Mathematisches Forschungsinstitut Oberwolfach

Report No. 16/2011

DOI: 10.4171/OWR/2011/16

Geometric Numerical Integration

Organised by
Ernst Hairer, Geneve
Marlis Hochbruck, Karlsruhe
Arieh Iserles, Cambridge
Christian Lubich, Tübingen

March 20th – March 26th, 2011

Abstract. The subject of this workshop was numerical methods that preserve geometric properties of the flow of an ordinary or partial differential equation. This was complemented by the question as to how structure preservation affects the long-time behaviour of numerical methods.

Mathematics Subject Classification (2000): 65xx.

Introduction by the Organisers

The subject of this workshop was numerical methods that preserve geometric properties of the flow of an ordinary or partial differential equation: symplectic and multisymplectic integrators for Hamiltonian systems, symmetric integrators for reversible systems, methods preserving first integrals and numerical methods on manifolds, including Lie group methods and integrators for constrained mechanical systems, and methods for problems with highly oscillatory solutions. The unifying theme is structure preservation: not just the "how?" but also "why?", "where?" and "what for?".

The motivation for developing structure-preserving algorithms for special classes of problems arises independently in such diverse areas of research as astronomy, molecular dynamics, mechanics, control theory, theoretical physics, electrical and electronic engineering and numerical analysis with important contributions from other areas of both applied and pure mathematics. Moreover, it turns out that the preservation of geometric properties of the flow not only produces an improved qualitative behaviour, but also allows for a significantly more accurate long-time integration than with general-purpose methods.

Geometric numerical integration has been an active and interdisciplinary research area in the last two decades. While the core of the subject is well documented in the monographs

> E. Hairer, Ch. Lubich, G. Wanner, Geometric Numerical Integration. Springer, Berlin, 2nd edition, 2006,
>
> K. Feng, M. Qin, Symplectic Geometric Algorithms for Hamiltonian systems. Springer, Berlin, 2010

and

> B. Leimkuhler, S. Reich, Simulating Hamiltonian Dynamics. Cambridge Univ. Press, 2004,

the area in its wider sense has in recent years undergone substantial and exciting developments:

1. Backward error analysis for infinite-dimensional systems, using modulated Fourier expansions and normal forms. This has, for the first time, brought partial differential equations to within the realm of backward error analysis, thereby helping to understand the long-term behaviour of numerical methods;
2. The understanding of numerical methods using tools from Hopf algebra and graph theory. This has clarified the sort of invariants that B-series-based numerical methods can respect. Thus, for example, such methods cannot conserve volume but they can be designed to respect Hamiltonian energy;
3. Understanding of symplectic structure of exponential integrators and the design of exponential integrators best capable of dealing with highly oscillatory solutions;
4. The design of asymptotic numerical approaches to numerical integration of highly oscillatory ODEs and DAEs, applications in electronic engineering and in control theory;
5. The analysis of multiscale algorithms, not least of the Heterogeneous Multiscale method, with modulated Fourier series;
6. The understanding how symmetries in differential equations and in discretized systems can be exploited by means of group theory and harmonic analysis, to reduce the cost of computations;
7. Geometric integrators for Maxwell and wave equations, applications in astrophysics, plasma physics and nano-optics;
8. Application of geometric integrators to the Schrödinger equation and reduced models in quantum mechanics;
9. Better understanding of geometric integrators in the context of molecular dynamics and macro-molecule modelling.

The intention of this second Oberwolfach workshop on geometric numerical integration (a first one was held in 2006) was to address these recent developments. Geometric numerical integration is by its very nature a multidisciplinary topic and this was well reflected in the workshop.

Workshop: Geometric Numerical Integration

Table of Contents

Abstracts

Hybrid Monte-Carlo sampling on Hilbert spaces and Geometric Integration

J.M. Sanz-Serna
(joint work with A. Beskos, F.J. Pinski, A.M. Stuart)

We address the problem of sampling from a probability distribution π defined on a Hilbert space $\mathcal{H}$. In a typical application one may have to deal with an evolutionary partial differential equation whose initial datum f is unknown and has to be estimated statistically by using solution values u_i obtained experimentally. In a Bayesian approach one assumes a prior distribution π_0 for f in a suitable space $\mathcal{H}$ and then the available measurements u_i determine, via Bayes theorem, the corresponding posterior distribution π. This posterior is analytically intractable and any desired information —such as its expectation or variance— has to be determined through Monte Carlo techniques: hence the need to sample from π. The study of conditioned diffusion also requires sampling from distributions defined on a Hilbert space.

We assume that (as it is the case in the applications we have just mentioned) the target distribution π is defined by prescribing its negative log-density Φ with respect to a reference Gaussian measure π_0:

$$\frac{d\pi}{d\pi_0}(q) \propto \exp\bigl(-\Phi(q)\bigr) \ .$$

In turn π_0 is specified by its mean and its covariance operator $\mathcal{C}$.

For probability distributions defined on $\mathbb{R}^N$, the hybrid Monte Carlo method (HMC), introduced in [3], provides a useful technique because it generates samples $f^{(n)}$ that do not lie in the neighbourhood of $f^{(n-1)}$: in this way the correlations in the chain are small and the algorithm has the possibility of exploring the whole state space with a moderate computational effort. HMC makes proposals by integrating with the Verlet method a Hamiltonian systems of differential equations; a Metropolis-Hastings accept/reject rule is employed to ensure that the generated chain is reversible with respect to π. The choice of Verlet as an integrator is dictated by its good geometric properties: it is essential that the time-stepper be time-reversible and volume-preserving.

Since standard HMC handles (finite-dimensional) distributions through their densities with respect to the standard Lebesgue measure in $\mathbb{R}^N$ and those densities have no counterpart in $\mathcal{H}$, it is clear that HMC cannot be *directly* applied to the task of sampling from the target π on $\mathcal{H}$. An *indirect* use would consist of first discretizing $\mathcal{H}$ and π and then applying HMC to the resulting finite-dimensional problem. (For instance in the partial differential equation case mentioned above one may represent the real function f by its values at N points in a spatial grid and replace $\mathcal{H}$ by $\mathbb{R}^N$.) However such indirect approach is unsatisfactory: the

performance of HMC is known [1] to degrade with increasing N and the time-step Δt in the Verlet algorithm has to be chosen smaller and smaller as $N \uparrow \infty$ if the acceptance probability is to remain bounded away from zero. There is no need to say that with Δt very small the computational cost becomes prohibitive. The situation is therefore akin to that found in the time-integration of, say, the heat equation with an explicit method, say Euler's rule. The method does not make sense if the heat equation is seen as an evolution equation on a Hilbert space $\mathcal{H}$ (the relevant infinitesimal generator —the Laplacian— is unbounded). One may first discretize in space to replace the unbounded operator by a large $N \times N$ differentiation matrix and then apply Euler to integrate in time the resulting system of ordinary differential equations. However for N large the matrices are necessarily ill-conditioned (they approximate an unbounded operator) and a very small value of Δt is required to ensure stability.

The preceding considerations show the advisability of introducing variants of HMC that may be applied on $\mathcal{H}$ (or —if one prefers— that may be applied on discretized state spaces with a probability of acceptance independent of the discretization parameter N). One algorithm that fulfils those requirements has recently been introduced in [2] and may be written purely in terms of the elements Φ and π_0 that define π without invoking any finite-dimensional approximations. Reference [2] shows (i) how to define a suitable infinite-dimensional Hamiltonian dynamics to guide the proposal mechanism, (ii) how to numerically integrate the Hamiltonian problem and (iii) how to formulate a suitable accept/reject rule.

For simplicity [2] only considers a simple time-integrator *à la Strang*. There is much scope for improvements by means of more sophisticated split-step methods and any help in that connection would be most welcome.

References

[1] A. Beskos, N.S. Pillai, G.O. Roberts, J.M. Sanz-Serna, A.M. Stuart, *Optimal tuning of Hybrid Monte-Carlo*, submitted.

[2] A. Beskos, F.J. Pinski, J.M. Sanz-Serna, A.M. Stuart, *Hybrid Monte-Carlo on Hilbert spaces*, submitted.

[3] S. Duane, A.D. Kennedy, B. Pendleton, D. Roweth, *Hybird Monte Carlo*, Phys. Lett B. **195** (1987), 216–222.

Numerical stroboscopic averaging for ODEs and DAEs

Mari Paz Calvo

(joint work with Ph. Chartier, A. Murua and J.M. Sanz-Serna)

We consider differential systems of the form

$$\frac{dy}{dt} = f(y, \frac{t}{\epsilon}; \epsilon), \quad t_0 \le t \le t_0 + L, \tag{1}$$

with initial condition $y(t_0) = y_0 \in \mathbb{R}^d$, where $f(y, \tau; \epsilon)$ is 2π-periodic in $\tau = t/\epsilon$, ϵ is a small parameter and, as $\epsilon \downarrow 0$, $L = \mathcal{O}(1)$. Our attention is restricted to cases where $f = \mathcal{O}(1/\epsilon)$ and the corresponding Poincaré map $\Psi_{t_0;\epsilon}$, which advances the

solution over one period $2\pi\epsilon$ starting from $t = t_0$, is an $\mathcal{O}(\epsilon)$ perturbation of the identity as $\epsilon \downarrow 0$. For a list of families of problems fitting in this framework, see [3].

In what follows we describe the stroboscopic averaging method (SAM) introduced in [2]. If $\varphi_{t_0,t;\epsilon}$ denotes the solution operator of (1), then $\Psi_{t_0;\epsilon} = \varphi_{t_0,t_0+2\pi\epsilon;\epsilon}$. One important observation is that $\Psi_{t_0;\epsilon}$ depends on t_0 in a $2\pi\epsilon$-periodic manner and, consequently, at stroboscopic times $t_n = t_0 + 2\pi\epsilon n$, $n = 0, \pm 1, \pm 2, \ldots$ it holds

$$y(t_n) = (\Psi_{t_0;\epsilon})^n y_0, \quad n = 0, \pm 1, \pm 2, \ldots.$$

Since the map $\Psi_{t_0;\epsilon}$ is close to the identity, standard results from backward error analysis of numerical integrators imply the existence of an autonomous differential system

$$\frac{dY}{dt} = F(Y;\epsilon) = F_1(Y) + \epsilon F_2(Y) + \epsilon^2 F_3(Y) + \cdots, \tag{2}$$

whose formal solution $Y(t)$ exactly coincides with $y(t)$ at the stroboscopic times $t_n = t_0 + 2\pi\epsilon n$, if $Y(t_0) = y(t_0)$. The series (2) does not converge in general, and in order to get rigorous results one has to consider a truncated version.

In order to integrate the highly oscillatory problem (1) we (approximately) compute the smooth interpolant $Y(t)$ by integrating the averaged equation (2) with a numerical method (macro-solver) with macro-step size H that should be much larger than the fast period $2\pi\epsilon$. In the spirit of the Heterogeneous Multiscale Methods of E and Engquist [4], our algorithm does not require the explicit knowledge of the analytic form of F. The information on F is gathered on the fly by integrating (with micro-step size h) the original system (1) in small time-windows of length $\mathcal{O}(\epsilon)$. There is much freedom in the choice of the macro-solver and micro-solver, including *standard variable-step/order codes.*

If the macro-solver is a linear multistep or Runge-Kutta method, then the only information on (2) required by the solver are function evaluations $F(Y^*;\epsilon)$ at given values of the argument Y^*. If $\Phi_{t;\epsilon}$ denotes the flow of the averaged system (2), it is clear that

$$F(Y^*;\epsilon) = \left.\frac{d}{dt}\Phi_{t;\epsilon}(Y^*)\right|_{t=t_0}$$

and, approximating the time-derivative by central differences,

$$F(Y^*;\epsilon) = \frac{1}{2\eta}[\Phi_{\eta;\epsilon}(Y^*) - \Phi_{-\eta;\epsilon}(Y^*)] + O(\eta^2).$$

Choosing now $\eta = 2\pi\epsilon$, and using that at stroboscopic times the solution of (1) coincides with the solution of (2) if the initial conditions are the same, we conclude that the formula

$$\widetilde{F}(Y^*;\epsilon) = \frac{1}{4\pi\epsilon}[\Psi_{t_0;\epsilon}(Y^*) - \Psi^{-1}_{t_0;\epsilon}(Y^*)], \tag{3}$$

provides an $\mathcal{O}(\epsilon^2)$ approximation $\widetilde{F}(Y^*;\epsilon)$ to $F(Y^*;\epsilon)$. To find $\Psi_{t_0;\epsilon}(Y^*)$ and $\Psi^{-1}_{t_0;\epsilon}(Y^*)$ one has to micro-integrate (1) with initial condition $y(t_0) = Y^*$ first forward from $t = t_0$ to $t = t_0 + 2\pi\epsilon$ and then backward from $t = t_0$ to $t = t_0 - 2\pi\epsilon$.

Difference formulae of higher order can be used but then, micro-integrations over wider windows are required.

In order to analyze the errors in SAM, it is important to notice that the method consists in integrating (2) with a chosen macro-integrator using inexact values of the right-hand side F. There are then three sources of errors:

(i) The error from the approximation of the true values of F by a finite-difference approximation $\widetilde{F}$, which is $\mathcal{O}(\epsilon^\delta)$, if δ denotes the order of the finite-difference formula ($\delta = 2$ if (3) is used).

(ii) The error in the difference formula due to the replacement of the Poincaré map Ψ by numerical approximations $\widetilde{\Psi}$ obtained via micro-integrations. If the micro-integrator has order p and the micro-step size is h/ϵ (in the variable τ), the implied error in the approximated value of F is $\mathcal{O}(\epsilon^{-1}(h/\epsilon)^p)$. However, choosing the micro-integrator suitably, the errors due to micro-integration may behave as $\epsilon^\nu(h/\epsilon)^p)$, with $\nu > 0$, leading to errors $\mathcal{O}(\epsilon^{\nu-1}(h/\epsilon)^p)$ in the approximation to F.

(iii) The error due to the macro-integrator used to solve the averaged equation, which is $\mathcal{O}(H^P)$, if H is the macro-step size and P is the order of the macro-integrator.

Combining (i)–(iii), one concludes that SAM provides approximations with error

$$\mathcal{O}\left(\epsilon^\delta + H^P + \epsilon^{\nu-1}\big(\frac{h}{\epsilon}\big)^p\right), \tag{4}$$

where the constant implied in the $\mathcal{O}$ notation is independent of ϵ, h and H.

The efficiency of SAM and the sharpness of the error estimate (4) are tested with different numerical experiments.

When integrating the perturbed Kepler problem as proposed in [5] using SAM with second order differencing and the classical fourth order Runge-Kutta method as macro and micro-integrator, one observes that halving ϵ leads to the same computational work but doubles the error, while the standard Runge-Kutta integrator works twice and doubles the error too ($\nu = 0$ in (4)). If the micro-integrations are performed using an appropriate splitting scheme, when ϵ is halved, the error is halved too, keeping the same computational work. The standard splitting scheme halves the error but doubles the computational effort. This situation corresponds to $\nu = 2$ in (4).

The second test problem we have considered is the well known van der Pol system, which has been efficiently integrated using SAM with the variable step-size code `ode45` of MATLAB as macro-integrator and Strang splitting for micro-integrations. Once the limit cycle has been reached, SAM increases the macro-step and, therefore, the variable-step macro-integration is much cheaper than its fixed step-size counterpart and both are much cheaper than the standard integration with the splitting scheme.

Finally, SAM has been applied to the vibrated double pendulum formulated in cartesian coordinates using the GGL approach, which leads to an index 2 differential algebraic system. As macro- and micro-integrator the third order, three stages,

half-explicit Runge-Kutta method in [1] has been used. In this case the error estimator (4) applies with $\nu = 1$, which means that if the dominant error in (4) is the one coming from micro-integrations, when halving ϵ, both, the errors and the computational cost for SAM are the same. For the standard integration with the half-explicit scheme although the error almost does not change, the computational cost doubles.

References

[1] V. Brasey and E. Hairer, *Half-explicit Runge-Kutta methods for differential-algebraic systems of index 2*, SIAM J. Numer. Anal. **30** (1993), 538–552.

[2] M. P. Calvo, Ph. Chartier, A. Murua and J. M. Sanz-Serna, *A stroboscopic numerical method for highly oscillatory problems*. In "Numerical Analysis of Multiscale Computations", B. Engquist, O. Runborg and R. Tsai eds., Lecture Notes in Computational Science and Engineering **82**, Springer.

[3] M. P. Calvo, Ph. Chartier, A. Murua and J. M. Sanz-Serna, *Numerical stroboscopic averaging for ODEs and DAEs*, submitted.

[4] W. E and B. Engquist, *The heterogeneous multiscale methods*, Comm. Math. Sci. **1** (2003), 87–132.

[5] U. Kirchgraber, *An ODE-solver based on the method of averaging*, Numer. Math. **53** (1988), 621–652.

Quasi-stroboscopic averaging: The non-autonomous case

Philippe Chartier

(joint work with Ander Murua and Jesus M. Sanz-Serna)

The talk considers non-autonomous oscillatory systems of ordinary differential equations with $d > 1$ nonresonant constant frequencies of the form

$$\frac{d}{dt}y \;=\; \epsilon f(y, t\omega), \quad y(0) \;=\; y_0 \in \mathbb{R}^D, \tag{1}$$

where $f(y, \theta)$ depends 2π-periodically on each of the scalar components $\theta_1, \ldots, \theta_d$ of the angular variable $\theta \in \mathbb{T}^d$ and $\omega \in \mathbb{R}^d$ is a constant vector of angular frequencies. Formal series like those used nowadays to analyze the properties of numerical integrators (B-series [7, 13, 9, 6]) are employed to construct higher-order averaged systems and the required changes of variables. With the new approach, the averaged system and the change of variables consist of vector-valued functions that may be written down immediately and scalar coefficients that are universal in the sense that they do not depend on the specific system being averaged and may therefore be computed once and for all. The new method may be applied to obtain a variety of averaged systems. In particular we study the quasi-stroboscopic averaged system characterized by the property that the true oscillatory solution and the averaged solution coincide at the initial time. We show that quasi-stroboscopic averaging is a geometric procedure because it is independent of the particular choice of co-ordinates used to write the given system. As a consequence, quasi-stroboscopic averaging of a canonical Hamiltonian (resp. of a divergence-free) system results in a canonical (resp. in a divergence-free) averaged system.

References

[1] V. I. Arnold, Geometrical Methods in the Theory of Ordinary Differential Equations, 2nd ed., Springer, New York, 1988.
[2] V. I. Arnold, Mathematical Methods of Classical Mechanics, 2nd ed., Springer, New York, 1989.
[3] M.P. Calvo, Ph. Chartier, A. Murua, J.M. Sanz-Serna, A stroboscopic method for highly oscillatory problems, in Numerical Analysis of Multiscale Computations (B. Engquist, O. Runborg, R. Tsai, eds.) in Lect. Notes Comput. Sci. Eng. 82, Springer, Berlin, to appear.
[4] M.P. Calvo, Ph. Chartier, A. Murua, J.M. Sanz-Serna, Numerical stroboscopic averaging for ODEs and DAEs, submitted.
[5] P. Chartier, A. Murua, J.M. Sanz-Serna, Higher-Order averaging, formal series and numerical integration I: B-series, *Found. Comput. Math.* **10**, 695–727 (2010).
[6] E. Hairer, Ch. Lubich and G. Wanner, Geometric Numerical Integration, 2nd ed., Springer, Berlin, 2006.
[7] E. Hairer, S. P. Nørsett and G. Wanner, Solving Ordinary Differential Equations I, Nonstiff Problems, 2nd ed., Springer, Berlin, 1993.
[8] P. Lochak, C. Meunier, Multiphase Averaging for Classical Systems, with Applications to Adiabatic Theorems, Springer, New York, 1988.
[9] A. Murua, Formal series and numerical integrators, part I: systems of ODEs and symplectic integrators, *Appl. Numer. Math.* **29**, 221–251 (1999).
[10] A. Murua, The Hopf algebra of rooted trees, free Lie algebras and Lie series, *Found. Comput. Math.* **6**, 387–426 (2006).
[11] L.M. Perko, Higher order averaging and related methods for perturbed periodic and quasi-periodic systems, *SIAM J. Appl. Math.* **17**, 698–724 (1968).
[12] J. A. Sanders, F. Verhulst and J. Murdock, Averaging Methods in Nonlinear Dynamical Systems (2nd. ed.), Springer, New York, 2007.
[13] J. M. Sanz-Serna and M. P. Calvo, Numerical Hamiltonian Problems, Chapman and Hall, London, 1994.

Quasi-stroboscopic averaging: The autonomous case

Ander Murua

(joint work with Philippe Chartier and Jesus M. Sanz-Serna)

B-series and related formal series expansions that are nowadays used to analyze numerical integrators [7], [13], [10], [6], provide a powerful means to study and implement the method of averaging (see e.g. [8], [12] and also [1], Chapter 4, [2], Chapter 10, [11]). We consider systems with $d \geq 1$ constant fast frequencies of the form

$$\frac{d}{dt}y \;=\; \epsilon f(y, t\omega), \quad y(0) \;=\; y_0 \in \mathbb{R}^D, \tag{1}$$

where $f(y,\theta)$ depends 2π-periodically on each of the scalar components $\theta_1,\ldots,\theta_d$ of the angular variable $\theta \in \mathbb{T}^d$ and $\omega \in \mathbb{R}^d$ is a constant vector of angular frequencies. Throughout the paper we assume that ω is *non-resonant*, i.e. $k \cdot \omega \neq 0$ for each multi-index $k \in \mathbb{Z}^d$, with $k \neq \mathbf{0}$. (Of course a problem with a resonant ω may be written in non-resonant form by lowering the number d of frequencies.) When the number d of frequencies is 1, quasi-stroboscopic averaging reduces to the stroboscopic averaging method investigated in [5]. (For the application of the idea

of stroboscopic averaging to the construction of numerical integrators the reader is referred to [3], [4].)

Our approach makes it possible to construct in an *explicit* way higher-order averaged systems and the associated changes of variables. In our new approach, the averaged system and the change of variables consist of vector-valued functions that may be written down immediately and scalar *coefficients* that are universal in the sense that they do not depend on the specific system being averaged and may therefore be computed once and for all for each vector of frequencies $\omega \in \mathbb{R}^d$. The new method may be applied to obtain a variety of averaged systems. In particular we study the *quasi-stroboscopic* averaged system characterized by the property that the true oscillatory solution and the averaged solution coincide at the initial time. We show that quasi-stroboscopic averaging is a geometric procedure because it is independent of the particular choice of co-ordinates used to write the given system. As a consequence, quasi-stroboscopic averaging of a canonical Hamiltonian (resp. of a divergence-free) system results in a canonical (resp. in a divergence-free) averaged system.

We also study the averaging of a family of near-integrable systems where our approach may be used to construct explicitly d formal first integrals for both the given system and its quasi-stroboscopic averaged version. As an application we construct three first integrals of a system that arises as a nonlinear perturbation of five coupled harmonic oscillators with one slow frequency and four resonant fast frequencies.

In this talk, we mainly focus on a family of autonomous problems where the vector field is an $\mathcal{O}(\epsilon)$ perturbation of an integrable system with constant frequencies. Such problems may be brought to the format (1) by means of a time-dependent change of variables. We describe how to obtain a quasi-stroboscopically averaged system with a number of favourable properties. If the original system is Hamiltonian, the averaged system will also be Hamiltonian and furthermore it is possible to construct explicitly d formal first integrals (of both the given and averaged systems) that are a $\mathcal{O}(\epsilon)$ perturbation of the first integrals of the unperturbed system. As an application we construct three first integrals of a system taken from [6] that arises as a nonlinear perturbation of five coupled harmonic oscillators with one slow frequency and four resonant fast frequencies.

References

[1] V. I. Arnold, Geometrical Methods in the Theory of Ordinary Differential Equations, 2nd ed., Springer, New York, 1988.

[2] V. I. Arnold, Mathematical Methods of Classical Mechanics, 2nd ed., Springer, New York, 1989.

[3] M.P. Calvo, Ph. Chartier, A. Murua, J.M. Sanz-Serna, A stroboscopic method for highly oscillatory problems, in Numerical Analysis of Multiscale Computations (B. Engquist, O. Runborg, R. Tsai, eds.) in Lect. Notes Comput. Sci. Eng. 82, Springer, Berlin, to appear.

[4] M.P. Calvo, Ph. Chartier, A. Murua, J.M. Sanz-Serna, Numerical stroboscopic averaging for ODEs and DAEs, submitted.

[5] P. Chartier, A. Murua, J.M. Sanz-Serna, Higher-Order averaging, formal series and numerical integration I: B-series, *Found. Comput. Math.* **10**, 695–727 (2010).

[6] E. Hairer, Ch. Lubich and G. Wanner, Geometric Numerical Integration, 2nd ed., Springer, Berlin, 2006.
[7] E. Hairer, S. P. Nørsett and G. Wanner, Solving Ordinary Differential Equations I, Nonstiff Problems, 2nd ed., Springer, Berlin, 1993.
[8] P. Lochak, C. Meunier, Multiphase Averaging for Classical Systems, with Applications to Adiabatic Theorems, Springer, New York, 1988.
[9] A. Murua, The Hopf algebra of rooted trees, free Lie algebras and Lie series, *Found. Comput. Math.* **6**, 387–426 (2006).
[10] A. Murua, Formal series and numerical integrators, part I: systems of ODEs and symplectic integrators, *Appl. Numer. Math.* **29**, 221–251 (1999).
[11] L.M. Perko, Higher order averaging and related methods for perturbed periodic and quasi-periodic systems, *SIAM J. Appl. Math.* **17**, 698–724 (1968).
[12] J. A. Sanders, F. Verhulst and J. Murdock, Averaging Methods in Nonlinear Dynamical Systems (2nd. ed.), Springer, New York, 2007.
[13] J. M. Sanz-Serna and M. P. Calvo, Numerical Hamiltonian Problems, Chapman and Hall, London, 1994.

Some topics in multisymplectic integration

Robert I McLachlan

Consider the two dichotomies ODEs vs PDEs and scalar vs differential conservation laws. These yield four classes of conservations laws that have striking similarities and differences both analytically and numerically. To review the ODE case first, a conserved quantity is a scalar H that obeys $H_t = 0$ on solutions of the ODE. Any such ODE can be written in coordinates in the form $z_t = K(z)\nabla H(z)$ where $K^\top = -K$ [1]. Unless the ODE has further structure, the conserved quantity merely constrains the motion to a hypersurface and little more can be said. Many numerical methods are known that conserve H (i.e., that obey $\Delta_t H = 0$) (e.g. projection methods and discrete gradient methods) but it is notable that H can be conserved by a B-series method if $K(z) \equiv$ const. The Average Vector Field method [2] is an example.

The next case is a differential conservation law for an ODE, of which the two main cases are conservation of a 2-form and a volume form. In the Hamiltonian case, $z_t = K(z)\nabla H(z)$ and the differential conservation law is $\omega_t = 0$ when $d\omega = 0$, $\omega = dz \wedge K^{-1}dz$. By its nature this is harder to conserve. Yet, once again, if $K(z) \equiv$ const. then ω can be conserved, i.e. $\Delta_t \omega = 0$, by a B-series method—the midpoint rule is an example. No symplectic integrators are known for general K.

The natural analogue of conserved quantities for PDEs are *conservation laws*. The scalar case in $1 + 1$ dimensions is $H_t + F_x = 0$. These can be found by Lagrangian, Hamiltonian, and direct methods, but the theory is different both analytically and dynamically from the ODE case. There is no general theory on how to preserve them under discretization, although discrete conservation laws are widely used when one component of the PDE itself is a conservation law. The discrete conservation law typically takes the form $\Delta_t \overline{H} + \Delta_x \overline{F} = 0$ where $\overline{H} \approx H$ depends locally on the fields.

Lagrangian and multiHamiltonian systems $Kz_t + Lz_x = \nabla S(z)$ have differential conservation laws like $\omega_t + \kappa_x = 0$, $\kappa = dz \wedge Kdz$, $\kappa = dz \wedge Ldz$. Discretizations of

$\omega_t + \kappa_x = 0$ are preserved by applying a symplectic discretization (e.g. symplectic Runge–Kutta) in x and in t [3]. A major open question concerns the implications of the multisymplectic conservation law for the dynamics. Notice that discrete conservation laws are necessarily different from continuous ones, unlike the ODE case, and may themselves depend on the method.

As so little is known about numerical methods for both types of conservation laws, it seems advantageous to simply find some more methods. We present a new class of methods for each type.

A discrete gradient on a vector space V with inner product $a^\top b$ is a function satisfying the discrete gradient axiom

$$(z_1 - z_0)^\top \overline{H_z}(z_0, z_1) = H(z_1) - H(z_0)$$

for all smooth functions $H : V \to \mathbb{R}$ and all z_0, $z_1 \in \mathbb{R}$. For example, $\overline{H_z}(z_0, z_1) = \int_0^1 \nabla H(\xi z_1 + (1-\xi)z_0)\,\mathrm{d}\xi$ is the average value discrete gradient.

A discrete gradient method for the ODE $z_t = K(z)\nabla H(z)$ is

$$\frac{z_1 - z_0}{\Delta t} = \overline{K}(z_0, z_1)\overline{H_z}(z_0, z_1), \quad \overline{K}^\top = -\overline{K}, \quad \overline{K}(z, z) = K(z).$$

We have $H(z_1) - H(z_0) = \overline{H_z}(z_0, z_1)^\top (z_1 - z_0) = \Delta t \overline{H_z}(z_0, z_1)^\top K \overline{H_z}(z_0, z_1) = 0$, i.e. H is conserved.

For Hamiltonians of the form $\mathcal{H} = \int H(z, z_x)\,dx$, $K^\top = -K$ a constant matrix, and equations of motion

$$z_t = K\frac{\delta \mathcal{H}}{\delta z} = K(H_z - \partial_x H_{z_x}),$$

the energy conservation law can be found as follows.

$$\begin{aligned} H_t &= H_z^\top z_t + H_{z_x}^\top z_{xt} \\ &= H_z^\top K(H_z - \partial_x H_{z_x}) + H_{z_x}^\top K(H_z - \partial_x H_{z_x})_x \\ &= \left(H_{z_x}^\top K H_z - H_{z_x}^\top K \partial_x H_{z_x}\right)_x \end{aligned}$$

Discrete gradient methods are a practical and general way to construct energy-preserving integrators for Hamiltonian PDEs. I now show that discrete gradient methods preserve the energy conservation law (ECL) as well.

Write $\overline{H}_y$ for the y-components of $\overline{H_z}$. If $z = (x, y)$, then the discrete gradient axiom becomes

$$(x_1 - x_0)^\top \overline{H}_x + (y_1 - y_0)^\top \overline{H}_y = H(x_1, y_1) - H(x_0, y_0).$$

We apply a discrete gradient method in (z, z_x) to the PDE giving

$$\frac{z_1 - z_0}{\Delta t} = K\left(\overline{H_z} - \partial_x \overline{H_{z_x}}\right)$$

The change in energy density over one time step is

$$\begin{aligned}\frac{H(z_1, z_{1x}) - H(z_0, z_{0x})}{\Delta t} &= \overline{H_z}^\top \frac{z_1 - z_0}{\Delta t} + \overline{H_{z_x}}^\top \frac{z_{1x} - z_{0x}}{\Delta t} \\ &= \overline{H_z}^\top K(\overline{H_z} - \partial_x \overline{H_{z_x}}) + \overline{H_{z_x}}^\top K(\overline{H_z} - \partial_x \overline{H_{z_x}})_x \\ &= \left(\overline{H_{z_x}}^\top K \overline{H_z} - \overline{H_{z_x}}^\top K \partial_x \overline{H_{z_x}}\right)_x\end{aligned}$$

A similar result holds for the Hamiltonian system $z_t = \mathcal{K}\frac{\delta H}{\delta z}$ where $\mathcal{K}$ is any constant skew-adjoint differential operator and $\overline{H} = H(x, z, z_x, z_{xx}, \dots)$ is any differential function of z; for the (e.g. multiHamiltonian) case $\mathcal{K}z_t = \frac{\delta H}{\delta z}$ where $\mathcal{K}$ may be singular; and for the fully discrete case, although this requires that the semidiscretization itself has an energy conservation law.

I now describe a new class of integrators for $Kz_t + Lz_x = \nabla S(z)$, the *diamond schemes*, that are linear and only locally implicit. These provide the first multisymplectic integrators that are well-defined for a variety of boundary conditions when the PDE has a noncanonical structure, e.g. KdV.

Although symplectic partitioned Runge–Kutta methods formally provide multisymplectic integrators on tensor product grids, many of them are unconditionally unstable. Gauss RK is stable, but it is fully implicit and not well-defined for many boundary conditions. Lobatto PRK can be explicit for some K, L, and S and is then conditionally stable [4].

The diamond scheme uses a non-tensor product grid; the grid lines are $x \pm at =$ const. and the cells are diamonds. An affine map yields a square grid and a symplectic PRK method is applied to each cell. For example, the midpoint rule yields

$$K\Delta_t z + L\Delta_x z = \nabla S(M_t M_x z)$$

($M =$ average). The method is multisymplectic and is implicit in at most a single cell. For r-stage PRK, the data are the stage values on each cell edge and r^2 implicit equations have to be solved within each cell. The method is explicit in x and implicit in z. For the KdV equation, it turns out to be fully explicit. It is conditionally stable (e.g. $\Delta t \le \Delta x$ for the linear wave equation), with no parasitic waves for any Δt, Δx. Because it is a linear method, the dispersion and stability of diamond RK can be determined for all linear PDEs (which is not the case for PRK methods).

The simple affine transformation of space-time has radically altered the numerical properties of the method, but preserved the key one (multisymplecticity) of interest here. I hope that this broader class will lead to useful methods and greater understanding of the multisymplectic domain.

References

[1] R. I. McLachlan, G. R. W. Quispel, and N. Robidoux, *Unified approach to Hamiltonian systems, Poisson systems, gradient systems, and systems with Lyapunov functions or first integrals*, Phys. Rev. Lett. **81** (1998), 2399–2403.

[2] E. Celledoni, R. I. McLachlan, B. Owren, and G. R. W. Quispel, *Energy-preserving integrators and the structure of B-series*, Foundations Comput. Math. **10** (2010), 673–693.

[3] S. Reich, *Multi-symplectic Runge–Kutta collocation methods for Hamiltonian wave equations*, J. Comput. Phys. **157** (2000), 473–499.
[4] R. I. McLachlan, Y. Sun, and P. S. P. Tse, *Linear stability of partitioned Runge–Kutta methods* **49** (2011), 232–263.

Splitting and composition methods for the time dependent Schrödinger equation

Sergio Blanes
(joint work with Fernando Casas, Ander Murua)

To describe and understand the dynamics and evolution of many basic atomic and molecular phenomena, their time dependent quantum mechanical treatment is essential, and this usually requires to solve the time dependent Schrödinger equation. If a space discretization is applied, one ends up with a linear system of ODEs of the form

$$i\frac{d}{dt}u(t) = Hu(t), \qquad u(0) = u_0 \in \mathbb{C}^N, \tag{1}$$

where H is a real symmetric matrix. The exact solution of eq. (1) is given by

$$u(t) = e^{-itH}\, u_0, \tag{2}$$

but to compute the exponential of the $N \times N$ complex and full matrix $(-itH)$ (typically also of large norm) by diagonalizing H can be prohibitively expensive for large values of N. In practice, thus, one turns to time stepping methods advancing the approximate solution from time t_n to $t_{n+1} = t_n + \tau$.

Due to the nature of this problem, the wave function vanishes asymptotically for large values of the space coordinates, and one can assume that the problem is periodic on a sufficiently large interval. This usually allows one to compute the matrix-vector products Hu by using FFTs. It is then natural to approximate $e^{-i\tau t\, H}u_n$ by explicit polynomials. The Taylor series is the simplest one. It provides very accurate results for low frequencies, but it needs many terms in the expansion to approximate the high frequencies. To this purpose, one can use Chebyshev polynomial approximations (if one knows in advance an upper bound of the spectral radius, $\rho(H)$). The action of the polynomials on vectors is computed by the Horner or the Clenshaw algorithms which need to keep in memory only two and three complex vectors, respectively.

We propose an alternative scheme to compute the matrix-vector products which leads to more efficient algorithms. This can be illustrated by the first order Euler method, $u_{n+1} = u_n - i\tau H u_n$. If we write $u = q + ip$, it is immediate to see that the algorithm is

$$\begin{cases} q_{n+1} = q_n + \tau H p_n \\ p_{n+1} = p_n - \tau H q_n. \end{cases} \tag{3}$$

Computing Hu_n requires a FFT for complex vectors, whereas $Hq_n,\ Hp_n$ require a FFT for real vectors so, the computational cost of (3) remains the same as Hu_n if appropriately implemented. However, in (3) we can replace the second equation

by $p_{n+1} = p_n - \tau H q_{n+1}$ at no extra cost, leading to the symplectic Euler method (or the first-order Lie-Trotter method) which is considerably more efficient (in the Taylor and Chebyshed methods, q and p are computed simultaneously instead of sequentially). This can be extended to higher orders if one rewrites (1) as [6]

$$\frac{d}{dt}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & H \\ -H & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} \tag{4}$$

and approximate the corresponding solution (the rotation matrix) by

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = K(\tau H)\begin{pmatrix} q_n \\ p_n \end{pmatrix}, \qquad K(\tau H) = \begin{pmatrix} K_1(\tau H) & K_2(\tau H) \\ K_3(\tau H) & K_4(\tau H) \end{pmatrix}, \tag{5}$$

where the entries $K_1(y)$ and $K_4(y)$ (respect., $K_2(y)$ and $K_3(y)$) are even (repect., odd) polynomials in $y \in \mathbb{R}$, and $\det K(y) = K_1(y)K_4(y) - K_2(y)K_3(y) \equiv 1$.

We have built several methods which cover most frequent problems by choosing appropriately the polynomials K_i [3]. To compute the action of $K(\tau H)$ on a vector, $(q_n, p_n)^T$, one can decompose the matrix $K(\tau H)$ as

$$K(\tau H) = \begin{pmatrix} I & 0 \\ -b_m\tau H & I \end{pmatrix}\begin{pmatrix} I & a_m\tau H \\ 0 & I \end{pmatrix} \cdots \begin{pmatrix} I & 0 \\ -b_1\tau H & I \end{pmatrix}\begin{pmatrix} I & a_1\tau H \\ 0 & I \end{pmatrix}, \tag{6}$$

where the coefficients a_i, b_i are obtained from the coefficients of the polynomials K_i (this decomposition is unique if it exists [2]). The scheme involves $2m$ products with real vectors (notice that K_i are polynomials up to degree $2m$) which has the same cost as a mth-order Taylor or Chebyshev polynomial approximation and belongs to the class of splitting methods. Next, to reach the desired accuracy for a given problem, one has to adjust the scaled time step τ/m (for a given value of $\rho(H)$) such that

$$\frac{\tau}{m} < \frac{1}{\rho(H)}\theta', \tag{7}$$

where θ' is a parameter depending on the method. To get accurate and stable solutions by Taylor methods we found that $\theta' \leq 0.25$ approximately, whereas for Chebyshev it is usually required that $\theta' \leq 0.5$, i.e. it is for most cases about twice faster. We have built new methods with values $\theta' \sim 0.75 - 1.4$ (m is the number of stages in (6)). Basically, the new schemes are obtained as follows [3]: we first choose $\theta' \in [0.5, 1.4]$ (a large choice for θ' makes more difficult to obtain accurate methods). Next, we consider a matrix $K(y)$ with components K_i of sufficient high order (typically $m \sim 20 - 40$) and look for approximations to the exact solution at different orders of accuracy which are stable for $y < m\theta'$ (here y plays the role of $\tau\rho(H)$). Each method is characterised by some parameters, μ_r, ν_r, and the error bound [3]

$$\|u_n - u(t)\| \leq \frac{t\,\mu_r\,\|u_0\|_{r+1} + \nu_r\|u_0\|_r}{\rho(H)^r}. \tag{8}$$

Here r is the order of the method and $\|u_0\|_k = \|H^k u_0\|$. Then, given a tolerance, we can choose among our methods the one which provides such accuracy with the

largest value of θ', i.e. with the lowest computational cost. The new methods show in practice a superiority w.r.t. Chebyshev similar to the superiority of Chebyshev w.r.t. Taylor. This class of splitting methods, in addition, preserve symplecticity and are conjugate to unitary methods [1, 2].

On the other hand, for those problems where H is separable into kinetic and potential energy, $H = T + V$, and the solution is relatively smooth, it can be more appropriate to consider unitary splitting methods

$$e^{-ib_m\tau V}\, e^{-ia_m\tau T} \cdots e^{-ib_1\tau V}\, e^{-ia_1\tau T}. \tag{9}$$

The number of exponentials m (and therefore the number of coefficients $\{a_i, b_i\}_{i=1}^m$) has to be sufficiently large to solve all the equations required to achieve a given order. The number of order conditions depends on the Lie algebra generated by the operators T, V, which is considerably more involved than in the previous case.

In general, the methods from the literature are such that $a_i, b_i \in \mathbb{R}$. In some recent works, looking for splitting methods in semigroups, complex solutions have been explored [4, 5, 7]. It was observed that in most cases, given a composition (say, for example, (9)) one can find many sets of real and many more sets of complex solutions for a_i, b_i. Usually, the best (in a certain unspecified sense) complex solution leads to considerably more accurate results than the best real solution. Typically, this higher accuracy does not pay for the extra cost by using complex arithmetic (roughly four times more expensive in most cases) in problems where negative real coefficients are allowed. However, the computational cost to compute one step for the composition (9) does not increase when the coefficients a_i, b_i are complex because the main factor in the computational cost is the evaluation of FFTs. On the other hand, since T is an unbounded operator, it is necessary that $a_i \in \mathbb{R}$, otherwise $(-ia_j\tau T)$ would be unbounded for at least one a_j. Thus one may argue whether new methods such that $a_i \in \mathbb{R}$ and $b_i \in \mathbb{C}$ are worth to be considered, being this a subject under investigation at present.

References

[1] S. Blanes, F. Casas and A. Murua, *Symplectic splitting operator methods for the time-dependent Schrdinger equation*, J. Chem. Phys. **124** (2006), 234105.

[2] S. Blanes, F. Casas and A. Murua, *On the linear stabylity of splitting methods*, Found. Comp. Math. **8** (2008), 357-393.

[3] S. Blanes, F. Casas and A. Murua, *Error analysis of splitting methods for the time-dependent Schrdinger equation*, Submitted.

[4] S. Blanes, F. Casas,P. Chartier and A. Murua, *Splitting methods with complex coefficients for some classes of evolution equations*, Submitted. arXiv:1102.1622v1

[5] F. Castella and P. Chartier and S. Descombes and G. Vilmart, *Splitting methods with complex times for parabolic equations*, BIT Numerical Analysis, **49** (2009), 487-508.

[6] S. Gray and D.E. Manolopoulos. *Symplectic integrators tailored to the time-dependent Schrödinger equation*, J. Chem. Phys., **104** (1996), 7099–7112.

[7] E. Hansen and A. Ostermann, *High order splitting methods for analytic semigroups exist*, BIT Numerical Analysis, **49** (2009), 527-542.

Statistically consistent model reduction for point vortices

Jason Frank
(joint work with Svetlana Dubinkina, Benedict Leimkuhler)

The talk was based on a paper with Svetlana Dubinkina and Ben Leimkuhler that recently appeared in *SIAM Multiscale Modeling and Simulation* [4]. In it we present an approach to constructing a simple model reduction based on the statistical mechanical concept of a thermodynamic reservoir, and the use of thermostat device such as the Nosé and Hoover methods [1, 2, 3].

The reduced modeling approach can be summarized as follows:

(1) Given a Hamiltonian system $\dot{y} = J\nabla H(y)$, $y(t) \in \mathbb{R}^d$, where $J = -J^T$, consider a partition of the variables into resolved $y_A \in \mathbb{R}^{d_A}$ and unresolved $y_B \in \mathbb{R}^{d-d_A}$ components, i.e. $y = (y_A, y_B)$, and a corresponding Hamiltonian splitting $H(y) = H_A(y_A) + H_B(y_B) + H_{AB}(y_A, y_B)$. The dynamics due to H_B will be absorbed into the reservoir, and we model the dynamics due to H_{AB} using a thermostat.

(2) The reduced system $y_A(t) \in \mathbb{R}^{d_A}$ is distributed according to canonical statistical mechanics. Almost every trajectory ergodically samples the Gibbs distribution $\rho \propto \exp(-\beta H_A)$.

(3) Since the reduced dynamics $y_A' = J\nabla H_A$ cannot sample ρ, we introduce a thermostat, the Nosé-Hoover-Langevin method:

$$\dot{y_A} = J\nabla H_A(y_A) + \xi\, g(y_A),$$
$$\dot{\xi} = h(y_A) - \frac{\alpha\sigma^2}{2} + \sigma\dot{w},$$

where g is an appropriately chosen vector field on $\mathbb{R}^{d_A}$, w' denotes the increments of a scalar Wiener process, and α and σ are appropriately chosen constants. The function $h(y_A) : \mathbb{R}^{d_A} \to \mathbb{R}$ is chosen such that the augmented distribution $\tilde{\rho} = \rho\exp(-\alpha\xi^2/2)$ is the unique steady state solution of the associated Fokker-Planck equation.

We applied this reduced modeling approach to the problem of point vortices on a cylinder. In [5], Oliver Bühler considered a heterogeneous system of point vortices, strong and weak, with positive and negative circulations, to study the theory proposed by Lars Onsager [6]. The equations of motion for point vortices are Hamiltonian:

$$\Gamma_i \dot{X}_i = K\frac{\partial H}{\partial X_i}, \qquad X_i \in \mathbb{R}^2$$

where $K = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and $H = -\frac{1}{4\pi}\sum_{i<j}\Gamma_i\Gamma_j \ln(|X_i - X_j|^2)$. Additional terms are added to the Hamiltonian to model the cylinder domain: for each vortex, an image vortex of opposite orientation is placed outside the domain.

The point vortex problem has the unusual property that the whereas the phase space is bounded (it is just N copies of the disc for N vortices), the energy function is unbounded from above and below—collisions of like and oppositely signed vortices drive the energy to plus and minus infinity, respectively.

Bühler's experimental setup was as follows. Strong vortices: 2 each positively and negatively oriented (vortex strength $\pm 10\pi$). Weak vortices: 48 each positively and negatively oriented (vortex strength $\pm 2\pi$) on a disc of radius 5. From the point of view of the strong vortices, the weak ones form a reservoir for the strong vortex dynamics. Bühler proposed this point of view and recorded statistics on the strong vortices.

Because Bühler's reservoir is composed of 96 weak vortices, finite reservoir effects were observed. To correct for this, the Gibbs distribution was modified to include a second term. Our thermostat was constructed to sample the distribution $\rho = \exp(-\beta H - \gamma H^2)$. This turns out to require a trivial modification of the method.

Figure 1 summarizes the statistics. The top, middle and bottom rows of plots include histograms derived from the strong vortex time series: the distance between like signed (top row) and opposite signed (middle row) vortices, and the radial distances of each vortex from the origin (bottom row). The three columns contain results for positive, zero and negative inverse statistical temperatures β. As Onsager predicted, at negative temperatures, like-signed vortices tend to cluster, and at positive temperatures, oppositely signed vortices cluster, or vortices spend more time near the wall, in the neighborhood of their image vortex. Bühler's data is shown in black; his direct numerical simulations are denoted by a thin, solid black line. Our statistics (shown in blue) compare favorably, and reproduce the qualitative features of his histograms.

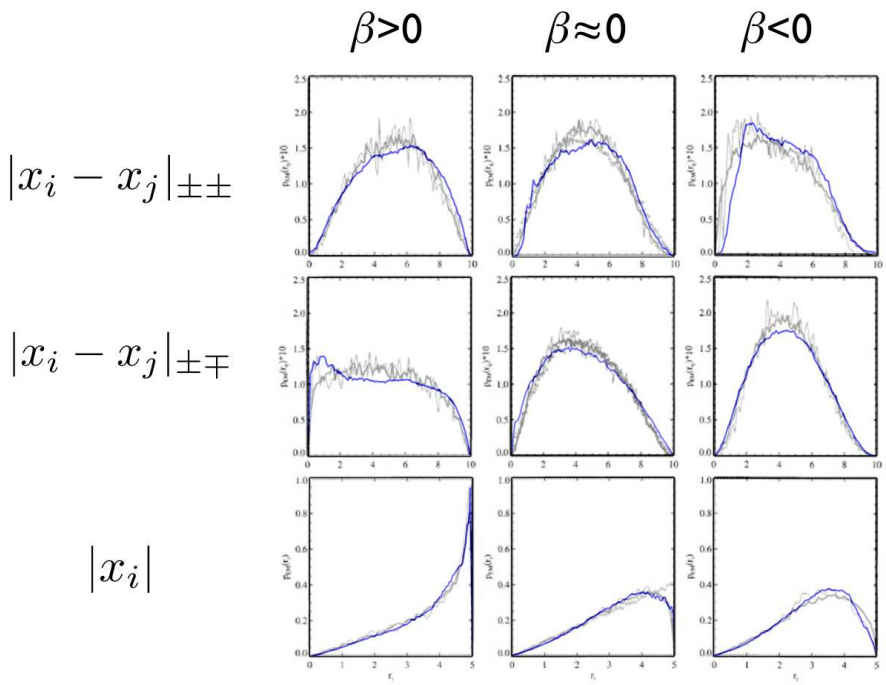


FIGURE 1. Histograms of strong vortex functions. See text for explanation.

We would like to extend this approach to grid-based methods for PDEs. The primary challenge to doing so is the correct interpretation of canonical statistical mechanics in this setting. For example, what constitutes the reservoir? Is it the set of unknowns on a fine grid, such that the reduced model is a subset of these? Another question we would like to address is the degree to which the trajectories of the thermostated variables (e.g. the strong point vortices) may still approximate, in an appropriate weak sense, the trajectories of full model.

References

[1] S. Nosé, *A molecular dynamics method for simulations in the canonical ensemble*, Mol. Phys., **52** (1984), 255–268.
[2] W.G. Hoover, *Canonical dynamics: Equilibrium phase space distributions*, Phys. Rev. A, **31** (1985), 1695–1697.
[3] A. Bulgac & D. Kusnezov, *Canonical ensemble averages from pseudomicrocanonical dynamics*, Phys. Rev. A **42** (1990), 5045–5048.
[4] S. Dubinkina, J. Frank, B. Leimkuhler, *Simplified modelling of energetic interactions with a thermal bath, with applications to a fluid vortex system*, Multiscale Model. Simul., (2010), to appear.
[5] O. Bühler, *Statistical mechanics of strong and weak vortices in a cylinder*, Phys. Fluids, **14** (2002), 22139–2149.
[6] L. Onsager, *Statistical hydrodynamics*, Nuovo Cimento (9), **6** (1949), 279–287.

Long term evolution of the Solar System

Jacques Laskar

As the Solar System motion is chaotic, with exponential divergence of nearby trajectories as

$$d = d_0\ 10^{T/10},$$

where T is in Myr, it is practically impossible to obtain a precise solution for its evolution over 100 Myr. Nevertheless, solutions of the Earth valid over extended times are needed for paleoclimate reconstructions. They are currently used in order to establish a geological time scale based on the correlation of sedimentary paleoclimate records with the computed variation of insolation on Earth resulting from planetary mutual perturbations.

In the past years, we have tried to match the need of geologists for a reference solution over 60 Myr, but this represents an improvement of 2 orders of magnitude in precision with respect to our best solution of 2004, valid over 40 Myr. The main limitation is on th model, but we also need to push the numerical symplectic scheme to its limits, and we are searching for optimal methods.

The same model can be used to answer the question of the stability of the Solar System.

In this case, only a statistical answer can be given. We have performed 2501 trajectories of the Solar System equations over 5 Gyr with close initial conditions,

compatible with our best knowledge of the present state. Most of the trajectories remain relatively stable over 5 Gyr, but for 1% of them, a resonance in the perihelion motion of Mercury and Jupiter increases in a large amount Mercury's excentricity, leading to a collision with Venus or the Sun.

In some cases, this induces a complete destabilisation of the full inner Solar System, with possible collisions of Venus and Mars with the Earth.

In a remarkable way, if the contribution of general relativity is removed, the 1% probability of strong instabilities raised to 60%.

Volume preserving splitting methods for divergence-free vector fields

Antonella Zanna
(joint work with Huiyan Xue)

In this talk we consider the problem of integrating divergence-free vector fields by explicit, volume preserving splitting methods. Our departure point is a splitting of a divergence-free vector field $\mathbf{f}(\mathbf{x})$ in a *diagonal* and *off-diagonal* part. The off-diagonal part consists of those terms such that $\partial_{x_i} f_i(\mathbf{x}) = 0$ (i.e. $\dot{x_i}$ does not depend explicitly on x_i), the diagonal part is the complement. By construction, the off-diagonal part is automatically divergence free and can be treated in a volume preserving manner by several methods already known in literature. Thus we focus on a vector field with a diagonal part only. The diagonal part appears explicitly in the divergence function and it is crucial to split it in an appropriate way, otherwise the resulting split vector fields will not be divergence free.

The main idea proposed in this talk is to consider a basis expansion of the divergence function. For each basis function, we reconstruct an Elementary Divergence Free Vector Field (EDFVF). This is obtained by collecting all the terms in the vector field that contribute to the basis function. We consider in detail the case of the monomial basis $\mathbf{x}^{\mathbf{j}} = x_1^{j_1} \cdots x_n^{j_n}$, which is relevant if the original vector field is polynomial (or if the given vector field can be easily approximated by a polynomial field). We show that the EDFVF for the monomial basis can be integrated exactly, because the basis functions always obey the same type (solvable) differential equation $\dot{\mathbf{x}^{\mathbf{j}}} = k\mathbf{x}^{\mathbf{2j}}$, where k is a constant depending on the coefficients of the EDFVF.

Thereafter we generalize our construction to the case when the basis consists of tensor products of 1D bases. We characterize the corresponding EDFVF and show that these always possess an integral. From the form of the EDFVF and the integral, we point at two possible ways to construct volume preserving methods: one (generic) is to use the integral to transform the EDFVF into a off-diagonal system, that can be treated by known methods; the other (specific to the basis) depends on whether we can solve for the basis function, as a function of time, as we were able in the polynomial case.

We show some numerical experiments in which we plot Poincaré sections for a cubic Stokes flow. The methods recover very well the sections, due to the intrinsic

volume preservation. The cost of the split methods (explicit) is comparable to the cost of an explicit one-step method.

In the future, we plan to extend this work to the reconstruction of particle flows in the context of partial differential equations.

References

[1] H. Xue and A. Zanna, *Explicit volume preserving splitting methods for polynomial divergence-free vector fields*, In preparation (2010).
[2] H. Xue and A. Zanna, *Explicit volume preserving splitting methods for divergence-free vector fields: tensor basis products*, In preparation (2011).

Resonances in long times integration of the nonlinear Schrödinger equation

Erwan Faou

We consider the nonlinear Schrödinger equation with cubic nonlinearity

$$i\partial_t u = -\Delta u + \varepsilon |u|^2 u, \tag{1}$$

where $u(t,x) \in \mathbb{C}$ is the wave function, and $x \in \mathbb{T}^d$ the d-dimensional torus. The coefficient $\varepsilon << 1$ is a small parameter. This equation is called *resonant* as the eigenvalues of the Laplace operator $-\Delta$ are all integers. We present here some results concerning the numerical simulation of this equation over long times.

If we decompose the exact solution in Fourier as $u(t,x) = \sum_{j\in\mathbb{Z}^d} \xi_j(t) e^{ij\cdot x}$, the equation (1) is written

$$\dot\xi_j = -i|j|^2\xi_j - i\varepsilon \sum_{j=k-\ell+m} \xi_k \bar\xi_\ell \xi_m = -i\frac{\partial H}{\partial \bar\xi_j}(\xi,\bar\xi),$$

where the Hamiltonian function is given by

$$H(\xi,\bar\xi) = \sum_{k\in\mathbb{Z}^d} |k|^2|\xi_k|^2 + \frac{\varepsilon}{2} \sum_{\substack{j,k,\ell,m\in\mathbb{Z}^d \\ k+m-\ell-j=0}} \xi_k\xi_m\bar\xi_\ell\bar\xi_j,$$

where $|k|^2 = k_1^2 + \cdots k_d^2$ for $k = (k_1,\dots,k_d) \in \mathbb{Z}^d$. Let us first consider the Fourier pseudo-spectral collocation approximation which consists in searching a Fourier polynomial

$$U^K(t,x) = \sum_{j\in B^K} \xi_j^K(t) e^{ij\cdot x}$$

satisfying (1) at the equidisant grid points $x_j = 2j\pi/K \in [-\pi,\pi]^d$, $j \in B^K$ (here B^K is a set of elements of $\mathbb{Z}^d$ depending on the parity of K). In this case, it can be shown that the coefficient ξ_j^K satisfy a Hamiltonian system associated with the Hamiltonian function

$$H^K(\xi,\bar\xi) = \sum_{k\in B^K} |k|^2|\xi_k|^2 + \frac{\varepsilon}{2} \sum_{\substack{k,m,\ell,j\in B^K \\ k+m-\ell-j\in K\mathbb{Z}}} \xi_k\xi_m\bar\xi_\ell\bar\xi_j =: T^K(\xi,\bar\xi) + P^K(\xi,\bar\xi).$$

We then consider the following time splitting method applied to this Hamiltonian: given a stepsize $\tau > 0$, we define the fully discrete Fourier coefficients $\xi^{K,n} = (\xi_j^{K,n})_{j\in\mathbb{Z}^d}$ by the recursion relation

$$\xi^{K,n+1} = \varphi^\tau_{P^K} \circ \varphi^\tau_{T^K}(\xi^{K,n}).$$

The exact evaluations of the flows $\varphi^\tau_{P^K}$ and $\varphi^\tau_{T^K}$ can be easily done using the Fast Fourier Transform algorithm. With this fully discrete solution, we can associated a modified energy in the sense of [4, 3]: For a fixed M, there exists C such that for all N, K, τ satisfying the CFL condition

$$\tau\frac{K^2}{2} < \frac{2\pi}{N+1},$$

there exists a polynomial Hamiltonian H^K_τ such that for all $\xi \in B_M = \{\, \xi \in \ell^1 \,|\, \|\xi\|_{\ell^1} \le M \,\}$, we have

$$\|\phi^\tau_{P^K} \circ \phi^\tau_{T^K}(\xi) - \phi^\tau_{H^K_\tau}(\xi)\|_{\ell^1} \le (\varepsilon\tau)^{N+1}(CN)^N.$$

Here, $\|\xi\|_{\ell^1} := \sum_{k\in\mathbb{Z}} |\xi_k|$ is the norm defining the Wiener algebra. Moreover, the first terms of the modified Hamiltonian are given by

$$H^K_\tau(\xi,\bar\xi) = \sum_{k\in B^K} |k|^2|\xi_k|^2 + \frac{\varepsilon}{2} \sum_{\substack{k,m,\ell,j\in B^K \\ k+m-\ell-j\in K\mathbb{Z}}} \frac{i\tau\Omega_{km\ell j}}{e^{i\tau\Omega_{km\ell j}}-1}\xi_k\xi_m\bar\xi_\ell\bar\xi_j + \mathcal{O}(\varepsilon^2\tau) \tag{2}$$

with $\Omega_{km\ell j} = |k|^2 + |m|^2 - |\ell|^2 - |m|^2$. The CFL condition ensures that $\tau\Omega_{km\ell j}$ is always smaller than 2π and the modified energy is hence well defined.

With this modified energy in hand, we aim at studying the long time behavior of the fully discrete approximations $\xi^{K,n}$. The situation differs in a significant way in dimension $d = 1$ or $d = 2$.

In dimension 1, the equation (1) is integrable [6] and it can be shown [5] that the *actions* $I_j(t) = |\xi_j(t)|^2$ of the exact solution are preserved for all times: we have for all $t \in \mathbb{R}$ and $j \in \mathbb{Z}$, $|I_j(t) - I_j(0)| \le C\varepsilon$ for some constant C independent on j. However, to reproduce this preservation property at the numerical level, we must assume that the number K is a prime number, see [3, Chapter VII]. More precisely, we can show that if $I_j^{K,n} = |\xi_j^{K,n}|^2$ are the numerical actions, we have

$$\forall\, j \in B^K, \quad |I_j^{K,n} - I_j^{K,0}| \le C\varepsilon, \quad \text{for} \quad n\tau \le \frac{T}{\varepsilon} \tag{3}$$

for some constant T independent on K, n and j, and under the condition that K is prime. If $K = 2P$ with P a prime number, the same relation holds for the *super actions* $J_j^{K,n} = I_j^{K,n} + I_{-j}^{K,n}$. An example of such numerical instabilities is given in Figure 1, where we perform the same simulation with different numbers of grid points: $K = 31$ and $K = 30 = 2 \times 3 \times 5$.

An open question is: does the preservation relation (3) holds for longer times, typically of order $\mathcal{O}(\varepsilon^{-r})$ with large r?

In dimension 2, it can be shown [2] that unlike the one-dimensional case, there exist initial data for which we can prove that there are energy exchanges between

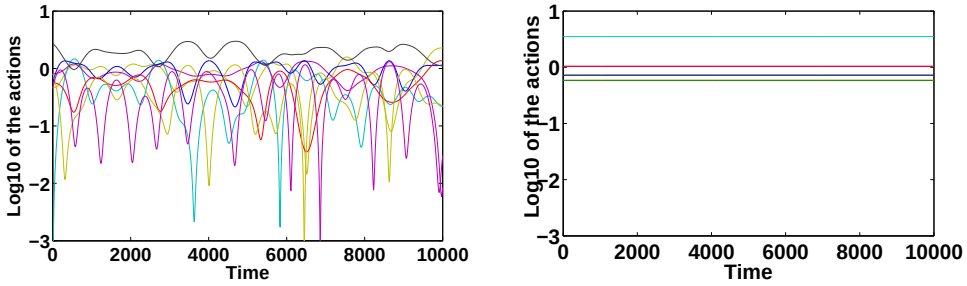
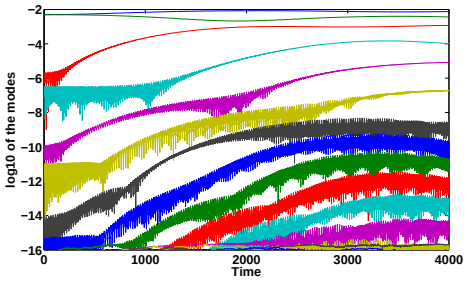


FIGURE 1. Evolution of the actions for $K = 30$ (left) and $K = 31$ (right)

the actions. The evolution of the actions of the exact solutions is represented in Figure 2 (in logarithmic scale), where the initial data $1 + 2\cos(x) + 2\cos(y)$ yields an *energy cascade* behavior. The analysis of this phenomenon relies on the



studying of the *resonant modulus*

$$\mathbb{K} = \{|j|^2 + |k|^2 - |\ell|^2 - |m|^2 = 0 \quad \text{and} \quad j + k - \ell - m = 0\},$$

which contains in particular quadruplets $(j, k, \ell, m) \in \mathbb{Z}^2$ for which the endpoints of the vectors j, k, ℓ, m form four corners of a non-degenerate rectangle with j and k opposing each other. Note that to prove the same result for the numerical solution, we have to study the corresponding *numerical resonant modulus*, and show that it contains the same rectangles as in the continuous case. In particular, it is not the case when using an implicit scheme and numerical simulations show that the correct energy exchanges are not reproduced, see [3, ChapterVII]. This is due to the fact that implicit schemes change the internal frequencies of the linear operator.

Finally, we emphasize that in this situation, it is conjectured in [1] that there exist initial data for (1) whose corresponding solutions $u(t)$ tend to infinity in H^s norm, for $s > 1$, when $t \to \infty$ (while all solutions have global bounds in H^1, as $\varepsilon > 0$). An open question would be to find - at least numerically - such solutions, and to exhibit the possible mechanisms that could produce such effects.

REFERENCES

[1] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Transfer of energy to high frequencies in the cubic defocusing nonlinear Schrdinger equation*, Invent. Math. 181 (2010), no. 1, 39-113.

[2] E. Faou and R. Carles, *Energy cascades for NLS on the torus.* http://arxiv.org/abs/1010.5173
[3] E. Faou, *Geometric numerical integration of Hamiltonian PDEs and applications to computational quantum mechanics.* http://www.irisa.fr/ipso/perso/faou/ETH/ETH.html
[4] E. Faou and B. Grébert, *Hamiltonian interpolation of splitting approximations for nonlinear PDEs.* To appear in Found. Comput. Math.
[5] B. Grébert, T. Kappeler and J. Pöschel, *Normal form theory for the NLS equation.* http://arxiv.org/abs/0907.3938
[6] V. E. Zakharov and 1. B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media.* Soviet Physics JETP 34 (1972), no. 1, 62–69.

Long-time analysis of Hamiltonian partial differential equations and their discretizations

Ludwig Gauckler

We discuss near-conservation properties of symplectic discretizations of Hamiltonian partial differential equations, for example nonlinear Schrödinger equations, in a weakly nonlinear setting. As a preparatory work for such a numerical analysis we study nonlinear perturbations of linear Hamiltonian partial differential equations.

We consider infinite dimensional Hamiltonian systems of the form

$$i\frac{d}{dt}\xi_j(t) = \omega_j\xi_j(t) + \frac{\partial P}{\partial \eta_j}(\xi(t), \overline{\xi(t)}), \quad j \in \mathcal{N} \subseteq \mathbb{Z}^d, \tag{1}$$

with real-valued frequencies ω_j and a nonlinearity given by a function $P(\xi, \eta)$ which is assumed to be (at least) cubic. Equations of this type describe for example the evolution of the Fourier coefficients of solutions of Hamiltonian partial differential equations such as the nonlinear Schrödinger equation

$$i\frac{\partial}{\partial t}\psi(x,t) = -\Delta\psi(x,t) + V(x) * \psi(x,t) + |\psi(x,t)|^2\psi(x,t)$$

with periodic boundary conditions in space and a potential V acting on ψ by convolution.

The actions along weakly nonlinear Hamiltonian pdes. Without the nonlinear term in (1) ($P = 0$), the *actions*

$$I_j(\xi, \overline{\xi}) = |\xi_j|^2$$

are exactly conserved along solutions of (1). In the presence of the nonlinear term in (1), but in a weakly nonlinear setting of small initial values

$$\|\xi(0)\|_s = \Big(\sum_{j\in\mathcal{N}} |j|^{2s}|\xi_j(0)|^2\Big)^{\frac{1}{2}} \leq \varepsilon$$

with $|j|^2 = \max(1, j_1^2 + \cdots + j_d^2)$, we expect near-conservation of actions on time intervals of length $\mathcal{O}(\varepsilon^{-1})$ since the nonlinearity in (1) is of order ε in variables of order 1.

By transforming the equation (1) to a *Birkhoff normal form* [1] or by expanding the solution of (1) as a *modulated Fourier expansion* [5] and [4], we can study much longer time intervals: We have *near-conservation of actions*

$$\sum_{j\in\mathcal{N}} |j|^{2s} \frac{|I_j(\xi(t),\overline{\xi(t)}) - I_j(\xi(0),\overline{\xi(0)})|}{\varepsilon^2} \le C_N \varepsilon^{\frac{1}{2}}$$

over *long times*

$$0 \le t \le \varepsilon^{-N}.$$

This can be shown under a non-resonance condition on the frequencies ω_j and a regularity assumption on the nonlinearity in (1). In the talk, the ideas of the proof of this result based on modulated Fourier expansions have been explained.

The actions along a numerical discretization. For the discretization of (1) a Lie–Trotter splitting or a Strang splitting of (1) in its linear and its nonlinear part is a popular choice. Along this numerical solution (and for small initial values) one has again *long-time near-conservation of actions* as along the exact solution. This result is obtained in a similar way as for the exact solution by considering a modulated Fourier expansion of the numerical solution [4]. Another possibility to prove this result is to use a Birkhoff normal form for the numerical method, [2] and [3].

The non-resonance condition used here does not only involve the frequencies ω_j but also the time step-size of the numerical discretization. The possibility of a numerical resonance was illustrated in the talk by a numerical experiment.

The energy along a numerical discretization. The *energy* or Hamiltonian function

$$H(\xi,\overline{\xi}) = \sum_{j\in\mathcal{N}} \omega_j |\xi_j|^2 + P(\xi,\overline{\xi})$$

is a conserved quantity of (1). For the same kind of discretization as before we have *near-conservation of energy*

$$\frac{H(\xi^n,\overline{\xi^n}) - H(\xi^0,\overline{\xi^0})}{\varepsilon^2} \le C_N \varepsilon^{\frac{1}{2}}$$

along the numerical solution $\xi^n \approx \xi(t_n)$ over *long times*

$$0 \le t_n = nh \le \varepsilon^{-N}.$$

This result is easily deduced from the near-conservation of actions.

References

[1] D. Bambusi, B. Grébert, *Birkhoff normal form for partial differential equations with tame modulus*, Duke Math. J. **135** (2006), 507–567.

[2] E. Faou, B. Grébert, E. Paturel, *Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. I. Finite-dimensional discretization*, Numer. Math. **114** (2010), 429–458.

[3] E. Faou, B. Grébert, E. Paturel, *Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. II. Abstract splitting*, Numer. Math. **114** (2010), 459–490.

[4] L. Gauckler, *Long-time analysis of Hamiltonian partial differential equations and their discretizations*, Dissertation (doctoral thesis), Universität Tübingen, 2010, http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-47540.
[5] E. Hairer, Ch. Lubich, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal. **38** (2000), 414–441.

Heterogeneous Multiscale Methods for Highly Oszillatory Dynamical Systems

Bjorn Engquist

The Heterogeneous Multiscale Method (HMM) is a framework for the numerical approximation of general classes of multiscale problems by coupling different methods on different scales in the same simulation. When HMM is applied to dynamical systems a microscale integrator supplies the average force to a macroscale method with long time steps. We discuss this technique and focus on the problem of finding appropriate macroscale variables.

Multiscale computation for oscillatory dynamical systems with more than two separated time scales

Richard Tsai
(joint work with Gil Ariel, Bjorn Engquist, and Seong Jun Kim)

We discuss several issues arising in designing hierarchical HMMs (Heterogeneous Multiscale Methods) for computing the effective behavior of highly oscillatory dynamical systems in a long time interval $[0, T]$. The number $0 < \epsilon \ll 1$ will be used to parametrize the time scales involved: the fastest time scale involves oscillations whose frequencies are at the order of $\frac{1}{\epsilon^2}$, and the $T \geq C\epsilon^{-k}$, $k \geq 0$. Constructively, hierarchical HMMs aim at resolving interactions of the oscillations in different scales.

When more than two separated time scales are considered, there exists a new type of slow variables whose time derivatives are not bounded (as $\epsilon \to 0$) along the flow of the dynamical system. Such type of slow variables do not appear for problems with only two separated time scales. In the literature, the time dependent function $x_1 = \sin(t)$ with $|\dot{x}_1| = \mathcal{O}(1)$ is naturally regarded as slow and $x_2 = \sin(t/\epsilon)$ with $|\dot{x}_2| = \mathcal{O}(\epsilon^{-1})$ is fast. Similarly $x_3 = \sin(t) + \epsilon \sin(t/\epsilon)$ is slow. In addition, we need to consider $x_4 = \sin(t) + \epsilon \sin(t/\epsilon^2)$ as slow even if $|\dot{x}_4| = \mathcal{O}(\epsilon^{-1})$. It will be regarded as slow because $|x_4 - \sin t| = \mathcal{O}(\epsilon)$ and $\sin(t)$ is slow. We show that the effective behavior described by this type of slow variables cannot be ignored in building multiscale algorithms.

The second topic of our discussion involves averaging over suitable tori at different time scales due to different types of near resonances in the system. The tori

in question are defined by the slow variables of the oscillatory dynamical systems. Consider the simple model problem:

$$\begin{cases} \frac{d}{dt}z_1 = 2\pi i \frac{1}{\epsilon} z_1, \\ \frac{d}{dt}z_2 = 2\pi i \frac{\lambda}{\epsilon} z_2, \\ \frac{d}{dt}w = g(w, z_1, z_2). \end{cases} \tag{1}$$

where $\lambda = 1 + \delta$, δ is a small irrational number, and hence $(z_1(t), z_2(t))$ stay on an invariant 2-torus. Depending on the strength of δ, the slow variable w can be approximated by different averaging procedures. We classify two kinds of near resonances – weak near resonances for the case $\delta = \epsilon^{\frac{1}{q}}$, $q > 1$, and strong near resonances for $\delta = \epsilon^p$, $p \geq 1$. This separation is related to the speed of the flow to cover an invariant torus.

When the system is in weak near resonance, the trajectories "cover" an invariant torus sufficiently fast and we need to average over a torus. On the other hand, when strong near resonance occurs, the system is effectively in resonance for any time interval that is independent of ϵ. Therefore, we only need to average the flow over suitable topological circles which are one dimensional periodic orbits on the invariant torus. For longer times, we need to average over the torus since time is long enough for the flow to cover the torus.

We propose an algorithm for efficiently computing averages over invariant tori for systems in weak near resonance. In an analogy to molecular dynamics, this algorithm combines time averaging with ensemble averaging. In short, our algorithm constructs a global orthogonal coordinate system on a torus, places a grid over the torus using the constructed coordinate system, and then computes suitable averages of the flows that start out from the grid nodes. This way, we can efficiently integrate over a torus via short time integration of the oscillatory system and iterative use of an efficient averaging kernel developed for averaging over circles. The orthogonal coordinate system on the torus is constructed by using a smooth invertible mapping of the dynamical system's flow direction and the normals of the torus to designated standard basis vectors in the embedding Euclidean space $\mathbb{R}^n$, and then properly pulling back the remaining basis vectors in $\mathbb{R}^n$ to the torus' tangle bundle.

If the dynamical system is ergodic on an invariant manifold and the corresponding invariant measure is non-uniform, the preceding method needs to be modified. Indeed, the flow density of such a dynamical system, is not explicitly available in general, so we must consider a way to statistically estimate the density by looking at the flow.

References

[1] G. Ariel, B. Engquist, H.-O. Kreiss, and R. Tsai. Multiscale computations for highly oscillatory problems. In *Multiscale modeling and simulation in science*, volume 66 of *Lect. Notes Comput. Sci. Eng.*, pages 237–287. Springer, Berlin, 2009.

[2] G. Ariel, B. Engquist, and R. Tsai. Oscillatory systems with three separated time scales – analysis and computation. In *Numerical Analysis of Multiscale Computations*, volume 82 of *Lect. Notes Comput. Sci. Eng.*

[3] G. Ariel, B. Engquist, and R. Tsai. A multiscale method for highly oscillatory ordinary differential equations with resonance. *Math. Comp.*, 78(266):929–956, 2009.
[4] B. Engquist and Y.-H. Tsai. Heterogeneous multiscale methods for stiff ordinary differential equations. *Math. Comp.*, 74(252):1707–1742, 2005.

On the efficiency of numerical homogenization methods

Assyr Abdulle

We study finite element (FE) discretizations of second-order elliptic problems of the form

$$-\nabla \cdot (a^\varepsilon(x)\nabla u^\varepsilon(x)) = f(x) \text{ in } \Omega, \qquad u^\varepsilon(x) = 0 \text{ on } \partial\Omega, \tag{1}$$

where Ω is a bounded convex polyhedron in $\mathbb{R}^d$, with $d \leq 3$. Zero Dirichlet boundary conditions are taken here for simplicity and we emphasize that other boundary conditions can be treated with the multiscale method described below. We also note that while we discuss linear elliptic problems, the multiscale method (and the a-priori analysis) can be derived for time-dependent problems (parabolic or hyperbolic) or nonlinear problems [7],[8],[9].

The $d \times d$ tensor $a^\varepsilon(x)$, assumed to be uniformly elliptic and bounded, is allowed to vary on a very small spatial scale denoted by ε. This behavior of a^ε makes a standard numerical approximation very costly if not impossible. Numerical homogenization methods are multiscale methods, inspired by homogenization theory [11],[14], aiming at computing an effective solution, $u^0(x)$, of the equation (1). This effective solution solves a homogenized problem, whose data are usually unknown in explicit form.

The FE-HMM method for computing a numerical approximation of the effective (homogenized) solution $u^0(x)$ is based on a macro FE space made of piecewise polynomial functions of degree ℓ defined on a macro partition $\mathcal{T}_H$ of Ω, and micro FEMs for the solution of so-called cell problems (involving the oscillating tensor of equation (1)) defined on sampling domains located at quadrature points within the macro partition. A proper averaging of the micro solutions allows then to define the effective bilinear form whose solution gives an approximation of the homogenized solution u^0. Supplemented with appropriate numerical correctors, the FE-HMM solution can also capture, in certain situation, the fine scale solution $u^\varepsilon(x)$ (see [1],[2],[3],[12],[13]).

A fully discrete a priori convergence analysis shows that the complexity of the FE-HMM (as any numerical homogenization method) is superlinear with respect to the macroscopic degrees of freedom [1],[2],[3]. In particular, macro and micro meshes have to be refined simultaneously to obtain optimal convergence rates with a minimal computational cost. For high dimensional problems or high order macro methods (for which a lot of sampling domains have to be used) the FE-HMM can become costly.

Three ways of reducing the complexity. We discuss three (non exclusive)

ways to reduce the complexity of numerical homogenization methods as the FE-HMM. To simplify the presentation we consider a tensor of the form $a^\varepsilon = a(x, x/\varepsilon)$.

Case of very regular micro oscillations of $a(x, \cdot)$. In this situation, the macro solver of the FE-HMM can be coupled with pseudo-spectral methods on the sampling domains. Taking advantage of the fast convergence of the micro solvers, this strategy can reduce significantly the computational cost provided enough regularity of the oscillating tensor at the micro scale [4].

Case of a regular dependence on the macro variable of $a(\cdot, x/\varepsilon)$. In such a situation, following the reduced basis framework, the idea is to compute in an offline stage a low dimensional subspace of the micro solutions used to define the effective bilinear form in the FE-HMM. These parametrized micro solutions are selected by a Greedy algorithm. Let $\mathcal{M}$ be the space of all cell problem solutions indexed by $\nu \in \Omega$, the barycenter of the sampling domains and $\eta = 1, \ldots, d$, the d solutions of the cell problem in each sampling domain. The goal is to find an $N-$dimensional subspace $\mathcal{M}_N$ of $\mathcal{M}$ that "minimizes" the distance $\sup_{\xi \in \mathcal{M}} \operatorname{dist}(\xi, \mathcal{M}_N)$ [10]. This precomputed set of selected solutions of cell problems (computed with high accuracy) can then be used in an online stage to obtain cheap micro solutions.

Case of low regularity of u^0. In this situation, explicit localized error indicators for robust and reliable adaptive mesh refinement can be derived. A (non-uniformly) refinement of the macromesh can be coupled to a refinement of the micromesh covering the sampling domains [5],[6]. As new microsolutions need not to be re-computed in non refined elements, adaptive mesh refinement allows for an important computational saving compared to uniform refinement strategies.

Acknowledgment. This work was supported in part by a Swiss National Science Foundation under Grant 200021_134716/1.

References

[1] A. Abdulle, *On a priori error analysis of fully discrete heterogeneous multiscale FEM*, SIAM Multiscale Model. Simul., 4, no. 2 (2005), 447–459.

[2] A. Abdulle, *The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs,* GAKUTO Int. Ser. Math. Sci., 31, (2009), 133–181.

[3] A. Abdulle, *A priori and a posteriori analysis for numerical homogenization: a unified framework*, to appear in Ser. Contemp. Appl. Math. CAM.

[4] A. Abdulle, B. Engquist, *Finite element heterogeneous multiscale methods with near optimal computational complexity*, SIAM Multiscale Model. Simul., 6, no. 4 (2007), 1059–1084.

[5] A. Abdulle, A. Nonnenmacher, *A posteriori error analysis of the heterogeneous multiscale method for homogenization problems,* C. R. Acad. Sci. Paris, Ser. I. 347, no. 17-18 (2009) 1081–1086.

[6] A. Abdulle and A. Nonnenmacher, *Adaptive FE heterogeneous multiscale method for homogenization problems*, Comput. Methods Appl. Mech. Engrg., to appear.

[7] A. Abdulle and M. Grote, *Finite element heterogeneous multiscale method for the wave equation,* to appear in SIAM Multiscale Model. Simul.

[8] A. Abdulle and G. Vilmart, *Analysis of the finite element heterogeneous multiscale method for nonmonotone elliptic homogenization problems*, preprint, http://infoscience.epfl.ch/record/163326.

[9] A. Abdulle and G. Vilmart, *Fully discrete finite element heterogeneous multiscale method for parabolic homogenization problems,* preprint.
[10] A. Abdulle and Y. Bai, in preparation.
[11] A. Bensoussan, J.-L. Lions and G. Papanicolaou, *Asymptotic analysis for periodic structures*, North Holland, Amsterdam, 1978.
[12] W. E and B. Engquist, *The Heterogeneous Multi-Scale Methods*, Commun. Math. Sci., 1 (2003), 87–132.
[13] W. E, P. Ming and P. Zhang, *Analysis of the heterogeneous multiscale method for elliptic homogenization problems*, J. Amer. Math. Soc. 18, no. 1 (2005), 121–156.
[14] V.V. Jikov, S.M. Kozlov and O.A. Oleinik, *Homogenization of differential operators and integral functionals*, Springer-Verlag, Berlin, 1994.

Data assimilation for dynamical systems

Sebastian Reich

(joint work with Georg Gottwald (University of Sydney), Kay Bergemann (Universität Potsdam), Eugenia Kalnay, Javier Amezcua, Kayo Ide (University of Maryland))

The basic task of data assimilation (nonlinear filtering) can be explained for second-order Langevin dynamics

$$dq = v dt, \tag{1}$$

$$dv = -V'(q)dt - \gamma v dt + \sqrt{\sigma} dw(t) \tag{2}$$

where $w(t)$ denotes standard Brownian motion. The parameters $\gamma > 0$ and $\sigma > 0$ are assumed to be known as well as the initial conditions $q(0)$ and $v(0)$. In addition it is assumed that one has "measurements" $Q(t)$ of the "true" positions $q_T(t)$ which satisfy the stochastic differential equation

$$dQ = v_T(t)dt + \sqrt{r} du(t) \tag{3}$$

where $u(t)$ is again Brownian motion, $(q_T(t), v_T(t))$, $t \geq 0$, is an unknown solution (the "truth"), and $r > 0$ is known.

Both naive approaches of either solving (1)-(2) with the given initial conditions or integrating (3) to obtain $Q(t)$ are able to track the reference solution $(q_T(t), v_T(t))$ over long periods of time. Instead one has to resort to filtering or smoothing techniques to combine (1)-(3).

In recent year, the ensemble Kalman filter [1] has emerged as a powerful nonlinear filter for intermittent data assimilation. We have extended this technique to continuous data assimilation problems as outlined above. In particular, the ensemble Kalman-Bucy filter [5, 7] leads to the following augmented system of stochastic differential equations. We first rewrite (1)-(3) in more abstract form as

$$dx = f(x,t)dt + \Sigma^{1/2} dw(t), \tag{4}$$

$$dy = Hx dt + R^{1/2} du(t) \tag{5}$$

Then the ensemble Kalman-Bucy filter equations for an ensemble of m members $x_i(t)$ are

$$dx_i = f(x_i,t)dt + \Sigma^{1/2}dw_i(t) - PH^TR^{-1}(Hx_idt - dy(t) + R^{1/2}du_i(t))$$

with empirical covariance matrix

$$P = \frac{1}{m-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})^T, \quad \bar{x} = \frac{1}{m}\sum_i x_i$$

and mutually independent Brownian motions $w_i(t)$, $u_i(t)$, $i = 1, \ldots, m$.

To make progress to more general and accuracte filters for nonlinear problems we consider intermittent data assimilation where (5) is replaced

$$y_q = Hx(t_q) + R^{1/2}\eta_q \tag{6}$$

at discrete times t_q, $q = 1, \ldots, K$ with $\eta_q \sim N(0,1)$. We also set $\Sigma = 0$ in (4). The ensemble Kalman-Bucy filter can also be applied to such filter problems and leads to efficient implementation of ensemble Kalman filters [2, 3, 4, 7].

On a more general level, one can reformulate (4) with $\Sigma = 0$ and (6) as a Vlasov-McKean system [5]

$$\dot{x} = f(x,t) + \sum_q \delta(t - t_q)M\nabla_x\psi(x,\rho) \tag{7}$$

$$\rho_t = -\nabla_x \cdot (\rho\dot{x}), \tag{8}$$

where $\delta(\cdot)$ denotes the Dirac delta function, M is a positive definite matrix, and the potential ψ is determined from

$$\nabla_x \cdot (\rho M\nabla_x\psi) = \rho(L - E_\rho[L]) \tag{9}$$

where

$$L(y_q;x) \propto \exp\left(-\frac{1}{2}(Hx - y_q)^TR^{-1}(Hx - y_q)\right)$$

is the likelihood associated with the measurement (6) and $E_\rho[L]$ denotes expectation of L with respect to the probability density ρ. Numerical approximations based on (7)-(8) including Gaussian mixture approximations for ρ have been discussed in [5, 6]. The key idea is to approximate (7)-(8) using ensemble Monte Carlo techniques. The ensemble at time t_q is converted into a statistical model yielding a density approximation $\tilde{\rho}$ which is then used in (9) in place of ρ to find the potential ψ in (7).

Interesting questions for further research include the importance of geometric integration methods for solving (7)-(8) and efficient numerical methods for solving (7)-(8) in the presence of highly oscillatory solution components.

References

[1] G. Evensen, *Data assimilation: The ensemble Kalman filter*, Springer-Verlag, New York, 2006.

[2] K. Bergemann, G. Gottwald, and S. Reich, *Ensemble propagation and continuous matrix factorization algorithms*, Q.J.R. Meteorolog. Soc., **135**, ,2009, 1560–1572.

[3] K. Bergemann and S. Reich, *A localization technique for ensemble Kalman filters*, Q.J.R. Meteorolog. Soc., **136**, 2010, 701–707.
[4] K. Bergemann and S. Reich, *A mollified ensemble Kalman filter*, Q.J.R. Meteorolog. Soc., **136**, 2010, 1636–1643.
[5] S. Reich, *A dynamical systems framework for intermittent data assimilation*, BIT Numerical Mathematics, **51**, 2011, 235–249.
[6] S. Reich, *A Gaussian mixture ensemble transform filter*, 2011, submitted.
[7] J. Amezcua, E. Kalnay, K. Ide, and S. Reich, *Using the Kalman-Bucy filter in an ensemble framework*, 2011, submitted.

Discrete Mechanics and Optimal Control: Structure preserving integration for the optimal control of mechanical systems

Sina Ober-Blöbaum

For the numerical solution of optimal control problems, direct methods are based on a discretization of the underlying differential equations which serve as equality constraints for the resulting finite dimensional nonlinear optimization problem. For the case of mechanical systems, the presented method, denoted by DMOC (Discrete Mechanics and Optimal Control), is based on the discretization of the variational structure of the system directly. The discretization of the Lagrange-d'Alembert principle leads to structure preserving time-stepping equations such that the resulting optimal control algorithm inherits the structure preserving properties of the mechanical system. Furthermore, the approximation order of the adjoint equations resulting from the necessary optimality conditions is the same as for the state system due to the symplecticity of the discretization scheme.

1. Discrete mechanics and optimal control

Consider a mechanical system described by a C^2-Lagrangian $L : TQ \to \mathbb{R}$ to be moved on a curve $q(t) \in Q \subseteq \mathbb{R}^n$ during the time interval $[0, T]$ from an initial state $(q^0, \dot{q}^0)$ to a final state $(q^T, \dot{q}^T)$. The motion is influenced via a control force $f_L(q(t), \dot{q}(t), u(t))$ with control parameter $u(t) \in U \subseteq \mathbb{R}^m$ such that a given objective functional

$$J(q,u) = \int_0^{\mathrm{T}} C(q(t), \dot{q}(t), u(t))\,\mathrm{d}t + \Phi(q(\mathrm{T}), \dot{q}(\mathrm{T})) \tag{1}$$

is minimized. Here $C : TQ \times U \to \mathbb{R}$ and $\Phi : TQ \to \mathbb{R}$ are continuously differentiable cost functions. The motion of the system is to satisfy the Lagrange-d'Alembert principle, which requires that

$$\delta \int_0^{\mathrm{T}} L(q(t), \dot{q}(t))\,\mathrm{d}t + \int_0^{\mathrm{T}} f_L(q(t), \dot{q}(t), u(t)) \cdot \delta q(t)\,\mathrm{d}t = 0 \tag{2}$$

for all variations δq with $\delta q(0) = \delta q(\mathrm{T}) = 0$ which is equivalent to the fulfillment of the Euler-Lagrange equations $\frac{d}{dt}\frac{\partial}{\partial \dot{q}} L(q, \dot{q}) - \frac{\partial}{\partial q} L(q, \dot{q}) = f(q, \dot{q}, u)$. The optimal control problem stated in (1) and (2) is transformed into a finite dimensional constrained optimization problem using a global discretization of the states and the controls. The state space TQ is replaced by $Q \times Q$ and the discretization

grid is defined by $\Delta t = \{t_k = kh \,|\, k = 0, \ldots, N\}$, $Nh = T$, where N is a positive integer and h is the step size. The path $q : [0, T] \to Q$ is replaced by a discrete path $q_d : \{t_k\}_{k=0}^{N} \to Q$, where $q_k = q_d(kh)$ is an approximation to $q(kh)$ [3, 4]. Similarly, the control path $u : [0, T] \to U$ is replaced by a discrete one $u_d = \{u_k\}_{k=0}^{N-1}$, where u_k is the discrete control parameter guiding the system from q_k to q_{k+1}.

The discrete Lagrange-d'Alembert principle (3), emerges using an approximation of the action integral in (2) by a discrete Lagrangian $L_d : Q \times Q \to \mathbb{R}$, $L_d(q_k, q_{k+1}) \approx \int_{kh}^{(k+1)h} L(q(t), \dot{q}(t))\, dt$, and discrete forces $f_k^- \cdot \delta q_k + f_k^+ \cdot \delta q_{k+1} \approx \int_{kh}^{(k+1)h} f(q(t), \dot{q}(t), u(t)) \cdot \delta q(t)\, dt$, where the left and right discrete forces $f_k^{\pm}$ now depend on (q_k, q_{k+1}, u_k). For the fulfillment of the discrete Lagrange-d'Alembert principle, it is necessary to consider discrete paths $\{q_k\}_{k=0}^{N}$ such that for all variations $\{\delta q_k\}_{k=0}^{N}$ with $\delta q_0 = \delta q_N = 0$, it is true that

$$\delta \sum_{k=0}^{N-1} L_d(q_k, q_{k+1}) + \sum_{k=0}^{N-1} \left(f_k^- \cdot \delta q_k + f_k^+ \cdot \delta q_{k+1} \right) = 0. \tag{3}$$

In a similar way, an approximation of the objective functional (1) generates the discrete objective function J_d involving the discrete cost functions C_d and Φ_d. The resulting discrete constrained optimization problem is stated as follows: Find discrete paths q_d and u_d that minimize the discrete objective function,

$$J_d(q_d, u_d) = \sum_{k=0}^{N-1} C_d(q_k, q_{k+1}, u_k) + \Phi_d(q_{N-1}, q_N, u_{N-1}), \tag{4}$$

subject to discretized boundary conditions and the forced discrete Euler-Lagrange (DEL) equations resulting from (3) as

$$D_2 L_d(q_{k-1}, q_k) + D_1 L_d(q_k, q_{k+1}) + f_{k-1}^+ + f_k^- = 0 \tag{5}$$

with $k = 1, ..., N-1$ and D_i denotes the derivative w.r.t. the i-th argument.

Note that (5) provides a discretization scheme for the Euler-Lagrange equations. Such a scheme is called variational integrator [3] since it is derived in a discrete variational way rather than based on a discretization of the ordinary differential equations. Variational integrators are structure-preserving integration methods, that are symplectic and momentum consistent, i.e. the symplectic structure and the momentum maps induced by symmetry groups are consistent with the control forces (or exactly preserved in absence of external forces) in the discrete solution independent of the step size h [4].

2. The adjoint system

The variational character of the discrete scheme plays an important role comparing the adjoint systems of the optimal control problem and its discrete version. The adjoint systems are determined by the necessary optimality conditions which are given by Pontryagin's maximum principle (PMP) and the Karush-Kuhn-Tucker equations (KKT) for the continuous and the discrete problem, respectively.

The optimal control problem stated in (1) and (2) can be rewritten in standard form as $\min_{x(\cdot),u(\cdot)} \bar{J}(x,u)$ s.t. $\dot{x} = f(x,u),\ x(0) = x^0$ with $x = (q, \dot{q})$. The PMP provides necessary optimality conditions in the following way (for simplicity we consider the case $C = 0$ and no final point constraint is involved). Let (x^*, u^*) be an optimal solution. Then, there exists a function $\lambda : [0,T] \to \mathbb{R}^n$ such that $\mathcal{H}(x^*(t), u^*(t), \lambda(t)) = \max_{u(t)\in U} \mathcal{H}(x(t), u(t), \lambda(t)) \forall t \in [0,T]$,

$$\dot{x}^*(t) = \nabla_\lambda \mathcal{H}(x^*(t), u^*(t), \lambda(t)),\ x^*(0) = x_0, \tag{6}$$

$$\dot{\lambda}(t) = -\nabla_x \mathcal{H}(x^*(t), u^*(t), \lambda(t)),\ \lambda(T) = \nabla_x \Phi(x^*(T)) \tag{7}$$

with the Hamiltonian $\mathcal{H}(x(t), u(t), \lambda(t)) = \lambda^T(t) \cdot f(x(t), u(t))$.

The discrete optimal control problem stated in (4) and (5) can be rewritten as the constrained optimization problem $\min_y \phi(y)$ s.t. $g(y) = 0$ with $y = (q_d, u_d)$, where $g(y) \in \mathbb{R}^p$ involves the forced DEL equations as well as the initial condition. If y^* is an optimal solution it has to fulfill the KKT equations given by

$$g(y^*) = 0 \tag{8}$$

$$\nabla\phi(y^*) - \nabla g(y^*)^T \lambda = 0 \tag{9}$$

with the Lagrange multiplier $\lambda \in \mathbb{R}^p$ assuming that $\nabla g(y^*)^T$ has full rank.

In optimal control theory an interesting question is, in which way the discrete necessary optimality conditions (8)-(9) approximate the continuous ones (6)-(7). Since the discretization (5) of the state system (6) was chosen, the consistency order is determined by the order of the variational integrator in use. In the same way, (9) can be seen as a discrete scheme for the adjoint system (7). Based on smoothness and coercivity conditions, the following was shown in [4]:

Theorem 2.1 (Consistency order of adjoint scheme)**.** *If the variational integrator scheme used for the discretization of the Lagrangian system is equivalent to a symplectic partitioned Runge-Kutta method, the resulting adjoint scheme is again a variational integrator scheme for the adjoint system, in particular both (state and adjoint) schemes are equivalent and of same order.*

The proof is based on existing results for standard Runge-Kutta methods [1, 2] applied to the corresponding Hamiltonian system and in addition, exploits the symplecticity conditions on the coefficients of the scheme.

3. Conclusion

Direct methods for the numerical solution of optimal control problems are based on a discretization of the state equation. It was shown that if a particular variational scheme (which is equivalent to a symplectic partitioned Runge-Kutta scheme) is used for this discretization, its symplecticity entails equivalent variational integrator schemes for state and adjoint system. For standard (non-symplectic) Runge-Kutta discretizations this is in general not true [2]. To obtain a general statement for any variational integrator, a proof directly on the Lagrangian side (rather than switching to the Hamiltonian side) is desirable.

References

[1] J. F. Bonnans and J. Laurent-Varin, *Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control*, Numerische Mathematik **103** (2006), 1-10.

[2] W. W. Hager, *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numerische Mathematik **87** (2000), 247-282.

[3] J. E. Marsden and M. West, *Discrete mechanics and variational integrators*, Acta Numerica **10** (2001), 357-514.

[4] S. Ober-Blöbaum, O. Junge and J. E. Marsden, *Discrete mechanics and optimal control: an analysis*, ESAIM: Control, Optimisation and Calculus of Variations. DOI: 10.1051/cocv/2010012 (2010).

Structure preserving integration of inequality constrained dynamics

Danny M. Kaufman

(joint work with Dinesh K. Pai, Eitan Grinspun)

Generalized equations of motion for a Hamiltonian system subject to combined unilateral and bilateral constraints can be expressed as a differential inclusion. When such constraints restrict system configuration, q, to a (generally nonconvex and nonsmooth) *admissible* set A, the Euler-Lagrange differential inclusion follows as

$$\mathsf{M}\ddot{\mathsf{q}} + \nabla V(\mathsf{q}) \in -\partial I_{\mathsf{A}}(\mathsf{q}) \; .$$

Here, and in the following, we reserve ∂ to denote the generalized gradient operator [4] while I_{A} gives the extended-value indicator function on the admissible set. Correspondingly the variational picture can be considered by composing the extended Lagrangian obtained by subtraction of I_A from the natural Lagrangian.

As in the standard smooth setting, these nonsmooth systems continue to preserve standard invariants, e.g., energy, momentum, phase-volume and the symplectic form. Structure-preserving numerical strategies have thus been applied to treat integration in this setting by considering both energy-momentum and symplectic-momentum preservation. In this latter vein Kane et al. [2] and Fetecau et al. [1] have observed that Discrete Variational Integration methods may be particularly well-suited for these purposes. In each such preliminary investigation, however, nonsmoothness has not been folded directly into the discrete action. Consequently it is observed that these methods experience drift and resultant instability [5, 3].

It is thus natural to consider a nonsmooth, discrete Variational Principle for inequality constrained systems. We begin by composing and extremizing a discrete, nonsmooth action that enforces hard constraints on all endpoints, $\{\mathsf{q}^0, ..., \mathsf{q}^k, ..., \mathsf{q}^N\}$, of the discrete trajectory with respect to discrete Lagrangian quadratures, L_d, obtaining

$$\delta_k \sum_{k=0}^{N-1} \Big(L_d(\mathsf{q}^k, \mathsf{q}^{k+1}) - I_{\mathsf{A}}(\mathsf{q}^{k+1}) \Big) \ni 0 \; .$$

Stationarity then gives us our constrained, Discrete Euler-Lagrange Inclusion (DELI),

$$D_2L_d(\mathsf{q}^{t-1},\mathsf{q}^t)+D_1L_d(\mathsf{q}^t,\mathsf{q}^{t+1})-\partial I_{\mathsf{A}}(\mathsf{q}^t)\ni 0\,,\quad \mathsf{q}^{t+1}\in\mathsf{A}\,. \tag{1}$$

We are then left with an interesting question of causality. We first observe that the above DELI system merely requires constraint forces to lie in the span of the normal cone generated by the configuration at time t and given by $-\partial I_{\mathsf{A}}(\mathsf{q}^t)$. Since this cone is independent of final configuration, (1) only specifies the span of possible constraint force directions but does not pin down a corresponding force magnitude.

Thus, in the inequality constrained setting, we observe that standard discrete variational structure is not sufficient, on its own, to define a well-posed integrator. We find that constraint forces must be applied along the directions positively spanned by the normal cone and, likewise, all inequalities must be satisfied, but fundamentally, nothing further is specified. In particular, an underdetermined system is composed with any number of solutions, most of which will not generate symplectic-momentum preserving maps. Our proposed DELI formulation thus provides a framework for numerical integration in addition to which further structure is required to compose a fully specified integration method.

To formulate the first such instantiation of a DELI-based method, we specifically note that the dichotomy between nonsmooth and smooth trajectories is not well-resolved by the discrete variational framework.

We start with the assumption that the admissible set is given explicitly by

$$\mathsf{A}:=\{\mathsf{q}:\ \ \mathsf{g}(\mathsf{q})=(g_0(\mathsf{q}),...,g_m(\mathsf{q}))^T\geq 0\},$$

and adopt a constraint subset notation. For each index subset, $\mathbb{K}\subset\{0,...,m\}$, define

$$\begin{aligned}
\mathsf{g}_{\mathbb{K}}(\mathsf{q})&:=\Big(\mathsf{g}_{k_0}(\mathsf{q}),...,\mathsf{g}_{k_l}(\mathsf{q})\Big)^T,\quad \{k_0,..,k_l\}\equiv\mathbb{K},\\
\mathsf{G}_{\mathbb{K}}(\mathsf{q})&:=\Big(\nabla\mathsf{g}_{k_0}(\mathsf{q})^T,...,\nabla\mathsf{g}_{k_l}(\mathsf{q})^T\Big)^T,\quad \{k_0,..,k_l\}\equiv\mathbb{K},\\
\mathsf{N}_{\mathbb{K}}(\mathsf{q})&:=\mathsf{G}_{\mathbb{A}\mathbb{K}}(\mathsf{q}),\quad \mathbb{A}\mathbb{K}=\big\{i:\ \ g_i(\mathsf{q})=0,\ \ i\in\mathbb{K}\big\}.
\end{aligned}$$

An absence of subscripting then defaults to the entire constraint set, e.g., $\mathsf{N}(\mathsf{q})=\mathsf{N}_{\{0,\dots,m\}}(\mathsf{q})$.

To formulate a DELI-based integrator we can consider the time-continuous structure of both smooth unilaterally constrained trajectories and the corresponding nonsmooth case. Imposing discrete analogues of both smooth-trajectory conditions and nonsmooth-jumps we obtain a DELI-based, Generalized Variational Integrator (GVI) whose time-forward, one-step map is given by sequencing through

a nonsmooth momentum update

$$\begin{aligned} \mathsf{p}^{t^+} &= \mathsf{p}^t + \mathsf{N}(\mathsf{q}^t)\lambda^+, \\ \mathsf{N}(\mathsf{q}^t)^T\mathsf{M}^{-1}\big(\mathsf{p}^t + \mathsf{N}(\mathsf{q}^t)\lambda^+\big) &\geq 0, \\ \lambda^+ &\geq 0, \\ E_d(\mathsf{q}^t, \mathsf{p}^t + \mathsf{N}(\mathsf{q}^t)\lambda^+) &= E_d(\mathsf{q}^t, \mathsf{p}^t), \end{aligned}$$

where E_d gives the numerical Hamiltonian; a smooth configuration update

$$\begin{aligned} D_1 L_d(\mathsf{q}^t, \mathsf{q}^{t+1}) + \mathsf{p}^{t^+} + \mathsf{N}_{\mathbb{S}(t)}(\mathsf{q}^t)\lambda^- &= 0, \\ 0 \leq \lambda^- \perp \ \mathsf{g}_{\mathbb{S}(t)}(\mathsf{q}^{t+1}) &\geq 0, \\ \mathbb{S}(t) = \Big\{ i : \quad \nabla g_i(\mathsf{q}^t)^T\mathsf{M}^{-1}\mathsf{p}^{t^+} = 0, \ \ g_i(\mathsf{q}^t) = 0, \ \ i \in \{0, .., m\} \Big\}; \end{aligned}$$

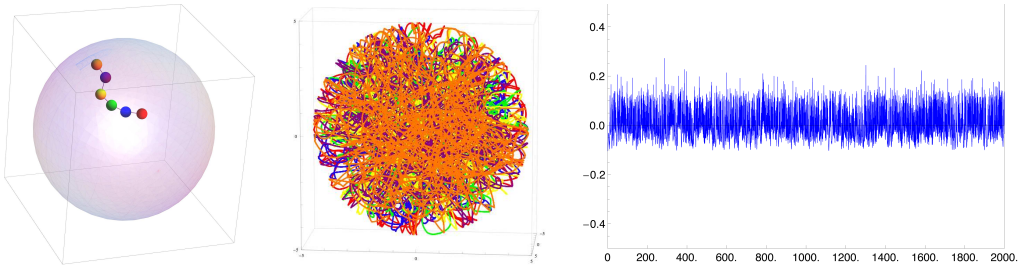
and a smooth momentum update

$$\begin{aligned} \mathsf{p}^{t+1} &= D_2 L_d(\mathsf{q}^t, \mathsf{q}^{t+1}) + \mathsf{N}_{\mathbb{S}(t)}(\mathsf{q}^{t+1}) \ \mu, \\ \mathsf{N}_{\mathbb{S}(t)}(\mathsf{q}^{t+1})^T\mathsf{M}^{-1}\mathsf{p}^{t+1} &= 0 \, . \end{aligned}$$

Definition (Active Set). At time t the *active set* is $\mathbb{A}(t) = \Big\{ i : g_i(\mathsf{q}^t) \leq 0, \ \ i \in \{0, .., m\} \Big\}$.

Definition (Discrete-Smooth Interval). An interval $[a, b]$ is *discrete-smooth* if for all $t \in [a, b]$ we have $\mathbb{S}(t) = \mathbb{A}(t)$.

Theorem. On discrete-smooth intervals GVI is symplectic-momentum preserving.



(a) (b) (c)

FIGURE 1. Numerical example of particles connected sequentially by rods interacting in a Lennard-Jones potential.

We then note that on intervals where discrete-smoothness does not hold we continue to observe good long-term behavior. To illustrate these properties we close by considering the numerical example of six spheres in $\mathbb{R}^3$, i.e., $\mathsf{q} = (\mathsf{q}_1^T, ..., \mathsf{q}_6^T)^T \in \mathbb{R}^{18}$, connected sequentially by bilateral rod constraints,

$$\mathsf{f}_i(\mathsf{q}) = \| \ \mathsf{q}_i - \mathsf{q}_{i+1} \ \| - \ l = 0.$$

Sphere-pairs are constrained by non-overlap conditions,

$$\mathsf{g}_{i,j}(\mathsf{q}) = \| \mathsf{q}_i - \mathsf{q}_j \| - 2r \geq 0,$$

interact in a Lennard-Jones potential,

$$V(\mathsf{q}) = \sum_{i \neq j} 4\epsilon \Big[\Big(\frac{\sigma}{\| \mathsf{q}_i - \mathsf{q}_j \|} \Big)^{12} - \Big(\frac{\sigma}{\| \mathsf{q}_i - \mathsf{q}_j \|} \Big)^{6} \Big],$$

and are further constrained to lie in a bounding sphere of radius r_B so that additionally,

$$\mathsf{g}_{i,b}(\mathsf{q}) = - \| \mathsf{q}_i \| - r + r_b \geq 0.$$

In our simulation we employ reduced time units, $(m\sigma^2/\epsilon)^{1/2}$, with the mass of all particles set to unity, $\sigma/2 = r = 0.3$, $r_b = 5$, and $l = 1$. Figure 1(a) shows a snapshot of the system configuration. Figure 1(b) traces out the entire trajectory computed by a Störmer-Verlet-based GVI of all particle centers over 2000 units of simulation time, stepped at $h = 10^{-2}$, while Figure 1(c) plots the change in energy over the same trajectory.

As illustrated in the above example, DELI can be further extended to simultaneously enforce bilateral constraints. For these details and additional materials see Kaufman and Pai [3].

References

[1] R. C. Fetecau, J. E. Marsden, M. Ortiz, and M. West, *Nonsmooth Lagrangian mechanics*, SIAM J. Appl. Dyn. Sys., **2**, (2003).

[2] C. Kane, E. Repetto, M. Ortiz, and J. Marsden, *Finite element analysis of nonsmooth contact*, Comp. Meth. Appl. Mech. Engrg., **180**, (1999).

[3] D. M. Kaufman and D. K. Pai, *Geometric numerical integration of inequality constrained, nonsmooth Hamiltonian systems*, arXiv:1007.2233, (2010).

[4] R. T. Rockafellar and R. J.B. Wets, *Variational Analysis*, Springer, (1998).

[5] D. E. Stewart, *Rigid-body dynamics with friction and impact*, SIAM Rev., **42**, (2000).

Discrete Dirac Mechanics and Discrete Dirac Geometry

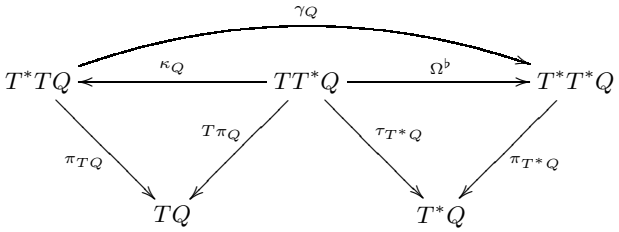
Melvin Leok

(joint work with Tomoki Ohsawa)

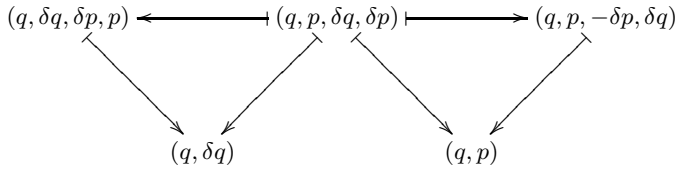
Dirac structures, which can be viewed as simultaneous generalizations of symplectic and Poisson structures, were introduced in [6, 7]. In the context of geometric mechanics [1, 2, 16], Dirac structures are of interest as they can directly incorporate Dirac constraints that arise in degenerate Lagrangian systems [10], LC circuits [4, 21, 24], interconnected systems [22], and nonholonomic systems [3], and thereby provide a unified geometric framework for studying such problems.

From the Hamiltonian perspective, these systems are described by implicit Hamiltonian systems [8, 21]. An implicit Hamiltonian system is defined by a Hamiltonian and a Dirac structure, which is a subbundle that satisfies certain

conditions. On the Lagrangian side, an implicit Lagrangian system [24] is defined by exploiting the geometric structure called the Tulczyjew's triple [19, 20] in addition to a Dirac structure. The Tulczyjew triple relates iterated tangent and cotangent bundles, and is given by,



or in local coordinates



Given a Lagrangian $L : TQ \to \mathbb{R}$, we define the *Dirac differential* $\mathfrak{D}L : TQ \to T^*T^*Q$ by

$$\mathfrak{D}L \equiv \gamma_Q \circ dL.$$

In local coordinates,

$$\mathfrak{D}L(q, v) = \left(q, \frac{\partial L}{\partial v}, -\frac{\partial L}{\partial q}, v \right).$$

Then, an *implicit Lagrangian system (ILS)* is defined by

$$(X, \mathfrak{D}L) \in D.$$

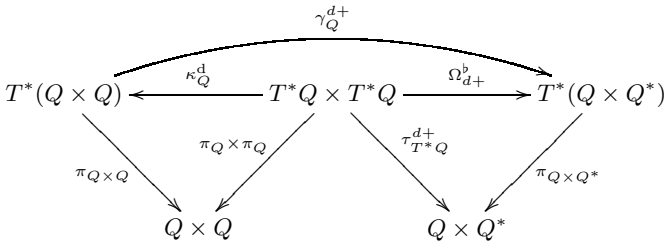
In particular, if D is the *induced Dirac structure* D_{Δ_Q}, given by,

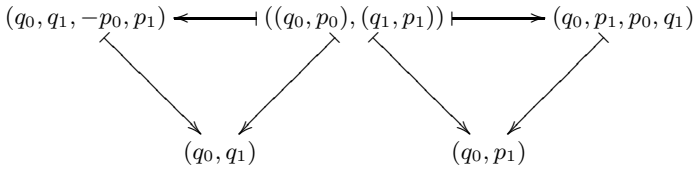
$$D_{\Delta_Q} \equiv \left\{ (v, \alpha) \in TT^*Q \oplus T^*T^*Q \,|\, v \in \Delta_{T^*Q}, \alpha - \Omega^\flat(v) \in \Delta^\circ_{T^*Q} \right\},$$

this reduces to

$$T\pi_Q(X) \in \Delta_Q, \qquad \Omega^\flat(X) - \mathfrak{D}L \in \Delta^\circ_{T^*Q},$$

A *discrete Tulczyjew's triple* is given by,





which provide the natural setting to define implicit discrete Lagrangian and Hamiltonian systems. The *discrete induced Dirac structure* is given by

$$D^{d+}_{\Delta_Q} \equiv \Big\{((z, z^+), \alpha_{\hat{z}}) \in (T^*Q \times T^*Q) \times T^*(Q \times Q^*) \Big| \\ (z, z^+) \in \Delta^d_{T^*Q},\ \alpha_{\hat{z}} - \Omega^\flat_{d+}((z, z^+)) \in \Delta^\circ_{Q\times Q^*} \Big\},$$

Then, an *implicit discrete Lagrangian system* is a triple (L_d, Δ^d_Q, X_d) with

$$\big(X^k_d, \mathfrak{D}^+ L_d(q_k, q^+_k)\big) \in D^{d+}_{\Delta_Q}.$$

In coordinates, we have the following *discrete implicit Lagrange–d'Alembert equations*:

$$(q_k, q_{k+1}) \in \Delta^d_Q, \qquad q_{k+1} = q^+_k,$$
$$p_{k+1} = D_2 L_d(q_k, q^+_k), \qquad p_k + D_1 L_d(q_k, q^+_k) \in \Delta^\circ_Q(q_k).$$

In particular, this recovers nonholonomic integrators [5, 18] that are typically derived from a discrete Lagrange–d'Alembert principle.

We also introduce discrete Lagrange–d'Alembert–Pontryagin and Hamilton–d'Alembert variational principles, that provide a variational characterization of implicit discrete Lagrangian and Hamiltonian systems that were described using discrete Dirac structures, and which reduce to the standard Lagrangian [17] and Hamiltonian [11, 14] variational integrators in the absence of constraints. Discrete Lagrangian, Hamiltonian, and nonholonomic mechanics have also been generalized to Lie groupoids [9, 15, 23]. A detailed exposition of the discrete geometric and variational structure of discrete Dirac mechanics can be found in [12, 13].

References

[1] R. Abraham and J. E. Marsden. *Foundations of Mechanics*. Addison–Wesley, 2nd edition, 1978.

[2] V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 1989.

[3] A. M. Bloch. *Nonholonomic Mechanics and Control*. Springer-Verlag, 2003.

[4] A. M. Bloch and P. E. Crouch. Representations of Dirac structures on vector spaces and nonlinear L-C circuits. In *Differential Geometry and Control Theory*, pages 103–117. American Mathematical Society, 1997.

[5] J. Cortés and S. Martínez. Non-holonomic integrators. *Nonlinearity*, 14(5):1365–1392, 2001.

[6] T. Courant. Dirac manifolds. *Transactions of the American Mathematical Society*, 319(2):631–661, 1990.

[7] T. Courant. Tangent Dirac structures. *Journal of Physics A: Mathematical and General*, 23(22):5153–5168, 1990.

[8] M. Dalsmo and A. J. van der Schaft. On representations and integrability of mathematical structures in energy-conserving physical systems. *SIAM Journal on Control and Optimization*, 37(1):54–91, 1998.

[9] D. Iglesias, J. C. Marrero, D. Martín Martín de Diego, and E. Martínez. Discrete nonholonomic Lagrangian systems on Lie groupoids. *J. Nonlinear Sci.*, 18(3):221–276, 2008.
[10] H. P. Künzle. Degenerate Lagrangean systems. *Annales de l'institut Henri Poincaré (A)*, 11(4):393–414, 1969.
[11] S. Lall and M. West. Discrete variational Hamiltonian mechanics. *Journal of Physics A: Mathematical and General*, 39(19):5509–5519, 2006.
[12] M. Leok and T. Ohsawa. Discrete dirac structures and implicit discrete lagrangian and hamiltonian systems. *AIP Conference Proceedings*, 1260(1):91–102, 2010.
[13] M. Leok and T. Ohsawa. Discrete Dirac structures and variational discrete Dirac mechanics. *Foundations of Computational Mathematics*, 2011. (accepted).
[14] M. Leok and J. Zhang. Discrete Hamiltonian variational integrators. *IMA J. Nuer. Anal*, 2011. (accepted).
[15] J. C. Marrero, D. Martín de Diego, and E. Martínez. Discrete Lagrangian and Hamiltonian mechanics on Lie groupoids. *Nonlinearity*, 19(6):1313–1348, 2006.
[16] J. E. Marsden and T. S. Ratiu. *Introduction to Mechanics and Symmetry*. Springer-Verlag, 1999.
[17] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, pages 357–514, 2001.
[18] R. McLachlan and M. Perlmutter. Integrators for nonholonomic mechanical systems. *Journal of Nonlinear Science*, 16(4):283–328, 2006.
[19] W. M. Tulczyjew. Les sous-variétés lagrangiennes et la dynamique hamiltonienne. *C. R. Acad. Sc. Paris*, 283:15–18, 1976.
[20] W. M. Tulczyjew. Les sous-variétés lagrangiennes et la dynamique lagrangienne. *C. R. Acad. Sc. Paris*, 283:675–678, 1976.
[21] A. J. van der Schaft. Implicit Hamiltonian systems with symmetry. *Reports on Mathematical Physics*, 41(2):203–221, 1998.
[22] A. J. van der Schaft. Port-Hamiltonian systems: an introductory survey. In *Proceedings of the International Congress of Mathematicians*, volume 3, 2006.
[23] A. Weinstein. Lagrangian mechanics and groupoids. In *Mechanics day (Waterloo, ON, 1992)*, volume 7 of *Fields Inst. Commun.*, pages 207–231. Amer. Math. Soc., Providence, RI, 1996.
[24] H. Yoshimura and J. E. Marsden. Dirac structures in Lagrangian mechanics Part I: Implicit Lagrangian systems. *Journal of Geometry and Physics*, 57(1):133–156, 2006.

Bilinear discretization of quadratic vector fields

Yuri B. Suris

This talk was devoted to the remarkable properties of the following discretization method applicable to any system of ODEs with quadratic vector fields:

$$\dot{x} = Q(x) + Bx + c \quad \rightsquigarrow \quad (x_{k+1} - x_k)/\epsilon = Q(x_k, x_{k+1}) + B(x_k + x_{k+1})/2 + c,$$

where $B \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$, each component of $Q : \mathbb{R}^n \to \mathbb{R}^n$ is a *quadratic* form, and $Q(x, y) = (Q(x+y) - Q(x) - Q(y))/2$ is the corresponding symmetric *bilinear* function. This discretization method goes back to W. Kahan [1].

It is important to observe that equations for x_{k+1} are linear, hence their solution by the Cramer's rule leads to a map $x_{k+1} = f(x_k, \epsilon)$ which is rational. Moreover, due to the obvious symmetry of the above difference equation with respect to the flip $x_k \leftrightarrow x_{k+1}$ accompanied by the change of sign $\epsilon \to -\epsilon$, one has the following reversibility:

$$f^{-1}(x, \epsilon) = f(x, -\epsilon).$$

so that the map f is birational.

Kahan illustrated his method by the famous Lotka-Volterra system, for which his discretization produces non-spiralling solutions, unlike most of the convenient integrators. This behavior was partly explained by J. Sanz-Serna [2] who demonstrated that Kahan's integrator for the Lotka-Volterra system is symplectic (Poisson).

In this talk, we are mainly interested in application of the Kahan's discretization method to algebraically integrable systems, which was initiated by R. Hirota and K. Kimura [3, 4] (for the Euler and the Lagrange tops). In this context, we refer to this method as the Hirota-Kimura (HK) discretization. My interest in this topic was stimulated by the talk given by T. Ratiu at the Oberwolfach Workshop "Geometric Integration" in March 2006, who claimed that HK-type discretizations for the Clebsch system and for the Kovalevsky top are also integrable. The HK discretization of the Euler top was demonstrated to be a bi-Hamiltonian system in [5]. Integrability of the HK discretization of the Clebsch system was actually proven in [6]. There are indications that the HK discretization of the Kovalevsky top is non-integrable, although a rigorous proof of such a statement seems to be a very difficult enterprize.

Our analysis of the work by Hirota-Kimura (which is not easy to put on a firm mathematical base) and our study of the HK discretization of the Clebsch system led us to the formulation of the following definition.

Definition. *For a given birational map $f : \mathbb{R}^n \to \mathbb{R}^n$, a set of functions $\Phi = (\varphi_1, \ldots, \varphi_l)$, linearly independent over $\mathbb{R}$, is called a* **HK-basis**, *if for every $x_0 \in \mathbb{R}^n$ there exists a vector $c = (c_1, \ldots, c_l) \neq 0$ such that*

$$c_1\varphi_1(f^i(x_0)) + \ldots + c_l\varphi_l(f^i(x_0)) = 0 \quad \forall i \in \mathbb{Z}.$$

For a given $x_0 \in \mathbb{R}^n$, the set of all vectors $c \in \mathbb{R}^l$ with this property will be denoted by $K_\Phi(x_0)$ and called the null-space of the basis Φ (at the point x_0). This set clearly is a vector space.

HK bases seem to be novel mathematical objects, closely related to but clearly different from integrals of motion. For instance, in the above definition we cannot claim that $h = c_1\varphi_1 + ... + c_l\varphi_l$ is an integral of motion, since vectors $c \in K_\Phi(x_0)$ vary from one initial point x_0 to another. However, existence of a HK-basis Φ with $\dim K_\Phi(x_0) = d$ has similar dynamical consequences as existence of d independent integrals of motion: in both cases the orbits of f are confined to $(n-d)$-dimensional invariant sets. Moreover, if Φ is a HK-basis for a map f, then $K_\Phi(f(x_0)) = K_\Phi(x_0)$ and thus the d-dimensional null-space $K_\Phi(x_0)$ is a $Gr(d, l)$-valued integral. Its Plücker coordinates are then scalar integrals. Especially simple is the situation when the null-space of a HK-basis has dimension $d = 1$:

Fact. *Let Φ be a HK-basis for f with $\dim K_\Phi(x_0) = 1$ for all $x_0 \in \mathbb{R}^n$. Let $K_\Phi(x_0) = [c_1(x_0) : \ldots : c_l(x_0)] \in \mathbb{RP}^{l-1}$. Then the functions c_j/c_k are integrals of motion for f.*

The main difficulties in the rigorous proof of integrability of HK discretizations, like for the Clebsch system in [6], come from the necessity of a symbolical solution of linear systems whose coefficients are composed of higher iterates of the map

f. However, the complexity of this map precludes any simple-minded attempts in this direction. For instance, for the Clebsch case the complexity of the map f is best described by the following data: the common denominator and the numerators of components of f are polynomials of degree 6 in six variables, the common denominator consists of 28 monomials, the numerators consist of up to 41 monomials. For the second iterate f^2, the common denominator of its components has total degree 27, with degree 24 in each of variables, while the numerators are of total degree 33, with degree 28 in each of variables. Thus, even a naive computation and storage of f^2 is nearly impossible with the current software and hardware. One can consult [6] for some tricks which allowed us to overcome these difficulties.

In [7], we give an overview of results on the integrability of HK discretizations available at present. An (incomplete) list of examples includes:

- Periodic Volterra chain of period $N = 3, 4$;
- Dressing chain with $N = 3$ particles;
- Euler and Lagrange cases of the heavy top;
- Kirchhof and Clebsch cases of the rigid body motion in an ideal fluid;
- Some cases of the Zhukovsky-Volterra gyrostat;
- Three-wave interaction system;
- System of two interacting Euler tops.

For the most complicated cases, our proofs are computer assisted. We did not find a general structure, which would provide us with less computational proofs and with more insight. In particular, nothing like a Lax representation has been found. Nothing is known about the existence of an invariant Poisson structure for these maps, although an invariant volume measure has been found for all of them.

In [6], we pushed forward the following conjecture:

Conjecture. *For any algebraically completely integrable system with a quadratic vector field, its Hirota-Kimura discretization remains algebraically completely integrable .*

However, at present we have a number of apparent counterexamples (it is extremely difficult to prove non-integrability). Nevertheless, the HK discretization preserves integrability much more often than a mere coincidence would allow.

The full story still waits to be clarified.

References

[1] W. Kahan, *Unconventional numerical methods for trajectory calculations*, Unpublished lecture notes (1993).

[2] J.M. Sanz-Serna, *An unconventional symplectic integrator of W.Kahan,* Applied Numer. Math. **16** (1994) 245–250.

[3] R.Hirota, K.Kimura, *Discretization of the Euler top,* J. Phys. Soc. Japan **69** (2000) 627–630.

[4] K.Kimura, R.Hirota, *Discretization of the Lagrange top,* J. Phys. Soc. Japan **69** (2000) 3193–3199.

[5] M. Petrera, Yu. Suris, *On the Hamiltonian structure of the Hirota-Kimura discretization of the Euler top,* Math. Nachr., 2010, **283**, 1654–1663.

[6] M. Petrera, A. Pfadler, Yu. Suris, *On integrability of Hirota-Kimura type discretizations. Experimental study of the discrete Clebsch system,* Experimental Math., 2009, **18**, 223–247.
[7] M. Petrera, A. Pfadler, Yu. Suris, *On integrability of Hirota-Kimura type discretizations,* Regular Chaotic Dyn. (to appear), `arXiv:1008.1040 [nlin.SI]`.

Solving KdV (and Painlevé II) numerically using Riemann–Hilbert problems

Sheehan Olver
(joint work with Tom Trogden)

Integrable systems are of immense importance in mathematics, quantum physics, molecular dynamics and elsewhere. In addition to other important properties, many integrable systems can be reformulated as Riemann–Hilbert (RH) problems. A classical example is the Korteweg–de Vries (KdV) equation,

$$u_t + 6uu_x + u_{xxx} = 0, \qquad u(0,x) = u_0(x),$$

for x on the real line. Associated to the initial data is the scattering data: given initial condition u_0, we can compute the *reflection coefficient* $r(k)$ and the *soliton data* $(\kappa_1, \gamma_1), \dots (\kappa_N, \gamma_N)$. From the scattering data, we can construct an RH problem which is equivalent to solving KdV [6]: find a function Φ, meremorphic off $\mathbb{R}$ with only simple poles at $\pm i\kappa_1, \dots, \pm i\kappa_N$, which satisfies

$$\Phi^+(k) = \Phi^-(k) \begin{pmatrix} 1 - |r(k)|^2 & -\bar{r}(k) e^{-2i(4tk^3+xk)} \\ r(k) e^{2i(4tk^3+xk)} & 1 \end{pmatrix} \qquad \text{for } x \in \mathbb{R},$$

$$\operatorname{Res}_{i\kappa_j} \Phi(k) = \Phi(i\kappa_j) \begin{pmatrix} 0 & 0 \\ i\gamma_k^2 e^{-2(-4t\kappa_j^3+x\kappa_j)} & 0 \end{pmatrix} \qquad \text{for } j = 1, \dots, N,$$

$$\operatorname{Res}_{-i\kappa_j} \Phi(k) = \Phi(-i\kappa_j) \begin{pmatrix} 0 & -i\gamma_k^2 e^{-2(-4t\kappa_j^3+x\kappa_j)} \\ 0 & 0 \end{pmatrix},$$

$$\Phi(\infty) = (1,1),$$

where Φ^+ and Φ^- are the limits as k approaches the real line from above and below, respectively. The goal of this talk is to demonstrate that KdV can be solved accurately for all values of x and t, by solving this RH problem.

We begin by demonstrating the approach used in [8] for solving the RH problem associated with the Painlevé II transcendent. This RH problem consists of finding a 2×2 matrix function Φ which is analytic off a curve Γ consisting of six rays in the complex plane, satisfying

$$\Phi^+(k) = \Phi^-(k) G(k) \quad \text{for } k \in \Gamma \text{ and } \Phi(\infty) = I, \tag{1}$$

where G is a function depending on x and the three Stokes' constants s_1, s_2, s_3 which satisfy

$$s_1 - s_2 + s_3 + s_1 s_2 s_3 = 0.$$

Using the Cauchy transform

$$\mathcal{C}f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(x)}{x-z}dx, \tag{2}$$

we can reduce (1) to a problem on the contour Γ: letting $\Phi = I + \mathcal{C}U$, (1) implies that U satisfies

$$(\mathcal{C}U)^+ - (\mathcal{C}U)^- G = G - I \tag{3}$$

on Γ. Assuming $\mathcal{C}$ can be computed reliably, including its left and right limits on Γ, we can solve (3) using a collocation/spectral method: represent

$$U \approx \sum c_k \psi_k$$

for an appropriate basis ψ_k, determining the (matrix) coefficients c_k by ensuring that (3) is satisfied at a sequence of points x_j on Γ. In other words, we solve the linear system

$$\sum c_k \left\{(\mathcal{C}\psi_k)^+(x_j) - [(\mathcal{C}\psi_k)^-(x_j)]G(x_j)\right\} = G(x_j) - I.$$

The key tool we require is computation of Cauchy transforms. If Γ is the unit interval, then one can compute the Chebyshev expansion of f, reducing the problem to that of computing the moments $\mathcal{C}T_k$, where T_k is the kth Chebyshev polynomial. For these moments, we can use the closed form formula derived in [7], resulting in an approximation to (2) which converges uniformly throughout the complex plane, including along the branch cut itself. This can be extended to each of the constituent pieces of Γ via conformal maps, therefore, we choose the basis ψ_k to be piecewise conformally mapped Chebyshev polynomials. In other words, the computation of $\mathcal{C}\psi_k$ is exact.

One last aspect is crucial to the accuracy of the method: all junction points of Γ must be included in the collocation system. Though the Cauchy transform of a piecewise mapped Chebyshev polynomial explodes at its junction points, we know precisely what the contribution to the Cauchy transform would be if this blow up were somehow cancelled by Cauchy transforms over the connecting curves. We take this as the value of the Cauchy transform at such points in the collocation system. Fortunately, this convention ensures that the resulting approximation of U is bounded, which in turn justifies the special definition chosen for the Cauchy transform.

The resulting numerical method converges spectrally fast, i.e., the error decays faster than any inverse polynomial in the number of collocation points used. This compares favourably with previous attempts at solving RH problems numerically, such as [4], where exponentially many points were needed near junction points to simulate boundedness.

There are two issues with the numerical approximation: the jump curve G becomes increasingly oscillatory for x far away from the origin, and the collocation system becomes badly conditioned. Both issues can be avoided by using *nonlinear steepest descent* [3]: the contour Γ can be deformed in the complex plane to transform oscillations into exponential decay, the jump matrix G can be factored

and any jump matrix which is diagonal or skew-diagonal can be removed via a *parametrix*. By using these techniques, we can obtain an RH problem which is numerically useful, and the resulting approximation remains accurate for large and small x.

To adapt this approach for the KdV RH problem, we must be able to handle the poles at $\pm i\kappa_j$. This means the solution can no longer be represented by a Cauchy transform over $\mathbb{R}$. However, we can represent the solution by such a Cauchy transform as well as the functions

$$\frac{1}{z - i\kappa_j},$$

and append the conditions on the residues to the collocation system. In different regimes of x and t we obtain different deformations, however, because we are solving the resulting RH problem numerically (rather than asymptotically), the validity of the deformations actually overlaps. Thus, for any x and t, even very large t, we can compute the solution to KdV to arbitrary accuracy, and each evaluation is independent: we do not need to time-step. On the other hand, because KdV exhibits very harsh oscillations in the dispersive tail, any traditional numerical method would rapidly break down.

References

[1] M.J. Ablowitz, P.A. Clarkson *Solitons, Evolution Equations and Inverse Scattering*, Cambridge University Press, 1991.

[2] P. Deift, *Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert Approach*, AMS, 2000.

[3] P. Deift, X. Zhou, *Asymptotics for the Painlevé II equation*, Comm. Pure Appl. Maths, **48** (1995), 277.

[4] A. Dienstfrey *The Numerical Solution of a Riemann-Hilbert Problem Related to Random Matrices and the Painlevé V ODE*, Courant Institute of Mathematical Sciences, 1998.

[5] A.S. Fokas, A.R. Its, A.A. Kapaev, V.Y. Novokshenov, *Painlevé transcendents: the Riemann-Hilbert approach* AMS, 2006.

[6] K. Grunert, G. Teschl, *Long-time asymptotics for the Korteweg–de Vries equation via nonlinear steepest descent*, Math. Phys., Anal. Geom., **12** (2009), 287–324.

[7] S. Olver, *Computing the Hilbert transform and its inverse*, to appear in Maths Comp.

[8] S. Olver, *Numerical solution of Riemann–Hilbert problems: Painlevé II* Found. Comput. Maths **11** (2011), 120–140.

On energy preserving integrators for polynomial Hamiltonians

Elena Celledoni

(joint work with B. Owren and Y. Sun)

We shall be concerned with canonical Hamiltonian systems

$$y' = J^{-1}\nabla H(y) = f(y), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{1}$$

The numerical solution of problems of the this type has been treated extensively in the literature, we refer to the monographs [4, 8] and the references therein for

details. Important properties of the system (1) are symplecticity and reversibility of the flow map, conservation of the Hamiltonian $H(y)$ along any solution $y(t)$ and volume preservation. The circumstances under which various numerical integrators inherit these properties are by now fairly well understood. For example Runge–Kutta methods can preserve symplecticity and reversibility, but no Runge–Kutta method can be energy-preserving for any Hamiltonian function H and no B-series can be volume preserving for any system (1).

In the present work we focus on the preservation of the Hamiltonian itself, we study integrators generating a sequence of approximations $\{y_n\}$ to the solution of (1) such that $H(y_n) = H(y_0)$ for all $n \geq 1$, and in particular, what can be achieved when the Hamiltonian is polynomial and the integrator is a Runge–Kutta method. For linear Hamiltonians, the resulting ODE is constant and any consistent Runge–Kutta scheme will reproduce the exact solution. If the Hamiltonian is a quadratic, then the resulting ODE is linear, and the condition for preserving energy is that the stability function of the method satisfies $R(z)R(-z) = 1$. For polynomials of higher order it is not known to which extent Runge–Kutta methods can preserve the Hamiltonian. However, it was observed in [9] that the Averaged Vector Field (AVF) method, defined as

$$y_{n+1} = y_n + h \int_0^1 f((1-\xi)y_n + \xi y_{n+1})\,\mathrm{d}\xi \tag{2}$$

preserves the Hamiltonian for all problems of the form (1). The AVF method has second order convergence. In particular, when the Hamiltonian is a polynomial, the integral can be exactly resolved, the same result is obtained if the integral in (2) is replaced by a quadrature rule of sufficiently high order. This was observed in [1]. In fact, using a standard linear formula with abscissae $c = (c_1, \ldots, c_s)^T$ and weights $b = (b_1, \ldots, b_s)^T$, the result is a Runge–Kutta method in which the Butcher matrix is given as $A = cb^T$. This immediately shows that for any polynomial Hamiltonian system, there exist Runge–Kutta methods which exactly preserve the energy, see [6] for a different proof of this result. Note also that any choice of quadrature rule of sufficiently high order yields the same approximation, the AVF method is reproduced exactly.

As pointed out in [1] any energy-preserving integrator for (1) must obey all quadrature conditions, but for polynomial systems this can be relaxed. Letting the Hamiltonian be a polynomial of degree m, a necessary condition for the energy to be preserved is that the quadrature conditions hold up to order m, or in terms of Runge–Kutta coefficients

$$\sum_i b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, m. \tag{3}$$

Thus, in considering energy preserving Runge–Kutta methods for polynomial Hamiltonians of degree $\leq m$ one may immediately restrict the focus to schemes whose coefficients satisfy (3). If $m = 2s$ then the smallest possible number of stages in the scheme is s the resulting abscissae and weights are those of the Gauss–Legendre quadrature rule. If $m = 2s-1$ then the smallest possible number

of stages is still s, but the quadrature rule is not uniquely given. One can prove the following theorem.

Theorem 0.1. *Let $m \geq 3$. Among all Runge–Kutta methods which exactly preserve all polynomial Hamiltonians of degree at most m, those with the minimal number of stages coincide with the AVF method when applied to such problems. The number of stages in these methods is $\lfloor (m+1)/2 \rfloor$.*

In general, there are energy preserving Runge–Kutta methods for polynomial Hamiltonian systems which do not coincide with the AVF-integrator. There also exist such methods of arbitrarily high order [7], see also [5]. Examples are easily obtained as composition methods based on the AVF-integrator.

The proof makes use of results from [3] and [2] and the W-transform.

References

[1] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, G. R. W. Quispel, and W. M. Wright, *Energy-preserving Runge–Kutta methods*, M2AN (Mathematical Modelling and Numerical Analysis) **43** (2009), 645-649.

[2] P. Chartier, E. Faou, and A. Murua, *An algebraic approach to invariant preserving integrators: The case of quadratic and Hamiltonian invariants*, Numer. Math.**103** (2006), 575–590.

[3] E. Faou, E. Hairer, and T.–L. Pham, *Energy conservation with non-symplectic methods: examples and counter-examples*, BIT **44** (2004), 699–709.

[4] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, Berlin, 2nd ed., 2006.

[5] E. Hairer, *Energy-preserving variant of collocation methods*, J. Numer. Anal. Ind. Appl. Math. **5** (2010) 73-84.

[6] F. Iavernaro and B. Pace, *s-Stage trapezoidal methods for the conservation of Hamiltonian functions of polynomial type*, AIP Conf. Proc. **936** (2007).

[7] F. Iavernaro and D. Trigiante, *High-order symmetric schemes for energy conservation of polynomial Hamiltonian problems*, J. Numer. Anal. Ind. Appl. Math. **4** (2009), 87?101..

[8] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monographs on Applied and Computational Mathematics, **14**. Cambridge University Press, Cambridge, 2004.

[9] G. R. W. Quispel and D. I. McLaren, *A new class of energy-preserving numerical integration methods*, J. Phys A **41**, 2008, 7 pp.

Collocation-like methods with conservation properties for the numerical integration of Hamiltonian systems

Felice Iavernaro

(joint work with Luigi Brugnano, Donato Trigiante)

Summary. The present report sketches the theory of Hamiltonian Boundary Value Methods (HBVMs), a class of Runge–Kutta methods, able to preserve the energy function of polynomial Hamiltonian canonical systems. A documentation, Matlab codes, and a complete set of references is available at the url [5].

Definition of the problem. We consider canonical Hamiltonian systems

$$\dot{y} = J\nabla H(y), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix},$$

and we assume that the Hamiltonian function $H(y)$ is a polynomial. Our aim is to define one-step methods $y_{n+1} = \Phi_h(y_n)$ (h is the stepsize of integration) that conserve the Hamiltonian function: $H(y_{n+1}) = H(y_n)$ for all n and $h > 0$.

The genesis. The first instance of conservative Runge–Kutta methods dates back to 2007 [8] and takes the form of a generalization of the trapezoidal method[1]

$$y_{n+1} - y_n = \frac{1}{2} J \left(\nabla H(y_n) + \nabla H(y_{n+1}) \right).$$

Multiplication of both sides by $\left(\nabla H(y_n) + \nabla H(y_{n+1})\right)^T$ yields

$$\left(\nabla H(y_n) + \nabla H(y_{n+1})\right)^T (y_{n+1} - y_n) = 0. \tag{1}$$

For quadratic Hamiltonian functions this is tantamount to the conservation law $H(y_{n+1}) = H(y_n)$. To see this, consider the segment σ joining y_n to y_{n+1}, namely $\sigma(t_0 + ch) = (1-c)y_n + cy_{n+1}$, with $c \in [0,1]$, and the line integral

$$\begin{aligned} H(y_{n+1}) - H(y_n) = \int_{y_n \to y_{n+1}} \nabla H(y)\,\mathrm{d}y &= \int_0^1 \dot{\sigma}(t_0+ch)^T \nabla H(\sigma(t_0+ch))\,\mathrm{d}c \\ &= h(y_{n+1}-y_n)^T \int_0^1 \nabla H(\sigma(t_0+ch))\,\mathrm{d}c \qquad (2) \\ &= \frac{1}{2} h (y_{n+1}-y_n)^T \left(\nabla H(y_n) + \nabla H(y_{n+1})\right), \qquad (3) \end{aligned}$$

where (3) comes from the application of the trapezoidal rule for the exact evaluation of the integral in (2), $\nabla H(y)$ being linear. Thus (1) means $H(y_{n+1}) = H(y_n)$.

Now suppose that $\deg H(y) = \nu$: the integrand in (2) has degree $\nu - 1$ and can be solved by any quadrature formula, with abscissae $c_1 < c_2 < \cdots < c_k$ in $[0,1]$ and weights $b_1, \ldots, b_k$, having degree of precision $d \geq \nu - 1$. In place of (3) we now obtain

$$H(y_{n+1}) - H(y_n) = h(y_{n+1}-y_n)^T \sum_{i=1}^{k} b_i \nabla H(Y_i), \qquad Y_i = \sigma(t_0 + c_i h),\ i = 1, \ldots, k.$$

This suggests the method

$$y_{n+1} = y_n + h \sum_{i=1}^{k} b_i f(Y_i), \qquad Y_i = (1-c_i)y_n + c_i y_{n+1}, \quad i = 1, \ldots, k. \tag{4}$$

The stages Y_i are called *silent stages* since their presence does not affect the degree of nonlinearity of the resulting R-K method: (4) is a mono-implicit method. The Butcher tableau associated with these methods is $\begin{array}{c|c} c & cb^T \\ \hline & b^T \end{array}$, where c and b are

[1] Any other Runge–Kutta method yielding energy conservation when $H(y)$ is quadratic, such as the implicit midpoint method, may be used to sketch the idea.

the vectors of the abscissae and weights respectively. Routine computation shows that each method has order two and is symmetric. The table below summarizes the behavior of (4) for some choices of the abscissae distribution.

Abscissae distribution	Newton-Cotes	Lobatto	Gauss
Energy preserving when	$\deg H \le k, k+1$	$\deg H \le 2k-2$	$\deg H \le 2k$

Higher order methods. Higher order energy-preserving Runge–Kutta methods have been obtained by considering a polynomial curve σ of higher degree in the phase space, along which to evaluate the line integral [9, 10]. Starting from [1], it was understood that to maximize the order of the resulting methods, the use of the Legendre polynomial basis $\{P_j(c)\}_{j=1,\dots s}$ shifted on the interval $[0,1]$ was in order ($P_j(c)$ denotes the polynomial of degree $j-1$). A polynomial $\sigma(t)$ of degree s, with $\sigma(t_0) = y_0$, is defined on $[t_0, t_0+h]$ by means of the expansion

$$\dot\sigma(t_0+ch) = \sum_{j=1}^{s} \gamma_j P_j(c) \Longrightarrow Y_i \equiv \sigma(t_0+c_ih) = y_0 + h\sum_{j=1}^{s}\gamma_j \int_0^{c_i} P_j(x)\,\mathrm{d}x, \tag{5}$$

where the (vector) coefficients $\{\gamma_j\}$ are to be regarded as unknowns and $y_1 = \sigma(t_0+h)$ is meant to yield the approximation to $y(t_0+h)$. We have

$$H(y_1)-H(y_0) = \int_{t_0}^{t_0+h} (\dot\sigma(t))^T \nabla H(\sigma(t))\mathrm{d}t = h\sum_{j=1}^{s}\gamma_j^T \int_0^1 P_j(c)\nabla H(\sigma(t_0+ch))\mathrm{d}c$$

which vanishes by imposing, for $j = 1, \dots, s$, the following orthogonality conditions

$$\gamma_j = \eta_j \int_0^1 P_j(\tau) J\nabla H(\sigma(t_0+ch))\mathrm{d}c = \eta_j \sum_{\ell=1}^{k} b_\ell P_j(c_\ell) J\nabla H(\sigma(t_0+c_\ell h)). \tag{6}$$

where, after observing that the integrand function has degree $\nu s - 1$, to obtain the rightmost equality, we have assumed that $d \ge \nu s - 1$. The scalars η_j are suitable nonzero scaling factors that make the resulting method consistent: $\eta_j = \left(\int_0^1 P_j^2(x)\mathrm{d}x\right)^{-1}$, $j = 1, \dots, s$ (and hence $\eta_j = 2j-1$ in case of the Legendre basis). Substituting (6) into the right formula in (5) yields the following R-K method

(7)

$$\left\{\begin{array}{l} Y_i = y_0 + h\sum_{j=1}^{s}\eta_j \int_0^{c_i} P_j(x)\mathrm{d}x \sum_{\ell=1}^{k} b_\ell P_j(c_\ell)\, f(Y_\ell), \\ i = 1, \dots, s, \\ \\ y_1 = y_0 + h\sum_{i=1}^{k} b_i Y_i, \end{array}\right. \quad \Longleftrightarrow \quad \begin{array}{c|c} c_1 & \\ \vdots & \mathcal{I}\Lambda\mathcal{P}^T\Omega \\ c_k & \\ \hline & b_1 \ \dots \ b_k \end{array}$$

where the matrices $\mathcal{I}, \mathcal{P} \in \mathbb{R}^{k\times s}$ and Λ, Ω are defined as

$$\mathcal{I}_{ij} = \int_0^{c_i} P_j(x)\mathrm{d}x, \quad \mathcal{P}_{ij} = P_j(c_i), \qquad \begin{array}{l}\Lambda = \mathrm{diag}(\eta_1, \dots, \eta_s), \\ \Omega = \mathrm{diag}(b_1, \dots, b_k).\end{array}$$

The Runge–Kutta method (7) has order $2s$ for all $d \ge 2s-1$, is symmetric and precisely A-stable, becomes the Gauss–Legendre method of order $2s$ when $k = s$, is

energy preserving when applied to canonical polynomial Hamiltonian systems with Hamiltonian function $H(y)$ of degree $\nu \leq \frac{d+1}{s}$ (if the c_i are the Gauss abscissae then $\nu \leq \frac{2k}{s}$). An interesting link to standard collocation methods (see [3]) is that the Butcher array in (7) may be put in the form $A(\widehat{\mathcal{P}}\widehat{\mathcal{P}}^T\Omega)$, where $\widehat{\mathcal{P}} = \mathcal{P}\Lambda^{1/2}$ and A is the Butcher array associated with a standard collocation method defined on the same abscissae c_i and weights b_i, provided $d \geq 2s-1$. Methods such as (7) are denoted by HBVM(k,s), where k is the total number of stages and s is the degree of the underlying polynomial σ. The block-BVM notation uncouples the linear and nonlinear parts of (7) thus making the method suitable for implementation.

∞-HBVMs [2]. These are the limit of HBVM(k,s), as $k \to \infty$, and are obtained by simply retaining the integrals instead of the sums in (6) (see [2]). The ∞-HBVM of order 2 becomes the averaged vector field [11, 6], while limit formulae based upon the Lagrange basis were first proposed by E. Hairer [7].

HBVMs and Fourier expansions [4]. Expanding f along the orthonormal Legendre basis and truncating after the first s terms yields a continuous perturbed problem with respect to which the HBVMs are standard collocation methods.

References

[1] L. Brugnano, F. Iavernaro, D. Trigiante, *Analisys of Hamiltonian Boundary Value Methods (HBVMs): a class of energy-preserving Runge–Kutta methods for the numerical solution of polynomial Hamiltonian dynamical systems*, *Preprint* (2009), (`arXiv:0909.5659`).

[2] L. Brugnano, F. Iavernaro, D. Trigiante, *Hamiltonian Boundary Value Methods (Energy Preserving Discrete Line Integral Methods)*, Jour. of Numer. Anal., Industr. and Appl. Math. **5** (no. 1-2) (2010), 17–37, (`arXiv:0910.3621`).

[3] L. Brugnano, F. Iavernaro, D. Trigiante, *The Lack of Continuity and the Role of Infinite and Infnitesimal in Numerical Methods for ODEs: the Case of Symplecticity*, Applied Mathematics and Computation (to appear), (arXiv:1010.4538).

[4] L. Brugnano, F. Iavernaro, D. Trigiante, *A unifying framework for the derivation and analysis of effective classes of one-step methods for ODEs*, *Preprint*, 2010, (`arXiv:1009.3165`).

[5] L. Brugnano, F. Iavernaro, D. Trigiante, *The Hamiltonian BVMs (HBVMs) Homepage*, `arXiv:1002.2757` (URL: `http://web.math.unifi.it/users/brugnano/HBVM/`).

[6] E. Celledoni, R.I. McLachlan, D. McLaren, B. Owren, G.R.W. Quispel, W.M. Wright, *Energy preserving Runge–Kutta methods*, M2AN **43** (2009), 645–649.

[7] E. Hairer, Energy-preserving variant of collocation methods, *J. Numer. Anal. Ind. Appl. Math.*, **5**,1-2 (2010).

[8] F. Iavernaro, B. Pace, *s-Stage Trapezoidal Methods for the Conservation of Hamiltonian Functions of Polynomial Type*, AIP Conf. Proc. **936** (2007), 603–606.

[9] F. Iavernaro, B. Pace, *Conservative Block-Boundary Value Methods for the Solution of Polynomial Hamiltonian Systems*, AIP Conf. Proc. **1048** (2008), 888–891.

[10] F. Iavernaro, D. Trigiante, *High-order symmetric schemes for the energy conservation of polynomial Hamiltonian problems*, Jour. of Numer. Anal., Industr. and Appl. Math. **4** (no. 1-2) (2009), 87–101.

[11] G.R.W. Quispel, D.I. McLaren, *A new class of energy-preserving numerical integration methods* J. Phys. A **41** (045206), 2008.

Unitary transformations depending on a small parameter

Fernando Casas
(joint work with J.A. Oteo, J. Ros)

Quantum Mechanics (QM) presented since its very creation a view of physical phenomena radically different to that of Classical Mechanics (CM). In many ways, however, CM played a seminal and guiding role in the construction of the quantum formalism. This is evident in the contributions of some of the founding fathers of QM: for example in the famous "Drei männer arbeit" by Born, Heisenberg and Jordan [1] in which they built the quantum perturbation theory in strong analogy with the classical version. In addition, all three of them were able to make use of their through knowledge of the classical theory, especially of the perturbation schemes and the theory of canonical transformations, for extending their matrix scheme to general quantum systems.

Since then the mutual relationship between QM and CM has been fruitful, manyfold and has gone in both ways. A fact that is not always sufficiently emphasized. One of the reasons is that once QM was formulated in Hilbert space the essentially linear character of the formalism made more evident the use of algebraic techniques in this context than in CM. Perturbation theory is another area in which one can say that the quantum treatment is more popular than its classical counterpart. Here again the underlying linear structure of the quantum case allows for a simpler presentation. This is especially true if one adheres to the old-fashioned formulation of classical canonical perturbation theory with mixed (new and old) coordinates in phase space. However, since the late 60's there is a clear alternative to this procedure. It arose in Celestial Mechanics [2] and found early applications in plasma physics and accelerator studies. This approach is based on the use of Lie algebraic methods in CM and originates a perturbation theory that, while being equivalent to the classical Poincaré–Von Zeipel's, is simpler in presentation and richer in applications.

In this presentation we adapt the Lie–Deprit algorithm of classical mechanics as a perturbation theory for general quantum systems. This approach has the advantage that it allows a unifying view, in the following sense. On one hand, it establishes a direct connection between the classical and the quantum formalism. On the other hand, the same algorithm can be applied both to time-independent and time-dependent quantum systems. In addition, and contrarily to the usual time dependent perturbation theory, the scheme is unitary at any order of approximation.

Suppose we are interested in a quantum system which can be described by a time independent Hamiltonian H_0 perturbed by a time-dependent $H'(t,\epsilon)$ that depends on a small parameter ϵ in such a way that $H'(t,\epsilon=0)=0$. The Hamiltonian whose dynamics has to be solved reads then

$$H(t,\epsilon) = H_0 + H'(t,\epsilon) \equiv H_0 + \sum_{n=1}^{\infty} \epsilon^n H_n(t), \tag{1}$$

where we have assumed that $H'(t,\epsilon)$ is analytic in ϵ. The time evolution of the wave function $\Psi(t)$ may be described in terms of the evolution operator, $\Psi(t) = U(t,t_0)\Psi(t_0)$, which is unitary and obeys the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}U(t,t_0) = H(t)U(t,t_0), \quad U(t_0,t_0) = I. \tag{2}$$

Since the dynamics corresponding to H_0 has been solved, one has $U_{H_0}(t,t_0) = \exp(-i(t-t_0)H_0/\hbar)$. The goal is then to construct a unitary near-identity transformation $T(t,\epsilon)$ such that the transformed system

$$i\hbar\frac{\partial}{\partial t}U_K(t,t_0) = K(t,\epsilon)U_K(t,t_0), \qquad U_K(t_0,t_0) = I \tag{3}$$

is easier to solve than the original equation (2). Then it is easy to verify that $U(t,t_0)$ is factorized as

$$U(t,t_0) = T(t,\epsilon)U_K(t,t_0)T^\dagger(t_0,\epsilon), \tag{4}$$

and the new Hamiltonian K is given by

$$K(t,\epsilon) = T^\dagger(t,\epsilon)H(t,\epsilon)T(t,\epsilon) + i\hbar\frac{\partial T^\dagger(t,\epsilon)}{\partial t}T(t,\epsilon). \tag{5}$$

Vey often we will take $T(t_0 = 0,\epsilon) = I$ in (4). We guarantee that T is indeed unitary is by introducing a skew-Hermitian operator $L(t,\epsilon)$ such that $T(t,\epsilon)$ is the solution of the operator differential equation $\frac{\partial}{\partial\epsilon}T(t,\epsilon) = -T(t,\epsilon)L(t,\epsilon)$. Equivalently,

$$\frac{\partial}{\partial\epsilon}T^\dagger(t,\epsilon) = L(t,\epsilon)\,T^\dagger(t,\epsilon), \qquad T^\dagger(t,0) = I. \tag{6}$$

The formal solution of this equation can be obtained by applying the so-called Magnus expansion [3], so that we can write $T^\dagger(t,\epsilon) = \exp(\Omega(t,\epsilon))$, where Ω is a skew-Hermitian operator. Deriving equation (5) with respect to ϵ we arrive after some algebra at

$$\frac{\partial K}{\partial\epsilon} = [L,K] + \mathrm{e}^{\mathrm{ad}_\Omega}\frac{\partial H}{\partial\epsilon} + i\hbar\frac{\partial L}{\partial t}, \tag{7}$$

where $[L,K] \equiv LK - KL$ and $\mathrm{ad}_\Omega B \equiv [\Omega, B]$, with $\mathrm{ad}^n_\Omega B \equiv \mathrm{ad}^{n-1}_\Omega B$.

At this stage, three different issues have to be addressed:

(1) Choose the new Hamiltonian K such that equation (3) is easy to solve.
(2) Compute the skew-Hermitian generator L of the required transformation.
(3) Construct the unitary transformation T from the generator L, or equivalently, the operator Ω in $T = \exp(-\Omega)$.

It turns out that first two problems above enumerated can be solved perturbatively with equation (7), whereas the third can be treated independently. To proceed, we introduce in addition to (1), the following series expansions:

$$K(t,\epsilon) = \sum_{n=0}^{\infty}\epsilon^n K_n(t), \qquad L(t,\epsilon) = \sum_{n=0}^{\infty}\epsilon^n L_{n+1}(t). \tag{8}$$

Then, by applying the Magnus expansion, it is possible to determine $\Omega(t,\epsilon)$ as a power series in ϵ, $\Omega(t,\epsilon)=\sum_{n=1}^{\infty}\epsilon^n v_n(t)$ and the $v_n(t)$ can be expressed in terms of $L_j(t)$. In particular

$$v_1 = L_1, \quad v_2 = \frac{1}{2}L_2, \quad v_3 = \frac{1}{3}L_3 - \frac{1}{12}[L_1, L_2].$$

On the other hand, we have $\mathrm{e}^{\mathrm{ad}_\Omega}\frac{\partial H}{\partial \epsilon} = \sum_{n=0}^{\infty}\epsilon^n w_n(t)$ which can also be obtained algorithmically, its first terms being

$$w_0 = H_1, \;\; w_1 = 2H_2 + [L_1, H_1], \;\; w_2 = 3H_3 + 2[L_1, H_2] + \frac{1}{2}[L_2, H_1] + \frac{1}{2}[L_1, [L_1, H_1]].$$

Finally, inserting these series into (7) and collecting terms of the same power in ϵ, results in the following homological equation

$$i\hbar\frac{\partial L_n}{\partial t} + [L_n, H_0] = nK_n - \tilde{F}_n, \qquad n = 1, 2, \ldots \tag{9}$$

with

$$\tilde{F}_n = \sum_{j=1}^{n-1}[L_{n-j}, K_j] + w_{n-1} \tag{10}$$

in addition to $K_0 = H_0$. This equation admits the formal solution ($L_n(t_0 = 0) = 0$)

$$L_n(t) = -\frac{i}{\hbar}\int_0^t \mathrm{d}u\, \mathrm{e}^{-i(t-u)H_0/\hbar}\big(nK_n(u) - \tilde{F}_n(u)\big)\, \mathrm{e}^{i(t-u)H_0/\hbar}. \tag{11}$$

The election of a particular K is a degree of freedom of the method, and thus it can be adapted to any particular problem one is dealing with. Perhaps the simplest option is to take $K = H_0$ or equivalently $K_n = 0$ for $n \geq 1$. In this way one tries to construct a unitary transformation in such a way that in the new image there is no perturbation at all. In that case $U_K(t) = \exp\big(-iH_0(t-t_0)/\hbar\big)$ and

$$U(t) = T(t,\epsilon)\, \mathrm{e}^{-\frac{i}{\hbar}H_0(t-t_0)} = \mathrm{e}^{-\Omega(t,\epsilon)}\, \mathrm{e}^{-\frac{i}{\hbar}H_0(t-t_0)}. \tag{12}$$

Other options are of course valid. For instance, if H_0 has a pure non-degenerate point spectrum we can choose K_n diagonal. This is the natural choice when the original Hamiltonian (1) is time independent. It is also worth stressing that this procedure can be generalized to *any* linear differential equation and thus constitutes a novel approach to carry out perturbative analysis whereas preserving qualitative (geometric) properties of the exact solution.

References

[1] M. Born, W. Heisenberg, P. Jordan, *Zur Quantenmechanik II*, Z. Phys. **35** (1926), 557–615. Also in *Sources of Quantum Mechanics*. Ed: B.L. Van der Waerden, Dover (1968).

[2] A. Deprit, *Canonical transformations depending on a small parameter*, Celes. Mech. **1** (1969), 12–30.

[3] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Commun. Pure Appl. Math. **7** (1954), 649–673.

Trigonometric schemes for stiff second-order SDEs

David Cohen
(joint work with Magdalena Sigg)

Inspired by the construction of trigonometric schemes for highly oscillatory deterministic problems, see [4, Chapter XIII] and references therein, we propose a new approach for the numerical discretisation of stiff second-order SDEs of the form

$$\ddot{X}_t + \omega^2 X_t = g(X_t) + \dot{W}_t,$$

where $\omega \gg 1$ is a large parameter, the nonlinearity $g(x) = -U'(x)$ is smooth and W_t is a standard Wiener process.

For this type of problems, we obtain order one mean-square error estimates in the position, independent of the large parameter ω. We have also shown that the expected value of the energy $\mathbb{E}[\frac{1}{2}\big((\dot{X}_t)^2 + \omega^2(X_t)^2\big) + U(X_t)]$ along the numerical solution has almost the same linear drift as the exact solution of our problem.

An extension of the method to systems of SDEs

$$\ddot{X}_t + \Omega^2 X_t = g(X_t) + B\dot{W}_t,$$

where $\Omega = \frac{1}{\varepsilon}A^{1/2}$ with $A \in \mathbb{R}^{d\times d}$ symmetric positive definite and $\varepsilon \ll 1$, $B \in \mathbb{R}^{d\times m}$ is "nice", and W_t is an m-dimensional Wiener process, is then presented. Similar results as the one for the above scalar case are briefly presented. The proofs heavily relies on the main result given in [3] for deterministic problems. We mention that our results are closely related to the one by [5] for stiff Langevin systems.

Currently we are extending the above results to stiff SDEs with multiplicative noises.

Finally, we conclude the talk by showing numerical experiments for possible extensions (the Kubo oscillator, Langevin-type equations) and for a nonlinear stochastic wave equation. These are ongoing works.

References

[1] D. Cohen, *On the numerical discretisation of stochastic oscillators*, preprint (2010).
[2] D. Cohen & M. Sigg, *Convergence analysis of trigonometric methods for stiff second-order stochastic differential equations*, preprint (2011).
[3] V. Grimm & M. Hochbruck, *Error analysis of exponential integrators for oscillatory second-order differential equations*, J. Phys. A **39** (2006), 5495–5507.
[4] E. Hairer, C. Lubich & G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer Series in Computational Mathematics **31**, 2002. Second ed., 2006.
[5] M. Tao, H. Owhadi & J.E. Marsden, *Structure preserving Stochastic Impulse Methods for stiff Langevin systems with a uniform global error of order 1 or 1/2 on position*, preprint (2010).

Meshfree integrators

Marco Caliari, Alexander Ostermann, Stefan Rainer

1. Introduction

We consider the time-dependent partial differential equation

$$\frac{\partial}{\partial t}u(t,\xi) = F\Big(t,\xi,u(t,\xi),\frac{\partial}{\partial \xi}u(t,\xi),\dots\Big), \quad t\in[0,T], \quad \xi\in\Omega\subset\mathbb{R}^d \tag{1}$$

subject to appropriate initial and boundary conditions. We assume that the *essential support* of the solution, that is the closure of the set of points where the magnitude of the solution is greater than some given threshold, is small with respect to the domain of interest, and varying in time.

For the numerical solution of (1), we propose a meshfree integrator. The spatial discretisation relies on a stable and robust interpolation based on compactly supported radial basis functions; exponential integrators are employed for the time integration. Our meshfree integrator controls the errors both in space and time.

2. Meshfree integrators

In this section we briefly describe how we compute the numerical approximation $u_n(\xi)\approx u(t_n,\xi)$ for discrete times $0=t_0<t_1<\ldots<t_N=T$. For a detailed description of the method, we refer to [2]. Given an approximation $u_{n-1}(\xi)$ at time t_{n-1}, we interpolate it by compactly supported radial basis functions

$$u_{n-1}(\xi)\approx s(\xi)=\textstyle\sum_{\eta\in H}\lambda_\eta\,\phi\big(\|\xi-\eta\|\big)$$

using a set of *interpolation points* $H=\{\eta_1,\dots,\eta_m\}$. The coefficients λ_η are determined from the interpolation conditions $s(\eta_i)=u_{n-1}(\eta_i),\ i=1,\dots,m$. In order to control the spatial interpolation error, we update the set of interpolation points using a residual subsampling method. For this purpose, we measure the difference between u_{n-1} and its interpolant at a different set of *check points*. We update the set of interpolation points (by a *coarsening* and a *refinement* procedure) until the error at all check points lies between two given thresholds $\theta_{\mathrm{c}}<\theta_{\mathrm{r}}$.

Approximating the right-hand side in (1) we obtain a system of stiff ordinary differential equations

$$w'(t)=G_n(t,w(t)), \quad t\in[t_{n-1},t_n],$$

where the vector $w(t_{n-1})$ contains the values of $u_{n-1}(\xi)$ at the interpolation and check points. This system is solved with an *exponential integrator*. For a review of such integrators, we refer to [6]. The required actions of matrix functions are computed with the *Real Leja Points Method* (see, e.g., [5, 2]).

The numerical solution $u_n(\xi)$ is finally constructed from the numerical approximation to $w(t_n)$. In order to control the error in time we use an embedded method. The error in space is again controlled by a residual subsampling method. Unless both errors are sufficiently small, we repeat the time step by taking a smaller step size and/or a different set of interpolation points.

3. Numerical example

We consider the solution of the nonlinear Schrödinger equation

$$\begin{cases} \mathrm{i}\varepsilon\partial_t\psi = -\dfrac{\varepsilon^2}{2}\Delta\psi + V(x,y)\psi - |\psi|^{2p}\psi, \quad (x,y)\in\mathbb{R}^2,\ t>0 \\ \psi(0,x,y) = \psi_0(x,y), \end{cases} \tag{2}$$

where $0 < p < 1$. As initial value we take a two-bump solution

$$\psi_0(x,y) = \sum_{j=1}^{2} r\left(\frac{x-\bar{x}_j}{\varepsilon}, \frac{y-\bar{y}_j}{\varepsilon}\right),$$

where $\bar{x}_j, \bar{y}_j$ are given offset centres and $r(x,y)\mathrm{e}^{-\mathrm{i}\lambda t}$ is the *ground state* solution (see, e.g., [3]) of the associated nonlinear potential-free Schrödinger equation

$$\mathrm{i}\partial_t\varphi = -\frac{1}{2}\Delta\varphi - |\varphi|^{2p}\varphi, \quad \|\varphi\|_{L^2}^2 = m,$$

that is the solution $\varphi(t,x,y) = r(x,y)\mathrm{e}^{-\mathrm{i}\lambda t}$ minimising the energy

$$E(\varphi) = E(r) = \frac{1}{2}\int_{\mathbb{R}^2} |\nabla r|^2 \mathrm{d}x\mathrm{d}y - \frac{1}{p+1}\int_{\mathbb{R}^2} |r|^{2p+2}\mathrm{d}x\mathrm{d}y.$$

From *Newton's laws*

$$\begin{cases} [\ddot{x}_j(t), \ddot{y}_j(t)] = -\nabla V(x(t), y(t)), \\ [x_j(0), y_j(0)] = [\bar{x}_j, \bar{y}_j], \quad [\dot{x}_j(0), \dot{y}_j(0)] = [0,0] \end{cases} \quad \text{for } j = 1,2 \tag{3}$$

one infers (see [4]) that the solution of (2) behaves like

$$\sum_{j=1}^{2} r\left(\frac{x-x_j(t)}{\varepsilon}, \frac{y-y_j(t)}{\varepsilon}\right) \exp\left(\frac{\mathrm{i}}{\varepsilon}\Big(x\dot{x}_j(t) + y\dot{y}_j(t) + \theta_j^{\varepsilon}(t)\Big)\right),$$

where $\theta_j^{\varepsilon}\colon \mathbb{R}^+ \to [0, 2\pi)$, $j = 1,2$ are suitable shifts, up to an error of size $\mathcal{O}(\varepsilon)$. This dynamical behaviour, in which the shape of $\psi(t,x,y)$ remains close to that of the initial value $\psi_0(x,y)$, is typically known as *soliton dynamics*.

In order to solve (2), we apply the fourth-order splitting method SRKN_6^b by Blanes and Moan [1]. The first part of the equation, with the Laplacian, is approximated in space using Wendland's compactly supported radial basis function $\phi_{3,2}$ (see [7]) and exactly integrated in time using an exponential integrator. The second part, with the potential and the nonlinear term, has an analytic solution.

In Figure 1 we show the behaviour of the solution of (2) for

$$\varepsilon = 0.01, \quad p = 0.2, \quad V(x,y) = \frac{3}{2}x^2 + y^2,$$

with $\bar{x}_1 = \bar{y}_1 = -2.5$, $\bar{x}_2 = \bar{y}_2 = 2$ at different times t. In this case, the solutions of (3) are analytically known (they lie on *Lissajous* curves). We observe that the shape of the two bumps is well preserved during time integration and their centres of mass follow the Lissajous curves. Moreover, the location of the interpolation points is always well spread around the essential support of the solution.

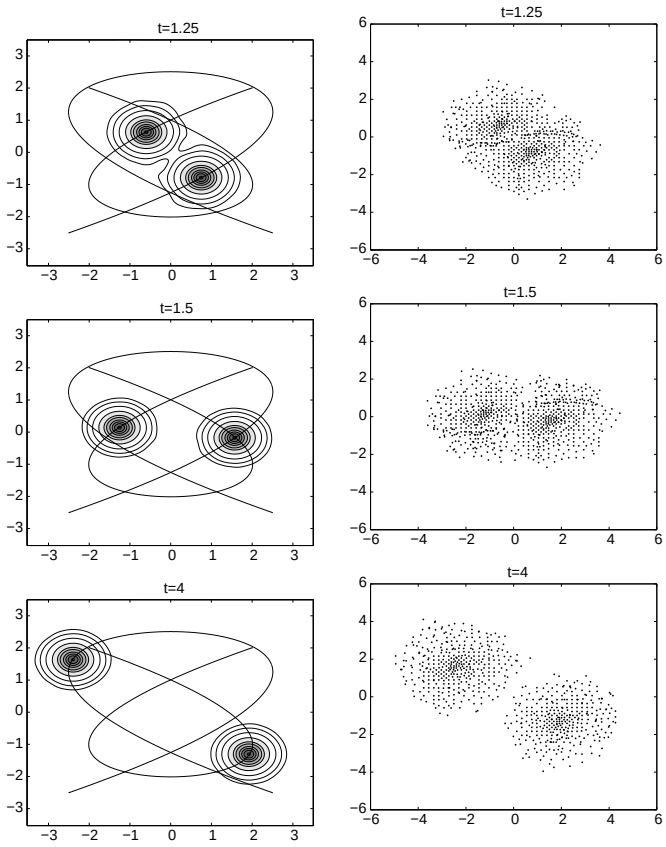


FIGURE 1. Contour levels of the solution (left) and location of interpolation points (right) at different times t.

REFERENCES

[1] S. Blanes and P. C. Moan, *Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, J. Comput. Appl. Math. **142** (2002), 313–330.

[2] M. Caliari, A. Ostermann, and S. Rainer, *Meshfree exponential integrators*, submitted (2010).

[3] M. Caliari, A. Ostermann, S. Rainer, and M. Thalhammer, *A minimisation approach for computing the ground state of Gross–Pitaevskii systems*, J. Comput. Phys. **228** (2009), 349–360.

[4] M. Caliari and M. Squassina, *Numerical computation of soliton dynamics for NLS equations in a driving potential*, Electron. J. Diff. Eqns. **89** (2010), 1–12.

[5] M. Caliari, M. Vianello, and L. Bergamaschi, *Interpolating discrete advection-diffusion propagators at Leja sequences*, J. Comput. Appl. Math. **172** (2004), 79–99.

[6] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numerica **19** (2010), 209–286.

[7] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math. **4** (1995), 389–396.

Superinterpolation in highly oscillatory quadrature

Daan Huybrechs
(joint work with Andreas Asheim, Sheehan Olver)

The computation of highly oscillatory integrals of the form

$$I[f] := \int_a^b f(x) e^{i\omega g(x)} \mathrm{d}x \tag{1}$$

is a recurring problem in many disciplines of science, in particular those involving wave phenomena. The evaluation of such integrals by standard numerical integration methods is considered to be computationally demanding for large values of the frequency parameter ω.

On the other hand, it is known from asymptotic analysis that the integral above has an asymptotic expansion in inverse powers of ω:

$$I[f] \sim \sum_{k=1}^{\infty} a_k \, \omega^{-b_k}, \qquad \omega \gg 1,$$

where the exponents b_k form an increasing sequence of values with positive real part. Moreover, it can be determined that the coefficients a_k are linear functionals of f that have explicit expression in terms of the derivatives of f at a small number of points. These points include the endpoints a and b of integral (1), as well as so-called *stationary points*. The latter points are those where $g'(x)$ vanishes.

A numerical approximation $Q[f]$ to $I[f]$ can be called *robust* for large ω if $Q[f]$ has the same asymptotic behaviour as the integral itself, i.e., if

$$Q[f] \sim \sum_{k=1}^{s} a_k \, \omega^{-b_k} + \mathcal{O}(\omega^{-b_{s+1}}), \qquad \omega \gg 1. \tag{2}$$

In this case, by construction the numerical error asymptotically decreases with ω at a potentially fast rate,

$$Q[f] - I[f] \sim \mathcal{O}(\omega^{-b_{s+1}}), \qquad \omega \gg 1. \tag{3}$$

Truncated asymptotic expansions trivially satisfy this property, but in general they have uncontrollable error and as such they are not desirable for use in applications where accurate approximations are required. Adding terms to the expansion does not necessarily improve accuracy and eventually leads to divergence. However, numerical methods satisfying (2) can be devised that do converge to $I[f]$. We conceptually introduce a new parameter n and focus on methods that satisfy

$$Q[f, n] \to I[f], \qquad n \to \infty,$$

in addition to (3).

Our first example of a numerical method with high asymptotic order that can be made to converge to $I[f]$ for all values of the parameter ω, large or small, is a Filon-type method [1]. The function f is replaced by a polynomial p and the result is integrated exactly, $Q[f] = I[p]$. Polynomial approximation is achieved using interpolation at certain quadrature points. It is observed in [1] that interpolating

derivatives at the critical points yields high asymptotic order of the approximation. Since the asymptotic expansions of $I[f]$ depends on precisely these values, the approximation $I[p]$ agrees asymptotically to $I[f]$ to an extent that depends on the number of derivatives being interpolated. In practice, Filon-type quadrature takes the form of an interpolatory quadrature rule using derivatives,

$$I[f] \approx Q[f] = \sum_{j=1}^{n} \sum_{i=0}^{s_i-1} w_{i,j} f^{(i)}(x_j).$$

Next, we ask ourselves the question what the maximal asymptotic order of accuracy of a numerical method can be, when s evaluations of f or its derivatives near the critical points are allowed. This is a topic studied in [2, 3]. The outcome of this study is Gaussian quadrature rules associated with functionals of the form

$$L[f] = \int_0^\infty f(x) e^{ix} \mathrm{d}x, \tag{4}$$

$$L[f] = \int_0^\infty f(x) e^{ix^r} \mathrm{d}x, \quad \text{or} \quad L[f] = \int_{-\infty}^\infty f(x) e^{ix^r} \mathrm{d}x.$$

In particular, the evaluation of f in s points that are (suitably scaled in an ω-dependent way) roots of polynomials with respect to these functionals leads to twice the asymptotic order of accuracy than that of an asymptotic expansions truncated after s terms. However, a considerable drawback of this approach is that these points lie in the complex plane, close to but away from the interval $[a, b]$. As such, the integrand has to be assumed to be analytic and deformation of the integration path into the complex plane has to be justified by Cauchy's theorem. The quadrature points in most cases lie on the *paths of steepest descent* of the oscillator $g(x)$. Letting the number of quadrature points s grow may or may not lead to convergence of the numerical approximation to $I[f]$: this depends on the behaviour of f in the complex plane.

A surprising observation made in [4] is that Filon-type methods and steepest descent methods can be combined. Given a set of s Gaussian quadrature points determined from a steepest descent analysis near each critical point of the integral, we augment this set with additional points on the interval $[a, b]$. Assume that p interpolates f in all these points and consider the numerical method $Q[f] = I[p]$. This is a Filon-type method with complex quadarature points. This method has the same asymptotic order as the steepest-descent based method, which is twice as high as a Filon-type method interpolating s derivatives of f at the critical points. The computational cost is the same and, moreover, the approximation can be made to converge to $I[f]$ by a judicious choice of the interpolation points on $[a, b]$. It is advocated in [4] to interpolate in the roots of the Chebyshev polynomial of the first kind, shifted to $[a, b]$. In this case, it is shown that the approximation error improves both with increasing frequency ω and with an increasing number of quadrature points on $[a, b]$.

Finally, we remark that Gaussian quadrature rules can be devised for functionals generalizing (4) to the form [5]

$$L[f] = \int_0^\infty f(x)h(x)\mathrm{d}x.$$

As before, the quadrature points are the roots of polynomials that are (formally) orthogonal with respect to the derived bilinear form

$$a(f, g) = L[fg],$$

in the sense that

$$a(p_n, p_m) = \delta_{m,n}.$$

The resulting quadrature rules are suitable for the evaluation of integrals of the form

$$\int_0^\infty f(x)h(\omega x)\mathrm{d}x.$$

Examples of practical interest are those where h is a trigonometric function, a Bessel function or the Airy function.

References

[1] A. Iserles and S. P. Nørsett, *Efficient quadrature of highly oscillatory integrals using derivatives*, Proc. R. Soc. Lond. A 461 (2005), 1383–1399.

[2] D. Huybrechs and S. Vandewalle, *On the evaluation of highly oscillatory integrals by analytic continuation*, SIAM J. Numer. Anal. 44 (2006), 1026-1048.

[3] A. Deaño and D. Huybrechs, *Complex Gaussian quadrature of oscillatory integrals*, Numer. Math. 113 (2009), 197-219.

[4] D. Huybrechs and S. Olver, *Superinterpolation in highly oscillatory quadrature*, Technical Report TW569 (2010), K.U.Leuven, Belgium.

[5] A. Asheim and D. Huybrechs, *Complex Gaussian quadrature for oscillatory integral transforms*, manuscript in preparation.

Asymptotic numerical methods for oscillatory systems of differential equations

Alfredo Deaño

(joint work with Marissa Condon and Arieh Iserles)

In this contribution we present a method to compute effectively solutions of systems of ordinary differential equations with highly oscillatory forcing terms. More precisely, we are concerned with systems of ODEs of the following form:

$$\boldsymbol{y}'(t) = \boldsymbol{h}(\boldsymbol{y}(t)) + g_\omega(t)\boldsymbol{f}(\boldsymbol{y}(t)), \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0, \tag{1}$$

where $\boldsymbol{y}(t) : \mathbb{R} \to \mathbb{R}^d$, $\boldsymbol{f}(\boldsymbol{y}), \boldsymbol{h}(\boldsymbol{y}) : \mathbb{R}^d \to \mathbb{R}^d$ are analytic functions (generally nonlinear), and the scalar term $g_\omega(t)$ can be written in the form of a modulated Fourier expansion (MFE), that is

$$g_\omega(t) = \sum_{m=-\infty}^{\infty} a_m(t)e^{im\omega t}. \tag{2}$$

Differential equations of this type are very relevant in the context of circuit modeling in electronic engineering, but (as we explain below) they also represent a numerical challenge because of the highly oscillatory behaviour of the solutions.

Some typical forcing terms in applications involve either one or two frequencies

$$g_\omega(t) = e^{i\omega t}, \qquad \omega \gg 1,$$

$$g_\omega(t) = \sin\omega_1 t \, \sin\omega_2 t, \qquad \omega_2 \gg \omega_1 \gg 1,$$

the second example being important in amplitude modulation, or full spectrum (in equations for diodes and transistors):

$$g_\omega(t) = e^{\eta\cos\omega t}, \qquad \omega \gg 1.$$

In this last case, a Fourier expansion can be obtained involving modified Bessel functions $I_n(\eta)$.

The form of the exact solution of such a problem can be analysed by writing the oscillatory ODE (1) as a perturbation of

$$\boldsymbol{z}'(t) = \boldsymbol{h}(\boldsymbol{z}(t)), \qquad \boldsymbol{z}(0) = \boldsymbol{y}_0, \tag{3}$$

with the same initial data. Then application of nonlinear variation of constants allows us to relate the solutions of the perturbed and unperturbed systems:

$$\boldsymbol{y}(t) = \boldsymbol{z}(t) + \int_0^t \boldsymbol{\Phi}(t-s)\boldsymbol{f}(\boldsymbol{y}(s))g_\omega(s)ds. \tag{4}$$

Here $\boldsymbol{\Phi}$ solves the so-called variational equation

$$\boldsymbol{\Phi}' = \frac{\partial \boldsymbol{h}(\boldsymbol{z}(t))}{\partial \boldsymbol{z}}\boldsymbol{\Phi}, \qquad \boldsymbol{\Phi}(0) = I.$$

The matrix $\boldsymbol{\Phi}$ may not be analytically available for general nonlinear equations, but if the integrand is smooth enough and $g_\omega(t)$ is a trigonometric function (see the examples cited before), then integration by parts gives that the integral in (4) is $\mathcal{O}(\omega^{-1})$ as $\omega \to \infty$.

This result shows that under quite general assumptions, the difference between both solutions is of order $\mathcal{O}(\omega^{-1})$. Moreover, the integral in (4) is highly oscillatory because of the factor $g_\omega(s)$. As a consequence of this, standard methods for numerical solutions of ODEs such as Runge–Kutta are not effective, since they will apply a classical quadrature rule to this integral at each step. It is known that this idea leads to a very small stepsize, of order $h\omega \approx 1$, which is both too expensive for implementation and also unreliable because of the large number of steps needed to integrate the ODE in a given interval.

As shown for instance in [6], integrals of Fourier type

$$I[f] = \int_a^b f(x)e^{i\omega g(x)}dx,$$

where $f(x)$ is smooth and $g'(x) \neq 0$ in $[a,b]$, can be expanded in inverse powers of ω, using integration by parts or Watson's lemma applied to oscillatory integrals, see for example [7]. In the spirit of the method of stationary phase, the

coefficients in such an expansion only depend on the functions $f(x)$ and $g(x)$ (and their derivatives) at the endpoints of the interval of integration. It follows that interpolation of $f(x)$ at these endpoints will result in an effective method for large values of ω, since in that case the error with respect to the quadrature rule $Q[f]$ will be

$$E[f] = \int_a^b f(x)e^{i\omega g(x)}dx - Q[f] = \mathcal{O}(\omega^{-2}).$$

The use of Hermite–type interpolation (including derivatives of $f(x)$ at the endpoints) leads to higher order methods. This is the idea behind *Filon-type methods*, which have been studied recently [6] as a very effective tool in the approximation of highly oscillatory integrals.

The method that we propose follows these lines, and proposes the following *ansatz* for the solution of the ODE, in terms of modulated Fourier expansions:

$$\boldsymbol{y}(t) = \sum_{n=0}^{\infty} \frac{\boldsymbol{\psi}_n(t)}{\omega^n}, \qquad \boldsymbol{\psi}_n(t) = \sum_{m=-\infty}^{\infty} \boldsymbol{p}_{n,m}(t)e^{im\omega t}, \qquad n \geq 0.$$

We expand everything formally in inverse powers of ω using the *ansatz*, and then substitute in the ODE. Separation first in orders of magnitude (powers of ω), and then in frequencies (values of m) leads either to *nonoscillatory* ODEs or to recursions for the coefficients $\boldsymbol{p}_{n,m}(t)$. In this way, oscillations are removed and are only added at the end, when assembling the modulated Fourier expansion.

For example, consider

$$\boldsymbol{y}'(t) = \boldsymbol{h}(\boldsymbol{y}(t)) + g_\omega(t)\boldsymbol{f}(\boldsymbol{y}(t)), \quad g_\omega(t) = \sum_{m=-\infty}^{\infty} \alpha_m e^{im\omega t}, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0.$$

If we equate $\mathcal{O}(1)$ terms, we have

$$\boldsymbol{p}'_{0,0}(t) = \boldsymbol{h}(\boldsymbol{p}_{0,0}(t)) + \alpha_0\boldsymbol{f}(\boldsymbol{p}_{0,0}(t)), \qquad \boldsymbol{p}_{0,0}(0) = \boldsymbol{y}(0) = \boldsymbol{y}_0,$$

which is nonoscillatory, and can be solved using standard methods. Additionally

$$\boldsymbol{p}_{1,m}(t) = \frac{\alpha_m}{im}\,\boldsymbol{f}(\boldsymbol{p}_{0,0}(t)), \qquad m \neq 0.$$

The $\mathcal{O}(\omega^{-1})$ level yields again a differential equation for $\boldsymbol{p}_{1,0}$:

$$\boldsymbol{p}'_{1,0} = \boldsymbol{J}\boldsymbol{h}(\boldsymbol{p}_{0,0})\boldsymbol{p}_{1,0} + \boldsymbol{J}\boldsymbol{f}(\boldsymbol{p}_{0,0}) \sum_{r=-\infty}^{\infty} \alpha_r\, \boldsymbol{p}_{1,-r},$$

together with $\boldsymbol{p}_{1,0}(0) = \boldsymbol{0}$ (because the full initial condition has been assigned to the zeroth term $\boldsymbol{\psi}_0(t)$), and a recursion for the next level:

$$\boldsymbol{p}_{2,m} = -\frac{i}{m}\left[-\boldsymbol{p}'_{1,m} + \boldsymbol{J}\boldsymbol{h}(\boldsymbol{p}_{0,0})\boldsymbol{p}_{1,m} + \boldsymbol{J}\boldsymbol{f}(\boldsymbol{p}_{0,0}) \sum_{r=-\infty}^{\infty} \alpha_r \boldsymbol{p}_{1,m-r}\right],$$

for $m \neq 0$. Here $\boldsymbol{J}\boldsymbol{h}$ and $\boldsymbol{J}\boldsymbol{f}$ are the Jacobian matrices of $\boldsymbol{h}$ and $\boldsymbol{f}$ respectively. Higher order terms can be computed following this scheme.

Other examples are considered in [3, 4], including second order non-linear equations of Van der Pol and Duffing type with oscillatory forcing, and also equations with ω dependence in the coefficients, like the classical Kapitsa inverted pendulum.

It is worth mentioning that this approach is also useful in equations with intrinsic oscillation of the form $\ddot{\boldsymbol{x}} + \Omega^2 \boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x})$, where

$$\Omega = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \omega I \end{array} \right),$$

and $\omega \gg 1$. These types of equations are of great importance in Geometric Numerical Integration, and have been extensively studied in [2, 5].

We also note that this approach is very closely related to stroboscopic and higher order averaging, as explained very recently in [1].

Ongoing and future work on this topic includes further exploration of these connections with related approaches and also extension to systems of DAEs, which are a more realistic model in electronic engineering, as well as delay differential equations (DDEs) and partial differential equations.

References

[1] Ph. Chartier, A. Murua, J. M. Sanz–Serna, *Higher-order averaging, formal series and numerical integration I: B-series*, Found. Comput. Math. **10** (2010), 695–727.

[2] D. Cohen, E. Hairer, C. Lubich, *Modulated Fourier expansions for highly oscillatory differential equations*, Found. Comput. Math. **3** (2003), 327–345.

[3] M. Condon, A. Deaño, A. Iserles, *On systems of differential equations with extrinsic oscillation*, Discrete Contin. Dynam. Systems **28** (2010), 1345–1367.

[4] M. Condon, A. Deaño, A. Iserles, *On second order differential equations with highly oscillatory forcing terms*, Proc. Royal Soc. A **466** (2010), 1809–1828.

[5] E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Verlag, 2006.

[6] A. Iserles, S. P. Nørsett, *Efficient quadrature of highly oscillatory integrals using derivatives*, Proc. Royal Soc. A **461** (2005), 1383-1399.

[7] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, 1974.

Shooting and operator determinant techniques for computing spectra

Simon J.A. Malham

(joint work with Issa Karambal, Veerle Ledoux, Robert Marangell, Jitse Niesen and Vera Thümmler)

Elliptic operators. For brevity, we focus on our applications of interest. These involve finding the pure-point spectra of elliptic operators $L := B\Delta + V(x)$, say of order n (even), on a domain $\Omega \subseteq \mathbb{R}^d$. Here B is a constant diagonal diffusion matrix, and the non-symmetric matrix potential $V = V(x)$ is the principal source of non-selfadjointness. Most of our development also applies for more general elliptic operators, diffusions and advection terms (the latter can often be scaled/weighted out, but this is not required). To compute the eigenvalues for

such operators, sophisticated projection algorithms have been developed. Such algorithms project the differential eigenvalue problem onto a finite dimensional basis (usually) of $\mathbb{L}^2(\Omega)$. Then further algorithms tailored to solving algebraic eigenvalue problems are applied. For example iteration techniques can be bolted on if particular eigenvalues are sought, and good initial guesses are available. One disadvantage of generic implementations is that spurious eigenvalues are generated. Typically though, as the projection dimension is increased, they converge towards the continuous spectrum, while genuine eigenvalues remain relatively static. We must however, carefully separate them.

Shooting. Suppose $\Omega = \mathbb{R}$, finite or semi-infinite intervals can be treated similarly. The elliptic eigenvalue problem for L with spectral parameter $\lambda \in \mathbb{C}$ is thus equivalent to finding localized homogeneous solutions of the nth order unbounded operator $\partial_x - A(x;\lambda)$ where A is a non-selfadjoint matrix. Under suitable (weak) assumptions on the far-field limits of V and using weights if necessary, there is a subregion of the spectral parameter plane, containing the right-half plane, that excludes the essential spectrum, but within which isolated eigenvalues may exist. Choose a region Λ of that subregion within which you seek eigenvalues. For $\lambda \in \Lambda$, the operator $\partial_x - A$ is Fredholm, with kernel and cokernel of the same dimension, say k, and thus index zero. This is equivalent to the existence of exponential dichotomies; see Sandstede [6]. The upshot is, there exist k solutions $Y = [Y_1 \ldots Y_k]$ satisfying natural boundary data as $x \to -\infty$, and $(n-k)$ solutions $Z = [Z_1 \ldots Z_{n-k}]$ for $x \to +\infty$. Setting $Y^{\wedge k} \coloneqq Y_1 \wedge \ldots \wedge Y_k$, the natural generalization of the Wronskian is the Evans function determinant $Y^{\wedge k} \wedge Z^{\wedge(n-k)}$ which is zero at eigenvalues. Using the correct exponential weighting it is independent of x and analytic in λ. Newton iteration can be used to determine the complex zeros/eigenvalues. Better yet we can apply the argument principle to globally count the number of eigenvalues inside $\partial\Lambda$. To each $(x,\lambda) \in \mathbb{R} \times \partial\Lambda$ attach $Y(x;\lambda) \oplus Z(x;\lambda)$. Cap off the cylinder $\mathbb{R} \times \partial\Lambda$ with the far-field solutions at $x = \pm\infty$ for $\lambda \in \Lambda^\circ$. We have generated a fibre bundle $\mathbb{E} \to \mathbb{S}^2$ where the fibres are solution sections in $\mathbb{C}^n$. One can compute the first Chern class $c_1(\mathbb{E}) \in H^2(\mathbb{S}^2;\mathbb{Z})$ and first Chern number which equals the winding number of the Evans function round $\partial\Lambda$ and thus the number of eigenvalues inside Λ; see Alexander, Gardner and Jones [1].

Schubert varieties. Shooting for $n \gg 1$ was a well-known difficulty until Humpherys and Zumbrun introduced their continuous orthogonalization method. In [3, 4] we proceeded instead as follows. To detect eigenvalues, it is sufficient to compute the subspace spanned by Y at (x,λ). This is an element in the Grassmannian $\mathrm{Gr}(n,k)$ of k-planes (or k-dimensional subspaces). Coordinate charts of $\mathrm{Gr}(n,k)$ are isomorphic to $\mathbb{C}^{(n-k)\times k}$. Another way to see this is that with respect to the polarization $\mathbb{C}^n = \mathbb{C}^k \oplus \mathbb{C}^{n-k}$ the graph of a linear operator $y \in \mathrm{Lin}(\mathbb{C}^k;\mathbb{C}^{n-k})$, i.e. $\{(\mathrm{id}_{\mathbb{C}^k} + y)(z) \colon z \in \mathbb{C}^k\}$, realizes such a k-plane. Hence $y \in \mathbb{C}^{(n-k)\times k}$ locally parameterizes $\mathrm{Gr}(n,k)$. For the spectral problem, the k-plane spanned by Y, represented in a given chart by $y \in \mathbb{C}^{(n-k)\times k}$, evolves according to the Riccati equation $y' = c + d\,y - y(a + b\,y)$, where a, b, c and d are appropriate sub-blocks

of $A(x;\lambda)$. Riccati solutions can become singular. This means we are leaving a 'good' representative coordinate patch. We could simply swap patches. However, singularities indicate a 'change of state' as follows. The Grassmannian is the disjoint union of Schubert cells C_μ, indexed by partitions μ of $k\cdot(n-k)$. Performing Gaussian elimination on any k-plane $Y \in \mathrm{Gr}(n,k)$ determines pivot positions and the unique cell Y lies in. The Schubert varieties X_μ are identified by the zero solution set of all $k\times k$ minors of C_μ. By identifying Schubert varieties X_μ with index μ obtainable from each other by $\mathrm{GL}(n)$ transformations, we determine the Schubert cycles $\sigma_\mu = [\mathrm{X}_\mu]$. A singularity indicates the solution has changed cell, variety and cycle. In $\mathbb{R}\times\Lambda$, there are singular curves which wind round and converge to eigenvalues as $x\to+\infty$. The Riccati flow evolves y along $\mathbb{R}$ on $\partial\Lambda$ (one hemisphere of $\mathbb{S}^2$) in a given top cycle. It carries all the spectral information with it for the partial spectral problem up to $x\in\mathbb{R}$. This is because, first, the far field data or 'state' is fixed, and second, the cohomological ring of the Grassmannian is completely known: the Schubert cycles give a basis, and intersections generate the cup-product. The Riccati flow tells us the current 'state' and the 'state' singularities correspond to. We can thus compute their intersection with the far field 'state'. Indeed the Riccati flow is enough to generate the determinant line bundle. As we expect, on the singularity curve the Evans function is zero (realized for a chosen flag). Punctures of $\mathbb{R}\times\partial\Lambda$ by singularity curves thus indicate entering eigenvalues for the partial problem, which if they do not exit, are eigenvalues for the full spectral problem. Note we can construct the explicit map $f\colon \mathbb{S}^2\to\mathrm{Gr}(n,k)$. Any vector bundle over $\mathbb{S}^2$ can be thought of as $E = f^*(\mathbb{T})$ with $\mathbb{T}$ the tautological bundle $\mathbb{T}\longrightarrow \mathrm{Gr}(n,k)\times\mathbb{C}^n\longrightarrow\mathbb{Q}$, with quotient bundle $\mathbb{Q}$. The Chern classes of $\mathbb{T}$ are $c_i(\mathbb{T}) = \sigma_{1^i}$, and of $\mathbb{Q}$ are $c_i(\mathbb{Q}) = \sigma_i$, the special Schubert cycles which generate $H^*(\mathrm{Gr}(n,k);\mathbb{Z})$. By naturality, we can determine the invariants of $\mathbb{E}$ using $c_i(f^*(\mathbb{T})) = f^*(c_1(\mathbb{T}))$. Computing the Riccati flow is relatively efficient. We use a Lie group integrator close to singularities. Further, for sectorial operators, the locus of singularity curves is significantly restricted—we can track eigenvalues by numerical continuation, so that computing eigenvalues becomes significantly more efficient.

Operator determinants. Let $\mathbb{H}$ be a separable Hilbert space over $\Omega\subseteq\mathbb{R}^d$ (now a general domain). The space of trace class operators K on $\mathbb{H}$, denoted $\mathbb{J}_1(\mathbb{H})$, is a Banach space with norm $\|K\|_{\mathbb{J}_1} \coloneqq \operatorname{tr}|K|$, while the space of compact Hilbert–Schmidt operators $\mathbb{J}_2(\mathbb{H})$ for which $\operatorname{tr}K^*K$ is finite, is a Hilbert space. If $K\in\mathbb{J}_1(\mathbb{H})$ then $\det_1(\mathrm{id}+K) \coloneqq \sum\operatorname{tr}K^{\wedge m}$, where $\operatorname{tr}K^{\wedge m}$ is the natural trace on $\mathbb{H}^{\wedge m}$. Indeed Fredholm computed an explicit formula for $\det_1(\mathrm{id}+K)$ as a series expansion in multiple integrals of determinants of the Green's kernel G associated with K. Essentially each term $\operatorname{tr}K^{\wedge m}$ generates a multiple integral in terms of G, and the determinants naturally arise out of the alternating algebra of $G^{\wedge m}$ on $\mathbb{H}^{\wedge m}$ (and the trace applied). If $K\in\mathbb{J}_2(\mathbb{H})$ then $\det_2(\mathrm{id}+K) \coloneqq \det_1(\mathrm{id}+K)\cdot\exp(-\operatorname{tr}K)$. Hilbert proved an analogous series expansion for $\det_2(\mathrm{id}+K)$ in terms of Green's kernels (exponentially scaling out $\operatorname{tr}K$ knocks out the diagonal kernel terms in the determinants).

We can solve the original eigenvalue problem by the perturbative approach. We decompose $L - \lambda$ into $L_0(\lambda) + \hat{L}$ where $\hat{L}$ contains part of the matrix potential such that $\hat{L} \to 0$ as $x \to \partial\Omega$ for an interior domain, or $|x| \to \infty$, say, for an exterior domain. The constant coefficient operator $L_0(\lambda)$ is associated with a background "free" state. If we can compute the Green's kernel corresponding to $K_0(\lambda) = L_0^{-1}(\lambda)$ then the eigenvalue problem is equivalent to $\big(\mathrm{id}+K_0(\lambda)\circ\hat{L}\big)\, u = 0$. In shooting we directly attempt to numerically invert $L - \lambda$. Here we invert the simpler operator $L_0(\lambda)$ (a big step) and then compute the operator determinant $\det\big(\mathrm{id} + K_0(\lambda) \circ \hat{L}\big)$. Bornemann [2] recently provided a tremendous impetus to this approach by demonstrating that simple quadrature approximations for 'id + $K_0(\lambda) \circ \hat{L}$' can lead to fast convergence results for smoother kernels. For practical problems, especially in higher dimensions, the kernels are not smooth and usually singular, nevertheless a generalization of Bornemann's approach can be applied and proves quite efficient compared to shooting (initial investigations with Karambal).

The talk presented these ideas and results we achieved thusfar using them.

References

[1] J.C. Alexander, R. Gardner and C.K.R.T. Jones, *A topological invariant arising in the stability analysis of traveling waves*, J. Reine Angew. Math. **410** (1990), 167–212.

[2] F. Bornemann, *On the numerical evaluation of Fredholm determinants*, Math. Comp. **79**(270) (2010), 871–915.

[3] V. Ledoux, S.J.A. Malham, J. Niesen and V. Thümmler, *Computing stability of multidimensional travelling waves*, SIADS **8**(1) (2009), 480–507.

[4] V. Ledoux, S.J.A. Malham and V. Thümmler, *Grassmannian spectral shooting*, Math. Comp. **79** (2010), 1585–1619.

[5] V. Ledoux, S.J.A. Malham and R. Marangell, *Computing spectra on Schubert varieties*, Preprint.

[6] B. Sandstede, *Stability of traveling waves*, In Handbook of Dyn. Sys. II, B. Fiedler, ed., Elsevier (2002), 983–1055.

Kepler, Newton and GNI

Gerhard Wanner

The emphasis of the talk is to describe the discovery of Kepler's Laws in the Astronomia Nova (1609), starting from Ptolemy–Copernicus–Brahe's model of the movement of the planets on excentric circles with the speed governed by a "punctum aequans". Kepler, trying to get rid of this puntum aequans, came slowly up with the area-law (in chapter XL) and, trying to adjust circular movement with Tycho Brahe's measurements, with the elliptic movement (in chapter LVI). Newton then, in Proposition I of the Principia (1687) explained the area-law by concentrating the continuous force to discrete shocks, which transforms the orbit into a polygon whose triangles explain the law from Eucl. I.41. Nowadays, this procedure is seen as the "syplectic Euler" method,one of the fundamental procedures for methods preserving geometric properties of the solution. In the next step (Proposition XI of the Principia) he derived the Law of Gravitation from

geometric properties of the ellipse. Only Euler, in 1747, was the first to formulate, what nowadays is called, Newton's differential equations of mechanics.

Reporter: Daniel Weiß