Probabilistic Numerics

III — Solving Ordinary Differential Equations

Philipp Hennig

Dobbiaco Summer School
20 June 2017

Research Group for Probabilistic Numerics
Max Planck Institute for Intelligent Systems
Tübingen, Germany

MAX-PLANCK-GESELLSCHAFT

Emmy Noether-Programm
DFG Deutsche Forschungsgemeinschaft

Some of the presented work was supported by
the Emmy Noether Programme of the DFG

Quick Recap:

- **Gauss-Markov** models (and their continous-time limit, **stochastic differential equations**) allow linear-time inference in Gaussian models, by **filtering** (and smoothing)
- **Gaussian quadrature** rules can be re-derived as **MAP** estimates in under Gauss-Markov priors, yielding **probabilistic functionality** in integration

Now:

- the same idea can be extended to **ordinary differential equations**

$$x : [t_0, T] \to \mathbb{R}^D \qquad \text{such that} \qquad \underbrace{x(t_0) = x_0 \in \mathbb{R}^D}_{\text{initial value}} \qquad \underbrace{x'(t) = f(x(t), t)}_{\text{ordinary differential equation}}$$

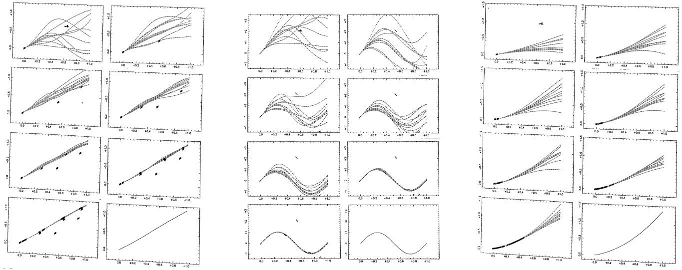
  if evaluations $f(\hat{x}, t)$ are treated as "noisy" observations of $x'(t)$, using *some* approximation $\hat{x}$.
- we construct a class of **fast** and **flexible** probabilistic ODE solvers
- analysis will be slightly more complicated

A cool idea

which needs a bit of work

[J. Skilling, *Bayesian solution to ODEs*, 1991]

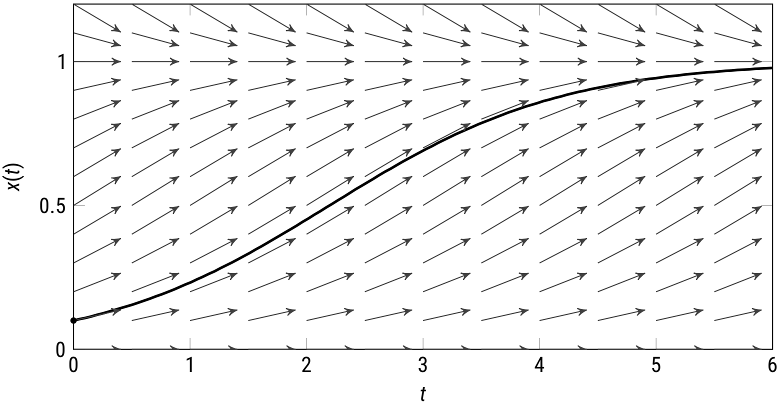
- use **Gaussian prior** $p(x) = \mathcal{GP}(\mu, k)$
- note $f(\tilde{x}, t)$ for $\tilde{x} \sim p(x)$ is uncertain **observation** of x'
- use them to iteratively **infer solution** $x(t)$



> *"Far from considering these suggestions to be the last word, the author considers them to be merely pointers to what could, and perhaps should, develop into an open field of research, complementing and if appropriate superseding the traditional algorithms."*
> John Skilling, 1991

A Sketch of What We're Looking For

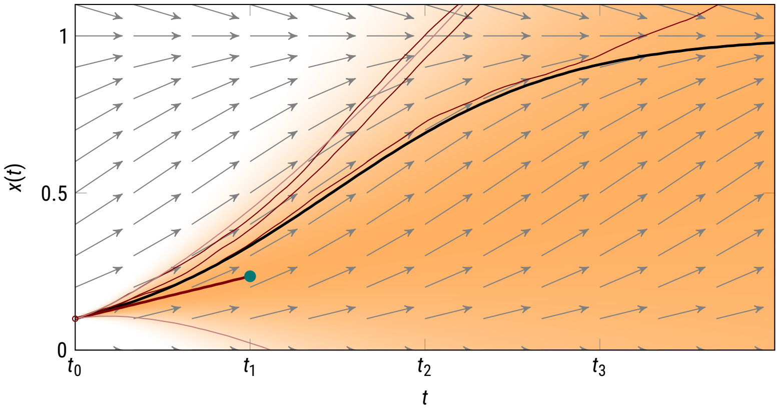
simple example: logistic ODE



$$x'(t) = f(x(t)) = x(t)(1 - x(t)) \qquad x_0 = 0.1$$

A Sketch of What We're Looking For

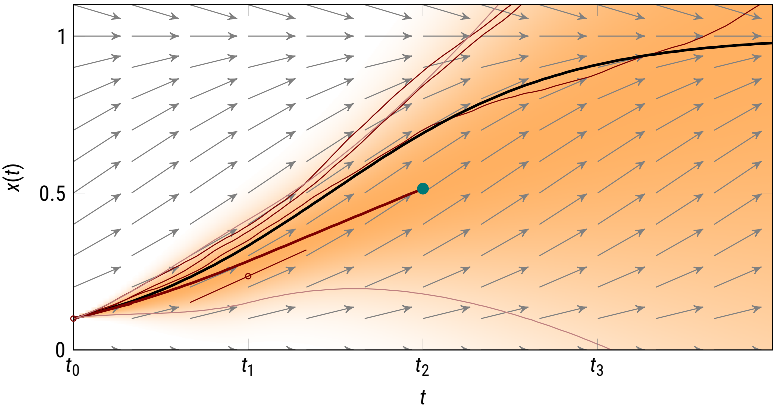
simple example: logistic ODE



$$y_0 = f(x_0, t_0) \qquad \hat{x}_1 = \mathbb{E}_{p(x|y_0,x_0)}(x(t_1))$$

A Sketch of What We're Looking For

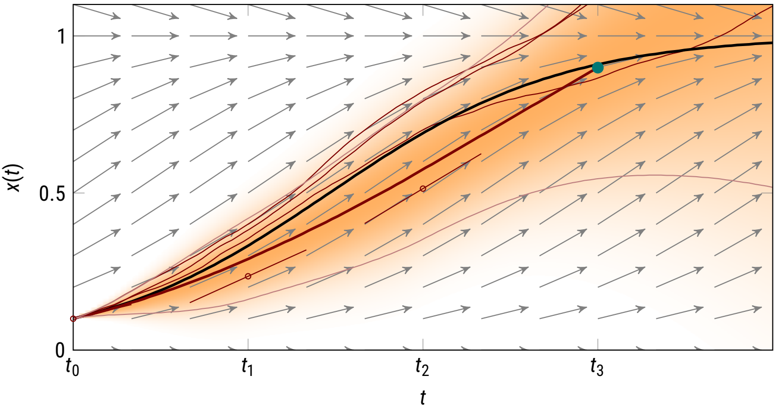
simple example: logistic ODE



$$y_i = f(x_i, t_i) \qquad \hat{x}_{i+1} = \mathbb{E}_{p(x \mid \mathbf{y}_{0:i-1}, x_0)}(x(t_i))$$

A Sketch of What We're Looking For

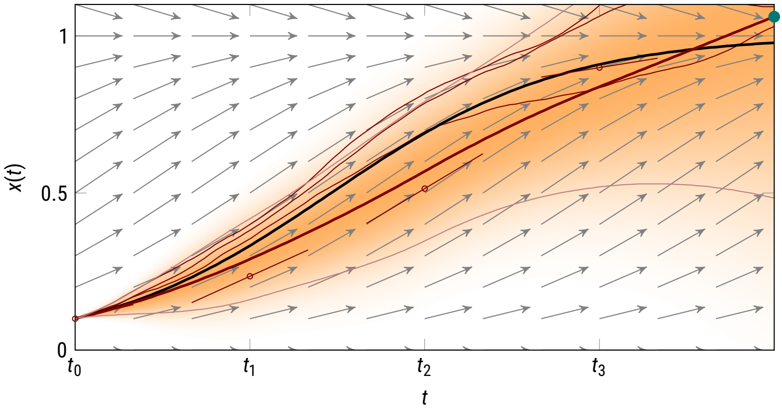
simple example: logistic ODE



$$y_i = f(x_i, t_i) \qquad \hat{x}_{i+1} = \mathbb{E}_{p(x|\mathbf{y}_{0:i-1}, x_0)}(x(t_i))$$

A Sketch of What We're Looking For

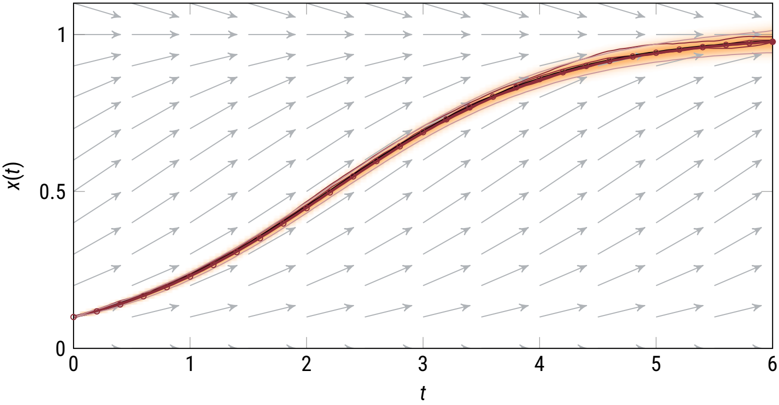
simple example: logistic ODE



$$y_i = f(x_i, t_i) \qquad \hat{x}_{i+1} = \mathbb{E}_{p(x \mid \mathbf{y}_{0:i-1}, x_0)}(x(t_i))$$

A Sketch of What We're Looking For

simple example: logistic ODE

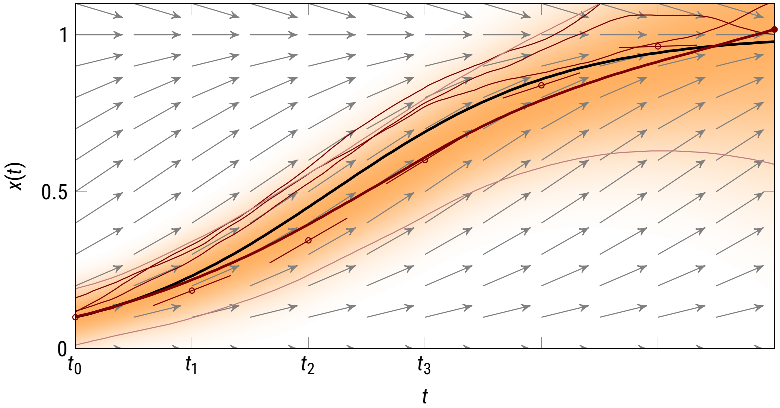


Desiderata:

- low cost: $\mathcal{O}(N)$
- consistency: $\|x(T) - \hat{x}(T)\| \leq KN^{-q}$
- calibration: $\|x(T) - \hat{x}(T)\|^2 \leq \mathrm{var}_{p(x|\mathbf{y},x_0)}(x(T))$
- PN features: uncertainty in inputs and outputs

A Sketch of What We're Looking For

simple example: logistic ODE



Desiderata:

low cost: $\mathcal{O}(N)$

consistency: $\|x(T) - \hat{x}(T)\| \leq KN^{-q}$

calibration: $\|x(T) - \hat{x}(T)\|^2 \leq \mathrm{var}_{p(x|\mathbf{y},x_0)}(x(T))$

PN features: uncertainty in inputs and outputs

A Proposal for a Probabilistic IVP Solver

Schober, Duvenaud, Hennig, NIPS 2014; Kersting & Hennig, UAI 2016; Schober, Särkkä, Hennig, 2017

$$p(x(t_0)) = \mathcal{N}(x(t_0), m_0, \rho_0) \qquad p(x'(t) \mid m_t, P_t) = \mathcal{N}(x'(t); @f(m_t, P_t, t), @R(m_t, P_t, t))$$

Model the solution of the **nonlinear** IVP as a **linear** ODE plus a Gaussian **stochastic** disturbance:

$$dz(t) = F(t)z(t)\,dt + L(t)d\omega \qquad \text{with} \qquad z(t) = \left[x(t) \quad x'(t) \quad z_3(t) \quad \dots z_{q+1}(t)\right]^\intercal \in \mathbb{R}^{q+1}$$

procedure IVP_PROB($@f, @R, t_0, T, N, m_0, \rho_0$)
  $[R, y_0] = \ell(f, m_0^-, P_0^-)$ // observe initial gradient
    $y = [x_0; y_0]$, $h = [I_2, \mathbf{0}_{2,q-1}]$, $H = [0, 1, \mathbf{0}_{1,q-2}]$ // joint first observation
    $R_0 = \operatorname{diag}(\rho_0, R)$ // initial observation uncertainty
    $K = P_0^- h^\intercal (hP_0^- h^\intercal + R_0)^{-1}$
    $m_0 = m_0^- + K(y - hm_0^-)$ // initial estimation mean
    $P_0 = (I - Kh)P_0^-$ // initial estimation covariance
  for i=1,2,… **do**
    $t_i = t_{i-1} + \min\{h, T - t_i\}$ // step
    $m_i^- = Am_{i-1}$ // predictive mean
    $P_i^- = AP_{i-1}A^\intercal + Q$ // predictive covariance
    $[R, y] = \ell(f, m_i^-, P_i^-)$ // **construct** observation
    $r = y - Hm_i^-$ // residual
    $K = P_i^- H^\intercal (HP_i^- H + R)^{-1}$ // gain
    $m_i = m_i^- + Kr$ // update mean
    $P_i = (I - KH)P_i^-$ // update covariance
    if $t_i = T$ **then return**
    end if
  end for
end procedure

Why this algorithm?

some initial observations

- for the moment, we will ignore how to adapt h and q
- we focus on estimating $x(T)$. If **dense output** at $x(0 \leq t \leq T)$ is required, add smoother run. This doubles the cost (also an issue in classic solvers).

This is not the only acceptable solver, but it has interesting properties:

- It is **fast** (computational cost $\mathcal{O}(N)$)
- It has natural “plug-ins” for **probabilistic functionality**.
  E.g.: external uncertainty on f and x_0.
- It **can** can have good analytic properties …

The First-Order Filter for ODEs

a first concrete probabilistic ODE solver

Schober, Duvenaud, Hennig, 2014

- Choose

$$F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 \\ \theta \end{bmatrix}, \quad \ell(f, m_i^-, P_i^-) = \begin{cases} y_i & = f(Hm_i^-, t_i) \\ R_i & = 0 \end{cases}.$$

- this gives

$$A = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad Q = \theta^2 \begin{bmatrix} h^3/3 & h^2/2 \\ h^2/2 & h \end{bmatrix}.$$

- given IV $x(t_0) = x_0, x'(t_0) = f(x_0, t_0)$, have $m_0 = [x_0, f(x_0, t_0)]^\intercal$ and $P_0 = \mathbf{0}_2$.
- **first** prediction:

$$m_1^- = \begin{bmatrix} x_0 + hf(x_0, t_0) \\ f(x_0, t_0) \end{bmatrix} \quad \text{and} \quad P_1^- = \mathbf{0} + Q = \theta^2 \begin{bmatrix} h^3/3 & h^2/2 \\ h^2/2 & h \end{bmatrix}.$$

- This is the prediction of **Euler's Method:**

$$\hat{x}(t_0 + h) = m_{1,1}^- = x_0 + hf(x_0, t_0)$$

The Trapezoid Rule — Again!

but the use in IVPs adds some complications

Schober, Duvenaud, Hennig, 2014

- first updated estimate

$$m_1 = \begin{bmatrix} x_0 + {}^1\!/_2 h(y_0 + y_1) \\ y_1 \end{bmatrix} \qquad \text{and} \qquad P_1 = \theta^2 \begin{bmatrix} {}^{h^3}\!/_{12} & 0 \\ 0 & 0 \end{bmatrix}.$$

- subsequent predictions, estimates:

$$m_i^- = \begin{bmatrix} m_{i-1,1} + hy_{i-1} \\ y_{i-1} \end{bmatrix} \qquad m_i = \begin{bmatrix} x_0 + \frac{h}{2} \sum_{j \leq i} (y_{j-1} + y_j) \\ y_i \end{bmatrix}$$

$$P_i^- = \theta^2 \begin{bmatrix} P_{(i-1),(11)} + {}^{h^3}\!/_3 & {}^{h^2}\!/_2 \\ {}^{h^2}\!/_2 & h \end{bmatrix} \qquad P_i = \theta^2 \begin{bmatrix} ({}^i\!/_{12})\, h^3 & 0 \\ 0 & 0 \end{bmatrix}$$

Classic Analysis

for the explicit trapezoid rule

Hairer, Nørsett & Wanner, 1993

- The initial mean $\boldsymbol{m_0}$ is correct. The first prediction is

$$m_{1,1} = x_0 + \frac{h}{2}(y_0 + y_1)$$
$$= x_0 + \frac{h}{2}(f(x_0, t_0) + f(x_0 + hf(x_0, t_0), t_0 + h))$$

- Consider Taylor expansion of this estimate around $\boldsymbol{t_0}$:

$$m_{1,1} = x_0 + hf(x_0, t_0) + \frac{h^2}{2}(f \cdot f_x + f_t)(x_0, t_0) + ...$$

- **If f has continuous partial derivatives**, the **true** solution has expansion

$$x(h) = x_0 + hx'(t_0) + \frac{h^2}{2}x''(t_0) + \cdots = x_0 + hf(x_0, t_0) + \frac{h^2}{2}(f_t + f \cdot f_x)(x_0, t_0) + ...$$

- …so the **leading local error** is of **third order**

$$x(h) - m_1 = \mathcal{O}(h^3) + ...$$

- **Calibration:** the estimation variance is

$$\mathrm{std}_p(x(t_1)) = \sqrt{P_{1,(11)}} = \frac{h^{3/2}}{\sqrt{12}}\theta \geq \mathcal{O}(h^3) \text{ for sufficiently small } h$$

- beyond first step, like above, replace x_0 with m_{i-1}, m^-_{i-1}:

$$\begin{aligned} m_i &= m_{i-1} + \frac{h}{2}(y_{i-1} + y_i) \\ &= m_{i-1} + \frac{h}{2}(f(m_{i-1}, t_{i-1}) + f(m_{i-1} + hf(m^-_{i-1}, t_{i-1}))). \end{aligned}$$

- if m^-_{i-1}, m_{i-1} are exact, then $m_i - x(t_i) = \mathcal{O}(h^3)$
- If f has no time derivative (e.g. it actually comes from a Wiener process), the leading error is $\mathcal{O}(h^2)$, and the local error estimate is a closer bound.

Classic Analysis

for the explicit trapezoid rule

Hairer, Nørsett & Wanner, 1993

Definition: (order)

An iterative IVP solver is said to have (local) **order** p if, for every sufficiently smooth ODE, there exists a constant $K \in \mathbb{R}$, such that

$$\|e_0\| := \|x(t_0 + h) - \hat{x}(t_0 + h)\| \leq Ch^{p+1}. \quad (1)$$

That is, if the Taylor expansions of the exact solution and the estimated solution at the first step of size h from the initial value coincide up to, and including, the p-th term. The term e_0 is called the **local error at t_0**. Analogously, the local error e_i at t_i is $x(t_i + h) - \hat{x}(t_i + h)$ under the assumption $\hat{x}(t_i) = x(t_i)$, i.e. that $e_{i-1} = 0$. The **global error** is

$$E = x(T) - \hat{x}(T), \quad (2)$$

the difference between true and estimated solution at the end-point T.

Global Error?

classic analysis

Theorem [cf. Hairer, Nørsett & Wanner, 1993, Thm. 3.4]

Let U be a neighborhood of $\{(t, x(t)) \mid t_0 \leq t \leq T\}$ (that is, a neighborhood of the true solution). Assume that in U, the Jacobian of f is bounded:

$$\|f_x\| \leq L, \tag{3}$$

and that $e_i \leq Ch^{p+1}$ holds in U. Then **the global error is bounded by**

$$E \leq h^p \frac{\tilde{C}}{L} \left(e^{L(T-t_0)} - 1\right), \tag{4}$$

where $h = \max h_i$ is the maximal step size, and $\tilde{C} = C$ if $L \geq 0$ and $\tilde{C} = Ce^{-Lh}$ otherwise, and h is sufficiently small for the numerical estimate to remain in U. In the limit of constant f, $L \to 0$, this bound is, as expected, the sum of the local bounds: $h^p C(T - t_0) = NCh^{p+1}$, where $N = (T - t_0)/h$ is the number of steps.

Global Error Calibration

The probabilistic explicit Trapezoid Rule can provide conservative error estimates

- from above, variance of $x(T)$ (after $N = (T - t_0)/h$ steps):

$$P_{N,(11)} = \frac{N}{12}\theta^2 h^3 = (T - t_0)\frac{\theta^2}{12}h^2.$$

- thus, for differentiable f

$$\theta \geq \sqrt{12}\tilde{C} \cdot h \cdot \frac{e^{L(T-t_0)} - 1}{L(T - t_0)} \qquad \Rightarrow \qquad E < \sqrt{P_{N,11}}.$$

- and for non-differentiable f (method is first order), one fixed value of θ may suffice.

Internal Consistency?

the probabilistic trapezoid rule is asymptotically self-consistent

Hennig, Osborne, Girolami, upcoming

- probabilistic trapezoid rule arises from **observation model**

$$\ell(f, m_i^-, P_i^-) = \begin{cases} y_i & = f(Hm_i^-, t_i) \\ R_i & = 0 \end{cases}.$$

- Option I: $R_i = 0$: **implicit** P(EC)$^\infty$ method iterating until

$$m_i = m_i + \frac{h}{2}(f(m_{i-1}, t_{i-1}) + f(m_i, t_i))$$

- Option II: As $f(x(t_i)) = f(Hm_i^-) + (x_i - Hm_i^-)f_x(Hm_i^-) + \dots$, set

$$R = \operatorname{var}_{p_i}(x'(t_i)) \approx f_x^2(Hm_i^-)P_{i,11}^- \sim \mathcal{O}(h^3).$$

- thus **the extrapolation error matches the discretization error**:

$$m_1 = m_1^- + Kr = m_1^- + \left(1 - \frac{R_1}{h + R_1}\right)\begin{bmatrix} h/2 \\ 1 \end{bmatrix} r = m_1^{\text{trapezoid}} + \begin{bmatrix} \mathcal{O}(h^3) \\ \mathcal{O}(h^2) \end{bmatrix}$$

Gaussian Filters as ODE solvers

- have linear computational cost
- offer natural “plug-in” for probabilistic functionality through observation likelihood
- can **at least** achieve non-trivial (globally linear) convergence order
- can have **calibrated** error estimates
- are **asymptotically Bayesian** (for more see Chris’ talk)

What is their relation to classic methods?

Runge-Kutta methods

single-step algorithms for IVPs

Runge, 1895; Kutta, 1901

procedure RKSOLVE$((f, x_0, t_0, T, h), (\mathbf{c}, \mathbf{a}, \mathbf{b}, h))$
  $\hat{x}_0 = x_0$
  for $i = 1, t_i < T, i{+}{+}$ **do**
    $t_i = t_{i-1} + h$
    $y_{i1} = f(\hat{x}_i, t_i)$
    for $j = 2, \dots, q$ **do**
      $y_{ij} = f\left(\hat{x}_i + h \sum_{\ell=1}^{j-1} a_{j\ell} y_{i\ell}, t_i + c_j h\right)$
    end for
    $\hat{x}_{i+1} = \hat{x}_i + h \sum_{j=0}^{q} b_j y_{ij}$
  end for
end procedure

$$
\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_q & a_{q1} & a_{q2} & \cdots & a_{q,q-1} & \\
\hline
 & b_1 & b_2 & \cdots & b_{q-1} & b_q
\end{array}
$$

Runge-Kutta methods are intricate

not every RK method has high order

📖 Hairer & Wanner

$$
\begin{array}{c|c}
0 & \\
\hline
 & 1
\end{array}
\qquad
\begin{array}{c|cc}
0 & & \\
\alpha & \alpha & \\
\hline
 & (1 - 1/2\alpha) & 1/2\alpha
\end{array}
$$

$$
\begin{array}{c|cccc}
0 & & & & \\
u & u & & & \\
v & v - \frac{v(v-u)}{u(2-3u)} & \frac{v(v-u)}{u(2-3u)} & & \\
\hline
 & 1 - \frac{2-3v}{(6u(u-v))} - \frac{2-3u}{6v(v-u)} & & \frac{2-3v}{6u(u-v)} & \frac{2-3u}{6v(v-u)}
\end{array}
$$

Runge-Kutta Methods of High Order Exist

but they are few and far between!

Dormand & Prince, 1980, Hairer & Wanner, 1996

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
b & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\tilde{b} & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}
$$

```
dopri5.f / ode45.m / scipy.integrate.ode.set_integrator('dopri5')
```

Connection Filter — Runge-Kutta

given no further info, in the first step, our filter **can** be an RK method

Theorem: (expanded from Schober, Duvenaud, Hennig, 2014. Hennig, Osborne, Girolami, upcoming)

Consider the integrated Wiener process model

$$z(t) = \begin{bmatrix} x(t) \\ x'(t) \\ \vdots \\ \frac{1}{q!}x^{(q)}(t) \end{bmatrix} \in \mathbb{R}^{q+1}, \quad F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \text{ and } L = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \theta \end{bmatrix}.$$

Assume the initial mean (before conditioning on the initial value) is $\boldsymbol{m}_0^- = \boldsymbol{0}_{q+1}$. The associated P_0^- has elements $[P_0^-]_{ij} = \alpha_{ij}^0 \varepsilon^{i+j-2q-3}$. with *starting time* $\tau = \varepsilon^{-1}$ with constants $\alpha_0 \in \mathbb{R}^{q+1 \times q+1}$. Then, in the "uninformative" limit of $\varepsilon \to 0$ (i.e. $\tau \to \infty$), the first $0 \leq k \leq q$ steps of IVP_PROB, **for any finite and constant sequence of observation noises** $R_1, \dots, R_k$, produce the predictive posterior mean function $\boldsymbol{m}_k^-(t)$ whose first state is the unique $\boldsymbol{k}$-th order polynomial such that its first derivative exactly matches all preceding observations $\boldsymbol{y}_i$ at $\boldsymbol{t}_i$:

$$[m_{k+1}^-]_1(t) = x_0 + \sum_{i=1}^{k+1} \frac{a_i}{i!}(t - t_0)^i, \text{ such that } \frac{\partial}{\partial t}[m_{k+1}^-]_1(t_i) = [m_{k+1}^-]_2(t_i) = y_i \quad \forall\, 0 \leq i \leq k.$$

For $q = 1$, for $q = 2$ with arbitrary $0 < t_1 \leq h$, for $q = 3$ with arbitrary $0 < t_1 < t_2$ and $t_2 = 2/3h$, and for $q = 4$ and $[t_1, t_2, t_3] = [1/3, 1/2, 1]h$, this first step is a Runge-Kutta step of orders 1,2,3, and 4, respectively.

- Explicit structure of initial covariance $[P_0^-]_{ij} = \alpha_{ij}^0 \varepsilon^{i+j-2q-3}$ with

$$\alpha_{ij}^0 := \frac{\theta^2}{(2q+3-i-j)(q+1-i)!(q+1-j)!} = \alpha_{ji}^0.$$

- begin with conditioning on initial value: $p(x_0 \mid z(t_0)) = \mathcal{N}(x_0; \eta z(t_0), \rho_0)$ (with $[\eta]_i := \delta_{1i}$) gives Kalman gain, updated mean, covariance,

$$[K]_i = \left[P_0^- \eta^\intercal \frac{1}{\eta P_0^- \eta^\intercal + \rho_0}\right]_i = \frac{\alpha_{1i}^0 \varepsilon^{i-1}}{\rho_0 \varepsilon^{2q+1} + \alpha_{11}^0} = \frac{\alpha_{1i}^0}{\alpha_{11}^0} \varepsilon^{i-1} + \mathcal{O}(\varepsilon^i)$$

$$[\bar{m}_0]_i = [m_0^- + K(x_0 - \eta m_0^-)]_i = x_0 \left(\frac{\alpha_{1i}^0}{\alpha_{11}^0} \varepsilon^{i-1} + \mathcal{O}(\varepsilon^i)\right) \quad \rightharpoonup \quad \begin{bmatrix} x_0 \\ \mathbf{0} \end{bmatrix}$$

$$[\bar{P}_0]_{ij} = \frac{(\alpha_{ij}^0 \alpha_{11}^0 - \alpha_{1i}^0 \alpha_{1j}^0) \varepsilon^{i+j-2q-3} + \alpha_{ij}^0 \rho_0 \varepsilon^{i+j-2}}{\rho_0 \varepsilon^{2q+1} + \alpha_{11}^0} \quad \rightharpoonup \quad \begin{bmatrix} \rho_0 & \mathbf{0} \\ \mathbf{0} & \mathcal{O}(\varepsilon^{i+j-2q-3}) \end{bmatrix}$$

- define new constants:

$$\alpha^1_{ij} = \alpha^1_{ji} := \begin{cases} \rho_0 \frac{\alpha^0_{1j}}{\alpha^0_{11}} & \text{for } i = 1 \text{ or } j = 1 \\ \alpha^0_{ij} - \frac{\alpha^0_{1i}\alpha^0_{1j}}{\alpha^0_{11}} & \text{else.} \end{cases}$$

- repeat as above for initial obervation $p(y_0 \mid z(t_0)) = \mathcal{N}(y_0; Hz(t_0), R_0)$

$$K_i = \frac{[\bar{P}_0]_{i2}}{R_0 + [\bar{P}_0]_{22}} = \frac{[\bar{P}_0]_{i2}\varepsilon^{2q-1}}{R_0\varepsilon^{2q-1} + \alpha^1_{22}} + \mathcal{O}(\varepsilon^{2q}) = \begin{cases} \mathcal{O}(\varepsilon^{2q}) & \text{for } i = 1 \\ \frac{\alpha^1_{i2}}{\alpha^1_{22}}\varepsilon^{i-2} + \mathcal{O}(\varepsilon^{i-1}) & \text{for } i > 1 \end{cases}$$

$$m_0 = \bar{m}_0 + K(y_0 - H\bar{m}_0) = \begin{bmatrix} x_0 + \mathcal{O}(\varepsilon) \\ y_0 + \mathcal{O}(\varepsilon) \\ \mathcal{O}(\varepsilon) \\ \vdots \\ \mathcal{O}(\varepsilon^{q-1}) \end{bmatrix} \rightarrow \begin{bmatrix} x_0 \\ y_0 \\ \mathbf{0}_{q-1} \end{bmatrix}$$

$$[P_0]_{ij} = \alpha^2_{ij} \begin{cases} 1 & \min\{i,j\} \leq 2 \\ \varepsilon^{j+i-2q-3} + \mathcal{O}(\varepsilon^{j+i-2q-2}) & \text{else} \end{cases} \rightarrow \begin{bmatrix} \rho_0 & 0 & \mathbf{0} \\ 0 & R_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{O}(\varepsilon^{-2q-3+i+j}) \end{bmatrix}$$

Lemma: (tedious proof omitted)

Assume the filter used in the above Theorem performs q finite steps of size $0 < \chi_k := c_k h < 1$. If at step $k < q - 1$, the filter has estimation covariance P_k whose elements to leading order in ε have the form (with constants α_{ij}^{k+2})

$$[P_k]_{ij} = \alpha_{ij}^{k+2} \cdot \begin{cases} 1 & \text{if } \min\{i, j\} \leq k+2 \\ \varepsilon^{i+j-2q-3} & \text{else} \end{cases}, \qquad (\star)$$

then in the $k+1$-th step, the gain on the second element is unit for all values of R_{k+1}

$$[K]_2 = \frac{[P_{k+1}^-]_{22}}{[P_{k+1}^-]_{22} + R_{k+1}} = 1,$$

and the updated estimation covariance P_{k+1} again has the structure of $(\star)$, with updated constants α_{ij}^{k+3}.

The Lemma completes the proof of the first claim of the theorem, because unit gain on z_2 at step $k+1$ implies $[m_{k+1}^-]_2 = y_{k+1}$. $\square$

Connection Filter — Runge-Kutta

explicit Butcher tableaus and posterior covariances

- The Theorem implies that, for all $k < q$, a_i in

$$[m^-_{k+1}]_1(t) = x_0 + \sum_{i=1}^{k+1} \frac{a_i}{i!}(t - t_0)^i$$

are given by solution of **van-der-Monde** system

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & (t_1 - t_0) & (t_1 - t_0)^2 & \cdots & (t_1 - t_0)^k \\ \vdots & & & & \vdots \\ 1 & (t_k - t_0) & (t_k - t_0)^2 & \cdots & (t_k - t_0)^k \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k/k! \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix}$$

Connection Filter — Runge-Kutta

explicit Butcher tableaus and posterior covariances

$q = 1$: Euler's method

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \qquad m^-(h) = \begin{bmatrix} x_0 + hy_0 \\ y_0 \end{bmatrix} \qquad P^-(h) = \theta^2 \begin{bmatrix} \rho_0 + R_0 h^2 + 1/3 h^3 & hR_0 + 1/2 h^2 \\ hR_0 + 1/2 h^2 & R_0 + h \end{bmatrix}$$

already known from above, reflected by filter's analysis:

- first order convergence
- tracking extrapolation error gives $R_0 \sim \mathcal{O}(h^3)$, dominated by $\mathcal{O}(h^3)$ discretization error
- external input uncertainty up to $R_0 \sim \mathcal{O}(h)$ can be absorbed without breaking convergence (reflected in P^-!):

$$m_1 - x(h) = m_0 + \frac{h}{2}(y_0 + (f(m_i^-) + \varepsilon)) = \mathcal{O}(h^2) + \frac{h}{2}\sqrt{R}$$

- initial value has to be known to variance $\rho_0 \sim \mathcal{O}(h^4)$ to not affect convergence (reflected in P^-!):

$$m_{1,1} - x(h) \leq m_{0,1} + hf(m_{0,1}, t_0) - x_0 - hf(x_0, t_0) + \mathcal{O}(h^2)$$

Connection Filter — Runge-Kutta

explicit Butcher tableaus and posterior covariances

$q = 2$: All 2nd-order RK methods: $(t_1 =: t_0 + \chi_1 = t_0 + c_2 h, \chi_2 := h - \chi_1)$

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2 & \\ \hline & 1 - \frac{1}{2c_2} & \frac{1}{2c_2} \end{array} \qquad m^-(h) = \begin{bmatrix} x_0 + hy_0 + \frac{y_1 - y_0}{2\chi_1} h^2 \\ y_0 + \frac{y_1 - y_0}{\chi_1} h \\ \frac{y_1 - y_0}{\chi_1} \end{bmatrix}.$$

$$[P^-(h)]_{11} = \theta^2 \left(\frac{\chi_2^5}{20} + \frac{\chi_1^5}{120} - \frac{\chi_1^3 \chi_2^2}{24} + \frac{\chi_1 \chi_2^4}{12} \right) + \rho_0 + R_0 \left(\frac{3}{4} \chi_1^2 + \frac{1}{4} \chi_2^2 + \frac{1}{4} \frac{\chi_2^4}{\chi_1^2} \right)$$
$$+ R_1 \left(\frac{\chi_1^2}{4} + \chi_1 \chi_2 + \chi_2^2 + \frac{\chi_2^2}{2} + \frac{\chi_2^3}{\chi_1} + \frac{1}{4} \frac{\chi_2^4}{\chi_1^2} \right)$$

- For $\rho_0 = R_0 = R_1 = 0$, using $\chi_1 = c_2 h$:

$$\lim_{\rho_0 = R_0 = R_1 = 0} [P^-(h)]_{11} = \frac{\theta^2 h^5}{120} \left(5c_2^3 - 20c_2^2 + 20c_1 - 6 \right).$$

- minimal error estimate for $c_2 = 2/3$, reflecting model's ignorance of f_x. Leading error at h is

$$x(t_0+h) - \hat{x}(t_0+h) = h^3 \left(1/6 (f_t f_x + f_x^2 f)(x_0, t_0) + (2/3 - c_2)\left(f_{tt} + 2 f_{tx} f + f_{xx} f^2\right)(x_0, t_0) \right) + \mathcal{O}(h^4).$$

Connection Filter — Runge-Kutta

explicit Butcher tableaus and posterior covariances

$q = 3$: Third order if $c_3 = 2/3$

$$
\begin{array}{c|cccc}
0 & & & & \\
c_2 & c_2 & & & \\
c_3 & c_3 - \frac{c_3^2}{2c_2} & \frac{c_3^2}{2c_2} & & \\
\hline
 & 1 - \frac{2-3c_3}{6c_2(c_2-c_3)} - \frac{2-3c_2}{6c_3(c_3-c_2)} & & \frac{2-3c_3}{6c_2(c_2-c_3)} & \frac{2-3c_2}{6c_3(c_3-c_2)}
\end{array}
$$

- choosing $c_3 = 2/3$ removes dependence on second derivatives in error for $\hat{x}(t_2)$
- tempting to say the algorithm automatically chooses the c_3 that gives $p = 3$. But it's arguably getting lucky

Connection Filter — Runge-Kutta

explicit Butcher tableaus and posterior covariances

$q = 4$: One unique choice for fourth order

$$
\begin{array}{c|cccc}
0 & & & & \\
c_2 & c_2 & & & \\
c_3 & c_3 - \frac{c_3^2}{2c_2} & \frac{c_3^2}{2c_2} & & \\
c_4 & c_4 - a_{42} - a_{43} & \frac{2c_4^3 - 3c_4^2 c_3}{6c_2(c_2 - c_3)} & \frac{2c_4^3 - 3c_4^2 c_2}{6c_3(c_3 - c_2)} & \\
\hline
 & 1 - b_2 - b_3 - b_4 & \frac{6c_3c_4 - 4(c_3+c_4)+3}{12c_2(c_2-c_3)(c_2-c_4)} & \frac{-6c_2c_4+4(c_2+c_4)-3}{12(c_2-c_3)c_3(c_3-c_4)} & \frac{6c_2c_3-4(c_2+c_3)+3}{12(c_2-c_4)(c_3-c_4)c_4}
\end{array}
$$

Classic analysis (Hairer, Nørsett & Wanner, 1993, Thm. 1.6): Fourth order iff

$$
\begin{aligned}
b_1 + b_2 + b_3 + b_4 &= 1 & c_4 &= 1 \\
b_2 c_2 + b_3 c_3 + b_4 c_4 &= 1/2 & b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 &= 1/3 \\
b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 &= 1/4 & b_3 c_3 a_{32} c_2 + b_4 c_4 (a_{42} c_2 + a_{43} c_3) &= 1/8 \\
b_3 a_{32} + b_4 a_{42} &= b_2 (1 - c_2) & b_4 a_{43} &= b_3 (1 - c_3)
\end{aligned}
$$

$$\Rightarrow \mathbf{c} = [0, 1/3, 1/2, 1]$$

- The very first q steps of our probabilistic filter, for an uninformative prior, can be a **Runge-Kutta** step of up to **fourth order**
  - the unique first-order method (Euler)
  - all second-order methods
  - a univariate family of all third-order methods
  - a single fourth-order method
- both external and internal (extrapolation) uncertainties can be tracked explicitly by the filter
- the extrapolation uncertainty is dominated by the discretization error: Claiming to observe exact derivatives is asymptotically ok
- external uncertainties are “ok” up to $R \sim \mathcal{O}(h^{2q-1})$, uncertainties in initial value up to $\rho_0 \sim \mathcal{O}(h^{2q+1})$

What happens after the first q steps?

Nordsieck Methods

perhaps a historical mishap?[1] 🗎 Nordsieck, 1962

Consider the **Nordsieck Vector**

$$z^{\mathrm{N}}_{t_i} = \begin{bmatrix} x(t_i) & hx'(t_i) & \frac{h^2}{2!}x''(t_i) & \dots & \frac{h^q}{q!}x^{(q)}(t_i) \end{bmatrix}^{\intercal}$$

and methods which recursively estimate (slightly simplified definition of f_{i+1})

$$z^{\mathrm{N}}_{t_{i+1}:=t_i+h} = Tz^{\mathrm{N}}_{t_i} + \ell(hf_{i+1} - HTz^{\mathrm{N}}_{t_i}) \text{ with } \ell \in \mathbb{R}^{q+1}, \quad f_{i+1} := f(HTz^{\mathrm{N}}_{t_i}, t_{i+1}) \qquad (\star)$$

with the **Pascal Triangle Matrix**

$$[T]_{ij} = \begin{cases} \binom{j-1}{i-1} & 1 \leq i \leq j \leq q+1 \\ 0 & \text{else.} \end{cases} \qquad \Rightarrow_{q=3} \qquad T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and the ODE requires $x'(t_{i+1}) = f_{i+1}$, thus $\ell_2 \overset{!}{=} 1$. For our filter:

$$z^{\mathrm{N}}_{t_i} = \mathrm{diag}_i(h^{i-1}/(i-1)!) \cdot z_{t_i}$$

[1]Kalman filter: 🗎 Swerling 1958, 🗎 Kalman 1960, 🗎 Kalman & Bucy, 1961

Nordsieck Methods

perhaps a historical mishap?[1]

Nordsieck, 1962

Equivalence to Linear Multistep Methods — Hairer, Nørsett, Wanner, 1993, §III.6

Consider the Nordsieck method ($\star$) with $\ell_2 = 1$ and assume $\ell_{q+1} \neq 0$.

- This method satisfies the linear multistep formula (Dahlquist, 1956)

$$\sum_{i=0}^{q} \alpha_i \hat{x}_{t_i} = h \sum_{i=0}^{q} \beta_i f(\hat{x}_{t_i}, t_i)$$

  where the map ($\alpha, \beta \leftrightarrow \ell$) is an isomorphism under certain technical constraints. (Hairer et al. Thms. 6.1–6.3)

- The method has order at least q, its global error constant is

$$C = -\frac{b^\top \ell}{(q+1)!\ell_{q+1}}$$

  where $b^\top = \left[B_0, B_1, \dots, B_q\right] = \left[1, -1/2, -1/6, 0, -1/30, \dots\right]$ are the Bernoulli numbers.

[1] Kalman filter: Swerling 1958, Kalman 1960, Kalman & Bucy, 1961

Steady State of Kalman Filters

the dynamics of Kalman filter covariances

$$P_t^- = AP_{t-1}A^\intercal + Q \qquad K = P_t^- H^\intercal S^{-1} \qquad P_t = (I - KH)P_t^-$$

- follows a **Discrete-time Algebraic Riccati Equation** (DARE)[2]

$$\begin{aligned} P_{i+1}^- &= AP_i^- A^\intercal - AP_i^- H^\intercal (HP_i^- H^\intercal + R)^{-1} HP_i^- A^\intercal + Q \\ &= Q + A((P_i^-)^{-1} + H^\intercal R^{-1} H)^{-1} A^\intercal . \end{aligned}$$

- Their solution can be found considering the symplectic matrix

$$Z = \begin{bmatrix} A^\intercal + H^\intercal R^{-1} H A^{-1} Q & -H^\intercal R^{-1} H A^{-1} \\ -A^{-1} Q & A^{-1} \end{bmatrix} \in \mathbb{R}^{2(q+1)\times 2(q+1)}.$$

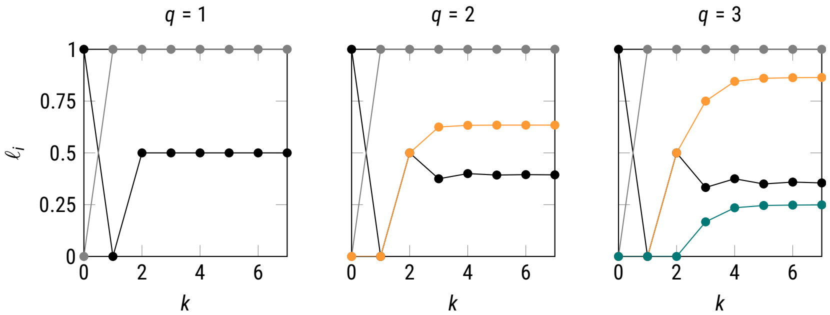
If Z has no eigenvalues on the unit circle, then half its eigenvalues are inside the unit circle. Let $U \in \mathbb{C}^{2(q+1)\times(q+1)}$ be the associated eigenvectors. Then the **steady-state** predictive covariance is

$$U = \begin{bmatrix} U_1 \in \mathbb{C}^{N\times N} \\ U_2 \in \mathbb{C}^{N\times N} \end{bmatrix} \qquad P_\infty^- = U_2 U_1^{-1}.$$

[2] The continuous-time limit (CARE) is $dP = AP + PA^\intercal - PH^\intercal R^{-1} HP + Q$. *Riccati equations* are ODEs of the form $x'(t) = q(t) + a(t)x(t) + b(t)x^2(t)$.

Connection Filter — Nordsieck

in the steady-state limit, our filter becomes a multi-step method



	ℓ_1	ℓ_2	ℓ_3	ℓ_4
$q = 1$	0.5	1		
$q = 2$	0.433	1	0.634	
$q = 3$	0.383	1	0.861	0.247

- In the **steady-state-limit**, IVP_PROB turns into a **Nordsieck Method**.
- (Nordsieck methods essentially capture all linear multistep methods)
- for $q = 1, 2, 3$ and $R = 0$ it has order q.
- the steady state is reached quickly if h, θ, R are kept constant
- in practice, however, these parameters are adapted continuously

Parameter Adaptation

Schober, Särkkä, Hennig, arXiv 1610.05261

just very briefly. For more, see op. cit. and

`https://github.com/ProbabilisticNumerics/pfos`

- use local error estimate to set θ (intensity of diffusion)
- e.g. by conjugate prior inference (see previous lecture) or by tracking Kalman residual, covariance

$$(HP_i^- H^\intercal + R)^{-1}(y - Hm_i^-)$$

  (lines 13/14 in IVP_PROB). The latter is related to **embedded formulae** in RK methods
- set h so predicted error $\sqrt{P_{11}^-} <$ abs_tol
- as everywhere, some care/hacking required

Uncertainty Quantification for ODE solvers

related work

- O. Chkrebtii, D.A. Campbell, B. Calderhead & M. Girolami
  Bayesian Uncertainty Quantification for Differential Equations
  Bayesian Analysis **11**/4 (2016), 1239–1267
- P.R. Conrad, M. Girolami, S. Särkkä, A.M. Stuart & K. Zygalakis
  Probability measures for numerical solutions of differential equations, arXiv 1506.04592
- H.C. Lie, A.M. Stuart, T.J. Sullivan
  Strong convergence rates of probabilistic integrators for ordinary differential equations arXiv 1703.03680

- perform ODE solver prediction (either classic single-step method or this filter), then add perturbation

$$\hat{x}(t_i) \leftarrow \mathbb{E}(x(t_i)) + \sqrt{\operatorname{var}(x(t_i))}u \qquad u \sim \mathcal{N}(0,1)$$

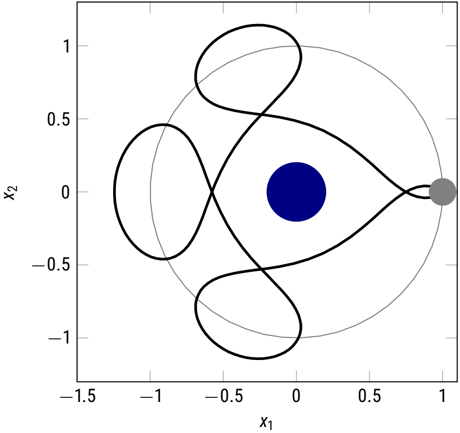
$$\Rightarrow \qquad \hat{x}(t_i) \sim \mathcal{N}(x(t_i); m_i, P_i)$$

- run once to get one (!) sample. Repeat many times to get **non-Gaussian** empirical measure
- more discussion in Chris Oates' lectures (?)

Example Problem

restricted 3-body problem

[from Hairer & Wanner, Chp. II, p. 129]



$$f(t,x) = \begin{bmatrix} x_3 \\ x_4 \\ x_1 + 2x_4 - \frac{\mu'(x_1+\mu)}{((x_1+\mu)^2+x_2^2)^{3/2}} - \frac{\mu x_2}{((x_1-\mu')^2+x_2^2)^{3/2}} \\ x_2 - 2x_3 - \frac{\mu' x_1}{((x_1+\mu)^2+x_2^2)^{3/2}} - \frac{\mu x_2}{((x_1-\mu')^2+x_2^2)^{3/2}} \end{bmatrix}$$

$$\mu = 0.012277471 \quad (♁ - ☽)$$

$$\mu' = 1 - \mu$$

$$x_0 = \begin{bmatrix} 0.994 \\ 0 \\ 0 \\ -2.00158510637908252240537862224 \end{bmatrix}$$

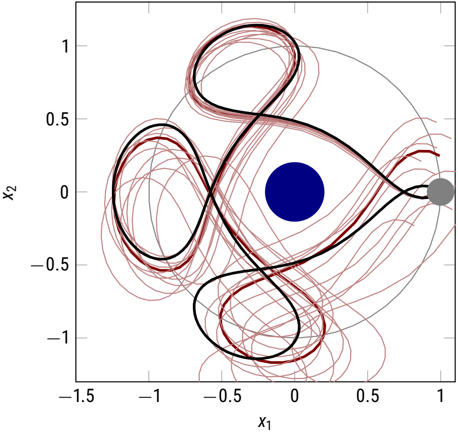
$$t_0 = 0$$

$$T = 17.0652165601579625588917206249$$

Example Problem

restricted 3-body problem

[from Hairer & Wanner, Chp. II, p. 129]

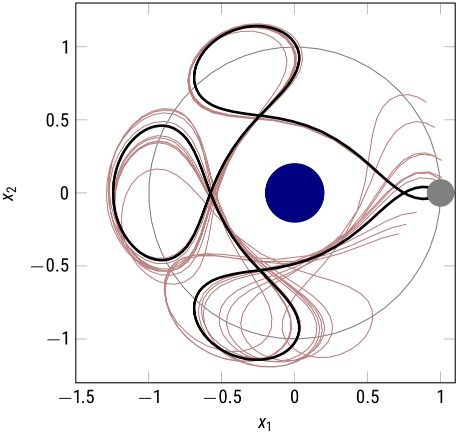


- solved by Kalman filter to first order

Example Problem

restricted 3-body problem

[from Hairer & Wanner, Chp. II, p. 129]

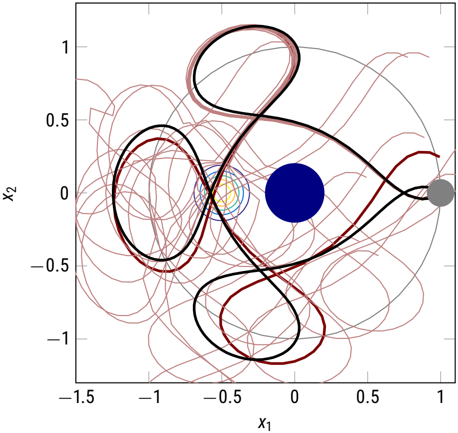


- solved by stochastic disturbances (first order)

Example Problem

restricted 3-body problem

[from Hairer & Wanner, Chp. II, p. 129]



- solved by Kalman filter, with heteroscedastic external uncertainty

$$R(t,x) = 0.1^2 \exp\left(-\frac{\sum_{i=1,2}(x_i - \mu_i)^2}{2 \cdot 0.1.^2}\right)$$

Summary

- Polynomial (integrated Wiener process) filters can be used to formulate **a** rich class of IVP solvers
- in the **first** step, under a non-informative prior, they (can) reproduce low-order **Runge-Kutta methods**
- in the steady-state **limit**, they approach **Nordsieck (multistep) methods**
- their posterior variance contracts at the **worst-case error** rate
- **non-Gaussian** extensions give more expressive, more expensive, estimates (analysis still in development)

These slides can be found at

`http://tinyurl.com/Dobbiaco-Hennig-3`