A Probabilistic Model for the Numerical Solution of Initial Value Problems

by Schober, Särkkä and Hennig

Jon Cockayne

July 26, 2017

Introduction

In Schober et al. [2016]...

- Schober et al. [2014] showed that for certain choices of GP prior, sequential conditioning led to a Runge-Kutta mean.

[1] Skeel [1979]

In Schober et al. [2016]...

- Schober et al. [2014] showed that for certain choices of GP prior, sequential conditioning led to a Runge-Kutta mean.
- This could only be applied over a single time step.

[1] Skeel [1979]

Introduction

In Schober et al. [2016]...

- Schober et al. [2014] showed that for certain choices of GP prior, sequential conditioning led to a Runge-Kutta mean.
- This could only be applied over a single time step.
- In this paper it is shown that for those priors, the predictive posterior corresponds to a linear multistep method for solving the ODE in Nordsieck form[1].

[1] Skeel [1979]

Introduction

In Schober et al. [2016]...

- Schober et al. [2014] showed that for certain choices of GP prior, sequential conditioning led to a Runge-Kutta mean.
- This could only be applied over a single time step.
- In this paper it is shown that for those priors, the predictive posterior corresponds to a linear multistep method for solving the ODE in Nordsieck form[1].
- This allows us to establish global error results in a consistent fashion.

[1] Skeel [1979]

Standard Solvers and Nordsieck Form

Problem Setup

We have an IVP:

$$y'(x) = f(t, y(t))$$

where

- $t \in \mathbb{T} := [t_0, T] \subset \mathbb{R}$
- $y : \mathbb{T} \to \mathbb{R}$
- $f : \mathbb{T} \times \mathbb{R} \to \mathbb{R}.$
- $y(t_0) = y_0.$

Problem Setup

We have an IVP:

$$y'(x) = f(t, y(t))$$

where

- $t \in \mathbb{T} := [t_0, T] \subset \mathbb{R}$
- $y : \mathbb{T} \to \mathbb{R}$
- $f : \mathbb{T} \times \mathbb{R} \to \mathbb{R}$.
- $y(t_0) = y_0$.

Produce a solution on a grid $\{t_n\}$, $n = 0, 1, 2, \ldots, N$. Let y_n be the approximation to y at time t_n, and $h_n = t_n - t_{n-1}$. Let $z_n = f(t_n, y_n)$.

Solution Methods

Only consider explicit methods.

Solution Methods

Only consider explicit methods.

The most basic: Explicit Euler

$$y_n = y_{n-1} + hz_{n-1}$$

Problems: global error $O(h)$, unstable.

We would like methods of order q - that is, with global error $O(h^q)$

(Explicit) Linear Multistep Methods approximate the solution as a linear combination of previous evaluations of f and estimates of y

$$\sum_{i=0}^{k} \alpha_i y_{n-i} = h \sum_{i=1}^{k} \beta_i z_{n-i} \tag{1}$$

Linear Multistep Methods

(Explicit) Linear Multistep Methods approximate the solution as a linear combination of previous evaluations of f and estimates of y

$$\sum_{i=0}^{k} \alpha_i y_{n-i} = h \sum_{i=1}^{k} \beta_i z_{n-i} \tag{1}$$

Two predominant methods:

- Adams-Bashforth
- Adams-Moulton

[See also: Teymur et al. [2016]]

Runge Kutta Methods evaluate f at multiple locations between time-steps:

$$k_i = f(t_n + c_i h, y + h \sum_{j<i} a_{ij} k_j)$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i$$

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \end{array}$$

Butcher Tableau for $q = s = 2$:

$$k_1 = f(t_n, y_n)$$
$$k_2 = f(t_n + \alpha h, y_n + \alpha h k_1)$$
$$y_{n+1} = y_n + h\left(1 - \frac{1}{2\alpha}\right)k_1 + \frac{h}{2\alpha}k_2$$

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

Midpoint method: $\alpha = \frac{1}{2}$

Heun's method: $\alpha = 1$

All Linear Multistep Methods[2] can be written in Nordsieck form:

$$\mathbf{x}_n = \left(y_n, hy'_n, \ldots, \frac{h^q y_n^{(q)}}{q!}\right)$$
$$\mathbf{x}_{n+1} = \left(\mathbf{I} - \mathbf{L}\mathbf{e}_1^T\right) \mathbf{P}\mathbf{x}_n + h\mathbf{L}z_n$$

where z_n solves

$$[\mathbf{P}\mathbf{x}_n]_1 + [\mathbf{L}]_1 z_n = hf(t_n + h, [\mathbf{P}\mathbf{x}_n]_0 + h[\mathbf{L}]_0 z_n)$$

$\mathbf{P}$ the Pascal Triangle matrix.

[2] and some Runge-Kutta Methods?

Nordsieck Form

For suitable choices of $\boldsymbol{L}$, Nordsieck methods can achieve local truncation error (at least) q.

A Probabilistic Model

Notation

For a probability space $(\Omega, \mathcal{F}, P)$...

- X_t the prior distribution for the solution $y(t)$[3], taking values in $\mathbb{R}^{q+1}$.
- Y the *true* solution
- Y' its first derivative.

[3] This is somewhat unclear from the text, but I believe it should be considered as the prior

Notation

For a probability space $(\Omega, \mathcal{F}, P)$...

- X_t the prior distribution for the solution $y(t)$[3], taking values in $\mathbb{R}^{q+1}$.
- Y the *true* solution
- Y' its first derivative.

We also let $\mathcal{F}_t$ denote the filtration generated by X_t.

[3] This is somewhat unclear from the text, but I believe it should be considered as the prior

SDE Formulation

Obervation: certain GPs can be written as a Linear Time Invariant (LTI) SDE:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{F}\boldsymbol{X}_t\mathrm{d}t + \boldsymbol{L}\mathrm{d}W_t$$

$\boldsymbol{F}$ the "state feedback matrix" and $\boldsymbol{L}$ the "diffusion vector". W_t a Wiener process, $\mathrm{d}W(t) \sim \mathcal{N}(0, \sigma^2\mathrm{d}t)$.

SDE Formulation

Obervation: certain GPs can be written as a Linear Time Invariant (LTI) SDE:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{F}\boldsymbol{X}_t \mathrm{d}t + \boldsymbol{L}\mathrm{d}W_t$$

$\boldsymbol{F}$ the "state feedback matrix" and $\boldsymbol{L}$ the "diffusion vector". W_t a Wiener process, $\mathrm{d}W(t) \sim \mathcal{N}(0, \sigma^2 \mathrm{d}t)$.

We assume Y_t and Y_t' are related to $\boldsymbol{X}_t$ by

$$Y_t = \boldsymbol{H}_0 \boldsymbol{X}_t$$
$$Y_t' = \boldsymbol{H}_1 \boldsymbol{X}_t$$

SDE Formulation

Given a $X_{t_*} \sim \mathcal{N}(m_*, C_*)$, for $t > t_*$ we have that X_t is also Gaussian.

SDE Formulation

Given a $X_{t_*} \sim \mathcal{N}(\boldsymbol{m}_*, \boldsymbol{C}_*)$, for $t > t_*$ we have that X_t is also Gaussian.

Let $\boldsymbol{A}(h) = \exp(\boldsymbol{F}h)$. Then

$$
\begin{aligned}
\boldsymbol{m}_t &= \boldsymbol{A}(t - t_*)\boldsymbol{m}_* \\
\operatorname{cov}(X_t, X_{t'}) &= \boldsymbol{A}(t' - t_*)\boldsymbol{C}_*\boldsymbol{A}(t - t_*)^\top \\
&\quad + \int_{t_*}^{\min(t,t')} \boldsymbol{A}(t - \tau)\boldsymbol{L}\sigma^2\boldsymbol{L}^\top\boldsymbol{A}(t' - \tau)^\top \mathrm{d}\tau
\end{aligned}
$$

SDE Formulation

Given a $X_{t_*} \sim \mathcal{N}(\boldsymbol{m}_*, \boldsymbol{C}_*)$, for $t > t_*$ we have that X_t is also Gaussian.

Let $A(h) = \exp(Fh)$. Then

$$\begin{aligned}\boldsymbol{m}_t &= A(t - t_*)\boldsymbol{m}_* \\ \operatorname{cov}(X_t, X_{t'}) &= A(t' - t_*)C_* A(t - t_*)^\top \\ &\quad + \int_{t_*}^{\min(t,t')} A(t - \tau) L \sigma^2 L^\top A(t' - \tau)^\top \mathrm{d}\tau\end{aligned}$$

For the practical algorithm we only ever need $\operatorname{cov}(X_{t+h}, X_{t+h})$, so let

$$Q(h) = \int_t^{t+h} A(h) L \sigma^2 L^\top A(h)^\top \mathrm{d}\tau$$

SDE Formulation

The advantage of this formulation is that if we view solution as a filtering problem, it allows us to express predictive means in Nordsieck form.

Assuming:

1. Observations $\boldsymbol{z}_t$ are linked to a hidden state $\boldsymbol{x}_t$ by a Linear operator:

$$\boldsymbol{z}_t = \boldsymbol{H}\,\boldsymbol{x}_t + \boldsymbol{\xi}_t$$

   where $\boldsymbol{\xi}_t \sim \mathcal{N}(0, R)$

Assuming:

1. Observations $\boldsymbol{z}_t$ are linked to a hidden state $\boldsymbol{x}_t$ by a Linear operator:

$$\boldsymbol{z}_t = \boldsymbol{H}\,\boldsymbol{x}_t + \boldsymbol{\xi}_t$$

   where $\boldsymbol{\xi}_t \sim \mathcal{N}(0, R)$

2. We have some Prediction equation

$$\boldsymbol{x}_{t+1}^{-} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\eta}_t$$

   where $\boldsymbol{\eta}_t \sim \mathcal{N}(0, Q)$

Assuming:

1. Observations $\boldsymbol{z}_t$ are linked to a hidden state $\boldsymbol{x}_t$ by a Linear operator:

$$\boldsymbol{z}_t = \boldsymbol{H}\,\boldsymbol{x}_t + \boldsymbol{\xi}_t$$

   where $\boldsymbol{\xi}_t \sim \mathcal{N}(0, R)$
2. We have some Prediction equation

$$\boldsymbol{x}^{-}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\eta}_t$$

   where $\boldsymbol{\eta}_t \sim \mathcal{N}(0, Q)$
3. We endow $\boldsymbol{x}_0$ with a Gaussian prior.

Assuming:

1. Observations $\boldsymbol{z}_t$ are linked to a hidden state $\boldsymbol{x}_t$ by a Linear operator:

$$z_t = H\, x_t + \xi_t$$

   where $\xi_t \sim \mathcal{N}(0, R)$

2. We have some Prediction equation

$$x_{t+1}^{-} = A x_t + \eta_t$$

   where $\eta_t \sim \mathcal{N}(0, Q)$

3. We endow $\boldsymbol{x}_0$ with a Gaussian prior.

the Kalman filter describes how to exactly update the posterior distribution as new observations are obtained.

More or less...

1. Predict: $\mathbf{x}_{t+1}^{-}|\mathbf{z}_t$
2. Update: $\mathbf{x}_{t+1}|\mathbf{z}_{t+1}$

Under the assumptions, distributions on $\mathbf{x}_t$ and $\mathbf{x}_t^{-}$ are Gaussian for all t.

The Kalman Filter

More or less...

1. Predict: $\mathbf{x}_{t+1}^{-}|\mathbf{z}_t$
2. Update: $\mathbf{x}_{t+1}|\mathbf{z}_{t+1}$

Under the assumptions, distributions on $\mathbf{x}_t$ and $\mathbf{x}_t^{-}$ are Gaussian for all t.

Kalman's equations tell us how to efficiently update the mean and covariance matrices for each new piece of information.

Filtering for ODEs

Start with a Gaussian distribution X_0. For $n = 0, \ldots, N$ do...

1. Compute the predictive distribution $X_{t_n}^- \mid z_{[n-1]}$
2. Compute $z_n = f(t_n, X_{t_n}^-)$ (noiseless)
3. Find $X_{t_n} \mid z_{[n]}$

Problem: Step 2 is not tractable.

Filtering for ODEs

Solution: replace step 2 with something explicitly computable:

1. Compute the predictive distribution $X_{t_n}^- \mid z_{[n-1]}$
2. Compute $z_n = f(t_n, \mathbb{E}(X_{t_n}^-))$ (noiseless)
3. Find $X_{t_n} \mid z_{[n]}$

The Algorithm

Runge Kutta Means

Runge Kutta Means

This formulation makes simpler the earlier result[4] regarding Runge Kutta Means.

[4] Schober et al. [2014]

Runge Kutta Means

This formulation makes simpler the earlier result[4] regarding Runge Kutta Means.

In what follows we will use an Integrated Wiener Process of order q (IWP(q)) form for $\boldsymbol{X}_t$:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{U}_{q+1}\boldsymbol{X}\mathrm{d}t + \boldsymbol{e}_q\mathrm{d}W$$

Where $\boldsymbol{U}_q$ is the upper shift matrix of size q. This gives a convenient closed-form for $\boldsymbol{A}(h)$, $\boldsymbol{Q}(h)$.

[4] Schober et al. [2014]

This formulation makes simpler the earlier result[4] regarding Runge Kutta Means.

In what follows we will use an Integrated Wiener Process of order q (IWP(q)) form for X_t:

$$\mathrm{d}X_t = U_{q+1}X\mathrm{d}t + e_q\mathrm{d}W$$

Where U_q is the upper shift matrix of size q. This gives a convenient closed-form for $A(h)$, $Q(h)$.

Choose the prior mean to be $m^-_{t_{-1}} \equiv 0$ and the prior covariance to be $C^-_{t_{-1}}$.

[4] Schober et al. [2014]

IWP(1)

The predictive mean at $t = 1$ is equivalent to Explicit Euler.

IWP(1)

The predictive mean at $t = 1$ is equivalent to Explicit Euler.

$$\boldsymbol{m}_{t_{-1}} = \begin{pmatrix} y_0 \\ m^-_{t_0,1} \end{pmatrix}$$

$$\boldsymbol{C}_{t_{-1}} = \begin{pmatrix} 0 & 0 \\ 0 & C^-_{t_0,11} \end{pmatrix}$$

for some $m^-_{t_0,1}$ and $C^-_{t_0,11}$

IWP(1)

The predictive mean at $t = 1$ is equivalent to Explicit Euler.

$$\boldsymbol{m}_{t_0} = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}$$

$$\boldsymbol{C}_{t_0} = \boldsymbol{0}$$

IWP(1)

The predictive mean at $t = 1$ is equivalent to Explicit Euler.

t_{-1} $\quad$ t_0 $\quad$ t_1

$$
\boldsymbol{m}_{t_1}^{-} = \begin{pmatrix} y_0 + hz_0 \\ z_0 \end{pmatrix}
$$

$$
\boldsymbol{C}_{t_1}^{-} = \boldsymbol{Q}(h)
$$

IWP(1)

The predictive mean at $t = 1$ is equivalent to Explicit Euler.

t_{-1} t_0 t_1

$$\boldsymbol{m}_{t_1} = \begin{pmatrix} y_0 + \frac{h}{2}(z_0 + z_1) \\ z_1 \end{pmatrix}$$

$$\boldsymbol{C}_{t_1} = \sigma^2 \begin{pmatrix} \frac{h^3}{12} & 0 \\ 0 & 0 \end{pmatrix}$$

This corresponds to an RK2 scheme. (Heun's Method, $\alpha = 1$).

Continuation

Continuing this past t_1 we see that the scheme no longer matches any known numerical method (not even Heun's method).

For an IWP(2) model we fix the prior covariance to be $\boldsymbol{C}_{t_{-1}}^{-} = \boldsymbol{Q}(\tau)$ for some $\tau > 0$.

For an IWP(2) model we fix the prior covariance to be $\boldsymbol{C}^{-}_{t_{-1}} = \boldsymbol{Q}(\tau)$ for some $\tau > 0$.

t_{-1} $\quad$ t_0 $\quad$ $t_0 + h\alpha$ $\quad$ t_1

$$\left[\boldsymbol{m}^{-}_{t_0+h\alpha}\right]_0 = y_0 + h\alpha z_0 + \frac{h^2\alpha^2}{2}\left(\frac{4z_0}{\tau} - \frac{20y_0}{3\tau}\right)$$

IWP(2)

For an IWP(2) model we fix the prior covariance to be $\boldsymbol{C}_{t_{-1}}^{-} = \boldsymbol{Q}(\tau)$ for some $\tau > 0$.

$$\left[\boldsymbol{m}_{t_0+h\alpha}^{-}\right]_0 = y_0 + h\alpha z_0 + \frac{h^2\alpha^2}{2}\left(\frac{4z_0}{\tau} - \frac{20y_0}{3\tau}\right)$$

For equivalence with RK2 we require

$$\left[\boldsymbol{m}_{t_0+h\alpha}^{-}\right]_0 = y_0 + h\alpha z_0$$

Solution (?): Send $\tau \to \infty$

IWP(2)

$\tau \to \infty$

- Equivalent to using an improper prior.
- More severe ramifications for continuation.
- Challenging computationally.

$\tau \to \infty$

- Equivalent to using an improper prior.
- More severe ramifications for continuation.
- Challenging computationally.

Difficult to justify given that the domain is restricted to $[t_0, T]$

Convergence Analysis and Nordiseck Methods

Connection to Nordsieck Methods

The posterior mean does not correspond to an RK method past t_1...

$\implies$ no global error results from this route!

Connection to Nordsieck Methods

The posterior mean does not correspond to an RK method past t_1...

$\implies$ no global error results from this route!

It does correspond to a general linear method in Nordsieck form.

Nordsieck Form

We re-scale the state vector:

$$\tilde{X} = \begin{pmatrix} Y_t \\ hY'_t \\ \frac{h^2}{2!} Y''_t \\ \vdots \\ \frac{h^q}{q!} Y_t^{(q)} \end{pmatrix} = \boldsymbol{B} X_t$$

yielding the new SDE

$$\mathrm{d}\tilde{X}_t = \boldsymbol{B} \boldsymbol{U}_{q+1} \boldsymbol{B}^{-1} \tilde{X}_t \mathrm{d}t + \boldsymbol{B} \boldsymbol{e}_q \mathrm{d}W$$

which makes $\tilde{A}(h)$ independent of h.

Proposition 1

The probabilistic Nordsieck method arising from the once-integrated Wiener process is equivalent in predictive posterior mean with the trapezoidal rule.

Proposition 1

The probabilistic Nordsieck method arising from the once-integrated Wiener process is equivalent in predictive posterior mean with the trapezoidal rule.

This means that the IWP(1) probabilistic Nordsieck method is actually of order 2 rather than order 1!

IWP(2)

Theorem 1

The predictive posterior mean of the IWP(2) with fixed step size h is a third order Nordsieck method.

Calibration and Experiments

Model Parameters

Three parameters to tune: q, σ^2, h. q is fixed to 2.

Model Parameters

Three parameters to tune: q, σ^2, h. q is fixed to 2.

As with standard solvers we adapt h_n at each step to not exceed a tolerance.

Model Parameters

Three parameters to tune: q, σ^2, h. q is fixed to 2.

As with standard solvers we adapt h_n at each step to not exceed a tolerance.

σ is chosen by maximising the marginal likelihood of the residual:

$$\hat{\sigma} := \arg\min_{\sigma} p\left(z_n - [m_{t_n}^{-}]_1 \mid \sigma\right)$$

Experiments

[To the paper!]

Conclusions

Criticisms

- The choice of $z_n = f(t_n, \mathbb{E}(X_{t_n}^-))$.

Criticisms

- The choice of $z_n = f(t_n, \mathbb{E}(X_{t_n}^-))$.
- The filtration assumption.

Further Work

- How does this relate to the fully Bayesian posterior distribution?

Further Work

- How does this relate to the fully Bayesian posterior distribution?
- What about implicit schemes?

Thanks!

References

Michael Schober, David K Duvenaud, and Philipp Hennig. Probabilistic ode solvers with runge-kutta means. In *Advances in neural information processing systems*, pages 739–747, 2014.

Michael Schober, Simo Särkkä, and Philipp Hennig. A probabilistic model for the numerical solution of initial value problems. *arXiv preprint arXiv:1610.05261*, 2016.

Robert D Skeel. Equivalent forms of multistep formulas. *Mathematics of Computation*, 33(148):1229–1250, 1979.

Onur Teymur, Kostas Zygalakis, and Ben Calderhead. Probabilistic linear multistep methods. In *Advances in Neural Information Processing Systems*, pages 4314–4321, 2016.