Representation of simulation errors in single step methods using state dependent noise

Edward Boje^{1,*}

¹Electrical Engineering, University of Cape Town, South Africa

Abstract. The local error of single step methods is modelled as a function of the state derivative multiplied by bias and zero-mean white noise terms. The deterministic Taylor series expansion of the local error depends on the state derivative meaning that the local error magnitude is zero in steady state and grows with the rate of change of the state vector. The stochastic model of the local error may include a constant, “catch-all” noise term.

A continuous time extension of the local error model is developed and this allows the original continuous time state differential equation to be represented by a combination of the simulation method and a stochastic term. This continuous time stochastic differential equation model can be used to study the propagation of the simulation error in Monte Carlo experiments, for step size control, or for propagating the mean and variance. This simulation error model can be embedded into continuous-discrete state estimation algorithms.

Two illustrative examples are included to highlight the application of the approach.

1 Introduction

In the solution of ordinary differential equations using numerical methods the problem of local and global errors arises because of finite order methods and finite step size. The problem has been considered exhaustively in the literature where, in simulation studies, the aim is usually to get a solution that has sufficiently small global error and over-estimating the error is wasteful but does no harm. In time- or resource-critical applications of numerical solution of ODEs, correct local error estimates are required and this motivates a more careful analysis of the local error. Of particular interest is the application of numerical simulation used to project state equations between measurement updates in continuous-discrete state estimation problems where a reliable estimate of the local error behaviour can be embedded into the state estimator to improve estimator optimality, for example, if the effect of the local error is over-estimated, the estimator will de-emphasise the state estimates in favour of the measurements and this will result in sub-optimal performance. (Two conundrums remain: (i) If one can estimate the local error, there is no reason not to add this to the computed solution to get an improved solution but then no local error estimate is available. (ii) Local errors accumulate to the global error making it hard to quantify the latter even with smoothness and stability assumptions on the underlying system model and good local error estimates.)

*e-mail: edward.boje@uct.ac.za

The development of the field of probabilistic solutions to differential equations is described in [1] and [2]. The ideas in [1] are developed in [3] to establish stronger convergence results for probabilistic integrators, and as their results apply to general additive noise models, they apply to the the state dependent noise with bias model used to model the local error in this paper. A general survey of probabilistic numerics is provided by [4].

In this paper, in Section 2 the local error for any single step method is modelled as a function of the state derivative multiplied by a bias term and a zero-mean white noise term, both appropriately scaled by the step size. Scaling the error model by the state derivative provides a richer model for the local error than used by [1], captures the fact that the local error depends on the rate of change of the state, and correctly gives zero local error effect in steady state. A stochastic difference equation is formed by combining the single step method with the local error model. To account for non-autonomous behaviour and as a “catch-all,” an unscaled noise term may also be added to the model.

Following [1], in 3 a continuous time extension of the discrete time local error model is developed so that the original continuous time state differential equation can be represented by a combination of the simulation method and a stochastic term. This continuous time stochastic differential equation can be used to propagate the simulation error via Monte Carlo analysis or by statistical moments; can be used for step size control; and has specific application to incorporating local simulation errors in continuous-discrete state estimation.

Section 4 presents two examples to illustrate the utility of the approach presented in the paper.

2 Stochastic, state dependent local error model for single step methods

Consider the autonomous n^{th} order dynamic system,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0. \quad (1)$$

Non-autonomous systems can be dealt with by appending an additional state variable, $x_{n+1}(t) = t$, with $f_{n+1} = 1$, which makes the extended system autonomous ([5], Chapter 4). Apply any single-step method with p^{th} order global error, with step size h_i per time step, i ,

$$\mathbf{x}_{i+1}^{\psi} = \mathbf{x}_i + h_i \psi(\mathbf{x}_i, h_i). \quad (2)$$

If (1) is Lipchitz along the state trajectory, to achieve a global error $\mathcal{O}(h^p)$, the local error must be $\mathcal{O}(h^{p+1})$. The local error (with index i to denote local error in making the step from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$) is,

$$\begin{aligned} \mathbf{e}_i &= \mathbf{x}(t_{i+1}) - \mathbf{x}_{i+1}^{\psi} \big|_{\mathbf{x}_i = \mathbf{x}(t_i)} \\ &= h^{p+1} \left(\mathbf{D}^{(p)}(\mathbf{x}_i) + \mathcal{O}_e(h) \right) \mathbf{f}(\mathbf{x}_i) \\ &= \mathcal{O}(h^{p+1}) \end{aligned} \quad (3)$$

where $\mathbf{D}^{(p)}(\mathbf{x}_i)\mathbf{f}(\mathbf{x}_i)$ represents the accumulation of the differences between $h^{(p+1)^{\text{th}}}$ order elementary differentials ([5], [6]) of the Taylor series expansions of the true solution and the method with a common term, $\mathbf{f}(\mathbf{x}_i)$, factored out. Because there may be product terms, e.g. $f_a(\mathbf{x}_i)f_b(\mathbf{x}_i)$, in the expansion, there may be ambiguity in elements of matrix $\mathbf{D}^{(p)}(\mathbf{x}_i)$ but not in the vector, $\mathbf{D}^{(p)}(\mathbf{x}_i)\mathbf{f}(\mathbf{x}_i)$. For a p^{th} order method, terms in the local error up to and including h^p cancel.

When applied to the linear time invariant (LTI) system, $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$, (3) gives the local error as $\mathbf{e}_i = h^{p+1}(d_p \mathbf{A}^p + \mathcal{O}_e(h))\mathbf{A}\mathbf{x}_i$, where d_p is a coefficient. ($d_p = 1/(p+1)!$ for explicit methods with $p \leq 4$ so that the stage number equals the order, [6].)

Although the local error is deterministic, based on the above observations and developing on previous work in this area ([1], [2],[3]), we propose to model the local error for each row, $r \in \{1 \dots n\}$, of the state differential equation as a combination of a deterministic term and a stochastic term,

$$\mathbf{e}_{r,i} \approx h^{p+1} \left(\mathbf{f}^T(\mathbf{x}_i) \mathbf{m}_r + \left[\mathbf{f}^T(\mathbf{x}_i), 1 \right] \xi_{r,i} \right) \quad (4)$$

where $\mathbf{m}_r$ is a (generally) non-zero mean vector that captures the average or deterministic component of the local error and $\xi_{r,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_r)$, a $(n+1) \times 1$, zero-mean white Gaussian random vector with constant covariance $\mathbf{Q}_r$ to capture the residual effect of the local error. (Equation (4) can be represented using a single, equivalent term, $\xi_{r,i} \sim \mathcal{N}(\mathbf{m}_r, \mathbf{Q}_r)$.) For LTI systems with p^{th} order method and a fixed step size, the local error multiplier, $\mathbf{m}_r(h)$ is constant (depending on h), as the local error is, $\mathbf{e}_i = h^{p+1} \mathbf{M}(h) \mathbf{A} \mathbf{x}_i$ with $\mathbf{M}(h) = \sum_{k=p+1}^{\infty} \frac{1}{k!} \mathbf{A}^{k-1} h^{k-(p+1)}$. For general non-linear systems or varying step size, $\mathbf{D}^{(p)}(\mathbf{x}_i) + \mathcal{O}(h)$ is not constant but may still have a non-zero mean over the operating region of the state space and range of step sizes. In (4), augmenting the state derivative vector with $f_{n+1} = 1$ takes care of time dependence for non-autonomous systems and provides a “catch-all” for errors that are not modelled via the state dependent terms. (This is the only term used in [1].) We do not recommend adding a constant offset to the deterministic part as this will change the steady state behaviour by adding bias to the equilibrium point.

The following approach is proposed for calibrating of the local error model.

1. Select a representative set of points over the typical operating state of the system, $\mathbf{x}_{[k]} \in \{\mathbf{X}\}$, $k = 1 \dots N$. Notice that the sample points selected do not need to form a solution to the differential equation, they only need to cover the points to be traversed in the state space adequately.
2. At each point, apply the intended numerical method to make a single step, $\mathbf{x}_{i+1,[k]}^{\psi} = \mathbf{x}_{i,[k]} + h_{i,[k]} \psi(\mathbf{x}_{i,[k]}, h_{i,[k]})$. Also record $\mathbf{f}_{i,[k]} = \mathbf{f}(\mathbf{x}_{i,[k]})$. Calculating the local error over a range of plausible step sizes, or making use of an automatic step size control algorithm to capture the step sizes appropriate to various points in the state space ensures that the local errors are properly represented in the data.
3. Use a higher order method or sub-steps (possibly with Richardson extrapolation) to get an accurate estimate of the state at the next time, $\mathbf{x}_{i+1,[k]} \approx \mathbf{x}_{[k]}(t_i + h)$, and hence estimate the local error, $\mathbf{e}_{i,[k]} = \mathbf{x}_{i+1,[k]} - \mathbf{x}_{i+1,[k]}^{\psi}$. This step is only performed with high precision during calibration but typically there would be some form of (cheap) step size control to limit the worst-case local error during normal operation.
4. Use linear least squares (or any other method) to find elements of a fixed $\mathbf{M}$ (with rows, $\mathbf{m}_r^T$) to minimise the mean square residual error, $\gamma_{r,i,[k]} = \mathbf{f}_{i,[k]}^T \mathbf{m}_r - \mathbf{e}_{r,i,[k]} / h_{i,[k]}^{(p+1)}$ over k . ($\mathbf{M}$ can be solved in a single least-squares step.) As $\mathbf{M} \mathbf{f}(\mathbf{x}_i) h^{p+1}$ is scaled by h^{p+1} it does not affect the asymptotic error and elements of $\mathbf{M}$ may be forced to zero, if statistical analysis of variance supports the hypothesis that individual elements are zero within acceptable confidence levels, or if modelling of the problem indicates that this choice is appropriate. The justification for weighting the local error with $1/h^{p+1}$ is to follow the asymptotic error behaviour and will penalise large errors incurred with smaller step sizes. Different weighting could be used to emphasise areas of the state space that require tighter modelling of the error.
5. Using the white noise model of the residual errors, $\xi_{r,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_r)$ to give $\vartheta_{r,i} = \left[\mathbf{f}^T(\mathbf{x}_i), 1 \right] \xi_{r,i}$, find $\mathbf{Q}_r$ to minimise $\|\gamma_{r,i}^2 - \mathcal{E}\{\vartheta_{r,i}^2\}\|_s$ over k for some norm, s . This problem can be solved in two steps: (i) For diagonal $\mathbf{Q}_r = \text{diag}\{q_{jj}\}$, $\mathcal{E}\{\vartheta_{r,i}^2\} = \sum_{j=1}^{n+1} q_{jj} f_j^2$

with $f_{n+1} = 1$) is linear in unknown q_{jj} , and can be found using a non-negative least squares solver (e.g. Matlab's `lsqnonneg`). (ii) If required, and having regard to the risk of over-fitting, find (e.g. upper) triangular Cholesky factor $\mathbf{R}$, ($\mathbf{R}^T \mathbf{R} = \mathbf{Q}_r$) to solve the optimisation problem. The diagonal $\mathbf{R} = \text{diag}\{\sqrt{q_{jj}}\}$ found above provide an appropriate starting value. The choice of metric, s (e.g. one- or two-norm) for the fit between $\gamma_{r,i}^2$ and $\vartheta_{r,i}^2$ will affect the result of the optimisation.

2.1 Variations on the theme

Step 5 can be interpreted/implemented in other ways:

1. It may be preferred to assume diagonal, $\mathbf{Q}_r$, both for simplicity and to avoid the risk of over-fitting. This assumes that the residual error covariance depends only on f_a^2 terms and not on product terms, $f_a f_b, a \neq b$ which may miss couplings that are physically relevant.
2. The residual error can be viewed/modelled per row as $\vartheta_{r,i} = \left[\mathbf{f}^T(\mathbf{x}_i), 1 \right] \mathbf{w}_r \xi_{r,i}$, with $\xi_{r,i} \sim \mathcal{N}(0, 1)$ a scalar white noise term with unit variance, and $\mathbf{w}_r$ a weighting vector, found to minimise $\|\gamma_{r,i}^2 - \mathcal{E}\{\vartheta_{r,i}^2\}\|_p$. The non-negative least squares solution discussed above for diagonal $\mathbf{Q}_r$ does not apply as $\mathcal{E}\{\vartheta_{r,i}^2\} = \left(\sum_{j=1}^{n+1} m_{r,j} f_{ij} \right)^2 \neq \sum_{j=1}^{n+1} r_{jj}^2 f_{ij}^2$.
3. As noted in Section 3 below, if the residual error vector over the entire simulation is modelled as, $\boldsymbol{\vartheta}_i = \mathbf{W} \begin{bmatrix} \mathbf{f}(\mathbf{x}_i) \\ 1 \end{bmatrix} \xi_i$ with $\xi_i \sim \mathcal{N}(0, 1)$ a scalar stochastic element, the equivalent, continuous time stochastic differential equation can be simulated using the strong first order Milstein method for the special case of scalar noise. This requires finding the $(n^2 + n)$ elements of the matrix $\mathbf{W}$ to minimise the matrix norm $\|\gamma_i \gamma_i^T - (\mathbf{W} \mathbf{f}_i)(\mathbf{W} \mathbf{f}_i)^T\|_s$ over k . The row solutions in variation (2) above (corresponding to the case that the rows of γ_i are uncorrelated) might be used as a starting value for numerical optimisation.
4. We have tried an alternative approach of constructing (term by term), a least-length pseudo noise term from the residual errors via, $\zeta_{r,i,[k]}^T = \gamma_{r,i,[k]} \begin{bmatrix} \mathbf{f}(\mathbf{x}_i) \\ 1 \end{bmatrix}^\dagger$ and then finding the covariance, $\text{cov}(\mathbf{Z}_r)$, with $\mathbf{Z}_r = [\zeta_{r,i,[1]}, \dots, \zeta_{r,i,[N]}]^T$. (Recall that the pseudo inverse of a vector, $\mathbf{z}$, is very simple, $\mathbf{z}^\dagger = \mathbf{z}^T / \|\mathbf{z}\|_2^2$). We have found that this approach is optimistic: While (by construction) the pseudo-noise generates the residual error, because of the way it allocates the residual error to individual pseudo-noise terms the scheme typically gives too low a variance and since $\mathcal{E}\{\xi_{r,i,[k]} \xi_{r,i,[k]}^T\} = \begin{bmatrix} \mathbf{f}_{i,[k]} & \mathbf{f}_{i,[k]}^\dagger \end{bmatrix}^T \mathbf{Q}_r \begin{bmatrix} \mathbf{f}_{i,[k]} & \mathbf{f}_{i,[k]}^\dagger \end{bmatrix}$, the resulting covariance cannot be expected to be constant. Finally, the dynamic range of $\mathbf{f}(\mathbf{x})$ depends on choice of units so weighting may be required.

The algorithm could be run on-line if computing resources allowed the local errors to be calculated accurately on the fly. Although this use-case seems to defeat the purpose of the exercise, the model parameters, $\mathbf{m}_r$, and $\mathbf{Q}_r$, could be adapted as new information becomes available while traversing of the operating region of the state space.

3 Continuous time extension of stochastic, local error model

There are applications where a continuous time equivalent of the local error model will be useful. These include continuous-discrete state estimation incorporating simulation error effects, cheap local and global error estimation, and step-size control. This section develops the continuous time extension of the results above by representing the continuous time system via a stochastic differential equation. [1] modelled the local error as a stochastic term added to the differential equation and calibrated this as a zero mean, iid Gaussian sequence, with a covariance scaled by h^{2p+1} . In this section, this idea is developed and applied to the more complicated local error model in (4).

If $\chi(t)$ is a zero mean, white Gaussian noise with covariance, $\mathbf{Q}_\chi(t)\delta(t)$, ($\delta(t)$ is the Dirac delta function) and $\mathbf{Q}_\chi(t)$ constant over the interval, $t \in (0, \tau]$. Then, the Wiener process, $\xi(\tau) = \int_0^\tau \chi(\lambda)d\lambda$, is zero mean with covariance, $\mathbf{Q}_\xi = \tau\mathbf{Q}_\chi$.

To develop a continuous time extension of the discrete time result in Section 2, the Taylor series expansions of the true solution and solution of the simulation method over the interval, $t = t_i + \tau$, $\tau \in (0, h]$, are written as follows. True solution:

$$\begin{aligned} \mathbf{x}(t_i + \tau) &= \mathbf{x}_i + \tau \mathbf{f}(\mathbf{x}_i) + \frac{\tau^2}{2} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \bigg|_{\mathbf{x}_i} \mathbf{f}(\mathbf{x}_i) + \frac{\tau^3}{3!} \mathbf{D}_{\mathbf{x}_i}^{(2)} \mathbf{f}(\mathbf{x}_i) + \dots \\ &= \mathbf{x}_i + \int_{t_i}^{t_i+\tau} \mathbf{f}_i(\lambda) d\lambda \end{aligned} \quad (5)$$

with all quantities evaluated at $(\mathbf{x}_i, t_i)$, e.g. $\mathbf{f}_i = \mathbf{f}(\mathbf{x}_i, t_i)$ and,

$$\mathbf{f}_i(\tau) = \mathbf{f}(\mathbf{x}_i) + \tau \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \bigg|_{\mathbf{x}_i} \mathbf{f}(\mathbf{x}_i) + \frac{\tau^2}{2} \mathbf{D}_{\mathbf{x}_i}^{(2)} \mathbf{f}(\mathbf{x}_i) + \dots \quad (6)$$

Simulation method: The single step method, (2), is written out in the interval, giving a continuous time function, $\mathbf{y}(t)$,

$$\begin{aligned} \mathbf{y}(t_i + \tau) &= \mathbf{y}_i + \tau \boldsymbol{\psi}(\mathbf{y}_i, \tau) \\ &= \mathbf{y}_i + \tau \boldsymbol{\psi}(\mathbf{y}_i, 0) + \tau^2 \frac{d\boldsymbol{\psi}(\mathbf{y}_i, \tau)}{d\tau} \bigg|_{\tau=0} + \frac{\tau^3}{2!} \frac{d^2 \boldsymbol{\psi}(\mathbf{y}_i, \tau)}{d\tau^2} \bigg|_{\tau=0} + \dots \\ &= \mathbf{y}_i + \int_{t_i}^{t_i+\tau} \mathbf{g}_i(\lambda) d\lambda \end{aligned} \quad (7)$$

$\mathbf{g}_i(\tau) = \boldsymbol{\psi}(\mathbf{y}_i, \tau) + \tau \frac{d\boldsymbol{\psi}(\mathbf{y}_i, \tau)}{d\tau}$ can be written as the Taylor series expansion of the time derivative of $\tau \boldsymbol{\psi}(\mathbf{y}_i, \tau)$ in the interval. If a deterministic correction is used in the simulation, the corresponding continuous time extension would be $\mathbf{y}(t_i + \tau) = \mathbf{y}_i + \tau \boldsymbol{\psi}(\mathbf{y}_i, \tau) + \tau^{p+1} \mathbf{M} \mathbf{f}(\mathbf{x}_i)$.

The continuous time extension of the local error is $\mathbf{e}(t) = \mathbf{x}(t) - \mathbf{y}(t)$,

$$\begin{aligned} \mathbf{e}(t_i + \tau) &= \mathbf{x}(t_i + \tau) - \mathbf{y}(t_i + \tau) |_{\mathbf{y}_i = \mathbf{x}_i} \\ &= \frac{\tau^{p+1}}{(p+1)!} (\mathbf{D}_{\mathbf{x}_i}^{(p)} - \mathbf{D}_{\mathbf{y}_i}^{(p)} + \mathbf{O}_e(\tau)) \mathbf{f}(\mathbf{x}_i) \\ &= \int_{t_i}^{t_i+\tau} \boldsymbol{\ell}_i(\lambda) d\lambda. \end{aligned} \quad (8)$$

where $\boldsymbol{\ell}_i(s)$ is the time derivative of the local error's continuous time expansion,

$$\boldsymbol{\ell}(\tau) = \frac{\tau^p}{p!} (\mathbf{D}_{\mathbf{e}_i}^{(p)} + \mathbf{O}_\ell(\tau)) \mathbf{f}(\mathbf{x}_i). \quad (9)$$

Table 1. Continuous time extension for Heun rule

Heun	$x_{i+1} = x_i + h/2 (f(x_i) + f(x_i + hf(x_i)))$
Continuous time extension	$g(t_i + \tau) = f + \tau \frac{\partial f}{\partial x} f + \frac{3\tau^2}{4} \left(\frac{\partial^2 f}{\partial x^2} f^2 \right) + \dots$
Local error	$e(t_i + \tau) = \tau^3 \left(-\frac{1}{12} \frac{\partial^2 f}{\partial x^2} f + \frac{1}{6} \left(\frac{\partial f}{\partial x} \right)^2 + O_e(\tau) \right) f$
Local error derivative	$\ell(t_i + \tau) = \tau^2 \left(-\frac{1}{4} \frac{\partial^2 f}{\partial x^2} f + \frac{1}{2} \left(\frac{\partial f}{\partial x} \right)^2 + O_\ell(\tau) \right) f$
LTI local error	$e(t_i + \tau) = \tau^3 \left(\frac{1}{6} A^2 + \frac{\tau}{4!} A^3 + O(\tau^2) \right) A x_i$
LTI local error derivative	$\ell(t_i + \tau) = \tau^2 \left(\frac{1}{2} A^2 + \frac{\tau}{3!} A^3 + O(\tau^2) \right) A x_i$

Table 2. Continuous time extension for linearly implicit Taylor-Heun

Heun	$x_{i+1} = x_i + h \left(1 - \frac{h}{2} \frac{\partial f}{\partial x} \Big _{x_i} \right)^{-1} f(x_i)$
Continuous time extension	$g(t_i + \tau) = f + \tau \frac{\partial f}{\partial x} f + \frac{3\tau^2}{4} \left(\frac{\partial f}{\partial x} \right)^2 f + \dots$
Local error	$e(t_i + \tau) = \tau^3 \left(\frac{1}{6} \frac{\partial^2 f}{\partial x^2} f - \frac{1}{12} \left(\frac{\partial f}{\partial x} \right)^2 + O_e(\tau) \right) f$
Local error derivative	$\ell(t_i + \tau) = \tau^2 \left(\frac{1}{2} \frac{\partial^2 f}{\partial x^2} f - \frac{1}{4} \left(\frac{\partial f}{\partial x} \right)^2 + O_\ell(\tau) \right) f$
LTI local error	$e(t_i + \tau) = \tau^3 \left(-\frac{1}{12} A^2 - \frac{\tau}{12} A^3 + O(\tau^2) \right) A x_i$
LTI local error derivative	$\ell(t_i + \tau) = \tau^2 \left(-\frac{1}{4} A^2 - \frac{\tau}{3} A^3 + O(\tau^2) \right) A x_i$

The above developments mean that, $\dot{x}(t) = f(t) = g_i(t) + \ell_i(t)$ in the interval $t = t_i + \tau$, $\tau \in (0, h]$. Both $e(t_i + \tau)$ and $\ell(t_i + \tau)$ depend on a principal term, $D_{e_i}^{(p)} = D_{x_i}^{(p)} - D_{y_i}^{(p)}$, scaled by $\frac{\tau^{p+1}}{(p+1)!} f(x_i)$ and $\frac{\tau^p}{p!} f(x_i)$ respectively, and on higher order terms (also scaled by $f(x_i)$ but dependent on the step size). The higher order expansions in the local error, $O_e(\tau)$, and local error derivative, $O_\ell(\tau)$, do not have the same coefficients, for example, an LTI system with an explicit method (truncated Taylor series) has, $\ell(t_i + \tau) = \left(\frac{\tau^p}{p!} A^p + O_\ell(\tau) \right) A x_i$.

As examples of the above discussion, the Heun (modified Euler) and linearly implicit Taylor-Heun schemes applied to the scalar, autonomous system, $\dot{x} = f(x)$, $x(t_0) = x_0$, with correct $x_i = x(t_i)$; and to the linear time invariant (LTI) system $\dot{x} = A x$ are summarised in Tables 1 and 2 respectively. For the LTI examples, the next higher order term in the error expansion is shown to illustrate that $O_e(\tau)$ and $O_\ell(\tau)$ terms are not the same.

The above calculations are possible because $\ell_i(t)$ is deterministic. If however, (after removing the deterministic component), $e_{r,i} = e_r(t_i + h)$ is modelled via the stochastic term, $h^{p+1} \vartheta_{r,i} = h^{p+1} [f_i^T, 1] \xi_{r,i}$, the stochastic version of $\ell_r(t_i + \tau) = [f_i^T, 1] \chi_r$ with $\chi_r \sim \mathcal{N}(0, h_i^{2p+1} Q_r)$ so that $\text{cov} \left(\int_0^{h_i} \chi(\tau) d\tau \right) = h_i^{2(p+1)} Q_r = \text{cov} (h_i^{p+1} \xi_{r,i})$ as required. Although this arrangement ensures that the local error has the required covariance at the end of each step, it does not model the error correctly during the step because of the difference between deterministic and stochastic integrals. For the deterministic version, the principal component of the local error is constant during the evolution of the continuous time extension, while for the stochastic version, the principal error effect (standard deviation) grows with $\sqrt{\tau}$. As most studies will end up with numerical simulation of each step, the under-estimation of the local error covariance during the initial part of the continuous time extension will not matter.

Defining the stochastic increment per row as, $d\beta_r = \chi_r dt$, the simulation of $\dot{x}_r(t) = f_r(x)$, can be approximated and studied via the stochastic differential equation,

$$dx_r(t) = g_{r,i}(t)dt + \left[f^T(x_i), 1 \right] d\beta_r \quad (10)$$

If (10) is simulated using the strong order 0.5 Euler-Maruyama approach ([7]), the original, generally higher order, deterministic method selected for the problem is retained as $\int_{t_i}^{t_i+h_i} g_{r,i}(\tau) d\tau = h_i \psi_{r,i}(x_i)$ from (7) and the stochastic terms are propagated with low order. This scheme is justified by noting that the noise term χ_r is pre-scaled by $h_i^{p+0.5}$ before integration. [3] emphasises the difference between solvers for stochastic differential equations and the present problem of probabilistic numerics where the variance of the local error model decays faster than h . No general claims about the resulting method order are made here and the cautions raised in [8] should be noted. The algorithm would be,

1. Draw independent random vectors, $\Delta\beta_{r,i}$ from the distributions, $\Delta\beta_{r,i} \sim \mathcal{N}(\mathbf{0}, h_i^{2(p+1)} Q_r)$.
2. Propagate the state vector via,

$$x_{r,i+1} = x_{r,i} + h_i \psi_r(x_i) + h_i^{p+1} f_i^T m_r + \left[f^T(x_i), 1 \right] \Delta\beta_{r,i} \quad (11)$$

If the local error is modelled with a scalar noise term that scales all rows of the SDE, $d\mathbf{x}(t) = \mathbf{g}_i(t)dt + \mathbf{W} \begin{bmatrix} f(x_i), \\ 1 \end{bmatrix} d\beta$, the state can be propagated using a modification of the first order Milstein method ([7]), giving,

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h_i \psi(\mathbf{x}_i) + \mathbf{W} \left(\begin{bmatrix} f(\mathbf{x}_i), \\ 1 \end{bmatrix} \Delta\beta_i + \frac{1}{2} \left[\frac{\partial f}{\partial \mathbf{x}} \right] \mathbf{W} \begin{bmatrix} f(\mathbf{x}_i), \\ 1 \end{bmatrix} (\Delta\beta_i^2 - q h_i) \right) \quad (12)$$

Notice that the Jacobian required in (12) for the stochastic component is already available for linearly implicit methods.

3.1 Application of results to numerical simulation

The results developed above may be applied to numerical simulation in several ways:

3.1.1 Simulation of an ensemble of solutions:

For each solution the underlying numerical method is modified and simulated via (11).

3.1.2 Simulation of mean and covariance of solutions:

The solution of the mean, $\bar{\mathbf{x}}(t) = \mathcal{E}\{\mathbf{x}(t)\}$, and covariance, $\mathbf{P}(t) = \mathcal{E}\{(\mathbf{x}(t) - \bar{\mathbf{x}}(t))(\mathbf{x}(t) - \bar{\mathbf{x}}(t))^T\}$ of the stochastic difference equation can be obtained using standard techniques of either linearizing the covariance equation, using unscented transforms or Monte Carlo particle approaches. The covariance solution provides an estimate of the uncertainty range of the global error. Ignoring high order terms the mean is,

$$\frac{d}{dt} \bar{\mathbf{x}} = \mathcal{E}\{f(\mathbf{x})\} \approx f(\bar{\mathbf{x}}) \quad (13)$$

giving,

$$\bar{\mathbf{x}}_{i+1} \approx \bar{\mathbf{x}}_i + h\boldsymbol{\psi}(\bar{\mathbf{x}}_i, h) + h^{p+1}\mathbf{M}\mathbf{f}(\bar{\mathbf{x}}_i) \quad (14)$$

The covariance is,

$$\begin{aligned} \frac{d}{dt}\mathbf{P} &= \mathcal{E}\{(\mathbf{x} - \hat{\mathbf{x}})\mathbf{f}^T(\mathbf{x}) + \mathbf{f}(\mathbf{x})(\mathbf{x} - \hat{\mathbf{x}})^T + \mathbf{L}(\mathbf{x})\mathbf{Q}\mathbf{L}^T(\mathbf{x})\} \\ &\approx \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}} \mathbf{P} + \mathbf{P} \left(\left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}} \right)^T + \mathbf{L}(\bar{\mathbf{x}})\mathbf{Q}\mathbf{L}^T(\bar{\mathbf{x}}) \end{aligned} \quad (15)$$

with $\mathbf{L}(\bar{\mathbf{x}}) = \text{diag}\{\{\mathbf{f}^T(\mathbf{x}_i), 1\}\}$ and $\mathbf{Q} = \text{diag}\{\mathbf{Q}_r\}$. The approximations in (13) and (15) assume sufficiently smooth behaviour of $\mathbf{f}(\mathbf{x})$ around the mean value - see for instance [9], Section 4.9 for second order expansions of the mean and covariance, or [7]. To discretize (15) with a simple (recall that the noise covariance is already $\mathcal{O}(h^{2p+1})$), A-stable method that preserves positive semi-definite behaviour of $\mathbf{P}$, the following Padé type approximation is recommended ([10]),

$$\mathbf{P}_{i+1} \approx \mathbf{M}_i \mathbf{P}_i \mathbf{M}_i^T + \mathbf{K}_i \mathbf{L}_i \mathbf{Q}_i \mathbf{L}_i^T \mathbf{K}_i^T \quad (16)$$

with $\mathbf{L}_i = \mathbf{L}(\bar{\mathbf{x}}_i)$, $\mathbf{K}_i = \left(\mathbf{I} - \frac{h}{2}\mathbf{A}_i\right)^{-1}$, $\mathbf{M}_i = \mathbf{K}_i \left(\mathbf{I} + \frac{h}{2}\mathbf{A}_i\right)$, and $\mathbf{A}_i = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}$. ([10] uses a mid-time-step calculation to achieve second order accuracy. Also see [11] for further discussion.)

3.1.3 Step size control based on stochastic model of local error:

If the local error, (4), is well-characterised by the noise term over the operating state of the system, this can be used for cheap local error estimation and control as $\mathbf{f}(\mathbf{x}_i)$ is calculated at the start of every step. For any step, the stochastic model of the local error for row r will have standard deviation, $\sigma_{r,i} = h^{p+1} \left\| \left[\mathbf{f}^T(\mathbf{x}_i), 1 \right] \mathbf{R}_r^T \right\|_2$.

If the error standard deviation is required to be less than $\hat{\sigma}_r = g_r h$, (the additional h to take care of the step size effect on the global error accumulation), the step size should be

$$h \leq \sqrt[p]{g_r / \left\| \left[\mathbf{f}^T(\mathbf{x}_i), 1 \right] \mathbf{R}_r^T \right\|_2} \quad (17)$$

to ensure that the step size is small whenever there is a high rate of change in a direction that causes high local error sensitivity. This idea is illustrated in Example 2.

4 Examples

Both examples in this section use inappropriately large step size to give visual illustrations of the results in the paper. Example 1 uses the first order Euler method and Example 2 uses the second order Heun (modified Euler) method.

4.1 Example 1: Explicit Euler Method

The explicit Euler method, $x_{i+1} = x_i + hf(x_i)$ is applied to the FitzHugh-Nagumo oscillator used as an example in [1]: $\frac{dx_1}{dt} = c(x_1 - x_1^3/3 + x_2)$, $\frac{dx_2}{dt} = -\frac{1}{c}(x_1 - a + bx_2)$ with $(a, b, c) = (0.2, 0.2, 3.0)$ and $x_0 = (-1, 1)^T$.

The principal term in the local error can be calculated readily as, $e(t_i + \tau) = \tau^2 \left(\frac{1}{2} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} + \mathcal{O}(\tau) \right) \mathbf{f}$ with $\frac{1}{2} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{pmatrix} 1.5(1 - x_1^2) & 1.5 \\ -0.167 & -0.033 \end{pmatrix}$. The approach outlined above is applied for $h = 0.01$ and $h = 0.1$ over the time interval $t \in (0, 20)$. Two cases of region of

Table 3. FitzHugh-Nagumo Oscillator

$h = 0.01$: On trajectory
$\mathbf{M} = \begin{pmatrix} 0.0321 \pm 0.0319 & 1.5705 \pm 0.0974 \\ -0.1667 \pm 0.0000 & -0.0351 \pm 0.0001 \end{pmatrix}$ $(\sigma_{\gamma_1} \quad \sigma_{\gamma_2}) = (2.0198 \quad 0.0022)$ $\mathbf{R}_1 = \text{diag} \left\{ \begin{bmatrix} 1.1096 & 2.9702 & 0.0000 \end{bmatrix}^T \right\}$ $\mathbf{R}_2 = \text{diag} \left\{ \begin{bmatrix} 0.0012 & 0.0033 & 0.0000 \end{bmatrix}^T \right\}$
$h = 0.01$: In disc
$\mathbf{M} = \begin{pmatrix} -0.1659 \pm 0.0846 & -0.2650 \pm 0.5369 \\ -0.1664 \pm 0.0001 & -0.0330 \pm 0.0006 \end{pmatrix}$ $(\sigma_{\gamma_1} \quad \sigma_{\gamma_2}) = (4.0711 \quad 0.0045)$ $\mathbf{R}_1 = \text{diag} \left\{ \begin{bmatrix} 0.9423 & 7.7670 & 0.0000 \end{bmatrix}^T \right\}$ $\mathbf{R}_2 = \text{diag} \left\{ \begin{bmatrix} 0.0010 & 0.0087 & 0.0000 \end{bmatrix}^T \right\}$
$h = 0.1$: On trajectory
$\mathbf{M} = \begin{pmatrix} -0.1021 \pm 0.0997 & 1.5282 \pm 0.3049 \\ -0.1656 \pm 0.0011 & -0.0504 \pm 0.0034 \end{pmatrix}$ $(\sigma_{\gamma_1} \quad \sigma_{\gamma_2}) = (1.9966 \quad 0.0223)$ $\mathbf{R}_1 = \text{diag} \left\{ \begin{bmatrix} 1.1463 & 2.4561 & 0.0000 \end{bmatrix}^T \right\}$ $\mathbf{R}_2 = \text{diag} \left\{ \begin{bmatrix} 0.0127 & 0.0289 & 0.0000 \end{bmatrix}^T \right\}$
$h = 0.1$: In disc
$\mathbf{M} = \begin{pmatrix} -0.2400 \pm 0.0763 & 0.9059 \pm 0.4839 \\ -0.1639 \pm 0.0009 & -0.0403 \pm 0.0055 \end{pmatrix}$ $(\sigma_{\gamma_1} \quad \sigma_{\gamma_2}) = (3.6691 \quad 0.0418)$ $\mathbf{R}_1 = \text{diag} \left\{ \begin{bmatrix} 0.9272 & 5.3044 & 1.4480 \end{bmatrix}^T \right\}$ $\mathbf{R}_2 = \text{diag} \left\{ \begin{bmatrix} 0.0105 & 0.0681 & 0.0098 \end{bmatrix}^T \right\}$

interest are considered: (i) points along the trajectory from a simulation and (ii) an ellipsoidal disc shaped region covering the anticipated trajectory, $0.65^2 \leq \left(\frac{x_1}{2.1}\right)^2 + \left(\frac{x_2}{1.3}\right)^2 \leq 1.0^2$. The results are tabulated in Table 3: For each linear least squares fit of the deterministic part of the local error term, $\mathbf{M}$, the square root of the diagonal elements of the covariance matrices of the solutions are shown to give an indication of the confidence regions. Also tabulated are the residual error standard deviations and the noise terms standard deviations required to solve the optimisation problem. Only diagonal noise terms are used because a full covariance matrix (Step (5.ii) in Section 2) only reduces the per-element residual error norms from (0.827 and 1.113×10^{-4}) to (0.704 and 9.328×10^{-5}) respectively in the (worst) case of disk region and $h = 0.1$ used below. Notice how the elements of $\mathbf{M}$ match $\frac{1}{2} \frac{\partial f}{\partial x}$ for the linear terms in the state space model, especially when the calculations are performed on trajectory points.

Fig. 1 illustrates the solution trajectory obtained with by comparing Matlab's ODE45 code with tight tolerances to the explicit Euler solution with $h = 0.1$, and using the diagonal-

Figure 1. FitzHugh-Nagumo oscillator, $h = 0.1$: Cyan - ensemble of solutions; Blue - accurate solution; Green - low order solution; Red dashed - 1σ bounds;

only results for the disc-shaped region. To illustrate further, an ensemble of solutions to the SDE and the explicit Euler solution with one standard deviation, $y_i \pm \sqrt{\text{diag}(P_i)}$, are shown. Fig. 2 illustrates the development of the global error. Notice that in this simulation, after multiple cycles, the predicted global error is optimistic compared to the actual value.

4.2 Example 2: Heun Method

The Heun method, $x_{i+1} = x_i + h\psi(x_i, h)$ with $\psi(x_i, h) = \frac{1}{2}(f(x_i) + f(x_i + hf(x_i)))$ is applied to the classic two body problem discussed in [5] and elsewhere. The model for an orbit in the x-y plane with the origin as one of the foci, is $\ddot{x} = -x/r^3$, $\ddot{y} = -y/r^3$, with $r^2 = x^2 + y^2$, $(\dot{x}_0, \dot{y}_0, x_0, y_0)^T = (0, \sqrt{\frac{1-e}{1+e}}, 0, 1-e)^T$ and large eccentricity, $e = 0.9$. Heun is a second order explicit method and the purpose of the example is to illustrate the results presented in this paper rather than to advocate Heun as the best numerical method for this problem. (Higher order methods can exploit the smoothness, and implicit methods and step size control be used to deal with the stiffness.)

The value of M obtained is quite consistent over a range of step sizes but it is averaging the local error effect over the whole of an orbit, which has significant velocity differences at different points due to the eccentricity. Even though the simulation method, $x_{i+1} = x_i + h\psi(x_i, h) + h^3 M f(x_i)$ is still globally second order this understanding suggests excluding the offset term (i.e. $M = 0$). As a result, the local error is modelled in its entirety via noise terms. The local error data is obtained by selecting points on the orbit found using Matlab's ODE23 with high tolerance.

To illustrate a simple step size control scheme, at each step, the noise covariance per row is calculated from (17) as $h = \min\left\{\left(\frac{p_{tol}}{\sigma_p}\right)^{1/2}, \left(\frac{p_{tol}}{\sigma_v}\right)^{1/3}\right\}$, with p_{tol} the allowed position tolerance per step and σ_p and σ_v the position and velocity standard deviation increments

Figure 2. Global error magnitudes for FitzHugh-Nagumo oscillator, $h = 0.1$: Cyan - error for ensemble of solutions; Blue - error between accurate and low order solutions; Red dashed - 1σ bounds;

predicted using the noise model and (16). (Practically, the step size selection is also jacketed between minimum and maximum values but these are not active in the simulation results presented below.)

Fig. 3 shows the position along a single orbit with $p_{tol} = 5 \times 10^{-4}$ used for step-size control. The accurate solution is shown along with Heun method and stochastic differential equation solutions are shown every 30 steps. The step size control algorithm varies the step size between 0.0010 and 0.0158 along the trajectory. Because of the way the step size is calculated this is less reactive than a regular step size control algorithm would be - for example, Matlab's ODE23 (albeit a third order method) used step sizes ranging from 0.0003 to 0.071 for the same problem with tighter tolerance. The displayed points are fairly uniformly distributed in space because larger step sizes are used when the body is far from the origin and moving slowly. The solution mostly remains within the predicted one standard deviation ellipses of the accurate solution but becomes hopelessly uncertain within one orbit at this resolution. The important point is that the error ellipses give insight into the loss of reliability of the solution.

5 Conclusions

This paper has introduced the idea of representing local errors in numerical simulation as stochastic variables formed by the product of the state derivative and white noise, with the possibility of including a constant stochastic term to take care of unmodelled effects and time dependence. This richer representation of local error allows a less conservative model of the local error that, for instance, correctly models zero local simulation error when the system is in steady state. A continuous time extension of the results allows numerical simulation to be modelled using the framework of stochastic differential equations with the time derivative of the single step method as the deterministic part and a state dependent noise to represent the

Figure 3. Position for two body problem: Blue - orbit, Red - Heun and uncertainty ellipses, Cyan - SDE solutions, every 30th step

continuous time effect of the local error. The stochastic term's covariance decays to zero with h^{2p+1} . Two low order examples have illustrated various aspects of the proposed approach.

References

- [1] Conrad, P.R., Girolami, M., Särkkä, S. Stuart, A. and Zygalakis, K., "Statistical analysis of differential equations: introducing probability measures on numerical solutions," *Statistics and Computing* 27, 1065–1082, (2017).
- [2] Schober, M., Särkkä, S. and Hennig, P., "A probabilistic model for the numerical solution of initial value problems," *Statistics and Computing* 29, 99–122, (2019).
- [3] Lie, H.C., Stuart, A.M. and Sullivan, T.J., "Strong convergence rates of probabilistic integrators for ordinary differential equations," *Statistics and Computing*, 29, 1265–1283 (2019).
- [4] Oats, C.J., Sullivan, T.J., "A modern retrospective on probabilistic numerics," *Statistics and Computing* 29, 1335–1351 (2019).
- [5] Shampine, L.F., *Numerical Solution of Ordinary Differential Equations* Chapman and Hall, (1994). ISBN 0-412-05151-6
- [6] Butcher, J.C., "Numerical methods for ordinary differential equations in the 20th century," *Journal of Computational and Applied Mathematics* 125, 1-29, (2000).
- [7] Särkkä, S. and Solin, A., *Applied Stochastic Differential Equations*, Cambridge University Press, (2019).
- [8] Burrage, K; Burrage, P.; Higham, D. J.; Kloeden, P. E. and Platen, E., "Comment on 'Numerical methods for stochastic differential equations'," *Physical Review Part E*, 74, 68701, (2006).
- [9] Jazwinski, A. *Stochastic Processes and Filtering Theory* Academic Press, (1970).
- [10] Mazzoni, T., "Computational aspects of continuous-discrete extended Kalman-filtering," *Computational Statistics*, 23, 519–539, (2008).
- [11] Paul Frogerais, P., Bellanger, J. and Senhadji, L., "Various Ways to Compute the Continuous-Discrete Extended Kalman Filter," *IEEE Transactions on Automatic Control*, 57 (4), 1000-1004, (2012).