiff Problems*. Springer, 1987.

S. Hauberg, M. Schober, M. Liptrot, P. Hennig, and A. Feragen. A random riemannian metric for probabilistic shortest-path tractography. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2015*. Springer, 2015.

P. Hennig. Probabilistic interpretation of linear solvers. *SIAM J on Optimization*, 25(1):210–233, 2015.

P. Hennig and M. Kiefel. Quasi-Newton Methods – a new direction. *Journal of Machine Learning Research*, 14:834–865, March 2013.

A. O'Hagan. Bayes–Hermite quadrature. *J of Statistical Planning and Inference*, 29(3):245–260, 1991.

A. O'Hagan. Some Bayesian Numerical Analysis. *Bayesian Statistics*, 4:345–363, 1992.

H. Poincaré. *Calcul des probabilités*. Gauthier-Villars, Paris, 1896.

S. Särkkä. *Recursive Bayesian Inference on Stochastic Differential Equations*. PhD thesis, Helsinki University of Technology, 2006.

M. Schober, D. Duvenaud, and P. Hennig. Probabilistic ODE Solvers with Runge-Kutta Means. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

M. Schober, N. Kasenburg, A. Feragen, P. Hennig, and S. Hauberg. Probabilistic shortest path tractography in DTI using Gaussian Process ODE solvers. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014.