Sampling Techniques for Computational Statistical Physics

Benedict Leimkuhler[1] and Gabriel Stoltz[2]
[1]Edinburgh University School of Mathematics,
Edinburgh, Scotland, UK
[2]Université Paris Est, CERMICS, Projet MICMAC
Ecole des Ponts, ParisTech – INRIA, Marne-la-Vallée,
France

Mathematics Subject Classification

82B05; 82-08; 65C05; 37M05

Short Definition

The computation of macroscopic properties, as predicted by the laws of statistical physics, requires sampling phase-space configurations distributed according to the probability measure at hand. Typically, approximations are obtained as time averages over trajectories of discrete dynamics, which can be shown to be ergodic in some cases. Arguably, the greatest interest is in sampling the canonical (constant temperature) ensemble, although other distributions (isobaric, microcanonical, etc.) are also of interest. Focusing on the case of the canonical measure, three important types of methods can be distinguished: (1) Markov chain methods based on the Metropolis–Hastings algorithm; (2) discretizations of continuous stochastic differential equations which are appropriate modifications and/or limiting cases of the Hamiltonian dynamics; and (3) deterministic dynamics on an extended phase space.

Description

Applications of sampling methods arise most commonly in molecular dynamics and polymer modeling, but they are increasingly encountered in fluid dynamics and other areas. In this article, we focus on the treatment of systems of particles described by position and momentum vectors q and p, respectively, and modeled by a Hamiltonian energy $H = H(q, p)$.

Macroscopic properties of materials are obtained, according to the laws of statistical physics, as the average of some function with respect to a probability measure μ describing the state of the system (▸ Calculation of Ensemble Averages):

$$\mathbb{E}_\mu(A) = \int_{\mathcal{E}} A(q, p)\, \mu(dq\, dp). \tag{1}$$

In practice, averages such as (1) are obtained by generating, by an appropriate numerical method, a sequence of microscopic configurations $(q^i, p^i)_{i \geq 0}$ such that

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} A(q^i, p^i) = \int_{\mathcal{E}} A(q, p)\, \mu(dq\, dp). \tag{2}$$

The Canonical Case

For simplicity, we consider the case of the canonical measure:

B. Engquist (ed.), *Encyclopedia of Applied and Computational Mathematics*,
DOI 10.1007/978-3-540-70529-1

$$\mu(dq\,dp) = Z_\mu^{-1}\mathrm{e}^{-\beta H(q,p)}\,dq\,dp,$$
$$Z_\mu = \int_{\mathcal{E}} \mathrm{e}^{-\beta H(q,p)}\,dq\,dp, \qquad (3)$$

where $\beta^{-1} = k_{\mathrm{B}}T$. Many sampling methods designed for the canonical ensemble can be extended or adapted to sample other ensembles.

If the Hamiltonian is separable (i.e., it is the sum of a quadratic kinetic energy and a potential energy), as is usually the case when Cartesian coordinates are used, the measure (3) has a tensorized form, and the components of the momenta are distributed according to independent Gaussian distributions. It is therefore straightforward to sample the kinetic part of the canonical measure. The real difficulty consists in sampling positions distributed according to the canonical measure

$$\nu(dq) = Z_\nu^{-1}\mathrm{e}^{-\beta V(q)}\,dq, \qquad Z_\nu = \int_{\mathcal{D}} \mathrm{e}^{-\beta V(q)}\,dq, \qquad (4)$$

which is typically a high dimensional distribution, with many local concentrated modes. For this reason, many sampling methods focus on sampling the configurational part ν of the canonical measure.

Since most concepts needed for sampling purposes can be used either in the configuration space or in the phase space, the following notation will be used: The state of the system is denoted by $x \in \mathcal{S} \subset \mathbb{R}^d$, which can be the position space $q \in \mathcal{D}$ (and then $d = 3N$), or the full phase space $(q,p) \in \mathcal{E}$ with $d = 6N$. The measure $\pi(dx)$ is the canonical distribution to be sampled (ν in configuration space, μ in phase space).

General Classification

From a mathematical point of view, most sampling methods may be classified as (see [2]):

1. "Direct" probabilistic methods, such as the standard rejection method, which generate identically and independently distributed (i.i.d) configurations
2. Markov chain techniques
3. Markovian stochastic dynamics
4. Purely deterministic methods on an extended phase-space

Although the division described above is useful to bear in mind, there is a blurring of the lines between the different types of methods used in practice, with Markov chains being constructed from Hamiltonian dynamics or degenerate diffusive processes being added to deterministic models to improve sampling efficiencies.

Direct probabilistic methods are typically based on a prior probability measure used to sample configurations, which are then accepted or rejected according to some criterion (as for the rejection method, for instance). Usually, a prior probability measure which is easy to sample should be used. However, due to the high dimensionality of the problem, it is extremely difficult to design a prior sufficiently close to the canonical distribution to achieve a reasonable acceptance rate. Direct probabilistic methods are therefore rarely used in practice.

Markov Chain Methods

Markov chain methods are mostly based on the Metropolis–Hastings algorithm [5, 13], which is a widely used method in molecular simulation. The prior required in direct probabilistic methods is replaced by a *proposal move* which generates a new configuration from a former one. This new configuration is then accepted or rejected according to a criterion ensuring that the correct measure is sampled. Here again, designing a relevant proposal move is the cornerstone of the method, and this proposal depends crucially on the model at hand.

The Metropolis–Hastings Algorithm

The Metropolis–Hastings algorithm generates a Markov chain of the system configurations $(x^n)_{n\geq 0}$ having the distribution of interest $\pi(dx)$ as a stationary distribution. The invariant distribution π has to be known only up to a multiplicative constant to perform this algorithm (which is the case for the canonical measure and its marginal in position). It consists in a two-step procedure, starting from a given initial condition x^0:

1. Propose a new state $\tilde{x}^{n+1}$ from x^n according to the proposition kernel $T(x^n,\cdot)$
2. Accept the proposition with probability $\min\left(1, \dfrac{\pi(\tilde{x}^{n+1})\,T(\tilde{x}^{n+1},x^n)}{\pi(x^n)\,T(x^n,\tilde{x}^{n+1})}\right)$, and set in this case $x^{n+1} = \tilde{x}^{n+1}$; otherwise, set $x^{n+1} = x^n$

It is important to count several times a configuration when a proposal is rejected.

The original Metropolis algorithm was proposed in [13] and relied on symmetric proposals in the configuration space. It was later extended in [5] to allow for nonsymmetric propositions which can bias proposals toward higher probability regions with respect to the target distribution π. The algorithm is simple to interpret in the case of a symmetric proposition kernel on the configuration space ($\pi(x) \propto \mathrm{e}^{-\beta V(q)}$ and $T(q,q') = T(q',q)$). The Metropolis–Hastings ratio is simply

$$r(q,q') = \exp\left[-\beta(V(q') - V(q))\right].$$

If the proposed move has a lower energy, it is always accepted, which allows to visit more frequently the states of higher probability. On the other hand, transitions to less likely states of higher energies are not forbidden (but accepted less often), which is important to observe transitions from one metastable region to another when these regions are separated by some energy barrier.

Properties of the Algorithm

The probability transition kernel of the Metropolis–Hastings chain reads

$$P(x,dx') = \min\left(1, r(x,x')\right) T(x,dx') + (1 - \alpha(x))\, \delta_x(dx'), \tag{5}$$

where $\alpha(x) \in [0,1]$ is the probability to accept a move starting from x (considering all possible propositions):

$$\alpha(x) = \int_{\mathcal{S}} \min\left(1, r(x,y)\right) T(x,dy).$$

The first part of the transition kernel corresponds to the accepted transitions from x to x', which occur with probability $\min(1, r(x,x'))$, while the term $(1 - \alpha(x))\delta_x(dx')$ encodes all the rejected steps.

A simple computation shows that the Metropolis–Hastings transition kernel P is reversible with respect to π, namely, $P(x,dx')\pi(dx) = P(x',dx)\pi(dx')$. This implies that the measure π is an invariant measure. To conclude to the pathwise ergodicity of the algorithm (2) (relying on the results of [14]), it remains to check whether the chain is (aperiodically) irreducible, i.e., whether any state can be reached from any other one in a finite number of steps. This property depends on the proposal kernel T, and should be checked for the model under consideration.

Besides determining the theoretical convergence of the algorithm, the proposed kernel is also a key element in devising efficient algorithms. It is observed in practice that the optimal acceptance/rejection rate, in terms of the variance of the estimator (a mean of some observable over a trajectory), for example, is often around 0.5, ensuring some balance between:

- Large moves that decorrelate the iterates when they are accepted (hence reducing the correlations in the chain, which is interesting for the convergence to happen faster) but lead to high rejection rates (and thus degenerate samples since the same position may be counted several times)
- And small moves that are less rejected but do not decorrelate the iterates much

This trade-off between small and large proposal moves has been investigated rigorously in some simple cases in [16,17], where optimal acceptance rates are obtained in a limiting regime.

Some Examples of Proposition Kernels

The most simple transition kernels are based on random walks. For instance, it is possible to modify the current configuration by a random perturbation applied to all particles. The problem with such symmetric proposals is that they may not be well suited to the target probability measure (creating very correlated successive configurations for small σ, or very unlikely moves for large σ). Efficient nonsymmetric proposal moves are often based on discretizations of continuous stochastic dynamics which use a biasing term such as $-\nabla V$ to ensure that the dynamics remains sufficiently close to the minima of the potential.

An interesting proposal relies on the Hamiltonian dynamics itself and consists in (1) sampling new momenta p^n according to the kinetic part of the canonical measure; (2) performing one or several steps of the Verlet scheme starting from the previous position q^n, obtaining a proposed configuration $(\widetilde{q}^{n+1}, \widetilde{p}^{n+1})$; and (3) computing $r^n = \exp[-\beta(H(\widetilde{q}^{n+1}, \widetilde{p}^{n+1}) - H(q^n, p^n))]$ and accepting the new position q^{n+1} with probability $\min(1, r^n)$. This algorithm is known as the Hybrid Monte Carlo algorithm (first introduced in [3] and analyzed from a mathematical viewpoint in [2,20]).

A final important example is parallel tempering strategies [10], where several replicas of the system are simulated in parallel at different temperatures, and sometimes exchanges between two replicas at different temperatures are attempted, the probability of such an exchange being given by a Metropolis–Hastings ratio.

Continuous Stochastic Dynamics

A variety of stochastic dynamical methods are in use for sampling the canonical measure. For simplicity of exposition, we consider here systems with the N-body Hamiltonian $H = p^T M^{-1} p/2 + V(q)$.

Brownian Dynamics

Brownian dynamics is a stochastic dynamics on the position variable $q \in \mathcal{D}$ only:

$$dq_t = -\nabla V(q_t)\, dt + \sqrt{\frac{2}{\beta}} dW_t, \qquad (6)$$

where W_t is a standard $3N$-dimensional Wiener process. It can be shown that this system is an ergodic process for the configurational invariant measure $\nu(dq) = Z_\nu^{-1} \exp(-\beta V(q))\, dq$, the ergodicity following from the elliptic nature of the generator of the process. The dynamics (6) may be solved numerically using the Euler–Maruyama scheme:

$$q^{n+1} = q^n - \Delta t \nabla V(q^n) + \sqrt{\frac{2\Delta t}{\beta}}\, G^n, \qquad (7)$$

where the $(G^n)_{n \geq 0}$ are independent and identically distributed (i.i.d.) centered Gaussian random vectors in $\mathbb{R}^{3N}$ with identity covariance matrix $\mathbb{E}\,(G^n \otimes G^n) = \mathrm{Id}_{3N}$. Although the discretization scheme does not exactly preserve the canonical measure, it can be shown under certain boundedness assumptions (see [12, 21]), that the numerical scheme is ergodic, with an invariant probability close to the canonical measure ν in a suitable norm. The numerical bias may be eliminated using a Metropolis rule, see, e.g., [16, 18].

Langevin Dynamics

Hamiltonian dynamics preserve the energy, while a sampling of the canonical measure requires visiting all the energy levels. Langevin dynamics is a model of a Hamiltonian system coupled with a heat bath, defined by the following equations:

$$\begin{cases} dq_t = M^{-1} p_t\, dt, \\ dp_t = -\nabla V(q_t)\, dt - \gamma(q_t)\, M^{-1} p_t\, dt + \sigma(q_t)\, dW_t, \end{cases} \qquad (8)$$

where W_t is a $3N$-dimensional standard Brownian motion, and σ and γ are (possibly position dependent) $3N \times 3N$ real matrices. The term $\sigma(q_t)\, dW_t$ is a fluctuation term bringing energy into the system, this energy being dissipated through the viscous friction term $-\gamma(q_t)\, M^{-1} p_t\, dt$. The canonical measure is preserved precisely when the "fluctuation-dissipation" relation $\sigma\sigma^T = \frac{2\gamma}{\beta}$ is satisfied. Many variants and extensions of Langevin dynamics are available.

Using the Hörmander conditions, it is possible to demonstrate ergodicity of the system provided $\sigma(q)$ has full rank (i.e., a rank equal to $3N$) for all q in position space. A spectral gap can also be demonstrated under appropriate assumptions on the potential energy function, relying on recent advances in hypocoercivity [22] or thanks to Lyapunov techniques.

Brownian motion may be viewed as either the noninertial limit ($m \to 0$) of the Langevin dynamics, or its overdamped limit ($\gamma \to \infty$) with a different timescaling.

The discretization of stochastic differential equations, such as Langevin dynamics, is still a topic of research. Splitting methods, which divide the system into deterministic and stochastic components, are increasingly used for this purpose. As an illustration, one may adopt a method whereby Verlet integration is supplemented by an "exact" treatment of the Ornstein–Uhlenbeck process, replacing

$$dp_t = -\gamma(q_t)\, M^{-1} p_t\, dt + \sigma(q_t)\, dW_t$$

by a discrete process that samples the associated Gaussian distribution. In some cases, it is possible to show that such a method is ergodic.

Numerical discretization methods for Langevin dynamics may be corrected in various ways to exactly preserve the canonical measure, using the Metropolis technique [5, 13] (see, e.g., the discussion in [9], Sect. 2.2).

Deterministic Dynamics on Extended Phase Spaces

It is possible to modify Hamiltonian dynamics by the addition of control laws in order to sample the canonical (or some other) distribution. The simplest example of such a scheme is the Nosé–Hoover method [6, 15] which replaces Newton's equations of motion by the system:

$$
\begin{aligned}
\dot{q} &= M^{-1}p,\\
\dot{p} &= -\nabla V(q) - \xi p,\\
\dot{\xi} &= Q^{-1}(p^T M^{-1} p - N k_B T),
\end{aligned}
$$

where $Q > 0$ is a parameter. It can be shown that this dynamics preserves the product distribution $e^{-\beta H(q,p)} e^{-\beta Q \xi^2/2}$ as a stationary macrostate. It is, in some cases (e.g., when the underlying system is linear), not ergodic, meaning that the invariant distribution is not unique [7]. Nonetheless, the method is still popular for sampling calculations. The best arguments for its continued success, which have not been founded rigorously yet, are that (a) molecular systems typically have large phase spaces and may incorporate liquid solvent, steep potentials, and other mechanisms that provide a strong internal diffusion property or (b) any inaccessible regions in phase space may not contribute much to the averages of typical quantities of interest.

The accuracy of sampling can sometimes be improved by stringing together "chains" of additional variables [11], but such methods may introduce additional and unneeded complexity (especially as there are more reliable alternatives, see below). When ergodicity is not a concern (e.g., when a detailed atomistic model of water is involved), an alternative to the Nosé–Hoover method is to use the Nosé–Poincaré method [1] which is derived from an extended Hamiltonian and which allows the use of symplectic integrators (preserving phase space volume and, approximately, energy, and typically providing better long-term stability; see ▶ Molecular Dynamics).

Hybrid Methods by Stochastic Modification

When ergodicity is an issue, it is possible to enhance extended dynamics methods by the incorporation of stochastic processes, for example, as defined by the addition of Ornstein–Uhlenbeck terms. One such method has been proposed in [19]. It replaces the Nosé–Hoover system by the highly degenerate stochastic system:

$$
\begin{aligned}
dq_t &= M^{-1} p_t \, dt,\\
dp_t &= (-\nabla V(q_t) - \xi p_t)\, dt,\\
d\xi_t &= \left[Q^{-1} \left(p_t^T M^{-1} p_t - \frac{N}{\beta} \right) - \gamma \right] dt + \sqrt{\frac{2\gamma}{\beta Q}}\, dW_t,
\end{aligned}
$$

which incorporates only a scalar noise process. This method has been called the Nosé–Hoover–Langevin method in [8], where also ergodicity was proved in the case of an underlying harmonic system (V quadratic) under certain assumptions. A similar technique, the "Langevin Piston" [4] has been suggested to control the pressure in molecular dynamics, where the sampling is performed with respect to the NPT (isobaric–isothermal) ensemble.

References

1. Bond, S., Laird, B., Leimkuhler, B.: The Nosé-Poincaré method for constant temperature molecular dynamics. J. Comput. Phys. **151**, 114–134 (1999)
2. Cancès, E., Legoll, F., Stoltz, G.: Theoretical and numerical comparison of sampling methods for molecular dynamics. Math. Model. Numer. Anal. **41**(2), 351–390 (2007)
3. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte-Carlo. Phys. Lett. B **195**(2), 216–222 (1987)
4. Feller, S., Zhang, Y., Pastor, R., Brooks, B.: Constant pressure molecular dynamics simulation: the langevin piston method. J. Chem. Phys. **103**(11), 4613–4621 (1995)
5. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
6. Hoover, W.: Canonical dynamics: equilibrium phase space distributions. Phys. Rev. A **31**, 1695–1697 (1985)
7. Legoll, F., Luskin, M., Moeckel, R.: Non-ergodicity of the Nosé-Hoover thermostatted harmonic oscillator. Arch. Ration. Mech. Anal. **184**, 449–463 (2007)
8. Leimkuhler, B., Noorizadeh, N., Theil, F.: A gentle stochastic thermostat for molecular dynamics. J. Stat. Phys. **135**(2), 261–277 (2009)
9. Lelièvre, T., Rousset, M., Stoltz, G.: Free Energy Computations: a Mathematical Perspective. Imperial College Press, London/Hackensack (2010)
10. Marinari, E., Parisi, G.: Simulated tempering – a new Monte-Carlo scheme. Europhys. Lett. **19**(6), 451–458 (1992)
11. Martyna, G., Klein, M., Tuckerman, M.: Nosé-Hoover chains: the canonical ensemble via continuous dynamics. J. Chem. Phys. **97**(4), 2635–2643 (1992)
12. Mattingly, J.C., Stuart, A.M., Higham, D.J.: Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. Stoch. Process. Appl. **101**(2), 185–232 (2002)
13. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1091 (1953)
14. Meyn, S.P., Tweedie, R.L.: Markov Chains and Stochastic Stability. Communications and Control Engineering Series. Springer, London/New York (1993)
15. Nosé, S.: A molecular-dynamics method for simulations in the canonical ensemble. Mol. Phys. **52**, 255–268 (1984)

16. Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. J. R. Stat. Soc. B **60**, 255–268 (1998)
17. Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**, 110–120 (1997)
18. Rossky, P.J., Doll, J.D., Friedman, H.L.: Brownian dynamics as smart Monte Carlo simulation. J. Chem. Phys. **69**, 4628–4633 (1978)
19. Samoletov, A., Chaplain, M.A.J., Dettmann, C.P.: Thermostats for "slow" configurational modes. J. Stat. Phys. **128**, 1321–1336 (2007)
20. Schütte, C.: Habilitation thesis. Freie Universität Berlin, Berlin (1999). http://publications.mi.fu-berlin.de/89/
21. Talay, D., Tubaro, L.: Expansion of the global error for numerical schemes solving stochastic differential equations. Stoch. Anal. Appl. **8**(4), 483–509 (1990)
22. Villani, C.: Hypocoercivity. Mem. Am. Math. Soc. **202**(950), 141 (2009)

Schrödinger Equation for Chemistry

Harry Yserentant
Institut für Mathematik, Technische Universität
Berlin, Berlin, Germany

Mathematics Subject Classification

81-02; 81V55; 35J10

Short Definition

The Schrödinger equation forms the basis of nonrelativistic quantum mechanics and is fundamental for our understanding of atoms and molecules. The entry motivates this equation and embeds it into the general framework of quantum mechanics.

Description

Introduction

Quantum mechanics links chemistry to physics. Conceptions arising from quantum mechanics form the framework for our understanding of atomic and molecular processes. The history of quantum mechanics began around 1900 with Planck's analysis of the black-body radiation, Einstein's interpretation of the photoelectric effect, and Bohr's theory of the hydrogen atom. A unified framework allowing for a systematic study of quantum phenomena arose, however, first in the 1920s. Starting point was de Broglie's observation of the wave-like behavior of matter, finally result- ing in the Schrödinger equation [8] and [3] for the multiparticle case. The purpose of this article is to motivate this equation from some basic principles and to sketch at the same time the mathematical structure of quantum mechanics. More information can be found in textbooks on quantum mechanics like Atkins and Friedman [1] or Thaller [11, 12]. The first one is particularly devoted to the understanding of the molecular processes that are important for chemistry. The second and the third one more emphasize the mathematical structure and contain a lot of impressive visualizations. The monograph [4] gives an introduction to the mathematical theory. A historically very interesting text, in which the mathematically framework of quantum mechanics has been established and which was at the same time a milestone in the development of spectral theory, is von Neumann's seminal treatise [13]. The present exposition is largely taken from Yserentant [14].

The Schrödinger Equation of a Free Particle

Let us first recall the notion of a plane wave, a complex-valued function

$$\mathbb{R}^d \times \mathbb{R} \to \mathbb{C} : (x,t) \to \mathrm{e}^{\mathrm{i}k\cdot x - \mathrm{i}\omega t}, \tag{1}$$

with $k \in \mathbb{R}^d$ the wave vector and $\omega \in \mathbb{R}$ the frequency. A dispersion relation $\omega = \omega(k)$ assigns to each wave vector a characteristic frequency. Such dispersion relations fix the physics that is described by this kind of waves. Most common is the case $\omega = c|k|$ which arises, for example, in the propagation of light in vacuum. When the wave nature of matter was recognized, the problem was to guess the dispersion relation for the matter waves: to guess, as this hypothesis creates a new kind of physics that cannot be deduced from known theories. A good starting point is Einstein's interpretation of the photoelectric effect. When polished metal plates are irradiated by light of sufficiently short wave length they may emit electrons. The magnitude of the electron current is as expected proportional to the intensity of the light source, but their energy surprisingly to the wave length or the frequency of the incoming light. Einstein's explanation was that light consists of single light quanta with energy and momentum

$$E = \hbar\omega, \quad p = \hbar k \tag{2}$$

depending on the frequency ω and the wave vector k. The quantity

$$\hbar = 1.0545716 \cdot 10^{-34}\,\mathrm{kg\,m^2\,s^{-1}}$$

is Planck's constant, an incredibly small quantity of the dimension energy × time called action, reflecting the size of the systems quantum mechanics deals with. To obtain from (2) a dispersion relation, Schrödinger started first from the energy-momentum relation of special relativity, but this led by reasons not to be discussed here to the wrong predictions. He therefore fell back to the energy-momentum relation

$$E = \frac{1}{2m}\,|p|^2$$

from classical, Newtonian mechanics. It leads to the dispersion relation

$$\omega = \frac{\hbar}{2m}\,|k|^2$$

for the plane waves (1). These plane waves can be superimposed to wave packets

$$\psi(x,t) = \left(\frac{1}{\sqrt{2\pi}}\right)^3 \int \mathrm{e}^{-\mathrm{i}\frac{\hbar}{2m}|k|^2 t}\, \widehat{\psi}_0(k)\, \mathrm{e}^{\mathrm{i}k\cdot x}\,\mathrm{d}k. \tag{3}$$

These wave packets are the solutions of the partial differential equation

$$\mathrm{i}\hbar\,\frac{\partial\psi}{\partial t} = -\,\frac{\hbar^2}{2m}\,\Delta\psi, \tag{4}$$

the Schrödinger equation for a free particle of mass m in absence of external forces.

The Schrödinger equation (4) is of first order in time. Its solutions, the wavefunctions of free particles, are uniquely determined by their initial state ψ_0. If ψ_0 is a rapidly decreasing function (in the Schwartz space) the solution possesses time derivatives of arbitrary order, and all of them are rapidly decreasing functions of the spatial variables. To avoid technicalities, we assume this for the moment. We further observe that

$$\int |\psi(x,t)|^2\,\mathrm{d}x = \int |\widehat{\psi}(k,t)|^2\,\mathrm{d}k$$

remains constant in time. This follows from Plancherel's theorem, a central result of Fourier analysis. We assume in the sequel that this value is normalized to 1, which is basic for the statistical interpretation of the wavefunctions ψ. The quantities $|\psi|^2$ and $|\widehat{\psi}|^2$ can then be interpreted as probability densities. The integrals

$$\int_{\Omega} |\psi(x,t)|^2\,\mathrm{d}x, \qquad \int_{\widehat{\Omega}} |\widehat{\psi}(k,t)|^2\,\mathrm{d}k$$

represent the probabilities to find the particle at time t in the region Ω of the position space, respectively, the region $\widehat{\Omega}$ of the momentum space. The quantity

$$\int \frac{\hbar^2}{2m}\,|k|^2\,|\widehat{\psi}(k,t)|^2\,\mathrm{d}k,$$

is the expectation value of the kinetic energy. With help of the Hamilton operator

$$H = -\,\frac{\hbar^2}{2m}\,\Delta, \tag{5}$$

this expectation value can be rewritten as

$$\int \psi\,\overline{H\psi}\,\mathrm{d}x = (\psi, H\psi).$$

The expectation values of the components of the momentum are in vector notation

$$\int \hbar k\,|\widehat{\psi}(k,t)|^2\,\mathrm{d}k.$$

Introducing the momentum operator

$$p = -\,\mathrm{i}\hbar\,\nabla \tag{6}$$

their position representation is the inner product

$$\int \psi\,\overline{p\psi}\,\mathrm{d}x = (\psi, p\psi).$$

The expectation values of the three components of the particle position are finally

$$\int x\,|\psi(x,t)|^2\,\mathrm{d}x = (\psi, q\psi),$$

with q the position operator given by $\psi \rightarrow x\psi$. This coincidence between observable physical quantities like energy, momentum, or position and operators acting upon the wavefunctions is in no way accidental. It forms the heart of quantum mechanics.

The Mathematical Framework of Quantum Mechanics

We have seen that the physical state of a free particle at a given time t is completely determined by a function in the Hilbert space L_2 that again depends uniquely on the state at a given initial time. In the case of more general systems, the space L_2 is replaced by another Hilbert space, but the general concept remains:

Postulate 1. *A quantum-mechanical system consists of a complex Hilbert space $\mathcal{H}$ with inner product $(\cdot,\cdot)$ and a one-parameter group $U(t)$, $t \in \mathbb{R}$, of unitary linear operators on $\mathcal{H}$ with*

$$U(0) = \mathrm{I}, \quad U(s+t) = U(s)U(t)$$

that is strongly continuous in the sense that for all $\psi \in \mathcal{H}$ in the Hilbert space norm

$$\lim_{t\to 0} U(t)\psi \;=\; \psi.$$

A state of the system corresponds to a normalized vector in $\mathcal{H}$. The time evolution of the system is described by the group of the propagators $U(t)$; the state

$$\psi(t) = U(t)\psi(0) \tag{7}$$

of the system at time t is uniquely determined by its state at time $t = 0$.

In the case of free particles considered so far, the solution of the Schrödinger equation and with that time evolution is given by (3). The evolution operators $U(t)$, or propagators, read therefore in the Fourier or momentum representation

$$\widehat{\psi}(k) \;\rightarrow\; \mathrm{e}^{-\mathrm{i}\frac{\hbar}{2m}|k|^2 t}\,\widehat{\psi}(k).$$

Strictly speaking, they have first only been defined for rapidly decreasing functions, functions in a dense subspace of L_2, but it is obvious from Plancherel's theorem that they can be uniquely extended from there to L_2 and have the required properties.

The next step is to move from Postulate 1 to an abstract version of the Schrödinger equation. For that we have to establish a connection between such strongly continuous groups of unitary operators and abstract Hamilton operators. Let $D(H)$ be the linear subspace of the given system Hilbert space $\mathcal{H}$ that consists of those elements ψ in $\mathcal{H}$ for which the limit

$$H\psi \;=\; \mathrm{i}\hbar \lim_{\tau\to 0} \frac{U(\tau)-\mathrm{I}}{\tau}\,\psi$$

exists in the sense of norm convergence. The mapping $\psi \rightarrow H\psi$ from the domain $D(H)$ into the Hilbert space $\mathcal{H}$ is then called the generator H of the group. The generator of the evolution operator of the free particle is the operator

$$H \;=\; -\frac{\hbar^2}{2m}\,\Delta \tag{8}$$

with the Sobolev space H^2 as domain of definition $D(H)$. In view of this observation, the following result for the general abstract case is unsurprising:

Theorem 1 *For all initial values $\psi(0)$ in the domain $D(H)$ of the generator of the group of the propagators $U(t)$, the elements (7) are contained in $D(H)$, too, depend continuously differentiable on t, and satisfy the differential equation*

$$\mathrm{i}\hbar\,\frac{\mathrm{d}}{\mathrm{d}t}\,\psi(t) \;=\; H\psi(t). \tag{9}$$

It should be noted once more, however, that the differential (9), the abstract Schrödinger equation, makes sense only for initial values in the domain of the generator H, but that the propagators are defined on the whole Hilbert space.

A little calculation shows that the generators of one-parameter unitary groups are necessarily symmetric. More than that, they are even selfadjoint. There is a direct correspondence between unitary groups and selfadjoint operators, Stone's theorem, a cornerstone in the mathematical foundation of quantum mechanics:

Theorem 2 *If $U(t)$, $t \in \mathbb{R}$, is a one-parameter unitary group as in Postulate 1, the domain $D(H)$ of its generator H is a dense subset of the underlying Hilbert space and the generator itself selfadjoint. Every selfadjoint operator H is conversely the generator of such a one-parameter unitary group, that is usually denoted as*

$$U(t) \;=\; \mathrm{e}^{-\frac{\mathrm{i}}{\hbar}Ht}.$$

Instead of the unitary group of the propagators, a quantum-mechanical system can be thus equivalently fixed by the generator H of this group, the Hamilton operator, or in the language of physics, the Hamiltonian of the system.

In our discussion of the free particle, we have seen that there is a direct correspondence between the expectation values of the energy, the momentum, and the position of the particle and the energy or Hamilton operator (5), the momentum operator (6), and the position operator $x \to x\psi$. Each of these operators is selfadjoint. This reflects the general structure of quantum mechanics:

Postulate 2. *Observable physical quantities, or observables, are in quantum mechanics represented by selfadjoint operators $A : D(A) \to \mathcal{H}$ defined on dense subspaces $D(A)$ of the system Hilbert space $\mathcal{H}$. The quantity*

$$\langle A \rangle \;=\; (\psi, A\psi) \tag{10}$$

is the expectation value of a measurement of A for the system in state $\psi \in D(A)$.

At this point, we have to recall the statistical nature of quantum mechanics. Quantum mechanics does not make predictions on the outcome of a single measurement of a quantity A but only on the mean result of a large number of measurements on "identically prepared" states. The quantity (10) has thus to be interpreted as the mean result that one obtains from a large number of such measurements. This gives reason to consider the standard deviation or uncertainty

$$\Delta A \;=\; \| A\psi \;-\; \langle A \rangle \psi \|$$

for states $\psi \in D(A)$. The uncertainty is zero if and only if $A\psi = \langle A \rangle \psi$, that is, if ψ is an eigenvector of A for the eigenvalue $\langle A \rangle$. Only in such eigenstates the quantity represented by the operator A can be sharply measured without uncertainty. The likelihood that a measurement returns a value outside the spectrum of A is zero.

One of the fundamental results of quantum mechanics is that, only in exceptional cases, can different physical quantities be measured simultaneously without uncertainty, the Heisenberg uncertainty principle. Its abstract version reads as follows:

Theorem 3 *Let A and B two selfadjoint operators and let ψ be a normalized state in the intersection of $D(A)$ and $D(B)$ such that $A\psi \in D(B)$ and $B\psi \in D(A)$. The product of the corresponding uncertainties is then bounded from below by*

$$\Delta A \, \Delta B \;\geq\; \frac{1}{2}\,|((BA - AB)\psi, \psi)|. \tag{11}$$

The proof is an exercise in linear algebra. As an example, we consider the components

$$q_k \;=\; x_k, \quad p_k \;=\; -\,\mathrm{i}\hbar\,\frac{\partial}{\partial x_k}$$

of the position and the momentum operator. Their commutators are

$$q_k p_k - p_k q_k \;=\; \mathrm{i}\hbar\,\mathrm{I}.$$

This results in the Heisenberg uncertainty principle

$$\Delta p_k \, \Delta q_k \;\geq\; \frac{1}{2}\,\hbar. \tag{12}$$

Position and momentum therefore can never be determined simultaneously without uncertainty, independent of the considered state of the system. The inequality (12) and with that also (11) are sharp as the instructive example

$$\psi(x) \;=\; \Big(\frac{1}{\sqrt{\vartheta}}\Big)^3 \psi_0\Big(\frac{x}{\vartheta}\Big)$$

of the rescaled three-dimensional Gauss functions

$$\psi_0(x) \;=\; \Big(\frac{1}{\sqrt{\pi}}\Big)^{3/2} \exp\Big(-\,\frac{1}{2}\,|x|^2\Big)$$

of arbitrary width demonstrates. For these wavefunctions, the inequality (12) actually turns into an equality. From

$$\widehat{\psi}(k) \;=\; (\sqrt{\vartheta})^3\,\psi_0(\vartheta k)$$

one recognizes that a sharp localization in space, that is, a small parameter ϑ determining the width of ψ, is combined with a loss of localization in momentum.

States with a well defined, sharp energy E play a particularly important role in quantum mechanics, that is, solutions $\psi \neq 0$ in $\mathcal{H}$ of the eigenvalue problem

$$H\psi \;=\; E\psi,$$

the stationary Schrödinger equation. The functions

$$t \to \mathrm{e}^{-\mathrm{i}\frac{E}{\hbar}t}\psi$$

represent then solutions of the original time-dependent Schrödinger equation. The main focus of quantum chemistry is on stationary Schrödinger equations.

The Quantum Mechanics of Multiparticle Systems

Let us assume that we have a finite collection of N particles of different kind with the spaces $L_2(\Omega_i)$ as system Hilbert spaces. The Hilbert space describing the system that is composed of these particles is then the tensor product of these Hilbert spaces or a subspace of this space, that is, a space of square integrable wavefunctions

$$\psi : \Omega_1 \times \ldots \times \Omega_N \to \mathbb{C}$$

with the N-tuples $(\xi_1, \ldots, \xi_N)$, $\xi_i \in \Omega_i$, as arguments. From the point of view of mathematics, this is of course another postulate that can in a strict sense not be derived from anything else, but is motivated by the statistical interpretation of the wavefunctions and of the quantity $|\psi|^2$ as a probability density. Quantum-mechanical particles of the same type, like electrons, can, however, not be distinguished from each other by any means or experiment. This is both a physical statement and a mathematical postulate that needs to be specified precisely. It has striking consequences for the form of the physically admissible wavefunctions and of the Hilbert spaces that describe such systems of indistinguishable particles.

To understand these consequences, we have to recall that an observable quantity like momentum or energy is described in quantum mechanics by a selfadjoint operator A and that the inner product $(\psi, A\psi)$ represents the expectation value for the outcome of a measurement of this quantity in the physical state described by the normalized wavefunction ψ. At least a necessary condition that two normalized elements or unit vectors ψ and ψ' in the system Hilbert space $\mathcal{H}$ describe the same physical state is surely that $(\psi, A\psi) = (\psi', A\psi')$ for all selfadjoint operators $A : D(A) \subseteq \mathcal{H} \to \mathcal{H}$ whose domain $D(A)$ contains both ψ and ψ', that is, that the expectation values of all possible observables coincide. This requirement fixes such states almost completely. Wavefunctions that describe the same physical state can differ at most by a constant phase shift $\psi \to \mathrm{e}^{\mathrm{i}\theta}\psi$, θ a real number. Wavefunctions that differ by such a phase shift lead to the same expectation values of observable quantities. The proof is again an exercise in linear algebra. In view of this discussion, the requirements on the wavefunctions describing a system of indistinguishable particles are rather obvious and can be formulated in terms of the operations that formally exchange the single particles:

Postulate 3. *The Hilbert space of a system of N indistinguishable particles with system Hilbert space $L_2(\Omega)$ consists of complex-valued, square integrable functions*

$$\psi : (\xi_1, \ldots, \xi_N) \to \psi(\xi_1, \ldots, \xi_N)$$

on the N-fold cartesian product of Ω, that is, is a subspace of $L_2(\Omega^N)$. For every ψ in this space and every permutation P of the arguments ξ_i, the function $\xi \to \psi(P\xi)$ is also in this space, and moreover it differs from ψ at most by a constant phase shift.

This postulate can be rather easily translated into a symmetry condition on the wavefunctions that governs the quantum mechanics of multiparticle systems:

Theorem 4 *The Hilbert space describing a system of indistinguishable particles either consists completely of antisymmetric wavefunctions, functions ψ for which*

$$\psi(P\xi) = \operatorname{sign}(P)\psi(\xi)$$

holds for all permutations P of the components $\xi_1, \ldots, \xi_N$ of ξ, that is, of the single particles, or only of symmetric wavefunctions, wavefunctions for which

$$\psi(P\xi) = \psi(\xi)$$

holds for all permutations P of the arguments.

Which of the two choices is realized depends solely on the kind of particles and cannot be decided in the present framework. Particles with antisymmetric wavefunctions are called fermions and particles with symmetric wavefunctions bosons.

Quantum chemistry is mainly interested in electrons. Electrons have a position in space and an internal property called spin that in many respects behaves like an angular momentum. The spin σ of an electron can attain the two values $\sigma = \pm 1/2$. The configuration space of an electron is therefore not the $\mathbb{R}^3$ but the cartesian product

$$\Omega = \mathbb{R}^3 \times \{-1/2, +1/2\}.$$

The space $L_2(\Omega)$ consists of the functions $\psi : \Omega \to \mathbb{C}$ with square integrable components $x \to \psi(x, \sigma)$, $\sigma = \pm 1/2$, and is equipped with the inner product

$$(\psi, \phi) = \sum_{\sigma = \pm 1/2} \int \psi(x, \sigma) \overline{\phi(x, \sigma)} \, \mathrm{d}x.$$

A system of N electrons is correspondingly described by wavefunctions

$$\psi : (\mathbb{R}^3)^N \times \{-1/2, 1/2\}^N \to \mathbb{C} \tag{13}$$

with square integrable components $x \to \psi(x, \sigma)$, where $x \in \mathbb{R}^{3N}$ and σ is a vector consisting of N spins $\sigma_i = \pm 1/2$. These wavefunctions are equipped with the inner product

$$(\psi, \phi) = \sum_{\sigma} \int \psi(x, \sigma) \overline{\phi(x, \sigma)} \, \mathrm{d}x,$$

where the sum now runs over the 2^N possible spin vectors σ.

Electrons are fermions, as all particles with half-integer spin. That is, the wavefunctions change their sign under a simultaneous exchange of the positions x_i and x_j and the spins σ_i and σ_j of electrons $i \neq j$. They are, in other words, antisymmetric in the sense that

$$\psi(Px, P\sigma) = \operatorname{sign}(P) \psi(x, \sigma)$$

holds for arbitrary simultaneous permutations $x \to Px$ and $\sigma \to P\sigma$ of the electron positions and spins. This is a general version of the Pauli principle, a principle that is of fundamental importance for the physics of atoms and molecules. The Pauli principle has stunning consequences. It entangles the electrons with each other, without the presence of any direct interaction force. A wavefunction (13) describing such a system vanishes at points (x, σ) at which $x_i = x_j$ and $\sigma_i = \sigma_j$ for indices $i \neq j$. This means that two electrons with the same spin cannot meet at the same place, a purely quantum-mechanical repulsion effect that has no counterpart in classical physics.

The Molecular Schrödinger Equation

Neglecting spin, the system Hilbert space of an atom or molecule consisting of N particles (electrons and nuclei) is the space $L_2(\mathbb{R}^3)^N = L_2(\mathbb{R}^{3N})$. The Hamilton operator

$$H = -\sum_{i=1}^{N} \frac{1}{2m_i} \Delta_i + \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{Q_i Q_j}{|x_i - x_j|}, \tag{14}$$

written down here in dimensionless form, is derived via the correspondence principle from its counterpart in classical physics, the Hamilton function

$$H(p, q) = -\sum_{i=1}^{N} \frac{1}{2m_i} |p_i|^2 + \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{Q_i Q_j}{|q_i - q_j|}$$

or total energy of a system of point-like particles in the potential field

$$V(q) = \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{Q_i Q_j}{|q_i - q_j|}.$$

The m_i are the masses of the particles in multiples of the electron mass and the Q_i the charges of the particles in multiples of the electron charge. As has first been shown by Kato [7], the Hamilton operator (14) can be uniquely extended from the space of the infinitely differentiable functions with bounded support to a selfadjoint operator H from its domain of definition $D(H) \subset L_2(\mathbb{R}^{3N})$ to $L_2(\mathbb{R}^{3N})$. It fits therefore into the abstract framework of quantum mechanics sketched above. The domain $D(H)$ of the extended operator is the Sobolev space H^2 consisting of the twice weakly differentiable functions with first and second order weak derivatives in L_2, respectively a subspace of this Sobolev space consisting of components of the full, spin-dependent wavefunctions in accordance with the Pauli principle if spin is taken into account. The resulting Schrödinger equation

$$\mathrm{i} \frac{\partial \psi}{\partial t} = H\psi$$

is an extremely complicated object, because of the high dimensionality of the problem but also because of the oscillatory character of its solutions and the many different time scales on which they vary and which can range over many orders of magnitude. Comprehensive survey articles on the properties of atomic and molecular Schrödinger operators are Hunziker and Sigal [6] and Simon [9].

Following Born and Oppenheimer [2], the full problem is usually split into the electronic Schrödinger equation describing the motion of the electrons in the field of given clamped nuclei, and an equation for the motion of the nuclei in a potential field that is determined by solutions of the electronic equation. The transition from the full Schrödinger equation taking also into account the motion of the nuclei to the electronic Schrödinger equation is a mathematically very subtle problem; see [10] and the literature cited therein or the article of Hagedorn (▸ Born–Oppenheimer Approximation, Adiabatic Limit, and Related Math. Issues) for more information. The intuitive idea behind this splitting is that the electrons move much more rapidly than the much heavier nuclei and almost instantaneously follow their motion. Most of quantum chemistry is devoted to the solution of the stationary electronic Schrödinger equation, the eigenvalue problem for the electronic Hamilton operator

$$H = -\frac{1}{2}\sum_{i=1}^{N}\Delta_i + V_0(x) + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\frac{1}{|x_i - x_j|}$$

again written down in dimensionless form, where

$$V_0(x) = -\sum_{i=1}^{N}\sum_{\nu=1}^{K}\frac{Z_\nu}{|x_i - a_\nu|}$$

is the nuclear potential. It acts on functions with arguments $x_1, \ldots, x_N$ in $\mathbb{R}^3$, which are associated with the positions of the considered electrons. The a_ν are the now fixed positions of the nuclei and the values Z_ν the charges of the nuclei in multiples of the electron charge. The equation has still to be supplemented by the symmetry constraints arising from the Pauli principle.

The spectrum of the electronic Schrödinger operator is bounded from below. Its essential spectrum is, by the Hunziker-van Winter-Zhislin theorem, a semi-infinite interval; see [4] for details. Of interest for chemistry are configurations of electrons and nuclei for which the minimum of the total spectrum is an isolated eigenvalue of finite multiplicity, the ground state energy of the system. The assigned eigenfunctions, the ground states, as well as all other eigenfunctions for eigenvalues below the essential spectrum decay then exponentially. That means that the nuclei can bind all electrons. More information on the mathematical properties of these eigenfunctions can be found in ▸ Exact Wavefunctions Properties. Chemists are mainly interested in the ground states. The position of the nuclei is then determined minimizing the ground state energy as function of their positions, a process that treats the nuclei as classical objects. It is called geometry optimization.

The Born-Oppenheimer approximation is only a first step toward the computationally feasible models that are actually used in quantum chemistry. The historically first and most simple of these models is the Hartree-Fock model in which the true wavefunctions are approximated by correspondingly antisymmetrized tensor products

$$u(x) = \prod_{i=1}^{N}\phi_i(x_i)$$

of functions ϕ_i of the electron positions $x_i \in \mathbb{R}^3$. These orbital functions are then determined via a variational principle. This intuitively very appealing ansatz often leads to surprisingly accurate results. Quantum chemistry is full of improvements and extensions of this basic approach; see the comprehensive monograph [5] for further information. Many entries in this encyclopedia are devoted to quantum chemical models and approximation methods that are derived from the Schrödinger equation. We refer in particular to the article (▸ Hartree–Fock Type Methods) on the Hartree-Fock method, to the contributions (▸ Post-Hartree-Fock Methods and Excited States Modeling) on post-Hartree Fock methods and (▸ Coupled-Cluster Methods) on the coupled cluster method, and to the article (▸ Density Functional Theory) on density functional theory. Time-dependent problems are treated in the contribution (▸ Quantum Time-Dependent Problems).

References

1. Atkins, P., Friedman, R.: Molecular Quantum Mechanics. Oxford University Press, Oxford (1997)
2. Born, M., Oppenheimer, R.: Zur Quantentheorie der Molekeln. Ann. Phys. **84**, 457–484 (1927)
3. Dirac, P.: Quantum mechanics of many electron systems. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. **123**, 714–733 (1929)
4. Gustafson, S., Sigal, I.: Mathematical Concepts of Quantum Mechanics. Springer, Berlin/Heidelberg/New York (2003)
5. Helgaker, T., Jørgensen, P., Olsen, J.: Molecular Electronic Structure Theory. Wiley, Chichester (2000)
6. Hunziker, W., Sigal, I.: The quantum N-body problem. J. Math. Phys. **41**, 3448–3510 (2000)
7. Kato, T.: Fundamental properties of Hamiltonian operators of Schrödinger type. Trans. Am. Math. Soc. **70**, 195–221 (1951)
8. Schrödinger, E.: Quantisierung als Eigenwertproblem. Ann. Phys. **79**, 361–376 (1926)
9. Simon, B.: Schrödinger operators in the twentieth century. J. Math. Phys. **41**, 3523–3555 (2000)
10. Teufel, S.: Adiabatic Perturbation Theory in Quantum Dynamics. Lecture Notes in Mathematics, vol. 1821. Springer, Berlin/Heidelberg/New York (2003)
11. Thaller, B.: Visual Quantum Mechanics. Springer, New York (2000)
12. Thaller, B.: Advanced Visual Quantum Mechanics. Springer, New York (2004)
13. von Neumann, J.: Mathematische Grundlagen der Quantenmechanik. Springer, Berlin (1932)
14. Yserentant, H.: Regularity and Approximability of Electronic Wave Functions. Lecture Notes in Mathematics, vol. 2000. Springer, Heidelberg/Dordrecht/London/New York (2010)

Schrödinger Equation: Computation

Shi Jin
Department of Mathematics and Institute of Natural Science, Shanghai Jiao Tong University, Shanghai, China
Department of Mathematics, University of Wisconsin, Madison, WI, USA

The Schrödinger Equation

The linear Schrödinger equation is a fundamental quantum mechanics equation that describes the complex-valued wave function $\Phi(t, \mathbf{x}, \mathbf{y})$ of molecules or atoms

$$i\hbar\partial_t\Phi(t,\mathbf{x},\mathbf{y}) = \mathcal{H}\Phi(t,\mathbf{x},\mathbf{y}), \qquad \mathbf{x}\in\mathbb{R}^N, \mathbf{y}\in\mathbb{R}^n, \tag{1}$$

where the vectors $\mathbf{x}$ and $\mathbf{y}$ denote the positions of N nuclei and n electrons, respectively, while $\hbar$ is the reduced Planck constant. The molecular Hamiltonian operator $\mathcal{H}$ consists of two parts, the kinetic energy operator of the nuclei and the electronic Hamiltonian $\mathcal{H}_e$ for fixed nucleonic configuration:

$$\mathcal{H} = -\sum_{j=1}^{N} \frac{\hbar^2}{2M_j}\Delta_{x_j} + \mathcal{H}_e(\mathbf{y},\mathbf{x}),$$

with,

$$\mathcal{H}_e(\mathbf{y},\mathbf{x}) = -\sum_{j=1}^{n} \frac{\hbar^2}{2m_j}\Delta_{y_j} + \sum_{j<k} \frac{1}{|y_j - y_k|} + \sum_{j<k} \frac{Z_j Z_k}{|x_j - x_k|} - \sum_{k=1}^{N}\sum_{j=1}^{n} \frac{Z_j}{|x_j - y_k|}.$$

Here m_j denotes mass of the j-th electron, and M_j, Z_j denote mass and charge of the j-th nucleus. The electronic Hamiltonian $\mathcal{H}_e$ consists of the kinetic energy of the electrons as well as the interelectronic repulsion potential, internuclear repulsion potential, and the electronic-nuclear attraction potential.

The Born-Oppenheimer Approximation

The main computational challenge to solve the Schrödinger equation is the high dimensionality of the molecular configuration space $\mathbb{R}^{N+n}$. For example, the carbon dioxide molecule CO_2 consists of 3 nuclei and 22 electrons; thus one has to solve the full time-dependent Schrödinger equation in space $\mathbb{R}^{75}$, which is a formidable task. The *Born-Oppenheimer approximation* [1] is a commonly used approach in computational chemistry or physics to reduce the degrees of freedom.

This approximation is based on the mass discrepancy between the light electrons, which move fast, thus will be treated quantum mechanically, and the heavy nuclei that move slower and are treated classically. Here one first solves the following time-independent *electronic* eigenvalue problems:

$$\mathcal{H}_e(\mathbf{y},\mathbf{x})\psi_k(\mathbf{y};\mathbf{x}) = E_k(\mathbf{x})\psi_k(\mathbf{y};\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^N, \quad k = 1,2,\ldots. \tag{2}$$

Assuming that the spectrum of $\mathcal{H}_e$, a self-adjoint operator, is discrete with a complete set of orthonormal eigenfunctions $\{\psi_k(\mathbf{y};\mathbf{x})\}$ called the *adiabatic* basis, over the electronic coordinates for every fixed nucleus coordinates $\mathbf{x}$, i.e.,

$$\int_{-\infty}^{\infty} \psi_j^*(\mathbf{y};\mathbf{x})\psi_k(\mathbf{y};\mathbf{x})d\mathbf{y} = \delta_{jk},$$

where δ_{jk} is the Kronecker delta. The electronic eigenvalue $E_k(\mathbf{x})$, called the *potential energy surface*, depends on the positions $\mathbf{x}$ of the nuclei.

Next the total wave function $\Phi(t,\mathbf{x},\mathbf{y})$ is expanded in terms of the eigenfunctions $\{\psi_k\}$:

$$\Phi(t,\mathbf{x},\mathbf{y}) = \sum_k \phi_k(t,\mathbf{x})\psi_k(\mathbf{y};\mathbf{x}). \tag{3}$$

Assume $m_j = m$, and $M_j = M$, for all j. We take the atomic units by setting $\hbar = 1, Z = 1$ and introduce $\varepsilon = \sqrt{m/M}$. Typically ε ranges between 10^{-2} and 10^{-3}. Insert ansatz (3) into the time-dependent Schrödinger equation (1), multiply all the terms from the left by $\psi_k^*(\mathbf{y};\mathbf{x})$, and integrate with respect to $\mathbf{y}$, then one obtains a set of coupled differential equations:

$$i\varepsilon\frac{\partial}{\partial t}\phi_k(t,\mathbf{x}) = \left[-\sum_{j=1}^{N}\frac{\varepsilon^2}{2}\Delta_{x_j} + E_k(\mathbf{x})\right]\phi_k(t,\mathbf{x}) + \sum_l C_{kl}\phi_l(t,\mathbf{x}), \tag{4}$$

where the coupling operator C_{kl} is important to describe quantum transitions between different potential energy surfaces.

As long as the potential energy surfaces $\{E_k(\mathbf{x})\}$ are well separated, all the coupling operators C_{kl} are ignored, and one obtains a set of decoupled Schrödinger equations:

$$i\varepsilon\frac{\partial}{\partial t}\phi_k(t,\mathbf{x}) = \left[-\sum_{j=1}^{N}\frac{\varepsilon^2}{2}\Delta_{x_j} + E_k(\mathbf{x})\right]\phi_k(t,\mathbf{x}), \quad (t,\mathbf{x}) \in \mathbb{R}^+ \times \mathbb{R}^N. \tag{5}$$

Thus the nuclear motion proceeds without the transitions between electronic states or energy surfaces. This is also referred to as the *adiabatic approximation*.

There are two components in dealing with a quantum calculation. First, one has to solve the eigenvalue problem (2). Variational methods are usually used [10]. However, for large n, this remains an intractable task. Various mean field theories have been developed to reduce the dimension. In particular, the *Hartree-Fock Theory* [9] and the *Density Function Theory* [8] aim at representing the $3n$-dimensional electronic wave function into a product of one-particle wave function in 3 dimension. These approaches usually yield nonlinear Schrödinger equations.

The Semiclassical Limit

The second component in quantum simulation is to solve the time-dependent Schrödinger equation (5). Its numerical approximation per se is similar to that of the parabolic heat equation. Finite difference, finite element or, most frequently, spectral methods can be used for the spatial discretization. For the time discretization, one often takes a time splitting of Strong or Trotter type that separates the kinetic energy from the potential operators in alternating steps. However, due to the smallness of $\hbar$ or ε, the numerical resolution of the wave function remains difficult. A classical method to deal with such an oscillatory wave problem is the WKB method, which seeks solution of the form $\phi(t,\mathbf{x}) = A(t,\mathbf{x})e^{iS(t,\mathbf{x})/\hbar}$ (in the sequel, we consider only one energy level in (5), thus omitting the subscript k and replacing E by V). If one applies this ansatz in (5), by ignoring the $O(\varepsilon)$ term, one obtains the *eikonal equation* for phase S and *transport equation* for amplitude A:

$$\partial_t S + \frac{1}{2}|\nabla S|^2 + V(x) = 0; \tag{6}$$

$$\partial_t A + \nabla S \cdot \nabla A + \frac{A}{2}\Delta S = 0\,. \tag{7}$$

The eikonal equation (6) is a typical *Hamilton-Jacobi* equation, which develops singularities in S, usually referred to caustics in the context of geometric optics. Beyond the singularity, one has to superimpose the solutions of S, each of which satisfying the eikonal equation (6), since the solution becomes *multivalued* [6].

This equation can be solved by the method of characteristics, provided that $V(x)$ is sufficiently smooth. Its characteristic flow is given by the following Hamiltonian system of ordinary differential equations, which is Newton's second law:

$$\frac{d\mathbf{x}}{dt}(t,\mathbf{y}) = \xi(t,\mathbf{y}); \quad \frac{d\xi}{dt}(t,\mathbf{y}) = -\nabla_{\mathbf{x}} V(\mathbf{x}(t,\mathbf{y}))\,. \tag{8}$$

Another approach to study the semiclassical limit is the *Wigner transform* [12]:

$$w^{\varepsilon}[\phi^{\varepsilon}](\mathbf{x},\xi) := \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \phi\left(\mathbf{x}+\frac{\varepsilon}{2}\eta\right)\overline{\phi}\left(\mathbf{x}-\frac{\varepsilon}{2}\eta\right)e^{i\xi\cdot\eta}d\eta\,, \tag{9}$$

which is a convenient tool to study the limit of $\phi(t,\mathbf{x})$ to obtain the classical Liouville equation:

$$\partial_t w + \xi\cdot\nabla_{\xi} w - \nabla V(\mathbf{x})\cdot\nabla_{\xi} w = 0\,. \tag{10}$$

Its characteristic equation is given by the Hamiltonian system (8).

Various Potentials

The above Wigner analysis works well if the potential V is smooth. In applications, E can be discontinuous (corresponding to potential barriers), periodic (for solid mechanics with lattice structure), random (with an inhomogeneous background), or even nonlinear (where the equation is a field equation, with applications to optics and water waves, or a mean field equation as an approximation of the original multiparticle linear Schrödinger equation). Different approaches need to be taken for each of these different cases.

- *V is discontinuous*. Through a potential barrier, the quantum tunnelling phenomenon occurs and one has to handle wave transmission and reflections [3].
- *V is periodic*. The Bloch decomposition is used to decompose the wave field into sub-fields along each of the Bloch bands, which are the eigenfunctions associated with a perturbed Hamiltonian that includes $\mathcal{H}$ plus the periodic potential [13].
- *V is random*. Depending on the space dimension and strength of the randomness, the waves can be *localized* [2] or *diffusive*. In the latter case, the Wigner transform can be used to study the high-frequency limit [7].
- *V is nonlinear*. The semiclassical limit (10) fails after caustic formation. The understanding of this limit for strong nonlinearities remains a major mathematical challenge. Not much is known except in the one-dimensional defocusing nonlinearity case ($V = |\phi|^2$) [4].

In (4), when different E_k intersect, or get close, one cannot ignore the quantum transitions between different energy levels. A semiclassical approach, known as the *surface hopping method*, was developed by Tully. It is based on the classical Hamiltonian system (8), with a Monte Carlo procedure to account for the quantum transitions [11].

For a survey of semiclassical computational methods for the Schrödinger equation, see [5].

References

1. Born, M., Oppenheimer, R.: Zur Quantentheorie der Molekeln. Ann. Phys. **84**, 457–484 (1927)
2. Fröhlich, J., Spencer, T.: Absence of diffusion in the Anderson tight binding model for large disorder or low energy. Commun. Math. Phys. **88**, 465–471 (1983)
3. Griffiths, D.J.: Introduction to Quantum Mechanics, 2nd edn. Prentice Hall, Upper Saddle River (2004)
4. Jin, S., Levermore, C.D., McLaughlin, D.W.: The semiclassical limit of the defocusing NLS hierarchy. Commun. Pure Appl. Math. **52**(5), 613–654 (1999)
5. Jin, S., Markowich P., Sparber, C.: Mathematical and computational methods for semiclassical Schrodinger equations. Acta Numer. **20**, 211–289 (2011)
6. Maslov, V.P., Fedoriuk M.V.: Semi-Classical Approximation in Quantum Mechanics. D. Reidel, Dordrecht/Hollan (1981)
7. Papanicolaou, G., Ryzhik, L.: Waves and transport. In: Caffarelli, L., Weinan, E. (eds.) Hyperbolic equations and frequency interactions (Park City, UT, 1995). Amer. Math. Soc. Providence, RI **5**, 305–382 (1999)
8. Parr, R.G., Yang W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, New York (1989)
9. Szabo, A., Ostlund, N.S.: Modern Quantum Chemistry. Dover, Mineola/New York (1996)
10. Thijssen, J.M.: Computational Physics. Cambridge University Press, Cambridge (1999)
11. Tully, J.: Molecular dynamics with electronic transitions. J. Chem. Phys. **93**, 1061–1071 (1990)
12. Wigner, E.: On the quantum correction for thermodynamic equilibrium. Phys. Rev. **40**, 749–759 (1932)
13. Wilcox, C.H.: Theory of Bloch waves. J. Anal. Math. **33**, 146–167 (1978)

Scientific Computing

Hans Petter Langtangen[1,2], Ulrich Rüde[3], and Aslak Tveito[1,2]
[1]Simula Research Laboratory, Center for Biomedical Computing, Fornebu, Norway
[2]Department of Informatics, University of Oslo, Oslo, Norway
[3]Department of Computer Science, University Erlangen-Nuremberg, Erlangen, Germany

Scientific Computing is about practical methods for solving mathematical problems. One may argue that the field goes back to the invention of Mathematics, but today, the term Scientific Computing usually means application of computers to solve mathematical problems. The solution process consists of several key steps, which form the so-called *simulation pipeline*:

1. Formulation of a mathematical model which describes the scientific problem and is suited for computer-based solution methods and some chosen hardware
2. Construction of algorithms which precisely describe the computational steps in the model
3. Implementation of the algorithms in software
4. Verification of the implementation
5. Preparation of input data to the model
6. Analysis and visualization of output data from the model
7. Validation of the mathematical model, with associated parameter estimation
8. Estimation of the precision (or uncertainty) of the mathematical model for predictions

In any of the steps, it might be necessary to go back and change the formulation of the model or previous steps to make further progress with other items on the list. When this process of iterative improvement has reached a satisfactory state, one can perform *computer simulations* of a process in nature, technological devices, or society. In essence, that means using the computer as a laboratory to mimic processes in the real world. Such a lab enables impossible, unethical, or costly real-world experiments, but often the process of developing a computer model gives increased scientific insight in itself. The disadvantage of computer simulations is that the quality of the results, or more precisely the quantitative prediction capabilities of the simulations, may not be well established.

Relations to Other Fields

A term closely related to Scientific Computing (and that Wikipedia actually treats as a synonym) is Computational Science, which we here define as solving a scientific problem with the aid of techniques from Scientific Computing. While Scientific Computing deals with solution techniques and tools, Computational Science has a stronger focus on the *science*, that is, a scientific question and the significance of the answer. In between these focal points, the craft of Scientific Computing is fundamental in order to *produce* an answer. Scientific Computing and Computational Science are developing into an independent scientific discipline; they combine elements from Mathematics and Computer Science to form the foundations for a new methodology of scientific discovery. Developing Computational Science and Scientific Computing may turn out to be as fundamental to the future progress in science as was the development of novel Mathematics in the times of Newton and Euler.

Another closely related term is *Numerical Analysis*, which is about the "development of practical algorithms to obtain *approximate* solutions of mathematical problems and the validation of these solutions through their *mathematical analysis*." The development of algorithms is central to both Numerical Analysis and Scientific Computing, and so is the validation of the computed solutions, but Numerical Analysis has a particular emphasis on mathematical analysis of the accuracy of approximate algorithms. Some narrower definitions of Scientific Computing would say that it contains all of Numerical Analysis, but in addition applies more experimental computing techniques to evaluate the accuracy of the computed results. Scientific Computing is not necessarily restricted to approximate solution methods, although those are the most widely used. Other definitions of Scientific Computing may include additional points from the list above, up to our definition which is wide and includes all the key steps in creating predictive computer simulations.

The term Numerical Analysis seems to have appeared in 1947 when the Institute for Numerical Analysis was set up at UCLA with funding from the National Bureau of Standards in the Office Naval Research.

A landmark for the term Scientific Computing dates back to 1980 when Gene Golub established SIAM Journal on Scientific Computing. Computational Science, and Computational Science and Engineering, became widely used terms during the late the 1990s. The book series *Lecture Notes in Computational Science and Engineering* was initiated in 1995 and published from 1997 (but the Norwegian University of Science and Technology proposed a professor in Computational Science as early as 1993). The Computational Science term was further coined by the popular conferences SIAM Conference on Computational Science and Engineering (from 2000) and the International Conference on Computational Science (ICCS, from 2001). Many student programs with the same names appeared at the turn of the century.

In Numerical Analysis the guiding principle is to perform computations based on a strong theoretical foundation. This foundation includes a proof that the algorithm under consideration is computable and how accurate the computed solution would be. Suppose, for instance, that our aim is to solve a system of algebraic equations using an iterative method. If we have an algorithm providing a sequence of approximations given by $\{x_i\}$, the basic questions of Numerical Analysis are (i) (existence) to prove that x_{i+1} can be computed provided that x_i already is computed and (ii) (convergence) how accurate is the i-th approximation. In general, these two steps can be extremely challenging. Earlier, the goal of having a solid theoretical basis for algorithms was frequently realistic since the computers available did not have sufficient power to address very complex problems. That situation has changed dramatically in recent years, and we are now able to use computers to study problems that are way beyond the realm of Numerical Analysis.

Scientific Computing is the discipline that takes over where a complete theoretical analysis of the algorithms involved is impossible. Today, Scientific Computing is an indispensable tool in science, and the models, methods, and algorithms under considerations are rarely accessible by analytical tools. This lack of theoretical rigor is often addressed by using extensive, carefully conducted computer experiments to investigate the quality of computed solutions. More standardized methods for such investigations are an important integral part of Scientific Computing.

Scientific Computing: Mathematics or Computer Science?

Universities around the world are organized in departments covering a reasonable portion of science or engineering in a fairly disjoint manner. This organizational structure has caused much headache amongst researchers in Numerical Analysis and Scientific Computing because these fields typically would have to find its place either in a Computer Science department or in a Mathematics department, and Scientific Computing and Numerical Analysis belong in part to both these disciplines. Heated arguments have taken place around the world, and so far no universal solution has been provided. The discussion may, however, be used to illustrate the validity of Sayre's law (From first lines of wikipedia.org/wiki/Sayres_law: Sayre's law states, in a formulation quoted by Charles Philip Issawi: "In any dispute the intensity of feeling is inversely proportional to the value of the issues at stake." By way of corollary, it adds: "That is why academic politics are so bitter.") which is often attributed to Henry Kissinger.

Formulation of Mathematical Models

The demand for a mathematical model comes from the curiosity or need to answer a scientific question. When the model is finally available for computer simulations, the results very often lead to reformulation of the model or the scientific question. This iterative process is the essence of doing science with aid of mathematical models.

Although some may claim that *formulation* of mathematical models is an activity that belongs to Engineering and classical sciences and that the models are *prescribed* in Scientific Computing, we will argue that this is seldom the case. The classical scientific subjects (e.g., Physics, Chemistry, Biology, Statistics, Mathematics, Computer Science, Economics) do formulate mathematical models, but the traditional focus targets models suitable for analytical insight. Models suitable for being run on computers require mathematical formulations adapted to the technical steps of the solution process. Therefore, experts on Scientific Computing will often go back to the model and reformulate it to improve steps in the solution process. In particular, it is important to formulate models that fit approximation, software, hardware, and parameter estimation constraints. Such aspects of formulating models have to a large extent been developed over the last decades through Scientific Computing research and practice. Successful Scientific Computing therefore demands a close interaction between understanding of the phenomenon under interest (often referred to as domain knowledge) and the techniques available in the solution process. Occasionally, intimate knowledge about the application and the scientific question enables the use of special properties that can reduce computing time, increase accuracy, or just simplify the model considerably.

To illustrate how the steps of computing impacts modeling, consider flow around a body. In Physics (or traditional Fluid Dynamics to be more precise), one frequently restricts the development of a model to what can be treated by pen and paper Mathematics, which in the current example means assumption of stationary laminar flow, that the body is a sphere, and that the domain is infinite. When analyzing flow around a body through computer simulations, a time-dependent model may be easier to implement and faster to run on modern parallel hardware, even if only a stationary solution is of interest. In addition, a time-dependent model allows the development of instabilities and the well-known oscillating vortex patterns behind the body that occur even if the boundary conditions are stationary. A difficulty with the discrete model is the need for a finite domain and appropriate boundary conditions that do not disturb the flow inside the domain (so-called Artificial Boundary Conditions). In a discrete model, it is also easy to allow for flexible body geometry and relatively straightforward to include models of turbulence. What kind of turbulence model to apply might be constrained by implementational difficulties, details of the computer architecture, or computing time feasibility. Another aspect that impacts the modeling is the estimation of the parameters in the model (usually done by solving ▶ Inverse Problems: Numerical Methods). Large errors in such estimates may favor a simpler and less accurate model over a more complex one where uncertainty in unknown parameters is greater. The conclusion on which model to choose depends on many factors and ultimately on how one defines and measures the accuracy of predictions.

Some common ingredients in mathematical models are Integration; Approximation of Functions (Curve and Surface Fitting); optimization of Functions or Functionals; Matrix Systems; Eigenvalue Problems; Systems of Nonlinear Equations; graphs and networks; ▶ Numerical Analysis of Ordinary Differential Equations; ▶ Computational Partial Differential Equations; Integral Equations; Dynamical System Theory; stochastic variables, processes, and fields; random walks; and Imaging Techniques. The entries on ▶ Numerical Analysis, ▶ Computational Partial Differential Equations, and ▶ Numerical Analysis of Ordinary Differential Equations provide more detailed overview of these topics and associated algorithms.

Discrete Models and Algorithms

Mathematical models may be continuous or discrete. Continuous models can be addressed by symbolic computing, otherwise (and usually) they must be made discrete through discretization techniques. Many physical phenomena leave a choice between formulating the model as continuous or discrete. For example, a geological material can be viewed as a finite set of small elements in contact (discrete element model) or as a continuous medium with prescribed macroscopic material properties (continuum mechanical model). In the former case, one applies a set of rules for how elements interact at a mesoscale level and ends up with a large system of algebraic equations that must be solved, or sometimes one can derive explicit formulas for how each element moves during a small time interval. Another discrete modeling approach is based on cellular automata, where physical relations are described between a fixed grid of cells. The Lattice Boltzmann method, for example, uses 2D or 3D cellular automata to model the dynamics of a fluid on a meso-scopic level. Here the interactions between the states of neighboring cells are derived from the principles of statistical mechanics.

With a continuous medium, the model is expressed in terms of partial differential equations (with appropriate initial and boundary conditions). These equations must be discretized by techniques like ▶ Finite Difference Methods, ▶ Finite Element Methods, or ▶ Finite Volume Methods, which lead to systems of algebraic equations. For a purely elastic medium, one can formulate mathematically equivalent discrete and discretized continuous models, while for complicated material behavior the two model classes have their pros and cons. Some will prefer a discrete element model because it often has fewer parameters to estimate than a continuum mechanical constitutive law for the material.

Some models are spatially discrete but continuous in time. Examples include ▸ Molecular Dynamics and planetary systems, while others are purely discrete, like the network of Facebook users.

The fundamental property of a discrete model, either originally discrete or a discretized continuous model, is its suitability for a computer. It is necessary to adjust the computational work, which is usually closely related to the accuracy of the discrete model, to fit the given hardware and the acceptable time for calculating the solution. The choice of discretization is also often dictated by software considerations. For example, one may prefer to discretize a partial differential equation by a finite difference method rather than a finite element method because the former is much simpler to implement and thus may lead to more efficient software and thus eventually more accurate simulation results.

The entries on ▸ Numerical Analysis, ▸ Computational Partial Differential Equations, ▸ Numerical Analysis of Ordinary Differential Equations, and Imaging present overviews of different discretization techniques, and more specialized articles go deeper into the various methods.

The accuracy of the discretization is normally the most important factor that governs the choice of technique. Discretized continuous models are based on approximations, and quantifying the accuracy of these approximations is a key ingredient in Scientific Computing, as well as techniques for assessing their computational cost. The field of Numerical Analysis has developed many mathematical techniques that can help establish a priori or a posteriori bounds on the errors in numerous types of approximations. The former can bound the error by properties of the exact solution, while the latter applies the approximate (i.e., the computed) solution in the bound. Since the exact solution of the mathematical problem remains unknown, predictive models must usually apply a posteriori estimates in an iterative fashion to control approximation errors.

When mathematical expressions for or bounds of the errors are not obtainable, one has to resort to experimental investigations of approximation errors. Popular techniques for this purpose have arisen from verification methods (see below).

With ▸ Symbolic Computing one can bring additional power to exact and approximate solution techniques based on traditional pen and paper Mathematics. One example is perturbation methods where the solution is expressed as a power series of some dimensionless parameter, and one develops a hierarchy of models for determining the coefficients in the power series. The procedure originally involves lengthy analytical computations by hand which can be automated using symbolic computing software such as Mathematica, Maple, or SymPy.

When the mathematical details of the chosen discretization are worked out, it remains to organize those details in algorithms. The algorithms are computational recipes to bridge the gap between the mathematical description of discrete models and their associated implementation in software. Proper documentation of the algorithms is extremely important such that others know all ingredients of the computer model on which scientific findings are based. Unfortunately, the details of complex models that are routinely used for important industrial or scientific applications may sometimes be available only through the actual computer code, which might even be proprietary.

Many of the mathematical subproblems that arise from a model can be broken into smaller problems for which there exists efficient algorithms and implementations. This technique has historically been tremendously effective in Mathematics and Physics. Also in Scientific Computing one often sees that the best way of solving a new problem is to create a clever glue between existing building blocks.

Implementation

Ideally, a well-formulated set of algorithms should easily translate into computer programs. While this is true for simple problems, it is not in the general case. A large portion of many budgets for Science Computing projects goes to software development. With increasingly complicated models, the complexity of the computer programs appears to grow even faster, because Computer Languages were not designed to easily express complicated mathematical concepts. This is one main reason why writing and maintaining scientific software is challenging.

The fundamental challenge to develop correct and efficient software for Scientific Computing is notoriously underestimated. It has an inherent complexity that cannot be addressed by automated procedures alone, but must be acknowledged as an independent scientific and engineering problem. Also, testing of the software quickly consumes even more resources. Software Engineering is a central field of Computer Science which addresses techniques for developing and testing software systems in general, but has so far had minor impact on scientific software. We can identify three reasons. First, the structure of the software is closely related to the mathematical concepts involved. Second, testing is very demanding since we for most relevant applications do not know the answer beforehand. Actually, much scientific software is written to explore new phenomena where neither qualitative nor quantitative properties of the computed results are known. Even when analytical insight is known about the solution, the computed results will contain unknown discretization errors. The third reason is that scientific software quickly consumes the largest computational resources available and hence employs special High-Performance Computing (HPC) platforms. These platforms imply that computations must run in parallel on heterogeneous architectures, a fact that seriously complicates the software development. There has traditionally been little willingness to adopt good Software Engineering techniques if they cause any loss of computational performance (which is normally the case).

In the first decades of Scientific Computing, FORTRAN was the dominating Computer Language for implementing algorithms and FORTRAN has still a strong position. Classical FORTRAN (with the dialects IV, 66, and 77) is ideal for mathematical formulas and heavy array computations, but lacks more sophisticated features like classes, namespaces, and modules to elegantly express complicated mathematical concepts. Therefore, the much richer C++ language attracted significant attention in scientific software projects from the mid-1990s. Today, C++ is a dominating language in new projects, although the recent FORTRAN 2003/2008 has many of the features that made C++ popular. C, C++, and FORTRAN enable the programmer to utilize almost the maximum efficiency of an HPC architecture. On the other hand, much less computationally efficient languages such as MATLAB, Mathematica, and Python have reached considerable popularity for implementing Scientific Computing algorithms. The reason is that these languages are more high level; that is, they allow humans to write computer code closer to the mathematical concepts than what is easily achievable with C, C++, and FORTRAN.

Over the last four decades, numerous high-quality libraries have been developed, especially for frequently occurring problems from numerical linear algebra, differential equations, approximation of functions, optimization, etc. Development of new software today will usually maximize the utilization of such well-tested libraries. The result is a heterogeneous software environment that involves several languages and software packages, often glued together in easy-to-use and easy-to-program applications in MATLAB or Python.

If we want to address complicated mathematical models in Scientific Computing, the software needs to provide the right abstractions to ease the implementation of the mathematical concepts. That is, the step from the mathematical description to the computer code must be minimized under the constraint of minor performance loss. This constrained optimization problem is the great challenge in developing scientific software.

Most classical mathematical methods are serial, but utilization of modern computing platforms requires algorithms to run in parallel. The development of algorithms for ▶ Parallel Computing is one of the most significant activities in Scientific Computing today. Implementation of parallel algorithms, especially in combination with high-level abstractions for complicated mathematical concepts, is an additional research challenge. Easy-to-use parallel implementations are needed if a broad audience of scientists shall effectively utilize Modern HPC Architectures such as clusters with multi-core and multi-GPU PCs. Fortunately, many of the well-known libraries for, e.g., linear algebra and optimization are updated to perform well on modern hardware.

Often scientific progress is limited by the available hardware capacity in terms of memory and computational power. Large-scale projects can require expensive resources, where not only the supercomputers per se but also their operational cost become limiting factors. Here it becomes mandatory that the accuracy provided by a Scientific Computing methodology is evaluated relative to its cost. Traditional approaches that just quantify the number of numerical operations turn often out to be misleading. Worse than that, theoretical analysis often provides only asymptotic bounds for the error with unspecified constants. This translates to cost assessments with unspecified constants that are of only little use for quantifying the real cost to obtain a simulation result. In Scientific Computing, such mathematical techniques must be combined with more realistic cost predictions to guide the development of effective simulation methods. This important research direction comes under names such as ▶ Hardware-Oriented Numerics for PDE or *systematic performance engineering* and is usually based on a combination of rigorous mathematical techniques with engineering-like heuristics. Additionally, technological constraints, such as the energy consumption of supercomputers are increasingly found to become critical bottlenecks. This in turn motivates genuinely new research directions, such as evaluating the numerical efficiency of an algorithm in terms of its physical resource requirements like the energy usage.

Verification

▶ Verification of scientific software means setting up a series of tests to bring evidence that the software solves the underlying mathematical problems correctly. A fundamental challenge is that the problems are normally solved approximately with an error that is quantitatively unknown. Comparison with exact mathematical results will in those cases yield a discrepancy, but the purpose of verification is to ensure that there are no additional nonmathematical discrepancies caused by programming mistakes.

The ideal tests for verification is to have exact solutions of the discrete problems. The discrepancy of such solutions and those produced by the software should be limited by (small) roundoff errors due to Finite Precision Arithmetic in the machine. The standard verification test, however, is to use exact solutions of the mathematical problems to compute observed errors and check that these behave correctly. More precisely, one develops exact solutions for the mathematical problem to be solved, or a closely related one, and establishes a theoretical model for the errors. Error bounds from Numerical Analysis will very often suggest error models. For each problem one can then vary discretization parameters to generate a data set of errors and see if the relation between the errors and the discretization parameters is as expected from the error model. This strategy constitutes the perhaps most important verification technique and demonstrates how dependent software testing is on results from Numerical Analysis.

Analytical insight from alternative or approximate mathematical models can in many physical problems be used to test that certain aspects of the solution behave correctly. For example, one may have asymptotic results for the solution far away from the locations where the main physics is generated. Moreover, principles such as mass, momentum, and energy balance for the whole system under consideration can also be checked. These types of tests are not as effective for uncovering software bugs as the tests described above, but add evidence that the software works.

Scientific software needs to compute correct numbers, but must also run fast. Tests for checking that the computational speed has not changed unexpectedly are therefore an integral part of any test suite. Especially on parallel computing platforms, this type of efficiency tests is as important as the correctness tests.

A fundamental requirement of verification procedures is that all tests are automated and can at any time be repeated. Version control systems for keeping track of different versions of the files in a software package can be integrated with automatic testing such that every registered change in the software triggers a run of the test suite, with effective reports to help track down new errors. It is also easy to roll back to previous versions of the software that passed all tests. Science papers that rely heavily on Scientific Computing should ideally point to web sites where the version history of the software and the tests are available, preferably also with snapshots of the whole computing environment where the simulation results were obtained. These elements are important for ▶ Reproducibility: Methods of the scientific findings.

Preparation of Input Data

With increasingly complicated mathematical models, the preparation of input data for such models has become a very resource consuming activity. One example is the computation of the drag (fuel consumption) of a car, which demands a mathematical description of the car's surface and a division of the air space outside the car into small hexahedra or tetrahedra. Techniques of ▶ Geometry Processing can be used to measure and mathematically represent the car's surface, while Meshing Algorithms are used to populate the air flow domain with small hexahedra or tetrahedra. An even greater challenge is met in biomedical or geophysical computing where Segmentation Methods must normally be accompanied by human interpretation when extracting geometries from noisy images.

A serious difficulty of preparing input data is related to lack of knowledge of certain data. This situation requires special methods for parameter estimation as described below.

Visualization of Output Data

Computer simulations tend to generate large amounts of numbers, which are meaningless to the scientists unless the numbers are transformed to informative pictures closely related to the scientific investigation at hand. This is the goal of ▸ Visualization. The simplest and often most effective type of visualization is to draw curves relating key quantities in the investigation. Frequently, curves give incomplete insight into processes, and one needs to visualize more complex objects such as big networks or time-dependent, three-dimensional scalar or vector fields. Even if the goal of simulating a car's fuel consumption is a single number for the drag force, any physical insight into enhancing geometric features of the car requires detailed visualization of the air velocities, the pressure, and vortex structures in time and 3D space. Visualization is partly about advanced numerical algorithms and partly about visual communication. The importance of effective visualization in Scientific Computing can hardly be overestimated, as it is a key tool in software debugging, scientific investigations, and communication of the main research findings.

Validation and Parameter Estimation

While verification is about checking that the algorithms and their implementations are done right, validation is about checking that the mathematical model is relevant for predictions. When we create a mathematical model for a particular process in nature, a technological device, or society, we think of a *forward model* in the meaning that the model requires a set of input data and can produce a set of output data. The input data must be known for the output data to be computed. Very often this is not the case, because some input data remains unknown, while some output data is known or can be measured. And since we lack some input data, we cannot run the forward model. This situation leads to a parameter estimation problem, also known as a model calibration problem, a parameter identification problem, or an ▸ Inverse Problems: Numerical Methods. The idea is to use some of the known output data to estimate some of the lacking input data with aid of the model. Forward models are for the most part well posed in the sense that small errors in input data are not amplified significantly in the output. Inverse problems, on the other hand, are normally ill posed: small errors in the measured output may have severe impact on the precision of our estimates of input data. Much of the methodology research is about reducing the ill posedness.

▸ Validation consists in establishing evidence that the computer model really models the real phenomena we are interested in. The idea is to have a range of test cases, each with some known output, usually measured in physical experiments, and checking that the model reproduces the known output. The tests are straightforwardly conducted if all input data is known. However, very often some input parameters in the model are unknown, and the typical validation procedure for a given test case is to tune those parameters in a way that reproduces the known output. Provided that the tuned parameters are within realistic regions, the model passes the validation test in the sense that there exists relevant input data such that the model predicts the observed output.

Understanding of the process being simulated can effectively guide manual tuning of unknown input parameters. Alternatively, many different numerical methods exist for automatically fitting input parameters. Most of them are based on constrained optimization, where one wants to minimize the squared distance between predicted and observed output with respect to the unknown input parameters, given the constraint that any predicted value must obey the model. Numerous specific solution algorithms are found in the literature on deterministic ▸ Inverse Problems: Numerical Methods. Usually, the solution of an inverse problems requires a large number of solutions of the corresponding forward problem.

Many scientific questions immediately lead to inverse problems. Seismic imaging is an example where one aims to estimate the spatial properties of the Earth's crust using measurements of reflected sound waves. Mathematically, the unknown properties are spatially varying coefficients in partial differential equations, and the measurements contain information about the solution of the equations. The primary purpose is to estimate the value of the coefficients, which in a forward model for wave motion constitute input data that must be known. When the focus is about solving the inverse problem itself, one can often apply simpler forward models (in seismic imaging, e.g., ordinary differential equations for ray tracing have traditionally replaced full partial differential equations for the wave motion as forward model).

Reliable estimation of parameters requires more observed data than unknown input data. Then we can search for the best fit of parameters, but there is no unique definition of what "best" is. Furthermore, measured output data is subject to measurement errors. It is therefore fruitful to acknowledge that the solution of inverse problems has a variability. Control of this variability gives parameter estimates with corresponding statistical uncertainties. For this purpose, one may turn to solving stochastic inverse problems. These are commonly formulated in a ▶ Bayesian Statistics: Computation where probability densities for the input parameters are suggested, based on prior knowledge, and the framework updates these probability densities by taking the forward model and the data into account. The inserted prior knowledge handles the ill posedness of deterministic inverse problems, but at a much increased computational cost.

Uncertainty Quantification

With stochastic parameter estimation we immediately face the question: How does the uncertainty in estimated parameters propagate through the model? That is, what is the uncertainty in the predicted output? Methods from ▶ Uncertainty Quantification: Computation can be used to answer this problem. If parameters are estimated by Bayesian frameworks, we have complete probability descriptions. With simpler estimation methods we may still want to describe uncertainty in the parameters in terms of assumed probability densities.

The simplest and most general method for uncertainty quantification is Monte Carlo simulation. Large samples of input data are drawn at random from the known probability densities and fed as input to the model. The forward model is run to compute the output corresponding to each sample. From the samples of output data, one can compute the average, variance, and other statistical measures of the quantities of interest. One must often use of the order 10^5–10^7 samples (and hence runs of the forward model) to compute reasonably precise statistics. Much faster but less general methods exist. During recent years, ▶ Polynomial Chaos Expansions have become popular. These assume that the mapping from stochastic input to selected output quantities is smooth such that the mapping can be effectively described by a polynomial expansion with few terms. The expansion may converge exponentially fast and reduce the number of runs of the forward model by several orders of magnitude.

Having validated the model and estimated the uncertainty in the output, we can eventually perform predictive computer simulations and calculate the precision of the predictions. At this point we have completed the final item in our list of key steps in the simulation pipeline.

We remark that although the stochastic parameter estimation framework naturally turns an originally deterministic model into a stochastic one, modelers may early in the modeling process assume that the details of some effect are not precisely known and therefore describe the effect as a stochastic quantity. Some input to the model is then stochastic and the question is how the statistical variability is propagated through the model. This basically gives the same computational problem as in uncertainty quantification. One example where stochastic quantities are used directly in the modeling is environmental forces from wind and waves on structures. The forces may be described as stochastic space-time fields with statistical parameters that must be estimated from measurements.

Laboratory and Field Experiments

The workflow in Scientific Computing to establish predictive simulation models is seen to involve knowledge from several subjects, clearly Applied and Numerical Mathematics, Statistics, and Computer Science, but the mathematical techniques and software work must be closely integrated with domain-specific modeling knowledge from the field where the science problem originates, say Physics, Mechanics, Geology, Geophysics, Astronomy, Biology, Finance, or Engineering disciplines. A less emphasized integration is with laboratory and field experiments, as Scientific Computing is often applauded to eliminate the need for experiments. However, we have explained that predictive simulations require validation, parameter estimation, and control of the variability of input and output data. The computations involved in these tasks cannot be carried out without access to carefully conducted experiments in the laboratory or the field. A hope is that an extensive amount of large-scale data sets from experiments can be made openly available to all computational scientists and thereby accelerate the integration of experimental data in Scientific Computing.

Self-Consistent Field (SCF) Algorithms

Eric Cancès
Ecole des Ponts ParisTech – INRIA, Université Paris Est, CERMICS, Projet MICMAC, Marne-la-Vallée, Paris, France

Definition

Self-consistent field (SCF) algorithms usually refer to numerical methods for solving the Hartree-Fock, or the Kohn-Sham equations. By extension, they also refer to constrained optimization algorithms aiming at minimizing the Hartree-Fock, or the Kohn-Sham energy functional, on the set of admissible states.

Discretization of the Hartree-Fock Model

As usual in electronic structure calculation, we adopt the system of atomic units, obtained by setting to 1 the values of the reduced Planck constant $\hbar$, of the elementary charge, of the mass of the electron, and of the constant $4\pi\epsilon_0$, where ϵ_0 is the dielectric permittivity of the vacuum.

The Hartree-Fock model reads, for a molecular system containing N electrons, as

$$E_0^{\mathrm{HF}}(N) = \inf\left\{ \mathcal{E}^{\mathrm{HF}}(\Phi),\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3_\Sigma))^N,\ \int_{\mathbb{R}^3_\Sigma} \phi_i \phi_j^* = \delta_{ij} \right\}, \quad (1)$$

where $\mathbb{R}^3_\Sigma = \mathbb{R}^3 \times \{\uparrow, \downarrow\}$, $\int_{\mathbb{R}^3_\Sigma} \phi_i \phi_j^* = \int_{\mathbb{R}^3_\Sigma} \phi_i(\mathbf{x}) \phi_j^*(\mathbf{x})\, d\mathbf{x} := \sum_{\sigma \in \{\uparrow,\downarrow\}} \int_{\mathbb{R}^3} \phi_i(\mathbf{r},\sigma)\, \phi_j(\mathbf{r},\sigma)^*\, d\mathbf{r}$,
and

$$\mathcal{E}^{\mathrm{HF}}(\Phi) = \frac{1}{2}\sum_{i=1}^N \int_{\mathbb{R}^3_\Sigma} |\nabla \phi_i|^2 + \int_{\mathbb{R}^3_\Sigma} \rho_\Phi V_{\mathrm{nuc}} + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\Phi(\mathbf{r})\, \rho_\Phi(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}' - \frac{1}{2}\int_{\mathbb{R}^3_\Sigma}\int_{\mathbb{R}^3_\Sigma} \frac{|\gamma_\Phi(\mathbf{x},\mathbf{x}')|^2}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{x}\, d\mathbf{x}'.$$

The function V_{nuc} denotes the nuclear potential. If the molecular system contains M nuclei of charges $z_1, \cdots, z_M$ located at positions $\mathbf{R}_1, \cdots, \mathbf{R}_M$, the following holds

$$V_{\mathrm{nuc}}(\mathbf{r}) = -\sum_{k=1}^M \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}.$$

The density ρ_Φ and the density matrix γ_Φ associated with Φ are defined by

$$\rho_\Phi(\mathbf{r}) = \sum_{i=1}^N \sum_{\sigma\in\{\uparrow,\downarrow\}} |\phi_i(\mathbf{r},\sigma)|^2 \quad \text{and}$$

$$\gamma_\Phi(\mathbf{x},\mathbf{x}') = \sum_{i=1}^N \phi_i(\mathbf{x})\phi_i(\mathbf{x}')^*.$$

The derivation of the Hartree-Fock model from the N-body Schrödinger equation is detailed in the contribution by I. Catto to the present volume.

The Galerkin approximation of the minimization problem (1) consists in approaching $E_0^{\mathrm{HF}}(N)$ by

$$E_0^{\mathrm{HF}}(N,\mathcal{V}) = \min\left\{\mathcal{E}^{\mathrm{HF}}(\Phi),\ \Phi=(\phi_1,\cdots,\phi_N) \in \mathcal{V}^N,\ \int_{\mathbb{R}^3_\Sigma}\phi_i\phi_j^* = \delta_{ij}\right\}, \tag{2}$$

where $\mathcal{V}\subset H^1(\mathbb{R}^3_\Sigma)$ is a finite dimensional approximation space of dimension N_b. Obviously, $E_0^{\mathrm{HF}}(N)\le E_0^{\mathrm{HF}}(N,\mathcal{V})$ for any $\mathcal{V}\subset H^1(\mathbb{R}^3_\Sigma)$.

In the sequel, we denote by $\mathbb{C}^{m,n}$ the vector space of the complex-valued matrices with m lines and n columns, and by $\mathbb{C}_h^{m,m}$ the vector space of the hermitian matrices of size $m\times m$. We endow these spaces with the Frobenius scalar product defined by $(A,B)_{\mathrm{F}} := \mathrm{Tr}\,(A^*B)$. Choosing a basis $(\chi_1,\cdots,\chi_{N_b})$ of $\mathcal{V}$, and expanding $\Phi=(\phi_1,\cdots,\phi_N)\in\mathcal{V}^N$ as

$$\phi_i(\mathbf{x}) = \sum_{i=1}^{N_b} C_{\mu i}\,\chi_\mu(\mathbf{x}),$$

problem (2) also reads

$$E_0^{\mathrm{HF}}(N,\mathcal{V}) = \min\left\{E^{\mathrm{HF}}(CC^*),\ C\in\mathbb{C}^{N_b\times\mathbb{N}},\ C^*SC = I_N\right\}, \tag{3}$$

where I_N is the identity matrix of rank N and where

$$E^{\mathrm{HF}}(D) = \mathrm{Tr}\,(hD) + \frac{1}{2}\mathrm{Tr}\,(G(D)D).$$

The entries of the overlap matrix S and of the one-electron Hamiltonian matrix h are given by

$$S_{\mu\nu} := \int_{\mathbb{R}^3_\Sigma}\chi_\mu^*\chi_\nu \tag{4}$$

and

$$h_{\mu\nu} := \frac{1}{2}\int_{\mathbb{R}^3_\Sigma}\nabla\chi_\mu^*\cdot\nabla\chi_\nu - \sum_{k=1}^{M} z_k\int_{\mathbb{R}^3_\Sigma}\frac{\chi_\mu(\mathbf{x})^*\chi_\nu(\mathbf{x})}{|\mathbf{r}-\mathbf{R}_k|}\,d\mathbf{x}. \tag{5}$$

The linear map $G\in\mathcal{L}(\mathbb{C}_h^{N_b\times N_b})$ is defined by

$$[G(D)]_{\mu\nu} := \sum_{\kappa,\lambda=1}^{N_b}\left[(\mu\nu|\kappa\lambda)-(\mu\lambda|\kappa\nu)\right]\,D_{\kappa\lambda},$$

where $(\mu\lambda|\kappa\nu)$ is the standard notation for the two-electron integrals

$$(\mu\nu|\kappa\lambda) := \int_{\mathbb{R}^3_\Sigma}\int_{\mathbb{R}^3_\Sigma}\frac{\chi_\mu(\mathbf{x})\chi_\nu(\mathbf{x})^*\chi_\kappa(\mathbf{x}')\chi_\lambda(\mathbf{x}')^*}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{x}\,d\mathbf{x}'. \tag{6}$$

We will make use of the symmetry property

$$\mathrm{Tr}\,(G(D)D') = \mathrm{Tr}\,(G(D')D). \tag{7}$$

The formulation (3) is referred to as the molecular orbital formulation of the Hartree-Fock model, by contrast with the density matrix formulation defined as

$$E_0^{\mathrm{HF}}(N,\mathcal{V}) = \min\left\{E^{\mathrm{HF}}(D),\ D\in\mathbb{C}_h^{N_b\times\mathbb{N}_b},\ DSD = D,\ \mathrm{Tr}\,(SD)=N\right\}. \tag{8}$$

It is easily checked that problems (3) and (8) are equivalent, remarking that the map $\{C\in\mathbb{C}^{N_b\times\mathbb{N}}\,|\,C^*SC = I_N\}\ni C\mapsto CC^*\in\{D\in\mathbb{C}_h^{N_b\times\mathbb{N}_b}\,|\,DSD = D, \mathrm{Tr}\,(SD) = N\}$ is onto. For any $\Phi = (\phi_1,\cdots,\phi_N)\in\mathcal{V}^N$ with $\phi_i(\mathbf{x}) = \sum_{\mu=1}^{N_b}C_{\mu i}\chi_\mu(\mathbf{x})$, the following holds

$$\gamma_\Phi(\mathbf{x},\mathbf{x}') = \sum_{\mu,\nu=1}^{N_b} D_{\mu\nu}\chi_\mu(\mathbf{x})\chi_\nu(\mathbf{x}')^* \quad\text{with}\quad D = CC^*.$$

We refer to the contribution by Y. Maday for the derivation of a priori and a posteriori estimates on the energy difference $E_0^{\mathrm{HF}}(N,\mathcal{V}) - E_0^{\mathrm{HF}}(N)$, and on the distance between the minimizers of (2) and those of (1), for L^2 and Sobolev norms.

Most Hartree-Fock calculations are performed in atomic orbital basis sets (see, e.g., [9] for details), and more specifically with Gaussian type orbitals. The latter are of the form

$$\chi_\mu(\mathbf{r},\sigma) = \sum_{g=1}^{N_g} P_{\mu,g}(\mathbf{r}-\mathbf{R}_{k(\mu)})e^{-\alpha_{\mu,g}|\mathbf{r}-\mathbf{R}_{k(\mu)}|^2}\,S_\mu(\sigma), \tag{9}$$

where $P_{\mu,g}$ is a polynomial, $\alpha_{\mu,g} > 0$, and $S_\mu = \alpha$ or β, with $\alpha(\sigma) = \delta_{\sigma,\uparrow}$ and $\beta(\sigma) = \delta_{\sigma,\downarrow}$. The main advantage of Gaussian type orbitals is that all the integrals in (4)–(6) can be calculated analytically [5].

In order to simplify the notation, we assume in the sequel that the basis $(\chi_1, \cdots, \chi_{N_b})$ is orthonormal, or, equivalently, that $S = I_{N_b}$. The molecular orbital and density matrix formulations of the discretized Hartree-Fock model then read

$$E_0^{\mathrm{HF}}(N, \mathcal{V}) = \min \left\{ E^{\mathrm{HF}}(CC^*),\ C \in \mathcal{C} \right\}, \tag{10}$$

$$\mathcal{C} = \left\{ C \in \mathbb{C}^{N_b \times N} \mid C^*C = I_N \right\},$$

$$E_0^{\mathrm{HF}}(N, \mathcal{V}) = \min \left\{ E^{\mathrm{HF}}(D),\ D \in \mathcal{P} \right\}, \tag{11}$$

$$\mathcal{P} = \left\{ D \in \mathbb{C}_h^{N_b \times N_b} \mid D^2 = D, \right.$$
$$\left. \mathrm{Tr}\,(D) = N \right\}.$$

Hartree-Fock-Roothaan-Hall Equations

The matrix

$$F(D) := h + G(D) \tag{12}$$

is called the Fock matrix associated with D. It is the gradient of the Hartree-Fock energy functional at D, for the Frobenius scalar product.

It can be proved (see again [9] for details) that if D is a local minimizer of (11), then

$$\begin{cases} D = \displaystyle\sum_{i=1}^{N} \Phi_i \Phi_i^* \\ F(D)\Phi_i = \epsilon_i\, \Phi_i \\ \Phi_i^* \Phi_j = \delta_{ij} \\ \epsilon_1 \le \epsilon_2 \le \cdots \le \epsilon_N \text{ are the lowest} \\ \qquad\qquad N \text{ eigenvalues } F(D). \end{cases} \tag{13}$$

The above system is a nonlinear eigenvalue problem: the vectors Φ_i are orthonormal eigenvectors of the hermitian matrix $F(D)$ associated with the lowest N eigenvalues of $F(D)$, and the matrix $F(D)$ depends on the eigenvectors Φ_i through the definition of D. The first three equations of (13) are called the Hartree-Fock-Roothaan-Hall equations. The property that $\epsilon_1, \cdots, \epsilon_N$ are the lowest N eigenvalues of the hermitian matrix $F(D)$ is referred to as the *Aufbau* principle. An interesting property of the Hartree-Fock model is that, for any local minimizer of (11), there is a positive gap between the Nth and $(N+1)$th eigenvalues of $F(D)$: $\gamma = \epsilon_{N+1} - \epsilon_N > 0$ [2, 9]. From a geometrical viewpoint, $D = \sum_{i=1}^N \Phi_i \Phi_i^*$ is the matrix of the orthogonal projector on the vector space spanned by the lowest N eigenvalues of $F(D)$. Note that (13) can be reformulated without any reference to the molecular orbitals Φ_i as follows:

$$D \in \operatorname{argmin} \left\{ \mathrm{Tr}\,(F(D)D'),\ D' \in \mathcal{P} \right\}, \tag{14}$$

and that, as $\gamma = \epsilon_{N+1} - \epsilon_N > 0$, the right-hand side of (14) is a singleton. This formulation is a consequence of the following property: for any hermitian matrix $F \in \mathbb{C}_h^{N_b \times N_b}$ with eigenvalues $\epsilon_1 \le \cdots \le \epsilon_{N_b}$ and any orthogonal projector D of the form $D = \sum_{i=1}^N \Phi_i \Phi_i^*$ with $\Phi_i^* \Phi_j = \delta_{ij}$ and $F\Phi_i = \epsilon_i \Phi_i$, the following holds $\forall D' \in \mathcal{P}$,

$$\mathrm{Tr}\,(FD') \ge \mathrm{Tr}\,(FD) + \frac{\epsilon_{N+1} - \epsilon_N}{2} \|D - D'\|_F^2. \tag{15}$$

Roothaan Fixed Point Algorithm

It is very natural to try and solve (13) by means of the following fixed point algorithm, originally introduced by Roothaan in [24]:

$$\begin{cases} F(D_k^{\mathrm{Rth}})\Phi_{i,k+1} = \epsilon_{i,k+1}\, \Phi_{i,k+1} \\ \Phi_{i,k+1}^* \Phi_{j,k+1} = \delta_{ij} \\ \epsilon_{1,k+1} \le \epsilon_{2,k+1} \le \cdots \le \epsilon_{N,k+1} \text{ are the lowest} \\ N \text{ eigenvalues } F(D_k^{\mathrm{Rth}}) \\ D_{k+1}^{\mathrm{Rth}} = \displaystyle\sum_{i=1}^{N} \Phi_{i,k+1} \Phi_{i,k+1}^*, \end{cases} \tag{16}$$

which also reads, in view of (15),

$$D_{k+1}^{\mathrm{Rth}} \in \operatorname{argmin} \left\{ \mathrm{Tr}\,(F(D_k^{\mathrm{Rth}})D'),\ D' \in \mathcal{P} \right\}. \tag{17}$$

Solving the *nonlinear* eigenvalue problem (13) then boils down to solving a sequence of *linear* eigenvalue problems.

It was, however, early realized that the above algorithm often fails to converge. More precisely, it can be proved that, under the assumption that

$$\inf_{k \in \mathbb{N}} (\epsilon_{N+1,k} - \epsilon_{N,k}) > 0, \tag{18}$$

which seems to be always satisfied in practice, the sequence $(D_k^{\rm Rth})_{k\in\mathbb{N}}$ generated by the Roothaan algorithm either converges to a solution D of the Hartree-Fock-Roothaan-Hall equations satisfying the *Aufbau* principle

$$\|D_k^{\rm Rth} - D\| \underset{k\to\infty}{\longrightarrow} 0 \qquad \text{with } D \text{ satisfying (13)}, \quad (19)$$

or asymptotically oscillates between two states $D_{\rm even}$ and $D_{\rm odd}$, none of them being solutions to the Hartree-Fock-Roothaan-Hall equations

$$\|D_{2k}^{\rm Rth} - D_{\rm even}\| \underset{k\to\infty}{\longrightarrow} 0 \quad \text{and} \quad \|D_{2k+1}^{\rm Rth} - D_{\rm odd}\| \underset{k\to\infty}{\longrightarrow} 0. \quad (20)$$

The behavior of the Roothaan algorithm can be explained mathematically, noticing that the sequence $(D_k^{\rm Rth})_{k\in\mathbb{N}}$ is obtained by minimizing by relaxation the functional

$$E(D, D') = {\rm Tr}\,(hD) + {\rm Tr}\,(hD') + {\rm Tr}\,(G(D)D').$$

Indeed, we deduce from (7), (12), and (17) that

$$\begin{aligned} D_1^{\rm Rth} &= \operatorname{argmin}\left\{{\rm Tr}\,(F(D_0^{\rm Rth})D'),\ D' \in \mathcal{P}\right\} \\ &= \operatorname{argmin}\left\{{\rm Tr}\,(hD_0^{\rm Rth}) + {\rm Tr}\,(hD') \right.\\ &\qquad \left. + {\rm Tr}\,(G(D_0^{\rm Rth})D'),\ D' \in \mathcal{P}\right\} \\ &= \operatorname{argmin}\left\{E(D_0^{\rm Rth}, D'),\ D' \in \mathcal{P}\right\}, \\ D_2^{\rm Rth} &= \operatorname{argmin}\left\{{\rm Tr}\,(F(D_1^{\rm Rth})D),\ D \in \mathcal{P}\right\} \\ &= \operatorname{argmin}\left\{{\rm Tr}\,(hD) + {\rm Tr}\,(hD_1^{\rm Rth}) \right.\\ &\qquad \left. + {\rm Tr}\,(G(D_1^{\rm Rth})D),\ D' \in \mathcal{P}\right\} \\ &= \operatorname{argmin}\left\{{\rm Tr}\,(hD) + {\rm Tr}\,(hD_1^{\rm Rth}) \right.\\ &\qquad \left. + {\rm Tr}\,(G(D)D_1^{\rm Rth}),\ D' \in \mathcal{P}\right\} \\ &= \operatorname{argmin}\left\{E(D, D_1^{\rm Rth}),\ D \in \mathcal{P}\right\}, \end{aligned}$$

and so on, and so forth. Together with (15) and (18), this leads to numerical convergence [6]: $\|D_{k+2}^{\rm Rth} - D_k^{\rm Rth}\|_{\rm F} \to 0$. The convergence/oscillation properties (19)/(20) can then be obtained by resorting to the Łojasiewicz inequality [17].

Oscillations of the Roothaan algorithm are called charge sloshing in the physics literature. Replacing $E(D, D')$ with the penalized functional $E(D, D') + \frac{b}{2}\|D - D'\|_{\rm F}^2$ $(b > 0)$ suppresses the oscillations when b is large enough, but the resulting algorithm

$$D_{k+1}^b \in \operatorname{argmin}\left\{{\rm Tr}\,(F(D_k^b - bD_k^b)D'),\ D' \in \mathcal{P}\right\}$$

often converges toward a critical point of the Hartree-Fock-Roothaan-Hall equations, which does not satisfy the *Aufbau* principle, and is, therefore, not a local minimizer. This algorithm, introduced in [25], is called the level-shifting algorithm. It has been analyzed in [6, 17].

Direct Minimization Methods

The molecular orbital formulation (10) and the density matrix formulation (11) of the discretized Hartree-Fock model are constrained optimization problems. In both cases, the minimization set is a (non-convex) smooth compact manifold. The set $\mathcal{C}$ is a manifold of dimension $NN_b - \frac{1}{2}N(N+1)$, called the Stiefel manifold; the set $\mathcal{P}$ of the rank-N orthogonal projectors in $\mathbb{C}^{N_b}$ is a manifold of dimension $N(N_b - N)$, called the Grassmann manifold. We refer to [12] for an interesting review of optimization methods on Stiefel and Grassmann manifolds.

From a historical viewpoint, the first minimization method for solving the Hartree-Fock-Roothaan-Hall equations, the so-called *steepest descent method*, was proposed in [20]. It basically consists in performing one unconstrained gradient step on the function $D \mapsto E(D)$ (i.e., $\widetilde{D}_{k+1} = D_k - t\nabla E(D_k) = D_k - tF(D_k)$), followed by a "projection" step $\widetilde{D}_{k+1} \to D_{k+1} \in \mathcal{P}$. The "projection" can be done using McWeeny's purification, an iterative method consisting in replacing at each step D with $3D^2 - 2D^3$. It is easily checked that if $\widetilde{D}_{k+1}$ is close enough to $\mathcal{P}$, the purification method converges quadratically to the point of $\mathcal{P}$ closest to $\widetilde{D}_{k+1}$ for the Frobenius norm. The steepest descent method has the drawback of any basic gradient method: it converges very slowly, and is therefore never used in practice.

Newton-like algorithms for computing Hartree-Fock ground states appeared in the early 1960s with Bacskay quadratic convergent (QC) method [3]. Bacskay's approach was to lift the constraints and use a standard Newton algorithm for *unconstrained* optimization. The local maps of the manifold $\mathcal{P}$ used in [3] are the following exponential maps: for any $C \in \mathbb{C}^{N_b \times N_b}$ such that $C^* S C = I_{N_b}$,

$$\mathcal{P} = \left\{ C e^{A} D_0 e^{-A} C^*, \ D_0 = \begin{bmatrix} I_N & 0 \\ 0 & 0 \end{bmatrix}, \right.$$
$$\left. A = \begin{bmatrix} 0 & -A_{\mathrm{vo}}^* \\ A_{\mathrm{vo}} & 0 \end{bmatrix}, \ A_{\mathrm{vo}} \in \mathbb{C}^{(N_b - N) \times N} \right\};$$

the suffix vo denotes the *virtual-occupied* off-diagonal block of the matrix A. Starting from some reference matrix C, Bacskay QC algorithm performs one Newton step on the *unconstrained* optimization problem

$$\min \left\{ E^C(A_{\mathrm{vo}}) := E^{\mathrm{HF}}(C e^{A} D_0 e^{-A} C^*), \right.$$
$$\left. A_{vo} \in \mathbb{C}^{(N_b - N) \times N} \right\},$$

and updates the reference matrix C by replacing C with $C e^{\widetilde{A}}$, where $\widetilde{A} = \begin{bmatrix} 0 & -\widetilde{A}_{\mathrm{vo}}^* \\ \widetilde{A}_{\mathrm{vo}} & 0 \end{bmatrix}$, $\widetilde{A}_{\mathrm{vo}}$ denoting the result of the Newton step. Newton methods being very expensive in terms of computational costs, various attempts have been made to build quasi-Newton versions of Bacskay QC algorithm (see, for e.g., [10, 13]).

A natural alternative to Bacskay QC is to use Newton-like algorithms for *constrained* optimization in order to directly tackle problems (10) or (11) (see, e.g., [26]). Trust-region methods for solving the constrained optimization problem (10) have also been developed by Helgaker and co-workers [27], and independently by Martínez and co-workers [14]. Recently, gradient flow methods for solving (10) [1] and (11) [17] have been introduced and analyzed from a mathematical viewpoint.

For molecular systems of moderate size, and when the Hartree-Fock model is discretized in atomic orbital basis sets, direct minimization methods are usually less efficient than the methods based on constraint relaxation or optimal mixing presented in the next two sections.

Lieb Variational Principle and Constraint Relaxation

We now consider the variational problem

$$\min \left\{ E^{\mathrm{HF}}(D), \ D \in \widetilde{\mathcal{P}} \right\}, \qquad (21)$$

$$\widetilde{\mathcal{P}} = \left\{ D \in \mathbb{C}_h^{N_b \times N_b}, \ 0 \le D \le 1, \ \mathrm{Tr}(D) = N \right\},$$

where $0 \le D \le 1$ means that $0 \le \Phi^* D \Phi \le \Phi^* \Phi$ for all $\Phi \in \mathbb{C}^{N_b}$, or equivalently, that all the eigenvalues of D lay in the range $[0, 1]$. It is easily seen that the set $\widetilde{\mathcal{P}}$ is convex. It is in fact the convex hull of the set $\mathcal{P}$. A fundamental remark is that all the local minimizers of (21) are on $\mathcal{P}$ [9]. This is the discretized version of a result by Lieb [18]. It is, therefore, possible to solve the Hartree-Fock model by relaxing the non-convex constraint $D^2 = D$ into the convex constraint $0 \le D \le 1$.

The orthogonal projection of a given hermitian matrix D on $\widetilde{\mathcal{P}}$ for the Frobenius scalar product can be computed by diagonalizing D [8]. The cost of one iteration of the usual projected gradient algorithm [4] is therefore the same at the cost of one iteration of the Roothaan algorithm.

A more efficient algorithm, the Optimal Damping Algorithm (ODA), is the following [7]

$$\begin{cases} D_{k+1} \in \operatorname{argmin} \left\{ \mathrm{Tr}\left(F(\widetilde{D}_k) D'\right), \ D' \in \mathcal{P} \right\} \\ \widetilde{D}_{k+1} = \operatorname{argmin} \left\{ E^{\mathrm{HF}}(\widetilde{D}), \ \widetilde{D} \in \mathrm{Seg}[\widetilde{D}_k, D_{k+1}] \right\}, \end{cases}$$

where $\mathrm{Seg}[\widetilde{D}_k, D_{k+1}] = \left\{ (1-\lambda)\widetilde{D}_k + \lambda D_{k+1}, \lambda \in [0,1] \right\}$ denotes the line segment linking $\widetilde{D}_k$ and D_{k+1}. As E^{HF} is a second degree polynomial in the density matrix, the last step consists in minimizing a quadratic function of λ on $[0, 1]$, which can be done analytically. The procedure is initialized with $\widetilde{D}_0 = D_0$, $D_0 \in \mathcal{P}$ being the initial guess. The ODA thus generates two sequences of matrices:

- The main sequence of density matrices $(D_k)_{k \in \mathbb{N}} \in \mathcal{P}^{\mathbb{N}}$ which is proven to numerically converge to an *Aufbau* solution to the Hartree-Fock-Roothaan-Hall equations [9]
- A secondary sequence $(\widetilde{D}_k)_{k \ge 1}$ of matrices belonging to $\widetilde{\mathcal{P}}$

The Hartree-Fock energy is a Lyapunov functional of ODA: it decays at each iteration. This follows from the fact that for all $D' \in \mathcal{P}$ and all $\lambda \in [0, 1]$,

$$E^{\mathrm{HF}}((1-\lambda)\widetilde{D}_k + \lambda D') = E^{\mathrm{HF}}(\widetilde{D}_k) + \lambda \mathrm{Tr}\left(F(\widetilde{D}_k)\right.$$
$$\left.(D' - \widetilde{D}_k)\right) + \frac{\lambda^2}{2} \mathrm{Tr}\left(G(D' - \widetilde{D}_k)(D' - \widetilde{D}_k)\right). \qquad (22)$$

The "steepest descent" direction, that is, the density matrix D for which the slope $s_{\widetilde{D}_k \to D} = \mathrm{Tr}\,(F(\widetilde{D}_k)(D - \widetilde{D}_k))$ is minimum, is precisely D_{k+1}.

In some sense, ODA is a combination of diagonalization and direct minimization. The practical implementation of ODA is detailed in [7], where numerical tests are also reported. The cost of one ODA iteration is approximatively the same as for the Roothaan algorithm. Numerical tests show that ODA is particularly efficient in the early steps of the iterative procedure.

Convergence Acceleration

SCF convergence can be accelerated by performing, at each step of the iterative procedure, a mixing of the previous iterates:

$$\begin{cases} D_{k+1} \in \operatorname{argmin}\left\{\mathrm{Tr}\,(\widetilde{F}_k D'),\ D' \in \mathcal{P}\right\} \\ \widetilde{F}_k = \displaystyle\sum_{j=0}^{k} c_{j,k} F(D_j), \quad \sum_{j=0}^{k} c_{j,k} = 1, \end{cases} \tag{23}$$

where the mixing coefficients $c_{j,k}$ are optimized according to some criterion. Note that in the Hartree-Fock setting, the mean-field Hamiltonian $F(D)$ is affine in D, so that mixing the $F(D_j)$'s amounts to mixing the D_j's:

$$\widetilde{F}_k = F(\widetilde{D}_k) \quad \text{where} \quad \widetilde{D}_k = \sum_{j=0}^{k} c_{j,k} D_j .$$

This is no longer true for Kohn-Sham models.

In Pulay's DIIS algorithm [22], the mixing coefficients are obtained by solving

$$\min\left\{\left\|\sum_{j=1}^{k} c_j [F(D_j), D_j]\right\|_{\mathrm{F}}^2,\ \sum_{j=1}^{k} c_j = 1\right\}.$$

The commutator $[F(D), D]$ is in fact the gradient of the functional $A \mapsto E^{\mathrm{HF}}(e^A D e^{-A})$ defined on the vector space of the $N_b \times N_b$ antihermitian matrices (note that $e^A D e^{-A} \in \mathcal{P}$ for all $D \in \mathcal{P}$ and A antihermitian); it vanishes when D is a critical point of E^{HF} on $\mathcal{P}$.

In the EDIIS algorithm [16], the mixing coefficients are chosen to minimize the Hartree-Fock energy of $\widetilde{D}_k$:

$$\min\left\{E^{\mathrm{HF}}\left(\sum_{j=1}^{k} c_j D_j\right),\ c_j \geq 0,\ \sum_{j=1}^{k} c_j = 1\right\}$$

(note that as the c_j's are chosen non-negative, $\widetilde{D}_k$ is the element of $\widetilde{\mathcal{P}}$ which minimizes the Hartree-Fock energy on the convex hull of $\{D_0, D_1, \cdots, D_k\}$).

The DIIS algorithm does not always converge. On the other hand, when it converges, it is extremely fast. This nice feature of the DIIS algorithm has not yet been fully explained by rigorous mathematical arguments (see however [23] for a numerical analysis of DIIS-type algorithms in an unconstrained setting).

SCF Algorithms for Kohn-Sham Models

After discretization in a finite basis set, the Kohn-Sham energy functional reads

$$E^{\mathrm{KS}}(D) = \mathrm{Tr}\,(hD) + \frac{1}{2}\mathrm{Tr}\,(J(D)D) + E_{\mathrm{xc}}(D),$$

where $[J(D)]_{\mu\nu} := \sum_{\kappa\lambda} (\mu\nu|\kappa\lambda) D_{\kappa\lambda}$ is the Coulomb operator, and where E_{xc} is the exchange-correlation energy functional [11]. In the standard Kohn-Sham model [15], E^{KS} is minimized on $\mathcal{P}$, while in the extended Kohn-Sham model [11], E^{KS} is minimized on the convex set $\widetilde{\mathcal{P}}$. The algorithms presented in the previous sections can be transposed mutatis mutandis to the Kohn-Sham setting, but as $E_{\mathrm{xc}}(D)$ is not a second order polynomial in D, the mathematical analysis is more complicated. In particular, no rigorous result on the Roothaan algorithm for Kohn-Sham has been published so far.

Note that the equality $F(\sum_i c_i D_i) = \sum_i c_i F(D_i)$ whenever $\sum_i c_i = 1$ is true for Hartree-Fock with $F(D) = \nabla E^{\mathrm{HF}}(D) = h + G(D)$, but not for Kohn-Sham with $F(D) = \nabla E^{\mathrm{KS}}(D) = h + J(D) + \nabla E_{\mathrm{xc}}(D)$. Consequently, in contrast with the situation encountered in the Hartree-Fock framework, mixing density matrices and mixing Kohn-Sham Hamiltonians are not equivalent procedures. This leads to a variety of acceleration schemes for Kohn-Sham that boil down to either DIIS or EDIIS in the Hartree-Fock setting. For the sake of brevity, and also because the situation is evolving fast (several new algorithms are proposed every year, and identifying the best algorithms is a matter of debate), we will not present these schemes here.

Let us finally mention that, if iterative methods based on repeated diagonalization of the mean-field Hamiltonian, combined with mixing procedures, are more efficient than direct minimization methods for moderate size molecular systems, and when the Kohn-Sham problem is discretized in atomic orbital basis sets, the situation may be different for very large systems, or when finite element or planewave discretization methods are used (see, e.g., [19,21] and references therein).

Cross-References

- ▶ A Priori and a Posteriori Error Analysis in Chemistry
- ▶ Density Functional Theory
- ▶ Hartree–Fock Type Methods
- ▶ Linear Scaling Methods
- ▶ Numerical Analysis of Eigenproblems for Electronic Structure Calculations

References

1. Alouges, F., Audouze, C.: Preconditioned gradient flows and applications to the Hartree-Fock functional. Numer. Methods PDE **25**, 380–400 (2009)
2. Bach, V., Lieb, E.H., Loss, M., Solovej, J.P.: There are no unfilled shells in unrestricted Hartree-Fock theory. Phys. Rev. Lett. **72**(19), 2981–2983 (1994)
3. Bacskay, G.B.: A quadratically convergent Hartree-Fock (QC-SCF) method. Application closed shell. Syst. Chem. Phys. **61**, 385–404 (1961)
4. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.: Numerical Optimization. Theoretical and Practical Aspects. Springer, Berlin/New York (2006)
5. Boys, S.F.: Electronic wavefunctions. I. A general method of calculation for the stationary states of any molecular system. Proc. R. Soc. A **200**, 542–554 (1950)
6. Cancès, E., Le Bris, C.: On the convergence of SCF algorithms for the Hartree-Fock equations. M2AN Math. Model. Numer. Anal. **34**, 749–774 (2000)
7. Cancès, E., Le Bris, C.: Can we outperform the DIIS approach for electronic structure calculations? Int. J. Quantum Chem. **79**, 82–90 (2000)
8. Cancès, E., Pernal, K.: Projected gradient algorithms for Hartree-Fock and density-matrix functional theory. J. Chem. Phys. **128**, 134108 (2008)
9. Cancès, E., Defranceschi, M., Kutzelnigg, W., Le Bris, C., Maday, Y.: Computational quantum chemistry: A primer. In: Handbook of Numerical Analysis, vol. X, pp. 3–270. North-Holland, Amsterdam (2003)
10. Chaban, G., Schmidt, M.W., Gordon, M.S.: Approximate second order method for orbital optimization of SCF and MCSCF wavefunctions. Theor. Chem. Acc. **97**, 88–95 (1997)
11. Dreizler, R.M., Gross, E.K.U.: Density Functional Theory. Springer, Berlin/New York (1990)
12. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthonormality constraints. SIAM J. Matrix Anal. Appl. **20**, 303–353 (1998)
13. Fischer, T.H., Almlöf, J.: General methods for geometry and wave function optimization. J. Phys. Chem. **96**, 9768–9774 (1992)
14. Francisco, J., Martínez, J.M., Martínez, L.: Globally convergent trust-region methods for self-consistent field electronic structure calculations, J. Chem. Phys. **121**, 10863–10878 (2004)
15. Kohn, K., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, A1133 (1965)
16. Kudin, K., Scuseria, G.E., Cancès, E.: A black-box self-consistent field convergence algorithm: one step closer, J. Chem. Phys. **116**, 8255–8261 (2002)
17. Levitt, A.: Convergence of gradient-based algorithms for the Hartree-Fock equations, preprint (2011)
18. Lieb, E.H.: Variational principle for many-fermion systems. Phys. Rev. Lett. **46**, 457–459 (1981)
19. Marks, L.D., Luke, D.R.: Robust mixing for ab initio quantum mechanical calculations. Phys. Rev. B **78**, 075114 (2008)
20. McWeeny, R.: The density matrix in self-consistent field theory. I. Iterative construction of the density matrix. Proc. R. Soc. Lond. A **235**, 496–509 (1956)
21. Mostofi, A.A., Haynes, P.D., Skylaris, C.K., Payne, M.C.: Preconditioned iterative minimization for linear-scaling electronic structure calculations. J. Chem. Phys. **119**, 8842–8848 (2003)
22. Pulay, P.: Improved SCF convergence acceleration. J. Comput. Chem. **3**, 556–560 (1982)
23. Rohwedder, T., Schneider, R.: An analysis for the DIIS acceleration method used in quantum chemistry calculations. J. Math. Chem. **49**, 1889–1914 (2011)
24. Roothaan, C.C.J.: New developments in molecular orbital theory. Rev. Mod. Phys. **23**, 69–89 (1951)
25. Saunders, V.R., Hillier, I.H.: A "level-shifting" method for converging closed shell Hartree-Fock wavefunctions. Int. J. Quantum Chem. **7**, 699–705 (1973)
26. Shepard, R.: Elimination of the diagonalization bottleneck in parallel direct-SCF methods. Theor. Chim. Acta **84**, 343–351 (1993)
27. Thøgersen, L., Olsen, J., Yeager, D., Jørgensen, P., Sałek, P., Helgaker, T.: The trust-region self-consistent field method: towards a black-box optimization in Hartree-Fock and Kohn-Sham theories. J. Chem. Phys. **121**, 16–27 (2004)

Semiconductor Device Problems

Ansgar Jüngel
Institut für Analysis und Scientific Computing,
Technische Universität Wien, Wien, Austria

Mathematics Subject Classification

82D37; 35Q20; 35Q40; 35Q79; 76Y05

Definition

Highly integrated electric circuits in computer processors mainly consist of semiconductor transistors which amplify and switch electronic signals. Roughly speaking, a semiconductor is a crystalline solid whose conductivity is intermediate between an insulator and a conductor. The modeling and simulation of semiconductor transistors and other devices is of paramount importance in the microelectronics industry to reduce the development cost and time. A semiconductor device problem is defined by the process of deriving physically accurate but computationally feasible model equations and of constructing efficient numerical algorithms for the solution of these equations. Depending on the device structure, size, and operating conditions, the main transport phenomena may be very different, caused by diffusion, drift, scattering, or quantum effects. This leads to a variety of model equations designed for a particular situation or a particular device. Furthermore, often not all available physical information is necessary, and simpler models are needed, helping to reduce the computational cost in the numerical simulation. One may distinguish four model classes: microscopic/mesoscopic and macroscopic semiclassical models and microscopic/mesoscopic and macroscopic quantum models (see Fig. 1).

Description

In the following, we detail only some models from the four model classes since the field of semiconductor device problems became extremely large in recent years. For instance, we ignore compact models, hybrid model approaches, lattice heat equations, transport in subbands and magnetic fields, spintronics, and models for carbon nanotube, graphene, and polymer thin-film materials. For technological aspects, we refer to [9].

Microscopic Semiclassical Models

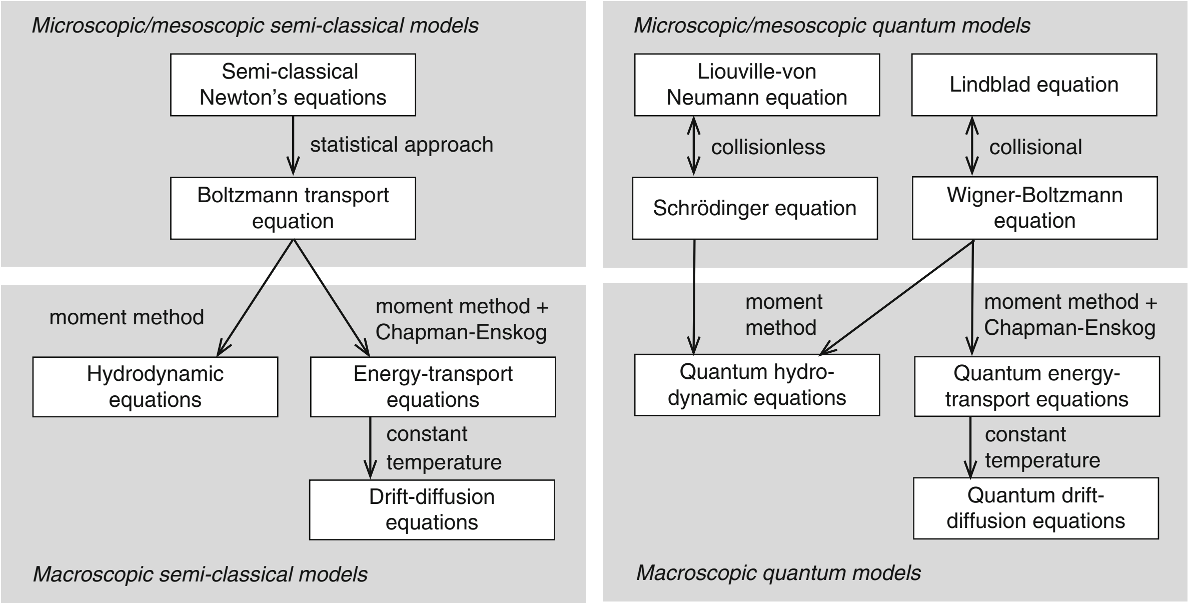
We are interested in the evolution of charge carriers moving in an electric field. Their motion can be modeled by Newton's law. However, in view of the huge number of electrons involved, the solution of the Newton equations is computationally too expensive. Moreover, we are not interested in the trajectory of each single particle. Hence, a statistical approach seems to be sufficient, introducing the distribution function (or "probability density") $f(x, v, t)$ of an electron ensemble, depending on the position $x \in \mathbb{R}^3$, velocity $v = \dot{x} = dx/dt \in \mathbb{R}^3$, and time $t > 0$. By Liouville's theorem, the trajectory of $f(x(t), v(t), t)$ does not change during time, in the absence of collisions, and hence,

$$0 = \frac{df}{dt} = \partial_t f + \dot{x}\cdot\nabla_x f + \dot{v}\cdot\nabla_v f \quad \text{along trajectories,} \tag{1}$$

where $\partial_t f = \partial f/\partial t$ and $\nabla_x f$, $\nabla_v f$ are gradients with respect to x, v, respectively.

Since electrons are quantum particles (and position and velocity cannot be determined with arbitrary accuracy), we need to incorporate some quantum mechanics. As the solution of the many-particle Schrödinger equation in the whole space is out of reach, we need an approximate approach. First, by Bloch's theorem, it is sufficient to solve the Schrödinger equation in a semiconductor lattice cell. Furthermore, the many-particle interactions are described by an effective Coulomb force. Finally, the properties of the semiconductor crystal are incorporated by the semiclassical Newton equations.

More precisely, let $p = \hbar k$ denote the crystal momentum, where $\hbar$ is the reduced Planck constant and k is the wave vector. For electrons with low energy, the velocity is proportional to the wave vector, $\dot{x} = \hbar k/m$, where m is the electron mass at rest. In the general case, we have to take into account the energy band structure of the semiconductor crystal (see [4, 7, 8] for details). Newton's third law is formulated as $\dot{p} = q\nabla_x V$, where q is the elementary charge and $V(x, t)$ is the electric potential. Then, using $\dot{v} = \dot{p}/m$ and $\nabla_k = (m/\hbar)\nabla_v$, (1) becomes the (mesoscopic) Boltzmann transport equation:


Microscopic/mesoscopic semi-classical models
Semi-classical Newton's equations
statistical approach
Boltzmann transport equation
moment method
moment method + Chapman-Enskog
Hydrodynamic equations
Energy-transport equations
constant temperature
Drift-diffusion equations
Macroscopic semi-classical models
Microscopic/mesoscopic quantum models
Liouville-von Neumann equation
Lindblad equation
collisionless
collisional
Schrödinger equation
Wigner-Boltzmann equation
moment method
moment method + Chapman-Enskog
Quantum hydro-dynamic equations
Quantum energy-transport equations
constant temperature
Quantum drift-diffusion equations
Macroscopic quantum models

Semiconductor Device Problems, Fig. 1 Hierarchy of some semiconductor models mentioned in the text

$$\begin{aligned}&\partial_t f + \frac{\hbar}{m} k \cdot \nabla_x f + \frac{q}{\hbar} \nabla_x V \cdot \nabla_k f \\ &= Q(f), \quad (x,k) \in \mathbb{R}^3 \times \mathbb{R}^3,\ t > 0,\end{aligned} \tag{2}$$

where $Q(f)$ models collisions of electrons with phonons, impurities, or other particles. The moments of f are interpreted as the particle density $n(x,t)$, current density $J(x,t)$, and energy density $(ne)(x,t)$:

$$\begin{aligned}n &= \int_{\mathbb{R}^3} f dk, \quad J = \frac{\hbar}{m} \int_{\mathbb{R}^3} k f dk, \\ ne &= \frac{\hbar^2}{2m} \int_{\mathbb{R}^3} |k|^2 f dk.\end{aligned} \tag{3}$$

In the self-consistent setting, the electric potential V is computed from the Poisson equation $\varepsilon_s \Delta V = q(n - C(x))$, where ε_s is the semiconductor permittivity and $C(x)$ models charged background ions (doping profile). Since n depends on the distribution function f, the Boltzmann–Poisson system is nonlinear.

The Boltzmann transport equation is defined over the six-dimensional phase space (plus time) whose high dimensionality makes its numerical solution a very challenging task. One approach is to employ the Monte Carlo method which consists in simulating a stochastic process. Drawbacks of the method are the stochastic nature and the huge computational cost. An alternative is the use of deterministic solvers, for example, expanding the distribution function with spherical harmonics [6].

Macroscopic Semiclassical Models

When collisions become dominant in the semiconductor domain, that is, the mean free path (the length which a particle travels between two consecutive collision events) is much smaller than the device size, a fluid dynamical approach may be appropriate. Macroscopic models are derived from (2) by multiplying the equation by certain weight functions, that is 1, k, and $|k|^2/2$, and integrating over the wave-vector space. Setting all physical constants to one in the following, for notational simplicity, we obtain, using the definitions (3), the balance equations:

$$\partial_t n + \operatorname{div}_x J = \int_{\mathbb{R}^3} Q(f) dk, \quad x \in \mathbb{R}^3,\ t > 0, \tag{4}$$

$$\partial_t J + \operatorname{div}_x \int_{\mathbb{R}^3} k \otimes k f dk - n \nabla_x V = \int_{\mathbb{R}^3} k Q(f) dk, \tag{5}$$

$$\begin{aligned}&\partial_t (ne) + \frac{1}{2} \operatorname{div}_x \int_{\mathbb{R}^3} k |k|^2 f dk - \nabla_x V \cdot \\ &\quad J = \frac{1}{2} \int_{\mathbb{R}^3} |k|^2 Q(f) dk.\end{aligned} \tag{6}$$

The higher-order integrals cannot be expressed in terms of the moments (3), which is called the closure problem. It can be solved by approximating f by the equilibrium distribution f_0, which can be justified by a scaling argument and asymptotic analysis. The equilibrium f_0 can be determined by maximizing the Boltzmann entropy under the constraints of given moments n, nu, and ne [4]. Inserting f_0 in (4)–(6) gives explicit expressions for the higher-order moments, yielding the so-called hydrodynamic model. Formally, there is some similarity with the Euler equations of fluid dynamics, and there has been an extensive discussion in the literature whether electron shock waves in semiconductors are realistic or not [10].

Diffusion models, which do not exhibit shock solutions, can be derived by a Chapman–Enskog expansion around the equilibrium distribution f_0 according to $f = f_0 + \alpha f_1$, where $\alpha > 0$ is the Knudsen number (the ratio of the mean free path and the device length) which is assumed to be small compared to one. The function f_1 turns out to be the solution of a certain operator equation involving the collision operator $Q(f)$. Depending on the number of given moments, this leads to the drift-diffusion equations (particle density given):

$$\partial_t n + \operatorname{div}_x J = 0, \quad J = -\nabla_x n + n\nabla_x V, \\ x \in \mathbb{R}^3,\ t > 0, \tag{7}$$

or the energy-transport equations (particle and energy densities given)

$$\partial_t n + \operatorname{div}_x J = 0, \quad J = -\nabla_x n + \frac{n}{T}\nabla_x V, \\ x \in \mathbb{R}^3,\ t > 0, \tag{8}$$

$$\partial_t (ne) + \operatorname{div}_x S + nu \cdot \nabla_x V = 0, \\ S = -\frac{3}{2}(\nabla_x (nT) - n\nabla_x V), \tag{9}$$

where $ne = \frac{3}{2}nT$, T being the electron temperature, and S is the heat flux. For the derivation of these models; we have assumed that the equilibrium distribution is given by Maxwell–Boltzmann statistics and that the elastic scattering rate is proportional to the wave vector. More general models can be derived too, see [4, Chap. 6].

The drift-diffusion model gives a good description of the transport in semiconductor devices close to equilibrium but it is not accurate enough for submicron devices due to, for example, temperature effects, which can be modeled by the energy-transport equations.

In the presence of high electric fields, the stationary equations corresponding to (7)–(9) are convection dominant. This can be handled by the Scharfetter–Gummel discretization technique. The key idea is to approximate the current density along each edge in a mesh by a constant, yielding an exponential approximation of the electric potential. This technique is related to mixed finite-element and finite-volume methods [2]. Another idea to eliminate the convective terms is to employ (dual) entropy variables. For instance, for the energy-transport equations, the dual entropy variables are $w = (w_1, w_2) = ((\mu - V)/T, -1/T)$, where μ is the chemical potential, given by $n = T^{3/2}\exp(\mu/T)$. Then (8) and (9) can be formulated as the system:

$$\partial_t b(w) - \operatorname{div}(D(w,V)\nabla w) = 0,$$

where $b(w) = (n, \frac{3}{2}nT)^\top$ and $D(w,V) \in \mathbb{R}^{2\times 2}$ is a symmetric positive definite diffusion matrix [4] such that standard finite-element techniques are applicable.

Microscopic Quantum Models

The semiclassical approach is reasonable if the carriers can be treated as particles. The validity of this description is measured by the de Broglie wavelength λ_B corresponding to a thermal average carrier. When the electric potential varies rapidly on the scale of λ_B or when the mean free path is much larger than λ_B, quantum mechanical models are more appropriate. A general description is possible by the Liouville–von Neumann equation:

$$i\varepsilon\partial_t\widehat{\rho} = [H, \widehat{\rho}] := H\widehat{\rho} - \widehat{\rho}H, \quad t > 0,$$

for the density matrix operator $\widehat{\rho}$, where $i^2 = -1$, $\varepsilon > 0$ is the scaled Planck constant, and H is the quantum mechanical Hamiltonian. The operator $\widehat{\rho}$ is assumed to possess a complete orthonormal set of eigenfunctions (ψ_j) and eigenvalues (λ_j). The sequence of Schrödinger equations $i\varepsilon\partial_t\psi_j = H\psi_j$ $(j \in \mathbb{N})$, together with the numbers $\lambda_j \geq 0$, is called a mixed-state Schrödinger system with the particle density $n(x,t) = \sum_{j=1}^{\infty}\lambda_j|\psi_j(x,t)|^2$. In particular, λ_j can be interpreted as the occupation probability of the state j.

The Schrödinger equation describes the evolution of a quantum state in an active region of a semiconductor device. It is used when inelastic scattering is sufficiently weak such that phase coherence can be assumed and effects such as resonant tunneling and quantum conductance can be observed. Typically, the device is connected to an exterior medium through access zones, which allows for the injection of charge carriers. Instead of solving the Schrödinger equation in the whole domain (self-consistently coupled to the Poisson equation), one wishes to solve the problem only in the active region and to prescribe transparent boundary conditions at the interfaces between the active and access zones. Such a situation is referred to as an open quantum system. The determination of transparent boundary conditions is a delicate issue since ad hoc approaches often lead to spurious oscillations which deteriorate the numerical solution [1].

Nonreversible interactions of the charge carriers with the environment can be modeled by the Lindblad equation:

$$i\varepsilon\partial_t\widehat{\rho}=[H,\widehat{\rho}]+i\sum_k\Big(L_k\widehat{\rho}L_k^*-\frac{1}{2}(L_k^*L_k\widehat{\rho}+\widehat{\rho}L_k^*L_k)\Big),$$

where L_k are the so-called Lindblad operators and L_k^* is the adjoint of L_k. In the Fourier picture, this equation can be formulated as a quantum kinetic equation, the (mesoscopic) Wigner–Boltzmann equation:

$$\partial_t w + p\cdot\nabla_x w + \theta[V]w = Q(w), \quad (x,p)\in\mathbb{R}^3\times\mathbb{R}^3,\ t>0, \tag{10}$$

where p is the crystal momentum, $\theta[V]w$ is the potential operator which is a nonlocal version of the drift term $\nabla_x V\cdot\nabla_p w$ [4, Chap. 11], and $Q(w)$ is the collision operator. The Wigner function $w = W[\widehat{\rho}]$, where W denotes the Wigner transform, is essentially the Fourier transform of the density matrix. A nice feature of the Wigner equation is that it is a phase-space description, similar to the semiclassical Boltzmann equation. Its drawbacks are that the Wigner function cannot be interpreted as a probability density, as the Boltzmann distribution function, and that the Wigner equation has to be solved in the high dimensional phase space. A remedy is to derive macroscopic models which are discussed in the following section.

Macroscopic Quantum Models

Macroscopic models can be derived from the Wigner–Boltzmann equation (10) in a similar manner as from the Boltzmann equation (2). The main difference to the semiclassical approach is the definition of the equilibrium. Maximizing the von Neumann entropy under the constraints of given moments of a Wigner function w, the formal solution (if it exists) is given by the so-called quantum Maxwellian $M[w]$, which is a nonlocal version of the semiclassical equilibrium. It was first suggested by Degond and Ringhofer and is related to the (unconstrained) quantum equilibrium given by Wigner in 1932 [3, 5]. We wish to derive moment equations from the Wigner–Boltzmann equation (10) for the particle density n, current density J, and energy density ne, defined by:

$$n=\int_{\mathbb{R}^3} M[w]dp, \quad J=\int_{\mathbb{R}^3} pM[w]dp, \quad ne=\frac{1}{2}\int_{\mathbb{R}^3}|p|^2M[w]dp.$$

Such a program was carried out by Degond et al. [3], using the simple relaxation-type operator $Q(w) = M[w]-w$. This leads to a hierarchy of quantum hydrodynamic and diffusion models which are, in contrast to their semiclassical counterparts, nonlocal.

When we employ only one moment (the particle density) and expand the resulting moment model in powers of ε up to order $O(\varepsilon^4)$ (to obtain local equations), we arrive at the quantum drift-diffusion (or density-gradient) equations:

$$\partial_t n + \mathrm{div}_x J = 0, \quad J = -\nabla_x n + n\nabla_x V + \frac{\varepsilon^2}{6}n\nabla_x\left(\frac{\Delta_x\sqrt{n}}{\sqrt{n}}\right), \quad x\in\mathbb{R}^3, \quad t>0.$$

This model is employed to simulate the carrier inversion layer near the oxide of a MOSFET (metal-oxide-semiconductor field-effect transistor). The main difficulty of the numerical discretization is the treatment of the highly nonlinear fourth-order quantum correction. However, there exist efficient exponentially fitted finite-element approximations, see the references of Pinnau in [4, Chap. 12].

Formally, the moment equations for the charge carriers and energy density give the quantum energy-transport model. Since its mathematical structure is less clear, we do not discuss this model [4, Chap. 13.2].

Employing all three moments n, nu, ne, the moment equations, expanded up to terms of order $O(\varepsilon^4)$, become the quantum hydrodynamic equations:

$$\begin{aligned}
&\partial_t n + \operatorname{div} J = 0, \quad \partial_t J + \operatorname{div}_x\left(\frac{J \otimes J}{n} + P\right)\\
&\qquad + n\nabla_x V = -\int_{\mathbb{R}^3} pQ(M[w])dp,\\
&\partial_t(ne) - \operatorname{div}_x((P + ne\mathbb{I})u - q) + \nabla_x V \cdot\\
&J = \frac{1}{2}\int_{\mathbb{R}^3} |p|^2 Q(M[w])dp, \quad x \in \mathbb{R}^3,\ t > 0,
\end{aligned}$$

where $\mathbb{I}$ is the identity matrix in $\mathbb{R}^{3\times 3}$, the quantum stress tensor P and the energy density ne are given by:

$$\begin{aligned}
P &= nT\mathbb{I} - \frac{\varepsilon^2}{12} n\nabla_x^2 \log n, \quad ne = \frac{3}{2}nT\\
&\qquad + \frac{1}{2}n|u|^2 - \frac{\varepsilon^2}{24} n\Delta_x \log n,
\end{aligned}$$

$u = J/n$ is the mean velocity, and $q = -(\varepsilon^2/24)n(\Delta_x u + 2\nabla_x \operatorname{div}_x u)$ is the quantum heat flux. When applying a Chapman–Enskog expansion around the quantum equilibrium, viscous effects are added, leading to quantum Navier–Stokes equations [5, Chap. 5]. These models are very interesting from a theoretical viewpoint since they exhibit a surprising nonlinear structure. Simulations of resonant tunneling diodes using these models give qualitatively reasonable results. However, as expected, quantum phenomena are easily destroyed by the occurring diffusive or viscous effects.

References

1. Antoine, X., Arnold, A., Besse, C., Ehrhardt, M., Schädle, A.: A review of transparent and artificial boundary conditions techniques for linear and nonlinear Schrödinger equations. Commun. Comput. Phys. **4**, 729–796 (2008)
2. Brezzi, F., Marini, L., Micheletti, S., Pietra, P., Sacco, R., Wang, S.: Discretization of semiconductor device problems. In: Schilders, W., ter Maten, W. (eds.) Handbook of Numerical Analysis. Numerical Methods in Electromagnetics, vol. 13, pp. 317–441. North-Holland, Amsterdam (2005)
3. Degond, P., Gallego, S., Méhats, F., Ringhofer, C.: Quantum hydrodynamic and diffusion models derived from the entropy principle. In: Ben Abdallah, N., Frosali, G. (eds.) Quantum Transport: Modelling, Analysis and Asymptotics. Lecture Notes in Mathematics 1946, pp. 111–168. Springer, Berlin (2009)
4. Jüngel, A.: Transport Equations for Semiconductors. Springer, Berlin (2009)
5. Jüngel, A.: Dissipative quantum fluid models. Revista Mat. Univ. Parma, to appear (2011)
6. Hong, S.-M., Pham, A.-T., Jungemann, C.: Deterministic Solvers for the Boltzmann Transport Equation. Springer, New York (2011)
7. Lundstrom, M.: Fundamentals of Carrier Transport. Cambridge University Press, Cambridge (2000)
8. Markowich, P., Ringhofer, C., Schmeiser, C.: Semiconductor Equations. Springer, Vienn (1990)
9. Mishra, U., Singh, J.: Semiconductor Device Physics and Design. Springer, Dordrecht (2007)
10. Rudan, M., Gnudi, A., Quade, W.: A generalized approach to the hydrodynamic model of semiconductor equations. In: Baccarani, G. (ed.) Process and Device Modeling for Microelectronics, pp. 109–154. Elsevier, Amsterdam (1993)

Shearlets

Gitta Kutyniok
Institut für Mathematik, Technische Universität
Berlin, Berlin, Germany

Mathematics Subject Classification

42C40; 42C15; 65T60

Synonyms

Shearlets; Shearlet system

Short Description

Shearlets are multiscale systems in $L^2(\mathbb{R}^2)$ which efficiently encode anisotropic features. They extend the framework of wavelets and are constructed by parabolic scaling, shearing, and translation applied to one or very few generating functions. The main application area of shearlets is imaging science, for example, denoising, edge detection, or inpainting. Extensions of shearlet systems to $L^2(\mathbb{R}^n)$, $n \geq 3$ are also available.

Description

Multivariate Problems

Multivariate problem classes are typically governed by anisotropic features such as singularities concentrated on lower dimensional embedded manifolds. Examples are edges in images or shock fronts of transport dominated equations. Since due to their isotropic nature wavelets are deficient to efficiently encode such functions, several directional representation systems were proposed among which are ridgelets, contourlets, and curvelets.

Shearlets were introduced in 2006 [10] and are to date the only directional representation system which provides optimally sparse approximations of anisotropic features while providing a unified treatment of the continuum and digital realm in the sense of allowing faithful implementations. One important structural property is their membership in the class of affine systems, similar to wavelets. A comprehensive presentation of the theory and applications of shearlets can be found in [16].

Continuous Shearlet Systems

Continuous shearlet systems are generated by application of *parabolic scaling* A_a, $\tilde{A}_a$, $a > 0$, *shearing* S_s, $s \in \mathbb{R}$ and *translation*, where

$$A_a = \begin{pmatrix} a & 0 \\ 0 & a^{1/2} \end{pmatrix}, \quad \tilde{A}_a = \begin{pmatrix} a^{1/2} & 0 \\ 0 & a \end{pmatrix},$$

$$\text{and} \quad S_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix},$$

to one or very few generating functions. For $\psi \in L^2(\mathbb{R}^2)$, the associated *continuous shearlet system* is defined by

$$\Big\{\psi_{a,s,t} = a^{-\frac{3}{4}}\psi(A_a^{-1}S_s^{-1}(\cdot - t)) : a > 0, s \in \mathbb{R},$$
$$t \in \mathbb{R}^2\Big\},$$

with a determining the *scale*, s the *direction*, and t the *position* of a *shearlet* $\psi_{a,s,t}$. The associated *continuous shearlet transform* of some function $f \in L^2(\mathbb{R}^2)$ is the mapping

$$L^2(\mathbb{R}^2) \ni f \mapsto \langle f, \psi_{a,s,t}\rangle, \quad a > 0, s \in \mathbb{R}, t \in \mathbb{R}^2.$$

The continuous shearlet transform is an isometry, provided that ψ satisfies some weak regularity conditions.

One common class of generating functions are *classical shearlets*, which are band-limited functions $\psi \in L^2(\mathbb{R}^2)$ defined by

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\,\hat{\psi}_2\left(\tfrac{\xi_2}{\xi_1}\right),$$

where $\psi_1 \in L^2(\mathbb{R})$ is a discrete wavelet, i.e., it satisfies $\sum_{j\in\mathbb{Z}} |\hat{\psi}_1(2^{-j}\xi)|^2 = 1$ for a.e. $\xi \in \mathbb{R}$ with $\hat{\psi}_1 \in C^\infty(\mathbb{R})$ and $supp\hat{\psi}_1 \subseteq [-\frac{1}{2}, -\frac{1}{16}] \cup [\frac{1}{16}, \frac{1}{2}]$, and $\psi_2 \in L^2(\mathbb{R})$ is a "bump function" in the sense that $\sum_{k=-1}^{1} |\hat{\psi}_2(\xi + k)|^2 = 1$ for a.e. $\xi \in [-1, 1]$ with $\hat{\psi}_2 \in C^\infty(\mathbb{R})$ and $supp\hat{\psi}_2 \subseteq [-1, 1]$. Figure 1 illustrates classical shearlets and the tiling of Fourier domain they provide, which ensures their directional sensitivity.

From a mathematical standpoint, continuous shearlets are being generated by a unitary representation of a particular semi-direct product, the *shearlet group* [2]. However, since those systems and their associated transforms do not provide a uniform resolution of all directions but are biased towards one axis, for applications cone-adapted continuous shearlet systems were introduced. For $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$, the *cone-adapted continuous shearlet system* $SH_{cont}(\phi, \psi, \tilde{\psi})$ is defined by

$$SH_{cont}(\phi, \psi, \tilde{\psi}) = \Phi_{cont}(\phi) \cup \Psi_{cont}(\psi) \cup \tilde{\Psi}_{cont}(\tilde{\psi}),$$

where

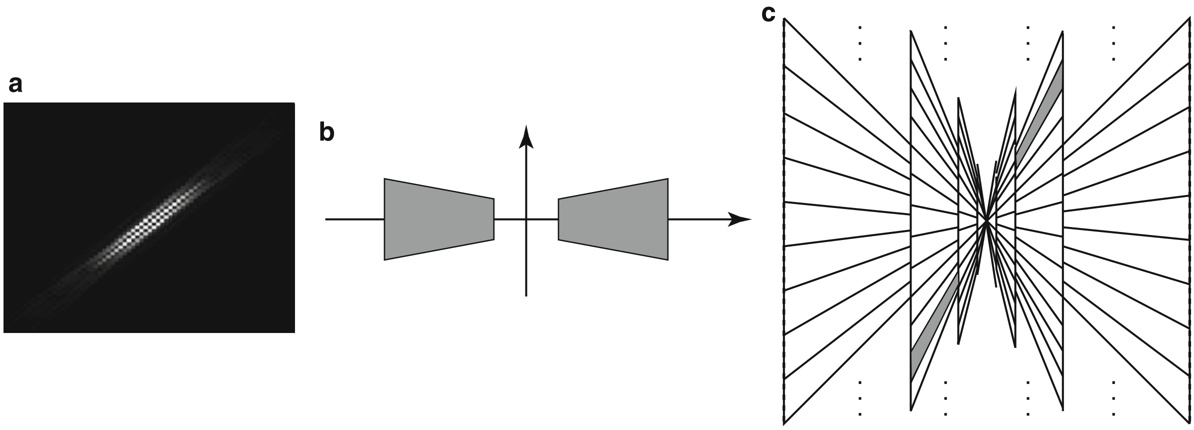
$$\Phi_{cont}(\phi) = \{\phi_t = \phi(\cdot - t) : t \in \mathbb{R}^2\},$$

$$\Psi_{cont}(\psi) = \{\psi_{a,s,t} = a^{-\frac{3}{4}}\psi(A_a^{-1}S_s^{-1}(\cdot - t))$$
$$: a \in (0, 1], |s| \leq 1 + a^{1/2}, t \in \mathbb{R}^2\},$$

$$\tilde{\Psi}_{cont}(\tilde{\psi}) = \{\tilde{\psi}_{a,s,t} = a^{-\frac{3}{4}}\tilde{\psi}(\tilde{A}_a^{-1}S_s^{-T}(\cdot - t))$$
$$: a \in (0, 1], |s| \leq 1 + a^{1/2}, t \in \mathbb{R}^2\}.$$

The associated transform is defined in a similar manner as before. The induced uniform resolution of all directions by a cone-like partition of Fourier domain is illustrated in Fig. 2.

The high directional selectivity of cone-adapted continuous shearlet systems is reflected in the result that they precisely resolve wavefront sets of



Shearlets, Fig. 1 Classical shearlets: (**a**) $|\psi_{a,s,t}|$ for exemplary values of a, s, and t. (**b**) Support of $\hat{\psi}$. (**c**) Approximate support of $\hat{\psi}_{a,s,t}$ for different values of a and s

distributions f by the decay behavior of $|\langle f, \psi_{a,s,t}\rangle|$ and $|\langle f, \tilde{\psi}_{a,s,t}\rangle|$ as $a \to 0$ [15].

Discrete Shearlet Systems

Discretization of the parameters a, s, and t by $a = 2^{-j}$, $j \in \mathbb{Z}$, $s = -k2^{-j/2}$, $k \in \mathbb{Z}$, and $t = A_{2^j}^{-1} S_k^{-1} m$, $m \in \mathbb{Z}^2$ leads to the associated discrete systems. For $\psi \in L^2(\mathbb{R}^2)$, the *discrete shearlet system* is defined by

$$\{\psi_{j,k,m} = 2^{\frac{3}{4}j}\, \psi(S_k A_{2^j} \cdot - m) : j, k \in \mathbb{Z}, m \in \mathbb{Z}^2\},$$

with j determining the *scale*, k the *direction*, and m the *position* of a *shearlet* $\psi_{j,k,m}$. The associated *discrete shearlet transform* of some $f \in L^2(\mathbb{R}^2)$ is the mapping

$$L^2(\mathbb{R}^2) \ni f \mapsto \langle f, \psi_{j,k,m}\rangle, \quad j, k \in \mathbb{Z}, m \in \mathbb{Z}^2.$$

Similarly, for $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$, the *cone-adapted discrete shearlet system* $SH_{disc}(\phi, \psi, \tilde{\psi})$ is defined by

$$SH_{disc}(\phi, \psi, \tilde{\psi}) = \Phi_{disc}(\phi) \cup \Psi_{disc}(\psi) \cup \tilde{\Psi}_{disc}(\tilde{\psi}),$$

where

$$\begin{aligned}
\Phi_{disc}(\phi) &= \{\phi_m = \phi(\cdot - m) : m \in \mathbb{Z}^2\},\\
\Psi_{disc}(\psi) &= \{\psi_{j,k,m} = 2^{\frac{3}{4}j}\, \psi(S_k A_{2^j} \cdot - m)\\
&\qquad : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\},\\
\tilde{\Psi}_{disc}(\tilde{\psi}) &= \{\tilde{\psi}_{j,k,m} = 2^{\frac{3}{4}j}\, \tilde{\psi}(S_k^T \tilde{A}_{2^j} \cdot - m)\\
&\qquad : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}.
\end{aligned}$$

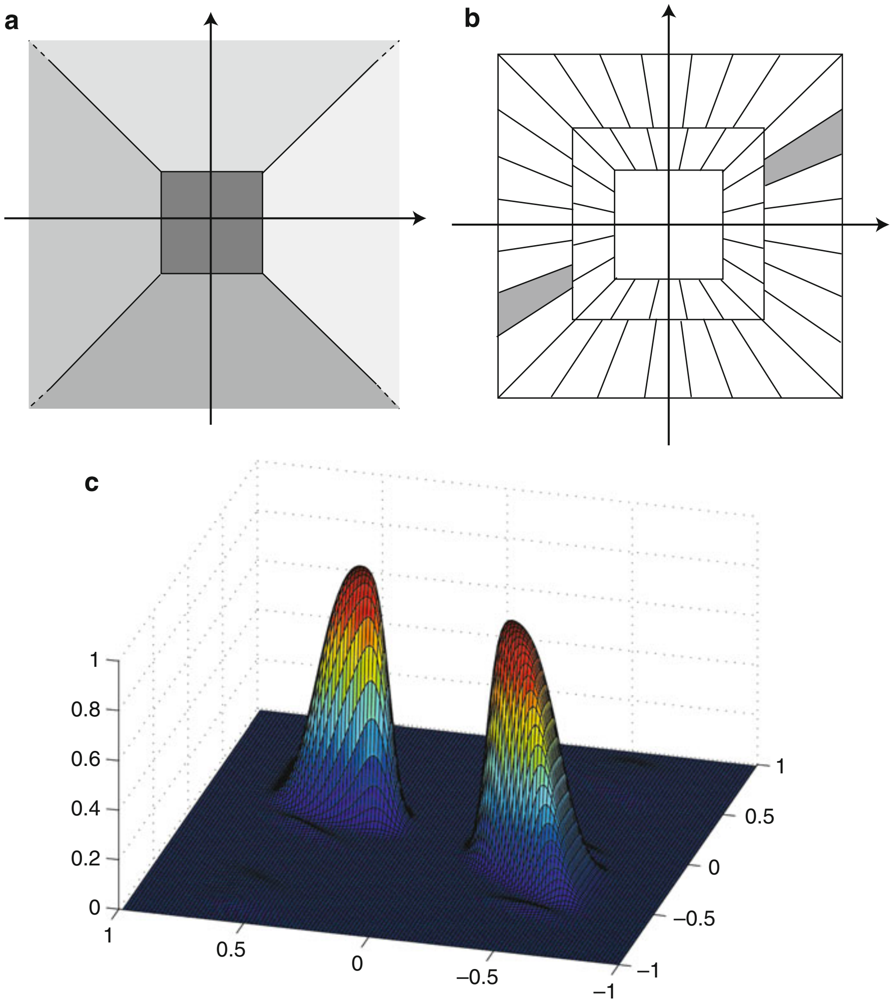
To allow more flexibility in the denseness of the positioning of shearlets, sometimes the discretization of the translation parameter t is performed by $t = A_{2^j}^{-1} S_k^{-1} \mathrm{diag}(c_1, c_2) m$, $m \in \mathbb{Z}^2, c_1, c_2 > 0$. A very general discretization approach is by coorbit theory which is however only applicable to the non-cone-adapted setting [4].

For classical (band-limited) shearlets as generating functions, both the discrete shearlet system and the cone-adapted discrete shearlet system – the latter one with a minor adaption at the intersections of the cones and suitable ϕ – form tight frames for $L^2(\mathbb{R}^2)$. A theory for compactly supported (cone-adapted) discrete shearlet systems is also available [14]. For a special class of separable generating functions, compactly supported cone-adapted discrete shearlet systems form a frame with the ratio of frame bounds being approximately 4.

Discrete shearlet systems provide optimally sparse approximations of anisotropic features. A customarily employed model are *cartoon-like functions*, i.e., compactly supported functions in $L^2(\mathbb{R}^2)$ which are C^2 apart from a closed piecewise C^2 discontinuity curve. Up to a log-factor, discrete shearlet systems based on classical shearlets or compactly supported shearlets satisfying some weak regularity conditions provide the optimal decay rate of the best N-term approximation of cartoon-like functions f [7, 17], i.e.,

$$\|f - f_N\|_2^2 \leq C\, N^{-2} (\log N)^3 \qquad \text{as } N \to \infty,$$

where here f_N denotes the N-term shearlet approximation using the N largest coefficients.



Shearlets, Fig. 2 Cone-adapted shearlet system: (**a**) Partitioning into cones. (**b**) Approximate support of $\hat{\psi}_{a,s,t}$ and $\widehat{\tilde{\psi}_{a,s,t}}$ for different values of a and s. (**c**) $|\hat{\psi}_{a,s,t}|$ for some shearlet ψ and exemplary values of a, s, and t

Fast Algorithms

The implementations of the shearlet transform can be grouped into two categories, namely, in Fourier-based implementations and in implementations in spatial domain.

Fourier-based implementations aim to produce the same frequency tiling as in Fig. 2b typically by employing variants of the Pseudo-Polar transform [5, 21]. Spatial domain approaches utilize filters associated with the transform which are implemented by a convolution in the spatial domain. A fast implementation with separable shearlets was introduced in [22], subdivision schemes are the basis of the algorithmic approach in [19], and a general filter approach was studied in [11].

Several of the associated algorithms are provided at www.ShearLab.org.

Extensions to Higher Dimensions

Continuous shearlet systems in higher dimensions have been introduced in [3]. In many situations, these systems inherit the property to resolve wavefront sets. The theory of discrete shearlet systems and their sparse approximation properties have been introduced and studied in [20] in dimension 3 with the possibility to extend the results to higher dimensions, and similar sparse approximation properties were derived.

Applications

Shearlets are nowadays used for a variety of applications which require the representation and processing of multivariate data such as imaging sciences. Prominent examples are deconvolution [23], denoising [6], edge detection [9], inpainting [13], segmentation [12], and separation [18]. Other application areas are sparse decompositions of operators such as the Radon operator [1] or Fourier integral operators [8].

References

1. Colonna, F., Easley, G., Guo, K., Labate, D.: Radon transform inversion using the shearlet representation. Appl. Comput. Harmon. Anal. **29**, 232–250 (2010)
2. Dahlke, S., Kutyniok, G., Maass, P., Sagiv, C., Stark, H.-G., Teschke, G.: The uncertainty principle associated with the continuous shearlet transform. Int. J. Wavelets Multiresolut. Inf. Process. **6**, 157–181 (2008)
3. Dahlke, S., Steidl, G., Teschke, G.: The continuous shearlet transform in arbitrary space dimensions. J. Fourier Anal. Appl. **16**, 340–364 (2010)
4. Dahlke, S., Steidl, G., Teschke, G.: Shearlet coorbit spaces: compactly supported analyzing shearlets, traces and embeddings. J. Fourier Anal. Appl. **17**, 1232–1255 (2011)
5. Easley, G., Labate, D., Lim, W-Q.: Sparse directional image representations using the discrete shearlet transform. Appl. Comput. Harmon. Anal. **25**, 25–46 (2008)
6. Easley, G., Labate, D., Colonna, F.: Shearlet based total Variation for denoising. IEEE Trans. Image Process. **18**, 260–268 (2009)
7. Guo, K., Labate, D.: Optimally sparse multidimensional representation using shearlets. SIAM J. Math. Anal. **39**, 298–318 (2007)
8. Guo, K., Labate, D.: Representation of Fourier integral operators using shearlets. J. Fourier Anal. Appl. **14**, 327–371 (2008)
9. Guo, K., Labate, D.: Characterization and analysis of edges using the continuous shearlet transform. SIAM J. Imaging Sci. **2**, 959–986 (2009)
10. Guo, K., Kutyniok, G., Labate, D.: Sparse multidimensional representations using anisotropic dilation and shear operators. In: Wavelets and Splines (Athens, 2005), pp. 189–201. Nashboro Press, Nashville (2006)
11. Han, B., Kutyniok, G., Shen, Z.: Adaptive multiresolution analysis structures and shearlet systems. SIAM J. Numer. Anal. **49**, 1921–1946 (2011)
12. Häuser, S., Steidl, G.: Convex multiclass segmentation with shearlet regularization. Int. J. Comput. Math. 90, 62–81 (2013)
13. King, E.J., Kutyniok, G., Zhuang, X.: Analysis of Inpainting via Clustered Sparsity and Microlocal Analysis. J. Math. Imaging Vis. (to appear)
14. Kittipoom, P., Kutyniok, G., Lim, W-Q.: Construction of compactly supported shearlet frames. Constr. Approx. **35**, 21–72 (2012)
15. Kutyniok, G., Labate, D.: Resolution of the wavefront set using continuous shearlets. Trans. Am. Math. Soc. **361**, 2719–2754 (2009)
16. Kutyniok, G., Labate, D. (eds.): Shearlets: Multiscale Analysis for Multivariate Data. Birkhäuser, Boston (2012)
17. Kutyniok, G., Lim, W-Q.: Compactly supported shearlets are optimally sparse. J. Approx. Theory **163**, 1564–1589 (2011)
18. Kutyniok, G., Lim, W-Q.: Image separation using wavelets and shearlets. In: Curves and Surfaces (Avignon, 2010). Lecture Notes in Computer Science, vol. 6920. Springer, Berlin, Heidelberg (2012)
19. Kutyniok, G., Sauer, T.: Adaptive directional subdivision schemes and shearlet multiresolution analysis. SIAM J. Math. Anal. **41**, 1436–1471 (2009)
20. Kutyniok, G., Lemvig, J., Lim, W-Q.: Optimally sparse approximations of 3D functions by compactly supported shearlet frames. SIAM J. Math. Anal. **44**, 2962–3017 (2012)
21. Kutyniok, G., Shahram, M., Zhuang, X.: ShearLab: a rational design of a digital parabolic scaling algorithm. SIAM J. Imaging Sci. **5**, 1291–1332 (2012)
22. Lim, W-Q.: The discrete shearlet transform: a new directional transform and compactly supported shearlet frames. IEEE Trans. Image Process. **19**, 1166–1180 (2010)
23. Patel, V.M., Easley, G., Healy, D.M.: Shearlet-based deconvolution. IEEE Trans. Image Process. **18**, 2673–2685 (2009)

Shift-Invariant Approximation

Robert Schaback
Institut für Numerische und Angewandte Mathematik
(NAM), Georg-August-Universität Göttingen,
Göttingen, Germany

Mathematics Subject Classification

41A25; 42C15; 42C25; 42C40

Synonyms

Approximation by integer translates

Short Definition

Shift-invariant approximation deals with functions f on the whole real line, e.g., *time series* and *signals*.

It approximates f by shifted copies of a single *generator* φ, i.e.,

$$f(x) \approx S_{f,h,\varphi}(x) := \sum_{k \in \mathbb{Z}} c_{k,h}(f)\varphi\left(\frac{x}{h} - k\right), \; x \in \mathbb{R}. \tag{1}$$

The functions $\varphi\left(\frac{\cdot}{h} - k\right)$ for $k \in \mathbb{Z}$ span a space that is *shift-invariant* wrt. integer multiples of h. Extensions [1, 2] allow multiple generators and multivariate functions. Shift-invariant approximation uses only a single scale h, while *wavelets* use multiple scales and *refinable* generators.

Description

Nyquist–Shannon–Whittaker–Kotelnikov sampling provides the formula

$$f(x) = \sum_{k \in \mathbb{Z}} f(kh)\,\text{sinc}\left(\frac{x}{h} - k\right)$$

for *band-limited* functions with frequencies in $[-\pi/h, +\pi/h]$. It is basic in Electrical Engineering for AD/DA conversion of *signals* after *low-pass filtering*. Another simple example arises from the *hat function* or order 2 *B-spline* $B_2(x) := 1 - |x|$ for $-1 \le x \le 1$ and 0 elsewhere. Then the “connect-the-dots” formula

$$f(x) \approx \sum_{k \in \mathbb{Z}} f(kh)B_2\left(\frac{x}{h} - k\right)$$

is a piecewise linear approximation of f by connecting the values $f(kh)$ by straight lines. These two examples arise from a generator φ satisfying the *cardinal* interpolation conditions $\varphi(k) = \delta_{0k},\; k \in \mathbb{Z}$, and then the right-hand side of the above formulas interpolates f at all integers. If the generator is a higher-order B-spline B_m, the approximation

$$f(x) \approx \sum_{k \in \mathbb{Z}} f(kh)B_m\left(\frac{x}{h} - k\right)$$

goes back to I.J. Schoenberg and is not interpolatory in general.

So far, these examples of (1) have very special coefficients $c_{k,h}(f) = f(kh)$ arising from *sampling* the function f at data locations $h\mathbb{Z}$. This connects shift-invariant approximation to *sampling* theory. If the shifts of the generator are orthonormal in $L_2(\mathbb{R})$, the coefficients in (1) should be obtained instead as $c_{k,h}(f) = (f, \varphi(\frac{\cdot}{h} - k))_2$ for any $f \in L_2(\mathbb{R})$ to turn the approximation into an optimal L_2 projection. Surprisingly, these two approaches coincide for the sinc case.

Analysis of shift-invariant approximation focuses on the error in (1) for various generators φ and for different ways of calculating useful coefficients $c_{k,h}(f)$. Under special technical conditions, e.g., if the generator φ is compactly supported, the *Strang–Fix conditions* [4]

$$\hat{\varphi}^{(j)}(2\pi k) = \delta_{0k},\; k \in \mathbb{Z},\; 0 \le j < m$$

imply that the error of (1) is $\mathcal{O}(h^m)$ for $h \to 0$ in Sobolev space $W_2^m(\mathbb{R})$ if the coefficients are given via L_2 projection. This holds for B-spline generators of order m.

The basic tool for analysis of shift-invariant L_2 approximation is the *bracket product*

$$[\varphi, \psi](\omega) := \sum_{k \in \mathbb{Z}} \hat{\varphi}(\omega + 2k\pi)\overline{\hat{\psi}(\omega + 2k\pi)},\; \omega \in \mathbb{R}$$

which is a 2π-periodic function. It should exist pointwise, be in $L_2[-\pi, \pi]$ and satisfy a *stability property*

$$0 < A \le [\varphi, \varphi](\omega) \le B,\; \omega \in \mathbb{R}.$$

Then the L_2 projector for $h = 1$ has the convenient Fourier transform

$$\hat{S}_{f,1,\varphi}(\omega) = \frac{[f, \varphi](\omega)}{[\varphi, \varphi](\omega)}\hat{\varphi}(\omega),\; \omega \in \mathbb{R},$$

and if $[\varphi, \varphi](\omega) = 1/2\pi$ for all ω, the integer shifts $\varphi(\cdot - k)$ for $k \in \mathbb{Z}$ are orthonormal in $L_2(\mathbb{R})$.

Fundamental results on shift-invariant approximation are in [1, 2], and the survey [3] gives a comprehensive account of the theory and the historical background.

References

1. de Boor, C., DeVore, R., Ron, A.: Approximation from shift–invariant subspaces of $L_2(R^d)$. Trans. Am. Math. Soc. **341**, 787–806 (1994)
2. de Boor, C., DeVore, R., Ron, A.: The structure of finitely generated shift–invariant spaces in $L_2(R^d)$. J. Funct. Anal. **19**, 37–78 (1994)

3. Jetter, K., Plonka, G.: A survey on L_2-approximation orders from shift-invariant spaces. In: Multivariate approximation and applications, pp. 73–111. Cambridge University Press, Cambridge (2001)
4. Strang, G., Fix, G.: A Fourier analysis of the finite element variational method. In: Geymonat, G. (ed.) Constructive Aspects of Functional Analysis. C.I.M.E. II Ciclo 1971, pp 793–840 (1973)

Simulation of Stochastic Differential Equations

Denis Talay
INRIA Sophia Antipolis, Valbonne, France

Synonyms

Numerical mathematics; Stochastic analysis; Stochastic numerics

Definition

The development and the mathematical analysis of stochastic numerical methods to obtain approximate solutions of deterministic linear and nonlinear partial differential equations and to simulate stochastic models.

Overview

Owing to powerful computers, one now desires to model and simulate more and more complex physical, chemical, biological, and economic phenomena at various scales. In this context, stochastic models are intensively used because calibration errors cannot be avoided, physical laws are imperfectly known (as in turbulent fluid mechanics), or no physical law exists (as in finance). One then needs to compute moments or more complex statistics of the probability distributions of the stochastic processes involved in the models. A stochastic process is a collection (X_t) of random variables indexed by the time variable t.

This is not the only motivation to develop stochastic simulations. As solutions of a wide family of complex deterministic partial differential equations (PDEs) can be represented as expectations of functionals of stochastic processes, stochastic numerical methods are derived from these representations.

We can distinguish several classes of stochastic numerical methods: Monte Carlo methods consist in simulating large numbers of independent paths of a given stochastic process; stochastic particle methods consist in simulating paths of interacting particles whose empirical distribution converges in law to a deterministic measure; and ergodic methods consist in simulating one single path of a given stochastic process up to a large time horizon. Monte Carlo methods allow one to approximate statistics of probability distributions of stochastic models or solutions to linear partial differential equations. Stochastic particle methods approximate solutions to deterministic nonlinear McKean–Vlasov PDEs. Ergodic methods aim to compute statistics of equilibrium measures of stochastic models or to solve elliptic PDEs. See, e.g., [2].

In all cases, one needs to develop numerical approximation methods for paths of stochastic processes. Most of the stochastic processes used as models or involved in stochastic representations of PDEs are obtained as solutions to stochastic differential equations

$$X_t(x) = x + \int_0^t b(X_s(x))\,ds \;+\; \int_0^t \sigma(X_s(x))\,dW_s, \tag{1}$$

where (W_t) is a standard Brownian motion or, more generally, a Lévy process. Existence and uniqueness of solutions, in strong and weak senses, are exhaustively studied, e.g., in [10].

Monte Carlo Methods for Linear PDEs

Set $a(x) := \sigma(x)\,\sigma(x)^t$, and consider the parabolic PDE

$$\frac{\partial u}{\partial t}(t,x) = \sum_{i=1}^{d} b^i(x)\,\partial_i u(x) + \frac{1}{2}\sum_{i,j=1}^{d} a^i_j(x)\,\partial_{ij} u(x) \tag{2}$$

with initial condition $u(0,x) = f(x)$. Under various hypotheses on the coefficients b and σ, it holds that $u(t,x) = \mathbb{E}u_0(t, X_t(x))$, where $X_t(x)$ is the solution to (1).

Let $h > 0$ be a time discretization step. Let (G_p) be independent centered Gaussian vectors with unit covariance matrix. Define the Euler scheme by $\bar{X}_0^h(x) = x$ and the recursive relation

$$\begin{aligned}\bar{X}^h_{(p+1)h}(x) =&\bar{X}^h_{ph}(x) + b\bar{X}^h_{ph}(x)\ h\\ &+\sigma(\bar{X}^h_{ph}(x))\ \sqrt{h}\ G_{p+1}.\end{aligned}$$

The simulation of this random sequence only requires the simulation of independent Gaussian random variables. Given a time horizon Mh, independent copies of the sequence $(G_p, 1 \leq p \leq M)$ provide independent paths $(\bar{X}^{h,k}_{ph}(x), 1 \leq p \leq M)$.

The global error of the Monte Carlo method with N simulations which approximates $u(ph, x)$ is

$$\begin{aligned}&\mathbb{E}u_0(X_{Mh}) - \frac{1}{N}\sum_{k=1}^{N} u_0\left(\bar{X}^{h,k}_{Mh}\right)\\ &= \underbrace{\mathbb{E}u_0(X_{Mh}) - \mathbb{E}u_0\left(\bar{X}^h_{Mh}\right)}_{=:\epsilon_d(h)}\\ &\quad \underbrace{+\,\mathbb{E}u_0\left(\bar{X}^h_{Mh}\right) - \frac{1}{N}\sum_{k=1}^{N} u_0\left(\bar{X}^{h,k}_{Mh}\right)}_{=:\epsilon_s(h,N)}.\end{aligned}$$

Nonasymptotic variants of the central limit theorem imply that the statistical error $\epsilon_s(h)$ satisfies

$$\forall M \geq 1,\ \exists C(M) > 0,\ \ \mathbb{E}|\epsilon_s(h)| \leq \frac{C(M)}{\sqrt{N}} \ \text{ for all } \ 0 < h < 1.$$

Using estimates on the solution to the PDE (2) obtained by PDE analysis or stochastic analysis (stochastic flows theory, Malliavin calculus), one can prove that the discretization error $e_d(h)$ satisfies the so-called Talay–Tubaro expansion

$$e_d(h) = C(T, x)\ h + Q_h(f, T, x)\ h^2,$$

where $|C(T, x)| + \sup_h |Q_h(u_0, T, x)|$ depend on b, σ, u_0, and T. Therefore, Romberg extrapolation techniques can be used:

$$\mathbb{E}\left\{\frac{2}{N}\sum_{k=1}^{N} u_0\left(\bar{X}^{h/2,k}_{Mh}\right) - \frac{1}{N}\sum_{k=1}^{N} u_0\left(\bar{X}^{h,k}_{Mh}\right)\right\} = \mathcal{O}(h^2).$$

For surveys of results in this direction and various extensions, see [8, 14, 17].

The preceding statistical and discretization error estimates have many applications: computations of European option prices, moments of solutions to mechanical systems with random excitations, etc.

When the PDE (2) is posed in a domain D with Dirichlet boundary conditions $u(t, x) = g(x)$ on ∂D, then $u(t, x) = \mathbb{E}f(X_t(x))\ \mathbb{I}_{t<\tau} + \mathbb{E}g(X_\tau(x))\ \mathbb{I}_{t\geq\tau}$, where τ is the first hitting time of ∂D by $(X_t(x))$. An approximation method is obtained by substituting $\bar{X}^h_{ph\wedge\bar{\tau}^h}(x)$ to $X_\tau(x)$, where $\bar{\tau}^h$ is the first hitting time of ∂D by the interpolated Euler scheme. For a convergence rate analysis, see, e.g., [7].

Let $n(x)$ denote the unit inward normal vector at point x on ∂D. When one adds Neumann boundary conditions $\nabla u(t, x) \cdot n(x) = 0$ on ∂D to (2), then $u(t, x) = \mathbb{E}f(X^\sharp_t(x))$, where $X^\sharp$:= is the solution to an SDE with reflection

$$\begin{aligned}X^\sharp_t(x) =& x + \int_0^t b(X^\sharp_s(x))\ ds + \int_0^t \sigma(X^\sharp_s(x))\ dW_s\\ &+ \int_0^{tn} (X_s) dL^\sharp_s(X),\end{aligned}$$

where $(L_t(X))$ is the local time of X at the boundary. Then one constructs the reflected Euler scheme in such a way that the simulation of the local time, which would be complex and numerically instable, is avoided. This construction and the corresponding error analysis have been developed in [4].

Local times also appear in SDEs related to PDEs with transmission conditions along the discontinuity manifolds of the coefficient $a(x)$ as in the Poisson–Boltzmann equation in molecular dynamics, Darcy law in fluid mechanics, etc. Specific numerical methods and error analyses were recently developed: see, e.g., [5].

Elliptic PDEs are interpreted by means of solutions to SDEs integrated from time 0 up to infinity or their equilibrium measures. Implicit Euler schemes often are necessary to get stability: see [12]. An alternative efficient methods are those with decreasing stepsizes introduced in [11].

Stochastic Particle Methods for Nonlinear PDEs

Consider the following stochastic particle system. The dynamics of the ith particle is as follows: given N independent Brownian motions $(W^{(i)}_t)$, multidimensional coefficients B and S, and McKean interaction kernels b and σ, the positions $X_t^{(i)}$ solve the stochastic differential system

$$dX_t^{(i)} = B\left(t, X_t^{(i)}, \tfrac{1}{N}\sum_{j=1}^{N} b\left(X_t^{(i)}, X_t^{(j)}\right)\right) dt \\ + S\left(t, X_t^{(i)}, \tfrac{1}{N}\sum_{j=1}^{N} \sigma\left(X_t^{(i)}, X_t^{(j)}\right)\right) dW_t^{(i)}. \tag{3}$$

Note that the processes $X_t^{(i)}$ are not independent. However, the propagation of chaos and nonlinear martingale problems theories developed in a seminal way by McKean and Sznitman allow one to prove that the probability distribution of the particles empirical measure process converges weakly when N goes to infinity. The limit distribution is concentrated at the probability law of the process (X_t) solution to the following stochastic differential equation which is nonlinear in McKean's sense (its coefficients depend on the probability distribution of the solution):

$$\begin{cases} dX_t = B(t, X_t, \int b(X_t, y)\nu_t(dy))dt + S(t, X_t, \int \sigma(X_t, y)\nu_t(dy))dW_t, \\ \nu_t(dy) := \text{probability distribution of } X_t. \end{cases} \tag{4}$$

In addition, the flow of the probability distributions ν_t solves the nonlinear McKean–Vlasov–Fokker–Planck equation

$$\frac{d}{dt}\nu_t = L_{\nu_t}^* \nu_t, \tag{5}$$

where, A denoting the matrix $S \cdot S^t$, L_ν^* is the formal adjoint of the differential operator

$$L_\nu := \sum_k B_k(t, x, \int b(x, y)\nu(dy))\partial_k \\ + \tfrac{1}{2}\sum_{j,k} A_{jk}(t, x, \int \sigma(x, y)\nu(dy))\partial_{jk}. \tag{6}$$

From an analytical point of view, the SDEs (4) provide probabilistic interpretations for macroscopic equations which includes, e.g., smoothened versions of the Navier–Stokes and Boltzmann equations: see, e.g., the survey [16] and [13].

From a numerical point of view, whereas the time discretization of (X_t) does not lead to an algorithm since ν_t is unknown, the Euler scheme for the particle system $\{(X_t^{(i)}), i = 1, \ldots, N\}$ can be simulated: the solution ν_t to (5) is approximated by the empirical distribution of the simulated particles at time t, the number N of the particles being chosen large enough. Compared to the numerical resolution of the McKean–Vlasov–Fokker–Planck equation by deterministic methods, this stochastic numerical approach is numerically relevant in the cases of small viscosities. It is also intensively used, for example, in Lagrangian stochastic simulations of complex flows and in molecular dynamics: see, e.g., [9, 15]. When the functions B, S, b, σ are smooth, optimal convergence rates have been obtained for finite time horizons, e.g., in [1, 3].

Other stochastic representations have been developed for backward SDEs related to quasi-linear PDEs and variational inequalities. See the survey [6].

References

1. Antonelli, F., Kohatsu-Higa, A.: Rate of convergence of a particle method to the solution of the McKean-Vlasov equation. Ann. Appl. Probab. **12**, 423–476 (2002)
2. Asmussen, S., Glynn, P.W.: Stochastic Simulation: Algorithms and Analysis. Stochastic Modelling and Applied Probability Series, vol. 57. Springer, New York (2007)
3. Bossy, M.: Optimal rate of convergence of a stochastic particle method to solutions of 1D viscous scalar conservation laws. Math. Comput. **73**(246), 777–812 (2004)
4. Bossy, M., Gobet, É., Talay, D.: A symmetrized Euler scheme for an efficient approximation of reflected diffusions. J. Appl. Probab. **41**(3), 877–889 (2004)
5. Bossy, M., Champagnat, N., Maire, S., Talay D.: Probabilistic interpretation and random walk on spheres algorithms for the Poisson-Boltzmann equation in Molecular Dynamics. ESAIM:M2AN **44**(5), 997–1048 (2010)
6. Bouchard, B., Elie, R., Touzi, N.: Discrete-time approximation of BSDEs and probabilistic schemes for fully

nonlinear PDEs. In: Advanced Financial Modelling. Radon Series on Computational and Applied Mathematics, vol. 8, pp. 91–124. Walter de Gruyter, Berlin (2008)

7. Gobet, E., Menozzi, S.: Stopped diffusion processes: boundary corrections and overshoot. Stochastic Process. Appl. **120**(2), 130–162 (2010)
8. Graham, C., Talay, D.: Mathematical Foundations of Stochastic Simulations. Vol. I: Stochastic Simulations and Monte Carlo Methods. Stochastic Modelling and Applied Probability Series, vol. 68. Springer, Berlin/New York (2013, in press)
9. Jourdain, B., Lelièvre, T., Roux, R.: Existence, uniqueness and convergence of a particle approximation for the Adaptive Biasing Force process. M2AN **44**, 831–865 (2010)
10. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Graduate Texts in Mathematics, vol. 113. Springer, New York (1991)
11. Lamberton, D., Pagès, G.: Recursive computation of the invariant distribution of a diffusion. Bernoulli **8**(23), 367–405 (2002)
12. Mattingly, J., Stuart, A., Tretyakov, M.V.: Convergence of numerical time-averaging and stationary measures via Poisson equations. SIAM J. Numer. Anal. **48**(2), 552–577 (2010)
13. Méléard, S.: A trajectorial proof of the vortex method for the two-dimensional Navier-Stokes equation. Ann. Appl. Probab. **10**(4), 1197–1211 (2000)
14. Milstein, G.N., Tretyakov, M.V.: Stochastic Numerics for Mathematical Physics. Springer, Berlin/New York (2004)
15. Pope, S.B.: Turbulent Flows. Cambridge University Press, Cambridge/New York (2003)
16. Sznitman, A.-S.: Topics in propagation of chaos. In: École d'Été de Probabilités de Saint-Flour XIX-1989. Lecture Notes Mathematics, vol. 1464. Springer, Berlin/New York (1991)
17. Talay, D.: Probabilistic numerical methods for partial differential equations: elements of analysis. In: Talay, D., Tubaro, L. (eds.) Probabilistic Models for Nonlinear Partial Differential Equations. Lecture Notes in Mathematics, vol. 1627, pp. 48–196, Springer, Berlin/New York (1996)

Singular Perturbation Problems

Robert O'Malley
Department of Applied Mathematics,
University of Washington, Seattle, WA, USA

Regular perturbation methods often succeed in providing approximate solutions to problems involving a small parameter ϵ by simply seeking solutions as a formal power series (or even a polynomial) in ϵ. When the regular perturbation approach fails to provide a uniformly valid approximation, one encounters a *singular* perturbation problem (using the nomenclature of Friedrichs and Wasow [4], now universal).

The prototype singular perturbation problem occurred as Prandtl's *boundary layer* theory of 1904, concerning the flow of a fluid of small viscosity past an object (cf. [15, 20]). Applications have continued to motivate the subject, which holds independent mathematical interest involving differential equations. Prandtl's Göttingen lectures from 1931 to 1932 considered the model

$$\epsilon y'' + y' + y = 0$$

on $0 \leq x \leq 1$ with prescribed endvalues $y(0)$ and $y(1)$ for a small *positive* parameter ϵ (corresponding physically to a large Reynolds number). Linearly independent solutions of the differential equation are given by

$$e^{-\sigma(\epsilon)x} \text{ and } e^{-\kappa(\epsilon)x/\epsilon}$$

where $\sigma(\epsilon) \equiv \frac{1-\sqrt{1-4\epsilon}}{2\epsilon} = 1 + O(\epsilon)$ and $\kappa(\epsilon) \equiv \frac{1+\sqrt{1-4\epsilon}}{2} = 1 - \epsilon + O(\epsilon^2)$ as $\epsilon \to 0$. Setting

$$y(x,\epsilon) = \alpha e^{-\sigma(\epsilon)x} + \beta e^{-\kappa(\epsilon)x/\epsilon},$$

we will need $y(0) = \alpha + \beta$ and $y(1) = \alpha e^{-\sigma(\epsilon)} + \beta e^{-\kappa(\epsilon)/\epsilon}$. The large decay constant κ/ϵ implies that $\alpha \sim y(1)e^{\sigma(\epsilon)}$, so

$$y(x,\epsilon) \sim e^{\sigma(\epsilon)(1-x)}y(1) + e^{-\kappa(\epsilon)x/\epsilon}(y(0) - e^{\sigma(\epsilon)}y(1))$$

and

$$y(x,\epsilon) = e^{(1-x)}y(1) + e^{-x/\epsilon}e^{x}(y(0) - ey(1)) + O(\epsilon).$$

The second term decays rapidly from $y(0) - ey(1)$ to zero in an $O(\epsilon)$-thick *initial layer* near $x = 0$, so the limiting solution

$$e^{1-x}y(1)$$

for $x > 0$ satisfies the *reduced* problem

$$Y_0' + Y_0 = 0 \text{ with } Y_0(1) = y(1).$$

Convergence of $y(x,\epsilon)$ at $x = 0$ is *nonuniform* unless $y(0) = ey(1)$. Indeed, to all orders ϵ^j, the *asymptotic* solution for $x > 0$ is given by the *outer expansion*

$$Y(x,\epsilon) = e^{\sigma(\epsilon)(1-x)}y(1) \sim \sum_{j\geq 0} Y_j(x)\epsilon^j$$

(see Olver [12] for the definition of an asymptotic expansion). It is supplemented by an *initial* (boundary) *layer correction*

$$\xi\left(\frac{x}{\epsilon},\epsilon\right) \equiv e^{-\kappa(\epsilon)x/\epsilon}(y(0) - e^{\sigma(\epsilon)}y(1))$$
$$\sim \sum_{j\geq 0} \xi_j\left(\frac{x}{\epsilon}\right)\epsilon^j$$

where terms ξ_j all decay exponentially to zero as the stretched *inner variable* x/ϵ tends to infinity.

Traditionally, one learns to complement regular outer expansions by local inner expansions in regions of nonuniform convergence. *Asymptotic matching* methods, generalizing Prandtl's fluid dynamical insights involving inner and outer approximations, then provide higher-order asymptotic solutions (cf. Van Dyke [20], Lagerstrom [11], and I'lin [7], noting that O'Malley [13] and Vasil'eva et al. [21] provide more efficient direct techniques involving boundary layer corrections). The Soviet A.N. Tikhonov and American Norman Levinson independently provided methods, in about 1950, to solve initial value problems for the slow-fast vector system

$$\begin{cases} \dot{x} = f(x,y,t,\epsilon) \\ \epsilon\dot{y} = g(x,y,t,\epsilon) \end{cases}$$

on $t \geq 0$ subject to initial values $x(0)$ and $y(0)$. As we might expect, the outer limit $\begin{pmatrix} X_0(t) \\ Y_0(t) \end{pmatrix}$ should satisfy the reduced (differential-algebraic) system

$$\begin{cases} \dot{X}_0 = f(X_0,Y_0,t,0), \; X_0(0) = x(0) \\ 0 = g(X_0,Y_0,t,0) \end{cases}$$

for an attracting root

$$Y_0 = \phi(X_0,t)$$

of $g = 0$, along which g_y remains a stable matrix. We must expect nonuniform convergence of the fast variable y near $t = 0$, unless $y(0) = Y_0(0)$. Indeed, Tikhonov and Levinson showed that

$$\begin{cases} x(t,\epsilon) = X_0(t) + O(\epsilon) \text{ and} \\ y(t,\epsilon) = Y_0(t) + \eta_0(t/\epsilon) + O(\epsilon) \end{cases}$$

(at least for t finite) where $\eta_0(\tau)$ is the *asymptotically stable* solution of the stretched problem

$$\frac{d\eta_0}{d\tau} = g(x(0),\eta_0 + Y_0(0),0,0),\; \eta_0(0) = y(0) - Y_0(0).$$

The theory supports practical numerical methods for integrating *stiff* differential equations (cf. [5]). A more inclusive geometric theory, using normally hyperbolic invariant manifolds, has more recently been extensively used (cf. [3, 9]).

For many linear problems, classical analysis (cf. [2, 6, 23]) suffices. *Multiscale* methods (cf. [8, 10]), however, apply more generally. Consider, for example, the two-point problem

$$\epsilon y'' + a(x)y' + b(x,y) = 0$$

on $0 \leq x \leq 1$ when $a(x) > 0$ and a and b are smooth. We will seek the solution as $\epsilon \to 0^+$ when $y(0)$ and $y(1)$ are given in the form

$$y(x,\eta,\epsilon) \sim \sum_{j\geq 0} y_j(x,\eta)\epsilon^j$$

using the fast variable

$$\eta = \frac{1}{\epsilon}\int_0^x a(s)ds$$

to provide boundary layer behavior near $x = 0$. Because

$$y' = y_x + \frac{a(x)}{\epsilon}y_\eta$$

and

$$y'' = y_{xx} + \frac{2}{\epsilon}a(x)y_{x\eta} + \frac{a'(x)}{\epsilon}y_\eta + \frac{a^2(x)}{\epsilon^2}y_{\eta\eta},$$

the given equation is converted to the partial differential equation

$$a^2(x)\left(\frac{\partial^2 y}{\partial\eta^2} + \frac{\partial y}{\partial\eta}\right) + \epsilon\left(2a(x)\frac{\partial^2 y}{\partial x\partial\eta} + a'(x)\frac{\partial y}{\partial\eta}\right.$$
$$\left. + a(x)\frac{\partial y}{\partial x} + b(x,y)\right) + \epsilon^2 y_{xx} = 0.$$

We naturally ask the leading term y_0 to satisfy

$$\frac{\partial^2 y_0}{\partial \eta^2} + \frac{\partial y_0}{\partial \eta} = 0,$$

so y_0 has the form

$$y_0(x, \eta) = A_0(x) + B_0(x)e^{-\eta}.$$

The boundary conditions moreover require that

$$A_0(0) + B_0(0) = y(0) \text{ and } A_0(1) \sim y(1)$$

(since $e^{-\eta}$ is negligible when $x = 1$). From the ϵ coefficient, we find that y_1 must satisfy

$$\begin{aligned} a^2(x)\left(\frac{\partial^2 y_1}{\partial \eta^2} + \frac{\partial y_1}{\partial \eta}\right) &= -2a(x)\frac{\partial^2 y_0}{\partial x \partial \eta} - a'(x)\frac{\partial y_0}{\partial \eta} \\ &\quad -a(x)\frac{\partial y_0}{\partial x} - b(x, y_0) \\ &= -a(x)A_0' \\ &\quad + \left(a(x)B_0' + a'(x)B_0\right)e^{-\eta} \\ &\quad -b(x, A_0 + B_0 e^{-\eta}). \end{aligned}$$

Consider the right-hand side as a power series in $e^{-\eta}$. Undetermined coefficient arguments show that its first two terms (multiplying 1 and $e^{-\eta}$) will resonate with the solutions of the homogeneous equation to produce unbounded or *secular* solutions (multiples of η and $\eta e^{-\eta}$) as $\eta \to \infty$ unless we require that

1. A_0 satisfies the reduced problem

$$a(x)A_0' + b(x, A_0) = 0, \;\; A_0(1) = y(1)$$

(and continues to exist throughout $0 \le x \le 1$)

2. B_0 satisfies the linear problem

$$\begin{aligned} &a(x)B_0' + \left(-b_y(x, A_0) + a'(x)\right)B_0 = 0, \\ &B_0(0) = y(0) - A_0(0). \end{aligned}$$

Thus, we have completely obtained the limiting solution $Y_0(x, \eta)$. We note that the numerical solution of restricted two-point problems is reported in Roos et al. [18]. Special complications, possibly *shock* layers, must be expected at *turning points* where $a(x)$ vanishes (cf. [14,24]).

Related *two-time scale* methods have long been used in celestial mechanics (cf. [17,22]) to solve initial value problems for nearly linear oscillators

$$\ddot{y} + y = \epsilon f(y, \dot{y})$$

on $t \ge 0$. Regular perturbation methods suffice on bounded t intervals, but for $t = O(1/\epsilon)$ one must seek solutions

$$y(t, \tau, \epsilon) \sim \sum_{j \ge 0} y_j(t, \tau)\epsilon^j$$

using the *slow time*

$$\tau = \epsilon t.$$

We must expect a boundary layer (i.e., nonuniform convergence) at $t = \infty$ to account for the cumulative effect of the small perturbation ϵf.

Instead of using two-timing or averaging (cf. [1] or [19]), let us directly seek an asymptotic solution of the initial value problem for the Rayleigh equation

$$\ddot{y} + y = \epsilon \dot{y}\left(1 - \frac{1}{3}\dot{y}^2\right)$$

in the form

$$\begin{aligned} y(t, \tau, \epsilon) = \mathcal{A}(\tau, \epsilon)e^{it} + \epsilon\mathcal{B}(\tau, \epsilon)e^{3it} + \epsilon^2\mathcal{C}(\tau, \epsilon)e^{5it} \\ + \ldots + \text{complex conjugate} \end{aligned}$$

for undetermined slowly varying complex-valued coefficients $\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$ depending on ϵ (cf. [16]). Differentiating twice and separating the coefficients of the odd harmonics $e^{it}, e^{3it}, e^{5it}, \ldots$ in the differential equation, we obtain

$$\begin{aligned} &2i\frac{d\mathcal{A}}{d\tau} - i\mathcal{A}\left(1 - |\mathcal{A}|^2\right) + \epsilon\left(\frac{d^2\mathcal{A}}{d\tau^2} - \mathcal{A}^2\frac{d\mathcal{A}^*}{d\tau}\right. \\ &\quad \left. -\frac{d\mathcal{A}}{d\tau}\left(1 - 2|\mathcal{A}|^2\right) - 3i(\mathcal{A}^*)^2\mathcal{B}\right) + \ldots = 0, \\ &-8\mathcal{B} - \frac{i}{3}\mathcal{A}^3 + \ldots = 0, \text{ and} \\ &-24\mathcal{C} - 3i\mathcal{A}^2\mathcal{B} + \cdots = 0. \end{aligned}$$

The resulting initial value problem

$$\frac{d\mathcal{A}}{d\tau} = \frac{\mathcal{A}}{2}\left((1-|\mathcal{A}|^2) + \frac{i\epsilon}{8}(|\mathcal{A}|^4 - 2) + \dots\right)$$

for the amplitude $\mathcal{A}(\tau,\epsilon)$ can be readily solved for finite τ by using polar coordinates and regular perturbation methods on all equations. Thus, we obtain the asymptotic solution

$$\begin{aligned} y(t,\tau,\epsilon) = {} & \mathcal{A}(\tau,\epsilon)e^{it} - \frac{i\epsilon}{24}\mathcal{A}^3(\tau,\epsilon)e^{3it} \\ & + \frac{\epsilon}{64}\left(\mathcal{A}^3(\tau,\epsilon)(3|\mathcal{A}(\tau,\epsilon)|^2 - 2)e^{3it}\right. \\ & - \left.\frac{\mathcal{A}^5(\tau,\epsilon)}{3}e^{5it}\right) + \dots + \text{complex} \\ & \text{conjugate} \end{aligned}$$

there. We note that the oscillations for related coupled van der Pol equations are of special current interest in neuroscience.

The reader who consults the literature cited will find that singular perturbations continue to provide asymptotic solutions to a broad variety of differential equations from applied mathematics. The underlying mathematics is also extensive and increasingly sophisticated.

References

1. Bogoliubov, N.N., Mitropolski, Y.A.: Asymptotic Methods in the Theory of Nonlinear Oscillations. Gordon and Breach, New York (1961)
2. Fedoryuk, M.V.: Asymptotic Analysis. Springer, New York (1994)
3. Fenichel, N.: Geometic singular perturbation theory for ordinary differential equations. J. Differ. Equ. **15**, 77–105 (1979)
4. Friedrichs, K.-O., Wasow, W.: Singular perturbations of nonlinear oscillations. Duke Math. J. **13**, 367–381 (1946)
5. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, 2nd revised edn. Springer, Berlin (1996)
6. Hsieh, P.-F., Sibuya Y.: Basic Theory of Ordinary Differential Equations. Springer, New York (1999)
7. Il'in, A.M.: Matching of Asymptotic Expansions of Solutions of Boundary Value Problems. American Mathematical Society, Providence (1992)
8. Johnson, R.S.: Singular Perturbation Theory, Mathematical and Analytical Techniques with Applications to Engineering. Springer, New York (2005)
9. Kaper, T.J.: An introduction to geometric methods and dynamical systems theory for singular perturbation problems. In: Cronin, J., O'Malley, R.E. (eds.) Analyzing Multiscale Phenomena Using Singular Perturbation Methods, pp. 85–131. American Mathematical Society, Providence (1999)
10. Kevorkian, J., Cole J.D.: Multiple Scale and Singular Perturbation Methods. Springer, New York (1996)
11. Lagerstrom, P.A.: Matched Asymptotic Expansions. Springer, New York (1988)
12. Olver, F.W.J.: Asymptotics and Special Functions. Academic, New York (1974)
13. O'Malley, R.E., Jr.: Singular Perturbation Methods for Ordinary Differential Equations. Springer, New York (1991)
14. O'Malley, R.E., Jr.: Singularly perturbed linear two-point boundary value problems. SIAM Rev. **50**, 459–482 (2008)
15. O'Malley, R.E., Jr.: Singular perturbation theory: a viscous flow out of Göttingen. Annu. Rev. Fluid Mech. **42**, 1–17 (2010)
16. O'Malley, R.E., Jr., Kirkinis, E.: A combined renormalization group-multiple scale method for singularly perturbed problems. Stud. Appl. Math. **124**, 383–410 (2010)
17. Poincaré, H.: Les Methodes Nouvelles de la Mecanique Celeste II. Gauthier-Villars, Paris (1893)
18. Roos, H.-G., Stynes, M., Tobiska L.: Robust Numerical Methods for Singularly Perturbed Differential Equations, 2nd edn. Springer, Berlin (2008)
19. Sanders, J.A., Verhulst, F., Murdock, J.: Averaging Methods in Nonlinear Dynamical Systems, 2nd edn. Springer, New York (2007)
20. Van Dyke, M.: Perturbation Methods in Fluid Dynamics. Academic, New York (1964)
21. Vasil'eva, A.B., Butuzov, V.F., Kalachev L.V.: The Boundary Function Method for Singular Perturbation Problems. SIAM, Philadelphia (1995)
22. Verhulst, F.: Nonlinear Differential Equations and Dynamical Systems, 2nd edn. Springer, New York (2000)
23. Wasow, W.: Asymptotic Expansions for Ordinary Differential Equations. Wiley, New York (1965)
24. Wasow, W.: Linear Turning Point Theory. Springer, New York (1985)

Solid State Physics, Berry Phases and Related Issues

Gianluca Panati
Dipartimento di Matematica, Universit di Roma “La Sapienza”, Rome, Italy

Synonyms

Dynamics of Bloch electrons; Theory of Bloch bands; Semiclassical model of solid-state physics

Definition/Abstract

Crystalline solids are solids in which the ionic cores of the atoms are arranged periodically. The dynamics of a test electron in a crystalline solid can be conveniently analyzed by using the *Bloch-Floquet transform*, while the localization properties of electrons are better described by using *Wannier functions*. The latter can also be obtained by minimizing a suitable localization functional, yielding a convenient numerical algorithm.

Macroscopic transport properties of electrons in crystalline solids are derived, by using adiabatic theory, from the analysis of a perturbed Hamiltonian, which includes the effect of external macroscopic or slowly varying electromagnetic potentials. The geometric *Berry phase* and its curvature play a prominent role in the corresponding effective dynamics.

The Periodic Hamiltonian

In a crystalline solid, the ionic cores are arranged periodically, according to a periodicity lattice $\Gamma = \left\{\gamma \in \mathbb{R}^d : \gamma = \sum_{j=1}^d n_j \gamma_j \text{ for some } n_j \in \mathbb{Z}\right\} \simeq \mathbb{Z}^d$, where $\{\gamma_1, \dots, \gamma_d\}$ are fixed linearly independent vectors in $\mathbb{R}^d$.

The dynamics of a test electron in the potential generated by the ionic cores of the solid and, in a mean-field approximation, by the remaining electrons is described by the Schrödinger equation $\mathrm{i}\partial_t \psi = H_{\text{per}}\psi$, where the Hamiltonian operator reads (in Rydberg units)

$$H_{\text{per}} = -\Delta + V_\Gamma(x) \quad \text{acting in } L^2(\mathbb{R}^d). \qquad (1)$$

Here, $\Delta = \nabla^2$ is the Laplace operator and the function $V_\Gamma : \mathbb{R}^d \to \mathbb{R}$ is periodic with respect to Γ, i.e., $V_\Gamma(x+\gamma) = V_\Gamma(x)$ for all $\gamma \in \Gamma$, $x \in \mathbb{R}^d$. A mathematical justification of such a model in the reduced Hartree-Fock approximation was obtained in Catto et al. [3] and Cancès et al. [4], see ▸ Mathematical Theory for Quantum Crystals and references therein.

To assure that H_{per} is self-adjoint in $L^2(\mathbb{R}^d)$ on the Sobolev space $W^{2,2}(\mathbb{R}^d)$, we make an usual Kato-type assumption on the Γ-periodic potential:

$$\begin{aligned} &V_\Gamma \in L^2_{\text{loc}}(\mathbb{R}^d) \text{ for } d \le 3, \\ &V_\Gamma \in L^p_{\text{loc}}(\mathbb{R}^d) \text{ with } p > d/2 \text{ for } d \ge 4. \end{aligned} \qquad (2)$$

Clearly, the case of a potential with Coulomb-like singularities is included.

The Bloch–Floquet Transform (Bloch Representation)

Since H_{per} commutes with the lattice translations, it can be decomposed as a direct integral of simpler operators by the (modified) Bloch–Floquet transform. Preliminarily, we define the dual lattice as $\Gamma^* := \left\{k \in \mathbb{R}^d : k \cdot \gamma \in 2\pi\mathbb{Z} \text{ for all } \gamma \in \Gamma\right\}$. We denote by Y (resp. Y^*) the centered fundamental domain of Γ (resp. Γ^*), namely,

$$Y^* = \left\{k \in \mathbb{R}^d : k = \sum\nolimits_{j=1}^d k'_j \gamma_j^* \text{ for } k'_j \in \left[-\frac{1}{2}, \frac{1}{2}\right]\right\},$$

where $\{\gamma_j^*\}$ is the dual basis to $\left\{\gamma_j\right\}$, i.e., $\gamma_j^* \cdot \gamma_i = 2\pi\delta_{j,i}$. When the opposite faces of Y^* are identified, one obtains the torus $\mathbb{T}_d^* := \mathbb{R}^d/\Gamma^*$.

One defines, initially for $\psi \in C_0(\mathbb{R}^d)$, the modified *Bloch–Floquet transform* as

$$(\tilde{\mathcal{U}}_{\text{BF}}\psi)(k,y) := \frac{1}{|Y^*|^{\frac{1}{2}}} \sum_{\gamma\in\Gamma} \mathrm{e}^{-\mathrm{i}k\cdot(y+\gamma)}\,\psi(y+\gamma),$$
$$y \in \mathbb{R}^d,\ k \in \mathbb{R}^d. \qquad (3)$$

For any fixed $k \in \mathbb{R}^d$, $\big(\tilde{\mathcal{U}}_{\text{BF}}\psi\big)(k,\cdot)$ is a Γ-periodic function and can thus be regarded as an element of $\mathcal{H}_{\text{f}} := L^2(\mathbb{T}_Y)$, $\mathbb{T}_Y$ being the flat torus $\mathbb{R}^d/\Gamma$. The map defined by (3) extends to a unitary operator $\tilde{\mathcal{U}}_{\text{BF}} : L^2(\mathbb{R}^d) \longrightarrow \int_{Y^*}^{\oplus} \mathcal{H}_{\text{f}}\, dk$, with inverse given by

$$\left(\tilde{\mathcal{U}}_{\text{BF}}^{-1}\varphi\right)(x) = \frac{1}{|Y^*|^{\frac{1}{2}}} \int_{Y^*} dk\ \mathrm{e}^{\mathrm{i}k\cdot x}\varphi(k,[x]),$$

where $[\,\cdot\,]$ refers to the decomposition $x = \gamma_x + [x]$, with $\gamma_x \in \Gamma$ and $[x] \in Y$.

The advantage of this construction is that the transformed Hamiltonian is a fibered operator over Y^*. Indeed, one checks that

$$\tilde{\mathcal{U}}_{\text{BF}}\, H_{\text{per}}\, \tilde{\mathcal{U}}_{\text{BF}}^{-1} = \int_{Y^*}^{\oplus} dk\ H_{\text{per}}(k)$$

with fiber operator

$$H_{\text{per}}(k) = \big(-\mathrm{i}\nabla_y + k\big)^2 + V_\Gamma(y)\,, \quad k \in \mathbb{R}^d, \qquad (4)$$

acting on the k-independent domain $W^{2,2}(\mathbb{T}_Y) \subset L^2(\mathbb{T}_Y)$. The latter fact explains why it is mathematically convenient to use the *modified* BF transform. Each fiber operator $H_{\text{per}}(k)$ is self-adjoint, has compact resolvent, and thus pure point spectrum accumulating at infinity. We label the eigenvalue increasingly, i.e., $E_0(k) \leq E_1(k) \leq E_2(k) \leq \ldots$. With this choice, they are Γ^*-periodic, i.e., $E_n(k+\lambda) = E_n(k)$ for all $\lambda \in \Gamma^*$. The function $k \mapsto E_n(k)$ is called the nth *Bloch band*.

For fixed $k \in Y^*$, one considers the eigenvalue problem

$$\begin{aligned} H_{\text{per}}(k)\, u_n(k,y) &= E_n(k)\, u_n(k,y), \\ \|u_n(k,\cdot)\|_{L^2(\mathbb{T}_Y)} &= 1. \end{aligned} \tag{5}$$

A solution to the previous eigenvalue equation (e.g., by numerical simulations) provides a complete solution to the dynamical equation induced by (1). Indeed, if the initial datum ψ_0 satisfies

$$(\tilde{\mathcal{U}}_{\text{BF}}\, \psi_0)(k,y) = \varphi(k)\, u_n(k,y) \text{ for some } \varphi \in L^2(Y^*),$$

(one says in jargon that "ψ_0 is concentrated on the nth band") then the solution $\psi(t)$ to the Schrödinger equation with initial datum ψ_0 is characterized by

$$(\tilde{\mathcal{U}}_{\text{BF}}\, \psi(t))(k,y) = \left(e^{-iE_n(k)t} \varphi(k)\right) u_n(k,y).$$

In particular, the solution is exactly concentrated on the nth band at any time. By linearity, one recovers the solution for any initial datum. Below, we will discuss to which extent this dynamical description survives when macroscopic perturbations of the operator (1) are considered.

Wannier Functions and Charge Localization

While the Bloch representation is a useful tool to deal with dynamical and energetic problems, it is not convenient to study the localization of electrons in solids. A related crucial problem is the construction of a basis of generalized eigenfunctions of the operator H_{per} which are exponentially localized in space. Indeed, such a basis allows to develop computational methods which scale linearly with the system size [6], makes possible the description of the dynamics by *tight-binding* effective Hamiltonians, and plays a prominent role in the modern theories of macroscopic polarization [9, 18] and of orbital magnetization [21].

A convenient system of localized generalized eigenfunctions has been proposed by Wannier [22]. By definition, a *Bloch function* corresponding to the nth Bloch band is any u satisfying (5). Clearly, if u is a Bloch function then $\tilde{u}$, defined by $\tilde{u}(k,y) = e^{i\vartheta(k)}\, u(k,y)$ for any Γ^*-periodic function ϑ, is also a Bloch function. The latter invariance is often called *Bloch gauge invariance*.

Definition 1 The *Wannier function* $w_n \in L^2(\mathbb{R}^d)$ corresponding to a Bloch function u_n for the Bloch band E_n is the preimage of u_n with respect to the Bloch-Floquet transform, namely

$$w_n(x) := \left(\tilde{\mathcal{U}}_{\text{BF}}^{-1} u_n\right)(x) = \frac{1}{|Y^*|^{\frac{1}{2}}} \int_{Y^*} dk\, e^{ik\cdot x} u_n(k,[x]).$$

The translated Wannier functions are

$$\begin{aligned} w_{n,\gamma}(x) &:= w_n(x-\gamma) \\ &= \frac{1}{|Y^*|^{\frac{1}{2}}} \int_{Y^*} dk\, e^{-ik\cdot\gamma}\, e^{ik\cdot x} u_n(k,[x]), \quad \gamma \in \Gamma. \end{aligned}$$

Thus, in view of the orthogonality of the trigonometric polynomials and the fact that $\tilde{\mathcal{U}}_{\text{BF}}$ is an isometry, the functions $\{w_{n,\gamma}\}_{\gamma\in\Gamma}$ are mutually orthogonal in $L^2(\mathbb{R}^d)$. Moreover, the family $\{w_{n,\gamma}\}_{\gamma\in\Gamma}$ is a complete orthonormal basis of $\tilde{\mathcal{U}}_{\text{BF}}^{-1}$ Ran P_*, where $P_*(k)$ is the spectral projection of $H_{\text{per}}(k)$ corresponding to the eigenvalue $E_n(k)$ and $P_* = \int_{Y^*}^{\oplus} P_*(k)\, dk$.

In view of the properties of the Bloch–Floquet transform, the existence of an exponentially localized Wannier function for the Bloch band E_n is equivalent to the existence of an analytic and Γ^*-pseudoperiodic Bloch function (recall that (3) implies that the Bloch function must satisfy $u(k+\lambda, y) = e^{-i\lambda\cdot y}\, u(k,y)$ for all $\lambda \in \Gamma^*$). A local argument assures that there is always a choice of the Bloch gauge such that the Bloch function is analytic around a given point. However, as several authors noticed [5, 13], there might be topological obstruction to obtain a global analytic Bloch function, in view of the competition between the analyticity and the pseudoperiodicity.

Hereafter, we denote by $\sigma_*(k)$ the set $\{E_i(k) : n \leq i \leq n+m-1\}$, corresponding to a physically relevant family of m Bloch bands, and we assume the following *gap condition*:

$$\inf_{k\in\mathbb{T}_d^*} \operatorname{dist}(\sigma_*(k), \sigma(H(k)) \setminus \sigma_*(k)) > 0. \qquad (6)$$

If a Bloch band E_n satisfies (6) for $m = 1$ we say that it is an single isolated Bloch band. For $m > 1$, we refer to a composite family of Bloch bands.

Single Isolated Bloch Band

In the case of a single isolated Bloch band, the problem of proving the existence of exponentially localized Wannier functions was raised in 1959 by W. Kohn [10], who solved it in dimension $d = 1$. In higher dimension, the problem has been solved, always in the case of a single isolated Bloch band, by J. des Cloizeaux [5] (under the nongeneric hypothesis that V_Γ has a center of inversion) and finally by G. Nenciu under general hypothesis [12], see also [8] for an alternative proof. Notice, however, that in real solids, it might happen that the interesting Bloch band (e.g., the conduction band in graphene) is not isolated from the rest of the spectrum and that $k \mapsto P_*(k)$ is not smooth at the degeneracy point. In such a case, the corresponding Wannier function decreases only polynomially.

Composite Family of Bloch Bands

It is well-known that, in dimension $d > 1$, the Bloch bands of crystalline solids are not, in general, isolated. Thus, the interesting problem, in view of real applications, concerns the case of composite families of bands, i.e., $m > 1$ in (6), and in this context, the more general notion of *composite Wannier functions* is relevant [1, 5]. Physically, condition (6) is always satisfied in semiconductors and insulators by considering the family of all the Bloch bands up to the Fermi energy.

Given a composite family of Bloch bands, we consider the orthogonal projector (in Dirac's notation)

$$P_*(k) := \sum_{i=n}^{n+m-1} |u_i(k)\rangle \langle u_i(k)| \,,$$

which is independent from the Bloch gauge, and we pose $P_* = \int_{Y^*}^{\oplus} P_*(k)\,dk$. A function χ is called a *quasi-Bloch function* if

$$P_*(k)\chi(k,\cdot) = \chi(k,\cdot) \ \text{ and } \ \chi(k,\cdot) \neq 0 \ \ \forall k \in Y^*. \qquad (7)$$

Although the terminology is not standard, we call *Bloch frame* a set $\{\chi_a\}_{a=1,\dots,m}$ of quasi-Bloch functions such that $\{\chi_1(k),\dots,\chi_m(k)\}$ is an orthonormal basis of Ran $P_*(k)$ at (almost-)every $k \in Y^*$. As in the previous case, there is a gauge ambiguity: a Bloch frame is fixed only up to a k-dependent unitary matrix $U(k) \in \mathcal{U}(m)$, i.e., if $\{\chi_a\}_{a=1,\dots,m}$ is a Bloch frame then the functions $\widetilde{\chi}_a(k) = \sum_{b=1}^m \chi_b(k) U_{b,a}(k)$ also define a Bloch frame.

Definition 2 The *composite Wannier functions* corresponding to a Bloch frame $\{\chi_a\}_{a=1,\dots,m}$ are the functions

$$w_a(x) := \left(\tilde{\mathcal{U}}_{\mathrm{BF}}^{-1}\chi_a\right)(x), \quad a \in \{1,\dots,m\}.$$

As in the case of a single Bloch band, the exponential localization of the composite Wannier functions is equivalent to the analyticity of the corresponding Bloch frame (which, in addition, must be Γ^*-pseudoperiodic). As before, there might be topological obstruction to the existence of such a Bloch frame. As far as the operator (1) is concerned, the existence of exponentially localized composite Wannier functions has been proved in Nenciu [12] in dimension $d = 1$; as for $d > 1$, the problem remained unsolved for more than two decades, until recently [2, 16]. Notice that for *magnetic* periodic Schrödinger operators the existence of exponentially localized Wannier functions is generically false.

The Marzari–Vanderbilt Localization Functional

To circumvent the long-standing controversy about the existence of exponentially localized composite Wannier functions, and in view of the application to numerical simulations, the solid-state physics community preferred to introduce the alternative notion of *maximally localized Wannier functions* [11]. The latter are defined as the minimizers of a suitable localization functional, known as the Marzari–Vanderbilt (MV) functional. For a single-band normalized Wannier function $w \in L^2(\mathbb{R}^d)$, the localization functional is

$$F_{MV}(w) = \int_{\mathbb{R}^d} |x|^2 |w(x)|^2 dx - \sum_{j=1}^{d} \left(\int_{\mathbb{R}^d} x_j |w(x)|^2 dx \right)^2, \qquad (8)$$

which is well defined at least whenever $\int_{\mathbb{R}^d} |x|^2 |w(x)|^2 dx < +\infty$. More generally, for a system of L^2-normalized composite Wannier functions $w = \{w_1, \dots, w_m\} \subset L^2(\mathbb{R}^d)$, the *Marzari–Vanderbilt localization functional* is

$$F_{MV}(w) = \sum_{a=1}^{m} F_{MV}(w_a) = \sum_{a=1}^{m} \int_{\mathbb{R}^d} |x|^2 |w_a(x)|^2 dx \\ - \sum_{a=1}^{m} \sum_{j=1}^{d} \left(\int_{\mathbb{R}^d} x_j |w_a(x)|^2 dx \right)^2. \qquad (9)$$

We emphasize that the above definition includes the crucial constraint that the corresponding Bloch functions $\varphi_a(k,\cdot) = (\tilde{\mathcal{U}}_{\mathrm{BF}}\, w_a)(k,\cdot)$, for $a \in \{1, \dots, m\}$, are a Bloch frame.

While such approach provided excellent results from the numerical viewpoint, the existence and exponential localization of the minimizers have been investigated only recently [17].

Dynamics in Macroscopic Electromagnetic Potentials

To model the transport properties of electrons in solids, one modifies the operator (1) to include the effect of the external electromagnetic potentials. Since the latter vary at the laboratory scale, it is natural to assume that the ratio ε between the lattice constant $a = |Y|^{1/d}$ and the length-scale of variation of the external potentials is small, i.e., $\varepsilon \ll 1$. The original problem is replaced by

$$\begin{aligned} &\mathrm{i}\varepsilon \partial_\tau \psi(\tau, x) \\ &= \left(\frac{1}{2} \left(-\mathrm{i}\nabla_x - A(\varepsilon x)\right)^2 + V_\Gamma(x) + V(\varepsilon x) \right) \psi(\tau, x) \\ &\equiv H_\varepsilon \psi(\tau, x) \end{aligned} \qquad (10)$$

where $\tau = \varepsilon t$ is the macroscopic time, and $V \in C_{\mathrm{b}}^\infty(\mathbb{R}^d, \mathbb{R})$ and $A_j \in C_{\mathrm{b}}^\infty(\mathbb{R}^d, \mathbb{R})$, $j \in \{1, \dots, d\}$ are respectively the external electrostatic and magnetic potential. Hereafter, for the sake of a simpler notation, we consider only $d = 3$.

While the dynamical equation (10) is quantum mechanical, physicists argued [1] that for suitable wavepackets, which are localized on the nth Bloch band and spread over many lattice spacings, the main effect of the periodic potential V_Γ is the modification of the relation between the momentum and the kinetic energy of the electron, from the free relation $E_{\mathrm{free}}(k) = \frac{1}{2}k^2$ to the function $k \mapsto E_n(k)$ given by the nth Bloch band. Therefore, the semiclassical equations of motion are

$$\begin{cases} \dot{r} = \nabla E_n(\kappa) \\ \dot{\kappa} = -\nabla V(r) + \dot{r} \times B(r) \end{cases} \qquad (11)$$

where $r \in \mathbb{R}^3$ is the macroscopic position of the electron, $\kappa = k - A(r)$ is the kinetic momentum with $k \in \mathbb{T}_d^*$ the Bloch momentum, $-\nabla V$ the external electric field and $B = \nabla \times A$ the external magnetic field.

In fact, one can derive also the first-order correction to (11). At this higher accuracy, the electron acquires an effective k-dependent electric moment $\mathcal{A}_n(k)$ and magnetic moment $\mathcal{M}_n(k)$. If the nth Bloch band is non-degenerate (hence isolated), the former is given by the *Berry connection*

$$\begin{aligned} \mathcal{A}_n(k) &= \mathrm{i}\, \langle u_n(k)\, ,\, \nabla_k u_n(k) \rangle_{\mathcal{H}_{\mathrm{f}}} \\ &= \mathrm{i} \int_Y u_n(k, y)^* \, \nabla_k u_n(k, y)\, dy, \end{aligned}$$

and the latter reads $\mathcal{M}_n(k) = \frac{\mathrm{i}}{2} \left\langle \nabla_k u_n(k), \times (H_{\mathrm{per}}(k) - E_n(k)) \nabla_k u_n(k) \right\rangle_{\mathcal{H}_{\mathrm{f}}}$, i.e., explicitly

$$\left[\mathcal{M}_n(k)\right]_i = \frac{\mathrm{i}}{2} \sum_{1 \le j, l \le 3} \epsilon_{ijl} \left\langle \partial_{k_j} u_n(k), (H_{\mathrm{per}}(k) \right. \\ \left. - E_n(k)) \partial_{k_l} u_n(k) \right\rangle_{\mathcal{H}_{\mathrm{f}}}$$

where ϵ_{ijl} is the totally antisymmetric symbol. The refined semiclassical equations read

$$\begin{cases} \dot{r} = \nabla_\kappa \left(E_n(\kappa) - \varepsilon B(r) \cdot \mathcal{M}_n(\kappa)\right) - \varepsilon \dot{\kappa} \times \Omega_n(\kappa) \\ \dot{\kappa} = -\nabla_r \left(V(r) - \varepsilon B(r) \cdot \mathcal{M}_n(\kappa)\right) + \dot{r} \times B(r) \end{cases} \qquad (12)$$

where $\Omega_n(k) = \nabla \times \mathcal{A}_n(k)$ corresponds to the curvature of the Berry connection. The previous equations have a hidden Hamiltonian structure [14]. Indeed, by introducing the semiclassical Hamiltonian function $H_{\mathrm{sc}}(r, \kappa) = E_n(\kappa) + V(r) - \varepsilon B(r) \cdot \mathcal{M}_n(\kappa)$, (12) become

$$\begin{pmatrix} \mathbb{B}(r) & -\mathbb{I} \\ \mathbb{I} & \varepsilon\, \mathbb{A}_n(\kappa) \end{pmatrix} \begin{pmatrix} \dot r \\ \dot\kappa \end{pmatrix} = \begin{pmatrix} \nabla_r H_{\text{sc}}(r,\kappa) \\ \nabla_\kappa H_{\text{sc}}(r,\kappa) \end{pmatrix} \tag{13}$$

where $\mathbb{I}$ is the identity matrix and $\mathbb{B}$ (resp. $\mathbb{A}_n$) is the 3×3 matrix corresponding to the vector field B (resp. Ω_n), i.e., $\mathbb{B}_{l,m}(r) = \sum_{1 \le j \le 3} \epsilon_{lmj} B_j(r) = (\partial_l A_m - \partial_m A_l)(r)$. Since the matrix appearing on the l.h.s corresponds to a symplectic form $\Theta_{B,\varepsilon}$ (i.e., a non-degenerate closed 2-form) on $\mathbb{R}^6$, (13) has Hamiltonian form with respect to $\Theta_{B,\varepsilon}$.

The mathematical derivation of the semiclassical model (12) from (10) as $\varepsilon \to 0$ has been accomplished in Panati et al. [14]. The first-order correction to the semiclassical (11) was previously investigated in Sundaram and Niu [19], but the heuristic derivation in the latter paper does not yield the term of order ε in the second equation. Without such a term, it is not clear if the equations have a Hamiltonian structure.

As for mathematically related problems, both the semiclassical asymptotic of the spectrum of H_ε and the corresponding scattering problem have been studied in detail (see [7] and references therein). The effective quantum Hamiltonians corresponding to (10) for $\varepsilon \to 0$ have also been deeply investigated [13].

The connection between (10) and (12) can be expressed either by an Egorov-type theorem involving quantum observables, or by using Wigner functions. Here we focus on the second approach.

First we define the Wigner function. We consider the space $\mathcal{C} = C_{\text{b}}^\infty(\mathbb{R}^{2d})$ equipped with the standard distance $d_{\mathcal{C}}$, and the subspace of Γ^*-periodic observables

$$\mathcal{C}_{\text{per}} = \{a \in \mathcal{C} : a(r, k + \lambda) = a(r,k) \ \forall\, \lambda \in \Gamma^*\}.$$

Recall that, according to the Calderon-Vaillancourt theorem, there is a constant C such that for $a \in \mathcal{C}$ its Weyl quantization $\widehat{a} \in \mathcal{B}(L^2(\mathbb{R}^3))$ satisfies

$$|\,\langle \psi, \widehat{a}\,\psi\rangle_{L^2(\mathbb{R}^3)}\,| \le C\, d_{\mathcal{C}}(a,0)\, \|\psi\|^2 .$$

Hence, the map $\mathcal{C} \ni a \mapsto \langle \psi, \widehat{a}\,\psi\rangle \in \mathbb{C}$ is linear continuous and thus defines an element W_ε^ψ of the dual space $\mathcal{C}'$, the Wigner function of ψ. Writing

$$\begin{aligned} \langle \psi, \widehat{a}\,\psi\rangle &=: \langle W_\varepsilon^\psi, a\rangle_{\mathcal{C}',\mathcal{C}} \\ &=: \int_{\mathbb{R}^{2d}} a(q,p)\, W_\varepsilon^\psi(q,p)\, dq\, dp \end{aligned}$$

and inserting the definition of the Weyl quantization for a one arrives at the formula

$$\begin{aligned} W_\varepsilon^\psi(q,p) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} d\xi\, \mathrm{e}^{\mathrm{i}\xi\cdot p}\, \psi^*(q + \varepsilon\xi/2) \\ \times \psi(q - \varepsilon\xi/2), \end{aligned} \tag{14}$$

which yields $W_\varepsilon^\psi \in L^2(\mathbb{R}^{2d})$. Although W_ε^ψ is real-valued, it attains also negative values in general, so it does not define a probability distribution on phase space.

After this preparation, we can vaguely state the link between (10) and (12), see [20] for the precise formulation. Let E_n be an isolated, nondegenerate Bloch band. Denote by $\overline{\Phi}_\varepsilon^\tau(r,k)$ the flow of the dynamical system (12) in canonical coordinates $(r,k) = (r, \kappa + A(r))$ (recall that the Weyl quantization, and hence the definition of Wigner function, is not invariant under non-linear changes of canonical coordinates). Then for each finite time-interval $I \subset \mathbb{R}$ there is a constant C such that for $\tau \in I$, $a \in \mathcal{C}_{\text{per}}$ and for ψ_0 "well-concentrated on the nth Bloch band" one has

$$\begin{aligned} &\left| \int_{\mathbb{R}^{2d}} a(q,p) \Big(W_\varepsilon^{\psi(\tau)}(q,p) - W_\varepsilon^{\psi_0} \circ \overline{\Phi}_\varepsilon^{-\tau}(q,p) \Big)\, dq\, dp \right| \\ &\quad \le \varepsilon^2\, C\, d_{\mathcal{C}}(a,0)\, \|\psi_0\|^2 , \end{aligned}$$

where $\psi(t)$ is the solution to the Schrödinger equation (10) with initial datum ψ_0.

Slowly Varying Deformations and Piezoelectricity

To investigate the contribution of the electrons to the macroscopic polarization and to the piezoelectric effect, it is crucial to know how the electrons move in a crystal which is strained at the macroscopic scale. Assuming the usual *fixed-lattice approximation*, the problem can be reduced to study the solutions to

$$\mathrm{i}\, \partial_t \psi(t,x) = \left(-\frac{1}{2}\Delta + V_\Gamma(x, \varepsilon t) \right) \psi(t,x) \tag{15}$$

for $\varepsilon \ll 1$, where $V_\Gamma(\cdot, t)$ is Γ-periodic for every $t \in \mathbb{R}$, i.e., the periodicity lattice does not depend on time. While a model with a fixed lattice might seem unrealistic at first glance, we refer to Resta [18] and

King-Smith and Vanderbilt [9] for its physical justification. The analysis of the Hamiltonian $H(t) = -\frac{1}{2}\Delta + V_{\Gamma}(x,t)$ yields a family of time-dependent Bloch functions $\{u_n(k,t)\}_{n\in\mathbb{N}}$ and Bloch bands $\{E_n(k,t)\}_{n\in\mathbb{N}}$.

Assuming that the relevant Bloch band is isolated from the rest of the spectrum, so that (6) holds true at every time, and that the initial datum is well-concentrated on the nth Bloch band, one obtains a semiclassical description of the dynamics analogous to (12). In this case, the semiclassical equations read

$$\begin{cases} \dot{r} = \nabla_k E_n(k,t) - \varepsilon\, \Theta_n(k,t) \\ \dot{k} = 0 \end{cases} \tag{16}$$

where

$$\Theta_n(k,t) = -\partial_t \mathcal{A}_n(k,t) - \nabla_k \phi_n(k,t)$$

with

$$\mathcal{A}_n(k,t) = \mathrm{i}\,\langle u_n(k,t)\,,\, \nabla_k u_n(k,t)\rangle_{\mathcal{H}_\mathrm{f}}$$
$$\phi_n(k,t) = -\mathrm{i}\,\langle u_n(k,t)\,,\, \partial_t u_n(k,t)\rangle_{\mathcal{H}_\mathrm{f}}\,.$$

The notation emphasizes the analogy with the electromagnetism: if $\mathcal{A}_n(k,t)$ and $\phi_n(k,t)$ are interpreted as the geometric analogous of the vector potential and of the electrostatic scalar potential, then $\Theta_n(k,t)$ and $\Omega_n(k,t)$ correspond, respectively, to the electric and to the magnetic field.

One can rigorously connect (15) and the semiclassical model (16), in the spirit of the result stated at the end of the previous section, see [15]. From (16) one obtains the *King-Smith and Vanderbilt formula* [9], which approximately predicts the contribution ΔP of the electrons to the macroscopic polarization of a crystalline insulator strained in the time interval $[0,T]$, namely,

$$\Delta P = \frac{1}{(2\pi)^d} \sum_{n\in N_{\mathrm{occ}}} \int_{Y*} (\mathcal{A}_n(k,T) - \mathcal{A}_n(k,0))\, dk\,, \tag{17}$$

where the sum runs over all the occupied Bloch bands, i.e., $N_{\mathrm{occ}} = \{n \in \mathbb{N} : E_n(k,t) \leq E_F\}$ with E_F the Fermi energy. Notice that (17) requires the computation of the Bloch functions only at the initial and at the final time; in view of that, the previous formula is the starting point of any *state-of-the-art* numerical simulation of macroscopic polarization in insulators.

Cross-References

- ▶ Born–Oppenheimer Approximation, Adiabatic Limit, and Related Math. Issues
- ▶ Mathematical Theory for Quantum Crystals

References

1. Blount, E.I.: Formalism of band theory. In: Seitz, F., Turnbull, D. (eds.) Solid State Physics, vol. 13, pp. 305–373. Academic, New York (1962)
2. Brouder, Ch., Panati, G., Calandra, M., Mourougane, Ch., Marzari, N.: Exponential localization of Wannier functions in insulators. Phys. Rev. Lett. **98**, 046402 (2007)
3. Catto, I., Le Bris, C., Lions, P.-L.: On the thermodynamic limit for Hartree-Fock type problems. Ann. Henri Poincaré **18**, 687–760 (2001)
4. Cancès, E., Deleurence, A., Lewin, M.: A new approach to the modeling of local defects in crystals: the reduced Hartree-Fock case. Commun. Math. Phys. **281**, 129–177 (2008)
5. des Cloizeaux, J.: Analytical properties of n-dimensional energy bands and Wannier functions. Phys. Rev. **135**, A698–A707 (1964)
6. Goedecker, S.: Linear scaling electronic structure methods. Rev. Mod. Phys. **71**, 1085–1111 (1999)
7. Gérard, C., Martinez, A., Sjöstrand, J.: A mathematical approach to the effective Hamiltonian in perturbed periodic problems. Commun. Math. Phys. **142**, 217–244 (1991)
8. Helffer, B., Sjöstrand, J.: Équation de Schrödinger avec champ magnétique et équation de Harper. In: Holden, H., Jensen, A. (eds.) Schrödinger Operators. Lecture Notes in Physics, vol. 345, pp. 118–197. Springer, Berlin (1989)
9. King-Smith, R.D., Vanderbilt, D.: Theory of polarization of crystalline solids. Phys. Rev. **B 47**, 1651–1654 (1993)
10. Kohn, W.: Analytic properties of Bloch waves and Wannier functions. Phys. Rev. **115**, 809–821 (1959)
11. Marzari, N., Vanderbilt, D.: Maximally localized generalized Wannier functions for composite energy bands. Phys. Rev. B **56**, 12847–12865 (1997)
12. Nenciu, G.: Existence of the exponentially localised Wannier functions. Commun. Math. Phys. **91**, 81–85 (1983)
13. Nenciu, G.: Dynamics of band electrons in electric and magnetic fields: rigorous justification of the effective Hamiltonians. Rev. Mod. Phys. **63**, 91–127 (1991)
14. Panati, G., Spohn, H., Teufel, S.: Effective dynamics for Bloch electrons: Peierls substitution and beyond. Commun. Math. Phys. **242**, 547–578 (2003)
15. Panati, G., Sparber, Ch., Teufel, S.: Geometric currents in piezoelectricity. Arch. Ration. Mech. Anal. **191**, 387–422 (2009)
16. Panati, G.: Triviality of Bloch and Bloch-Dirac bundles. Ann. Henri Poincaré **8**, 995–1011 (2007)

17. Panati, G., Pisante, A.: Bloch bundles, Marzari-Vanderbilt functional and maximally localized Wannier functions. preprint arXiv.org (2011)
18. Resta, R.: Theory of the electric polarization in crystals. Ferroelectrics **136**, 51–75 (1992)
19. Sundaram, G., Niu, Q.: Wave-packet dynamics in slowly perturbed crystals: gradient corrections and Berry-phase effects. Phys. Rev. B **59**, 14915–14925 (1999)
20. Teufel, S., Panati, G.: Propagation of Wigner functions for the Schrödinger equation with a perturbed periodic potential. In: Blanchard, Ph., Dell'Antonio, G. (eds.) Multiscale Methods in Quantum Mechanics. Birkhäuser, Boston (2004)
21. Thonhauser, T., Ceresoli, D., Vanderbilt, D., Resta, R.: Orbital magnetization in periodic insulators. Phys. Rev. Lett. **95**, 137205 (2005)
22. Wannier, G.H.: The structure of electronic excitation levels in insulating crystals. Phys. Rev. **52**, 191–197 (1937)

Source Location

Victor Isakov
Department of Mathematics and Statistics, Wichita State University, Wichita, KS, USA

General Problem

Let A be a partial differential operator of second order

$$Au = f \text{ in } \Omega. \tag{1}$$

In the inverse source problem, one is looking for the source term f from the boundary data

$$u = g_0,\ \partial_\nu u = g_1 \text{ on } \Gamma_0 \subset \partial\Omega, \tag{2}$$

where g_0, g_1 are given functions. In this short expository note, we will try to avoid technicalities, so we assume that (in general nonlinear) A is defined by a known C^2-function and f is a function of $x \in \Omega$ where Ω is a given bounded domain in $\mathbb{R}^n$ with C^2 boundary. ν denotes the exterior unit normal to the boundary of a domain. $H^k(\Omega)$ is the Sobolev space with the norm $\| \|_{(k)}(\Omega)$.

A first crucial question is whether there is enough data to (uniquely) find f. If A is a linear operator, then solution f of this problem is not unique. Indeed, let u_0 be a function in the Sobolev space $H^2(\Omega)$ with zero Cauchy data $u_0 = \partial_\nu u_0 = 0$ on Γ_0, and let $f_0 = Au_0$. Due to linearity, $A(u + u_0) = f + f_0$. Obviously, u and $u + u_0$ have the same Cauchy data on Γ_0, so f and $f + f_0$ produce the same data (2), but they are different in Ω. It is clear that there is a very large (infinite dimensional) manifold of solutions to the inverse source problem (1) and (2). To regain uniqueness, one has to restrict unknown distributions to a smaller but physically meaningful uniqueness class.

Inverse Problems of Potential Theory

We start with an inverse source problem which has a long and rich history. Let Φ be a fundamental solution of a linear second-order elliptic partial differential operator A in $\mathbb{R}^n$. The potential of a (Radon) measure μ supported in Ω is

$$u(x;\mu) = \int_\Omega \Phi(x,y)d\mu(y). \tag{3}$$

The general **inverse problem of potential theory** is to find $\mu, supp\mu \subset \Omega$, from the boundary data (2).

Since $Au(;\mu) = \mu$ (in generalized sense), the inverse problem of potential theory is a particular case of the inverse source problem. In the inverse problem of gravimetry, one considers $A = -\Delta$,

$$\Phi(x,y) = \frac{1}{4\pi|x-y|},$$

and the gravity field is generated by volume mass distribution $f \in L^1(\Omega)$. We will identify f with a measure μ. Since f with the data (2) is not unique, one can look for f with the smallest ($L^2(\Omega)$-) norm. The subspace of harmonic functions f_h is L^2-closed, so for any f, there is a unique f_h such that $f = f_h + f_0$ where f_0 is (L^2)-orthogonal to f_h. Since the fundamental solution is a harmonic function of y when x is outside Ω, the term f_0 produces zero potential outside Ω. Hence, the harmonic orthogonal component of f has the same exterior data and minimal L^2-norm. Applying the Laplacian to the both sides of the equation $-\Delta u(;f_h) = f_h$, we arrive at the biharmonic equation $\Delta^2 u(;f_h) = 0$ in Ω. When $\Gamma_0 = \partial\Omega$, we have a well-posed first boundary value problem for the biharmonic equation for $u(;f_h)$. Solving this problem, we find f_h from the previous Poisson equation.

However, it is hard to interpret f_h (geo)physically, knowing f_h does not help much with finding f.

A (geo)physical intuition suggests looking for a perturbing inclusion D of constant density, i.e., for $f = \chi_D$ (characteristic function of an open set D).

Since (in distributional sense) $-\Delta u(;\mu) = \mu$ in Ω, by using the Green's formula (or the definition of a weak solution), we yield

$$-\int_{\Omega} u^* d\mu = \int_{\partial\Omega} ((\partial_\nu u)u^* - (\partial_\nu u^*)u) \qquad (4)$$

for any function $u^* \in H^1(\Omega)$ which is harmonic in Ω. If $\Gamma_0 = \partial\Omega$, then the right side in (4) is known; we are given all harmonic moments of μ. In particular, letting $u^* = 1$, we obtain the total mass of μ, and by letting u^* to be coordinate (linear) functions, we obtain moments of μ of first order and hence the center of gravity of μ.

Even when one assumes that $f = \chi_D$, there is a nonuniqueness due to possible disconnectedness of the complement of D. Indeed, it is well known that if D is the ball $B(a, R)$ with center a of radius R, then its Newtonian potential $u(x, D) = M\frac{1}{4\pi|x-a|}$, where M is the total mass of D. So the exterior potentials of all annuli $B(a, R_2) \setminus B(a, R_1)$ are the same when $R_1^3 - R_2^3 = C$ where C is a positive constant. Moreover, by using this simple example and some reflections in $\mathbb{R}^n$, one can find two different domains with connected boundaries and equal exterior Newtonian potentials. Augmenting this construction by the condensation of singularities argument from the theory of functions of complex variables, one can construct a continuum of different domains with connected boundaries and the same exterior potential. So there is a need to have geometrical conditions on D.

A domain D is called star shaped with respect to a point a if any ray originated at a intersects D over an interval. An open set D is x_1 convex if any straight line parallel to the x_1-axis intersects D over an interval.

In what follows Γ_0 is a non-void open subset of $\partial\Omega$.

Theorem 1 *Let D_1, D_2 be two domains which are star shaped with respect to their centers of gravity or two x_1 convex domains in $\boldsymbol{R}^n$. Let u_1, u_2 be potentials of $D = D_1, D_2$.*

If $u_1 = u_2, \partial_\nu u_1 = \partial_\nu u_2$ on Γ_0, then $D_1 = D_2$.

Returning to the uniqueness proof, we assume that there are two x_1-convex D_1, D_2 with the same data. By uniqueness in the Cauchy problem for the Laplace equation, $u_1 = u_2$ near $\partial\Omega$. Then from (4) (with $d\mu = (\chi_{D_1} - \chi_{D_2})dm$, dm is the Lebesgue measure)

$$\int_{D_1} u^* = \int_{D_2} u^*$$

for any function u^* which is harmonic in Ω. Novikov's method of orthogonality is to assume that D_1 and D_2 are different and then to select u^* in such way that the left integral is less than the right one. To achieve this goal, u^* is replaced by its derivative, and one integrates by parts to move integrals to boundaries and makes use of the maximum principles to bound interior integrals.

The inverse problem of potential theory is a severely (exponentially) ill-conditioned problem of mathematical physics. The character of stability, conditional stability estimates, and regularization methods of numerical solutions of such problems are studied starting from pioneering work of Fritz John and Tikhonov in 1950–1960s.

To understand the degree of ill conditioning, one can consider harmonic continuation from the circle $\Gamma_0 = \{x : |x| = R\}$ onto the inner circle $\Gamma = \{x : |x| = \rho\}$. By using polar coordinates (r, ϕ), any harmonic function decaying at infinity can be (in a stable way) approximated by $u(r, \phi; M) = \sum_{m=1}^{M} u_m r^{-m} e^{im\phi}$. Let us define the linear operator of the continuation as $A(\partial_r(, R)) = \partial_r u(, \rho)$. Using the formula for $u(; M)$, it is easy to see that the condition number of the corresponding matrix is $(\frac{R}{\rho})^M$ which is growing exponentially with respect to M. If $\frac{R}{\rho} = 10$, then the use of computers is only possible when $M < 16$, and typical practical measurements errors of 0.01 allow meaningful computational results when $M < 3$.

The following logarithmic stability estimate holds and can be shown to be best possible. We denote by $||_2(S^2)$ the standard norm in the space $C^2(S^2)$.

Theorem 2 *Let D_1, D_2 be two domains given in polar coordinates (r, σ) by the equations $\partial D_j = \{r = d_j(\sigma)\}$ where $|d_j|_2(S^2) \leq M_2, \frac{1}{M_2} < d_j, j = 1, 2$. Let $\varepsilon = ||u_1 - u_2||_{(1)}(\Gamma_0) + ||\partial_\nu(u_1 - u_2)||_{(0)}(\Gamma_0)$.*

Then there is a constant C depending only on M_2, Γ_0 such that $|d_1 - d_2| \leq C(-log\varepsilon)^{-\frac{1}{C}}$.

A proof in [4] is using some ideas from the proof of Theorem 1 and stability estimates for harmonic continuation.

Moreover, while it is not possible to obtain (even local) existence results, a special local existence theorem is available [4], chapter 5. In more detail, if one assumes that u_0 is a potential of some C^3-domain D_0, that the Cauchy data for a function u are close to the Cauchy data of u_0, and that, moreover, u admits harmonic continuation across ∂D_0, as well as suitable behavior at infinity, then u is a potential of a domain D which is close to D_0.

The exterior gravity field of a polygon (polyhedron) D develops singularities at the corner points of D. Indeed, $\partial_j \partial_k u(x; \chi_D)$ where D is a polyhedron with corner at x_0 behaves as $-C log|x - x_0|$, [4], section 4.1. Since these singularities are uniquely identified by the Cauchy data, one has obvious uniqueness results under mild geometrical assumptions on D. Moreover, the use of singularities provides us with constructive identification tools, based on range type algorithms in the harmonic continuation, using, for example, the operator of the single layer potential.

For proofs and further results on inverse problems of potential theory, we refer to the work of V. Ivanov, Isakov, and Prilepko [4, 7].

An inverse source problem for nonlinear elliptic equations arises when detecting doping profile (source term in equations modeling semiconductors).

In the inverse problem of **magnetoencephalography**, A is defined to be Maxwell's system, and f is a first-order distribution supported in Ω (e.g., head of a patient). As above, there are difficulties due to nonuniqueness and severe instability. One of the simple cases is when $f = \sum_{m=1}^{M} a_m \partial_{d(m)} \delta(-x(m))$, where $\delta(-x(m))$ is the Dirac delta function with the pole $x(m)$ and $d(m)$ is a direction. Then uniqueness of f is obvious, and for not large M, the problem of determining $a_m, x(m)$ is well conditioned. However, such simplification is not satisfactory for medical diagnostics. For simplicity of exposition, we let now $A = -\Delta$. One of the more realistic assumptions is that f is a double layer distributed with density g over a so-called cortical surface Γ, i.e., $f = g\partial_\nu d\Gamma$. Γ can be found by using different methods, so one can assume that it is known. So one looks for a function $g \in L^1(\Gamma)$ on Γ from the Cauchy data (2) for the double layer potential

$$u(x; f) = \int_\Gamma g(y) \partial_{\nu(y)} \Phi(x, y) d\Gamma(y).$$

Uniqueness of g (up to a constant) is obvious, and stability is similar to the inverse problem of gravimetry. For biomedical inverse (source) problems, we refer to [1].

Finding Sources of Stationary Waves

Stationary waves of frequency k in a simple case are solutions to the Helmholtz equation, i.e., $A = -\Delta - k^2$. The radiating fundamental solution of this equation is

$$\Phi(x, y) = \frac{e^{ik|x-y|}}{4\pi|x-y|}.$$

The inverse source problem at a fixed k has many similarities with the case $k = 0$, except that maximum principles are not valid anymore. In particular, Theorem 1 is not true: potential of a ball $u(; \chi_B)$ can be zero outside Ω containing B for certain choices of k and the radius of B.

Looking for f supported in $\bar{D}$ (D is a subdomain of Ω) can be viewed as finding acoustical sources distributed over $\bar{D}$. Besides, this inverse source problem has immediate applications to so-called **acoustical holography**. This is a method to detect (mechanical) vibrations of $\Gamma = \partial D$ from measurements of acoustical pressure u on $\Gamma_0 \subset \partial\Omega$. In simple accepted models, the normal speed of Γ is $\partial_\nu u$ on Γ. By solving the exterior Dirichlet problem for the Helmholtz equation outside Ω, one can uniquely and in a stable way determine $\partial_\nu u$ on Γ_0. One can show that if k is not a Dirichlet eigenvalue, then any $H^1(\Omega)$ solution u to the Helmholtz equation can be uniquely represented by $u(; gd\Gamma)$, so we can reduce the continuation problem to the inverse source problem for $\mu = gd\Gamma$ (single layer distribution over Γ).

The continuation of solutions of the Helmholtz equation is a severely ill-posed problem, but its ill conditioning is decreasing when k grows, and if one is looking for the "low frequency" part of g, then stability is Lipschitz. This "low frequency" part is increasing with growing k.

As above, the inverse source problem at fixed k has the similar uniqueness features. However, if $f = f_0 + kf_1$, where f_0, f_1, depend only on x, one regains uniqueness. This statement is easier to understand considering u as the time Fourier transform of a solution of a wave equation with f_0, f_1 as the initial data. In transient (nonstationary) problems, one collects additional boundary data over a period of time.

Hyperbolic Equations

The inverse source problem in a wave motion is to find $(u, f) \in H^{(2)}(\Omega) \times L^2(\Omega)$ from the following partial differential equation

$$\partial_t^2 u - \Delta u = f,\ \partial_t^2 f = 0 \text{ in } \Omega = G \times (0, T),$$

with natural lateral boundary and initial conditions

$$\partial_\nu u = 0 \text{ on } \partial\Omega \times (0, T),\ u = \partial_t u = 0 \text{ on } \Omega \times \{0\}, \tag{5}$$

and the additional data

$$u = g \text{ on } \Gamma_0 = S_0 \times (0, T),$$

where S_0 is a part of ∂G. Assuming $\partial_t^2 u \in H^2(\Omega)$ and letting

$$v = \partial_t^2 u \tag{6}$$

and differentiating twice with respect to t transform this problem into finding the initial data in the following hyperbolic mixed boundary value problem

$$\partial_t^2 v - \Delta v = 0 \text{ in } \Omega, \tag{7}$$

with the lateral boundary condition

$$\partial_\nu v = 0 \text{ on } \partial G \times (0, T), \tag{8}$$

from the additional data

$$v = \partial_t^2 g \text{ on } \Gamma_0. \tag{9}$$

Indeed, one can find u from (6) and the initial conditions (5).

Theorem 3 *Let*

$$2dist(x, S_0; G) < T, x \in \partial G.$$

Then the data (9) *on Γ_0 for a solution v of* (7) *and* (8) *uniquely determine v on Ω.*

If, in addition, $S_0 = \partial G$, then

$$\|v(, 0)\|_{(1)}(G) + \|\partial_t v(, 0)\|_{(0)}(G) \leq C\|\partial_t^2 g\|_{(1)}(\Gamma_0). \tag{10}$$

Here, $d(x, S_0; G)$ is the (minimal) distance from x to S_0 inside G.

The statement about uniqueness for arbitrary S_0 follows from the sharp uniqueness of the continuation results for second-order hyperbolic equations and some geometric ideas [5], section 3.4. For hyperbolic equations with analytic coefficients, these sharp results are due to Fritz John and are based on the Holmgren Theorem. For C^1-space coefficients, the Holmgren Theorem was extended by Tataru. Stability of continuation (and hence in the inverse source problem) is (as for the harmonic continuation) at best of the logarithmic type (i.e., we have severely ill-conditioned inverse problem).

When $S_0 = \partial G$, one has a very strong (best possible) Lipschitz stability estimate (10). This estimate was obtained by Lop-Fat Ho (1986) by using the technique of multipliers; for more general hyperbolic equations by Klibanov, Lasiecka, Tataru, and Triggiani (1990s) by using Carleman-type estimates; and by Bardos, Lebeau, and Rauch (1992) by propagation of singularities arguments. Similar results are available for general linear hyperbolic equations of second order with time-independent coefficients. However, for Lipschitz stability, one has to assume the existence of a suitable pseudo-convex function or absence of trapped bicharacteristics. Looking for the source of the wave motion (in the more complicated elasticity system), in particular, can be interpreted as finding location and intensity of earthquakes. The recent medical diagnostic technique called **thermoacoustical tomography** can be reduced to looking for the initial displacement u_0. One of the versions of this problem is a classical one of looking for a function from its spherical means. In a limiting case when radii of spheres are getting large, one arrives at one of the most useful problems of **tomography** whose mathematical theory was initiated by Radon (1917) and Fritz John (1940s). For a recent advance in tomography in case of general attenuation, we refer to [2]. Detailed references are in [4, 5].

In addition to direct applications, the inverse source problems represent linearizations of (nonlinear) problems of finding coefficients of partial differential equations and can be used in the study of uniqueness and stability of identification of coefficients. For example, subtracting two equations $\partial_t^2 u_2 - a_2 \Delta u_2 = 0$ and $\partial_t^2 u_1 - a_1 \Delta u_1 = 0$ yields $\partial_t^2 u - a_2 \Delta u = \alpha f$ with $\alpha = \Delta u_1$ (as a known weight function) and unknown $f = a_2 - a_1$. A general technique to show uniqueness and stability of such inverse source problems by utilizing Carleman estimates was introduced in [3].

Acknowledgements This research was in part supported by the NSF grant DMS 10-07734 and by Emylou Keith and Betty Dutcher Distinguished Professorship at WSU.

References

1. Ammari, H., Kang, H.: An Introduction to Mathematics of Emerging Biomedical Imaging. Springer, Berlin (2008)
2. Arbuzov, E.V., Bukhgeim, A.L., Kazantsev, S.G.: Two-dimensional tomography problems and the theory of A-analytic functions. Siber. Adv. Math. **8**, 1–20 (1998)
3. Bukhgeim, A.L., Klibanov, M.V.: Global uniqueness of class of multidimensional inverse problems. Sov. Math. Dokl. **24**, 244–247 (1981)
4. Isakov, V.: Inverse Source Problems. American Mathematical Society, Providence (1990)
5. Isakov, V.: Inverse Problems for PDE. Springer, New York (2006)
6. Novikov, P.: Sur le probleme inverse du potentiel. Dokl. Akad. Nauk SSSR **18**, 165–168 (1938)
7. Prilepko, A.I., Orlovskii, D.G., Vasin, I.A.: Methods for Solving Inverse Problems in Mathematical Physics. Marcel Dekker, New York (2000)

Sparse Approximation

Holger Rauhut
Lehrstuhl C für Mathematik (Analysis),
RWTH Aachen University, Aachen, Germany

Definition

The aim of sparse approximation is to represent an object – usually a vector, matrix, function, image, or operator – by a linear combination of only few elements from a basis, or more generally, from a redundant system such as a frame. The tasks at hand are to design efficient computational methods for finding sparse representations and to estimate the approximation error that can be achieved for certain classes of objects.

Overview

Sparse approximations are motivated by several types of applications. An important source is the various tasks in signal and image processing tasks, where it is an empirical finding that many types of signals and images can indeed be well approximated by a sparse representation in an appropriate basis/frame. Concrete applications include compression, denoising, signal separation, and signal reconstruction (compressed sensing).

On the one hand, the theory of sparse approximation is concerned with identifying the type of vectors, functions, etc. which can be well approximated by a sparse expansion in a given basis or frame and with quantifying the approximation error. For instance, when given a wavelet basis, these questions relate to the area of function spaces, in particular, Besov spaces. On the other hand, algorithms are required to actually find a sparse approximation to a given vector or function. In particular, if the frame at hand is redundant or if only incomplete information is available – as it is the case in compressed sensing – this is a nontrivial task. Several approaches are available, including convex relaxation (ℓ_1-minimization), greedy algorithms, and certain iterative procedures.

Sparsity

Let $\mathbf{x}$ be a vector in $\mathbb{R}^N$ or $\mathbb{C}^N$ or $\ell_2(\Gamma)$ for some possibly infinite set Γ. We say that $\mathbf{x}$ is s-sparse if

$$\|\mathbf{x}\|_0 := \#\{\ell : x_\ell \neq 0\} \leq s.$$

For a general vector $\mathbf{x}$, the error of best s-term approximation quantifies the distance to sparse vectors,

$$\sigma_s(\mathbf{x})_p := \inf_{\mathbf{z}:\|\mathbf{z}\|_0 \leq s} \|\mathbf{x} - \mathbf{z}\|_p.$$

Here, $\|\mathbf{x}\|_p = (\sum_j |x_j|^p)^{1/p}$ is the usual ℓ_p-norm for $0 < p < \infty$ and $\|\mathbf{x}\|_\infty = \sup_j |x_j|$. Note that the vector $\mathbf{z}$ minimizing $\sigma_s(\mathbf{c})_p$ equals $\mathbf{x}$ on the indices corresponding to the s largest absolute coefficients of $\mathbf{x}$ and is zero on the remaining indices. We say that $\mathbf{x}$ is compressible if $\sigma_s(\mathbf{x})_p$ decays quickly in s, that is, for suitable s we can approximate $\mathbf{x}$ well by an s-sparse vector. This occurs for instance in the particular case when $\mathbf{x}$ is taken from the ℓ_q-unit ball $B_q = \{\mathbf{x} : \|\mathbf{x}\|_q \leq 1\}$ for small q. Indeed, an inequality due to Stechkin (see, e.g., [21, Lemma 3.1]) states that, for $0 < q < p$,

$$\sigma_s(\mathbf{x})_p \leq s^{1/p-1/q} \|\mathbf{x}\|_q. \qquad (1)$$

This inequality enlightens the importance of ℓ_q-spaces with $q < 1$ in this context.

The situation above describes sparsity with respect to the canonical basis. For a more general setup, consider a (finite- or infinite-dimensional) Hilbert space $\mathcal{H}$ (often a space of functions) endowed with an orthonormal basis $\{\psi_j, j \in J\}$. Given an element $f \in \mathcal{H}$, our aim is to approximate it by a finite linear combination of the ψ_j, that is, by

$$\sum_{j \in S} x_j \psi_j,$$

where $S \subset J$ is of cardinality at most s, say. In contrast to linear approximation, the index set S is not fixed a priori but is allowed to depend on f. Analogously as above, the error of best s-term approximation is then defined as

$$\sigma_s(f)_{\mathcal{H}} := \inf_{\mathbf{x}:\|\mathbf{x}\|_0 \le s} \|f - \sum_{j \in J} x_j \psi_j\|_{\mathcal{H}}$$

and the element $\sum_j x_j \psi_j$ with $\|\mathbf{x}\|_0 \le s$ realizing the infimum is called a best s-term approximation to f. Due to the fact that the support set of $\mathbf{x}$ (i.e., the index set of nonzero entries of $\mathbf{x}$) is not fixed a priori, the set of such elements does not form a linear space, so that one sometimes simply refers to nonlinear approximation [12, 30].

One may generalize this setup further. For instance, instead of requiring that $\{\psi_j : j \in J\}$ forms an orthonormal basis, one may assume that it is a frame [7, 19, 23], that is, there are constants $0 < A \le B < \infty$ such that

$$A\|f\|_{\mathcal{H}}^2 \le \sum_{j \in J} |\langle \psi_j, f\rangle|^2 \le B\|f\|_{\mathcal{H}}^2.$$

This definition includes orthonormal bases but allows also redundancy, that is, the coefficient vector $\mathbf{x}$ in the expansion $f = \sum_{j \in J} x_j \psi_j$ is no longer unique. Redundancy has several advantages. For instance, since there are more possibilities for a sparse approximation of f, the error of s-sparse approximation may potentially be smaller. On the other hand, it may get harder to actually find a sparse approximation (see also below).

In another direction, one may relax the assumption that $\mathcal{H}$ is a Hilbert space and only require it to be a Banach space. Clearly, then the notion of an orthonormal basis also does not make sense anymore, so that $\{\psi_j, j \in J\}$ is then just some system of elements spanning the space – possibly a basis.

Important types of systems $\{\psi_j, j \in J\}$ considered in this context include the trigonometric system $\{e^{2\pi i k \cdot}, k \in \mathbb{Z}\} \subset L^2[0,1]$, wavelet systems [9, 36], or Gabor frames [23].

Quality of a Sparse Approximation

One important task in the field of sparse approximation is to quantify how well an element $f \in \mathcal{H}$ or a whole class $B \subset \mathcal{H}$ of elements can be approximated by sparse expansions. An abstract way [12, 20] of describing good approximation classes is to introduce

$$B_p := \{f \in \mathcal{H} : f = \sum_{j \in J} x_j \psi_j, \|\mathbf{x}\|_p < \infty\}$$

with norm $\|f\|_{B_p} = \inf\{\|\mathbf{x}\|_p : f = \sum_j x_j \psi_j\}$. If $\{\psi_j : j \in J\}$ is an orthonormal basis, then it follows directly from (1) that, for $0 < p < 2$,

$$\sigma_s(f)_{\mathcal{H}} \le s^{1/2-1/p}\|f\|_{B_p}. \tag{2}$$

In concrete situations, the task is then to characterize the spaces B_p. In the case, that $\{\psi_j : j \in J\}$ is a wavelet system, then one obtains Besov spaces [32, 36], and in the case of the trigonometric system, this results in the classical Fourier algebra when $p = 1$. If $\{\psi_j : j \in J\}$ is a frame, then (2) remains valid up to a multiplicative constant. In the special case of Gabor frames, the space B_p coincides with a class of modulation spaces [23].

Algorithms for Sparse Approximation

For practical purposes, it is important to have algorithms for computing optimal or at least near-optimal sparse approximations. When $\mathcal{H} = \mathbb{C}^N$ is finite dimensional and $\{\psi_j, j = 1, \ldots, N\} \subset \mathbb{C}^N$ is an orthonormal basis, then this is easy. In fact, the coefficients in the expansion $f = \sum_{j=1}^N x_j \psi_j$ are given by $x_j = \langle f, \psi_j\rangle$, so that a best s-term approximation to f in $\mathcal{H}$ is given by

$$\sum_{j\in S}\langle f,\psi_j\rangle\psi_j$$

where S is an index set of s largest absolute entries of the vector $(\langle f,\psi_j\rangle)_{j=1}^N$.

When $\{\psi_j, j=1,\ldots,M\}\subset\mathbb{C}^N$ is redundant, that is, $M>N$, then it becomes a nontrivial problem to find the sparsest approximation to a given $f\in\mathbb{C}^N$. Denoting by Ψ the $N\times M$ matrix whose columns are the vectors ψ_j, this problem can be expressed as finding the minimizer of

$$\min\|\mathbf{x}\|_0 \quad \text{subject to} \quad \|\Psi\mathbf{x}-f\|_2\le\varepsilon, \qquad (3)$$

for a given threshold $\varepsilon>0$. In fact, this problem is known to be NP hard in general [11, 25]. Several tractable alternatives have been proposed. We discuss the greedy methods matching pursuit and orthogonal matching pursuit as well as the convex relaxation method basis pursuit (ℓ_1-minimization) next. Other sparse approximation algorithms include iterative schemes, such as iterative hard thresholding [1] and iteratively reweighted least squares [10].

Matching Pursuits

Given a possibly redundant system $\{\psi_j, j\in J\}\subset\mathcal{H}$ – often called a dictionary – the greedy algorithm matching pursuit [24,27,31,33] iteratively builds up the support set and the sparse approximation. Starting with $r_0=f$, $S_0=\emptyset$ and $k=0$ it performs the following steps:

1. $j_k := \operatorname{argmax}\left\{\frac{|\langle r_k,\psi_j\rangle|}{\|\psi_j\|} : j\in J\right\}$.
2. $S_{k+1} := S_k\cup\{j_k\}$.
3. $r_{k+1}=r_k-\frac{\langle r_k,\psi_{j_k}\rangle}{\|\psi_{j_k}\|_2}\psi_{j_k}$.
4. $k\mapsto k+1$.
5. Repeat from step (1) with $k\mapsto k+1$ until a stopping criterion is met.
6. Output $\widetilde{f}=\widetilde{f}_k=\sum_{\ell=1}^k\frac{\langle r_\ell,\psi_{j_\ell}\rangle}{\|\psi_{j_\ell}\|_2}\psi_{j_\ell}$.

Clearly, if s steps of matching pursuit are performed, then the output $\widetilde{f}$ has an s-sparse representation with respect to $\widetilde{f}$. It is known that the sequence $\widetilde{f}_k$ converges to f when k tends to infinity [24]. A possible stopping criterion for step (5) is a maximal number of iterations, or that the residual norm $\|r_k\|\le\epsilon$ for some prescribed tolerance $\epsilon>0$.

Matching pursuit has the slight disadvantage that an index k may be selected more than once. A variation on this greedy algorithm which avoids this drawback consists in the orthogonal matching pursuit algorithm [31, 33] outlined next. Again, starting with $r_0=f$, $S_0=\emptyset$ and $k=0$, the following steps are conducted:

1. $j_k := \operatorname{argmax}\left\{\frac{|\langle r_k,\psi_j\rangle|}{\|\psi_j\|} : j\in J\right\}$.
2. $S_{k+1} := S_k\cup\{j_k\}$.
3. $x^{(k+1)} := \operatorname{argmin}_{z:\operatorname{supp}(z)\subset S_{k+1}}\|f-\sum_{j\in S_{k+1}} z_j\psi_j\|_2$.
4. $r_{k+1} := f-\sum_{j\in S_{k+1}} x_j^{(k+1)}\psi_j$.
5. Repeat from step (1) with $k\mapsto k+1$ until a stopping criterion is met.
6. Output $\widetilde{f}=\widetilde{f}_k=\sum_{j\in S_k} x_j^{(k)}\psi_j$.

The essential difference to matching pursuit is the orthogonal projection step in (3). Orthogonal matching pursuit may require a smaller number of iterations than matching pursuit. However, the orthogonal projection makes an iteration computationally more demanding than an iteration of matching pursuit.

Convex Relaxation

A second tractable approach to sparse approximation is to relax the ℓ_0-minimization problem to the convex optimization problem of finding the minimizer of

$$\min\|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\Psi\mathbf{x}-f\|_2\le\varepsilon. \qquad (4)$$

This program is also known as basis pursuit [6] and can be solved using various methods from convex optimization [2]. At least in the real-valued case, the minimizer $\mathbf{x}^*$ of the above problem will always have at most N nonzero entries, and the support of $\mathbf{x}^*$ defines a linear independent set $\{\psi_j : x_j^*\neq 0\}$, which is a basis of $\mathbb{C}^N$ if $\mathbf{x}^*$ has exactly N nonzero entries – thus, the name basis pursuit.

Finding the Sparsest Representation

When the dictionary $\{\psi_j\}$ is redundant, it is of great interest to provide conditions which ensure that a specific algorithm is able to identify the sparsest possible representation. For this purpose, it is helpful to define the coherence μ of the system $\{\psi_j\}$, or equivalently of the matrix Ψ having the vectors ψ_j as its columns. Assuming the normalization $\|\psi_j\|_2=1$, it is defined as the maximal inner product between different dictionary elements,

$$\mu=\max_{j\neq k}|\langle\psi_j,\psi_k\rangle|.$$

Suppose that f has a representation with s terms, that is, $f = \sum_j x_j \psi_j$ with $\|\mathbf{x}\|_0 \leq s$. Then s iterations of orthogonal matching pursuit [3, 33] as well as basis pursuit (4) with $\varepsilon = 0$ [3, 14, 34] find the sparsest representation of f with respect to $\{\psi_j\}$ provided that

$$(2s - 1)\mu < 1. \tag{5}$$

Moreover, for a general f, both orthogonal matching pursuit and basis pursuit generate an s-sparse approximation whose approximation error is bounded by the error of best s-term approximation up to constants; see [3, 18, 31] for details.

For typical "good" dictionaries $\{\psi_j\}_{j=1}^M \subset \mathbb{C}^N$, the coherence scales as $\mu \sim \sqrt{N}$ [29], so that the bound (5) implies that s-sparse representations in such dictionaries with small enough sparsity, that is, $s \leq c\sqrt{N}$, can be found efficiently via the described algorithms.

Applications of Sparse Approximation

Sparse approximation find a variety of applications. Below we shortly describe compression, denoising, and signal separation. Sparse representations play also a major role in adaptive numerical methods for solving operator equations such as PDEs. When the solution has a sparse representation with a suitable basis, say finite elements or wavelets, then a significant acceleration with respect to standard linear methods can be achieved. The algorithms used in this context are of different nature than the ones described above. We refer to [8] for details.

Compression

An obvious application of sparse approximation is image and signal compression. Once a sparse approximation is found, one only needs to store the nonzero coefficients of the representation. If the representation is sparse enough, then this requires significantly less memory than storing the original signal or image. This principle is exploited, for instance, in the JPEG, MPEG, and MP3 data compression standards.

Denoising

Often acquired signals and images are corrupted by noise, that is, the observed signal can be written as $\widetilde{f} = f + \eta$, where f is the original signal and η is a vector representing the noise. The additional knowledge that the signal at hand can be approximated well by a sparse representation can be exploited to clean the signal by essentially removing the noise. The essential idea is to find a sparse approximation $\sum_j x_j \psi_j$ of $\widetilde{f}$ with respect to a suitable dictionary $\{\psi_j\}$ and to use it as an approximation to the original f. One algorithmic approach is to solve the ℓ_1-minimization problem (4), where ϵ is now a suitable estimate of the ℓ_2-norm of the noise η. If Ψ is a wavelet basis, this principle is often called wavelet thresholding or wavelet shrinkage [16] due to connection of the soft-thresholding function [13].

Signal Separation

Suppose one observes the superposition $f = f_1 + f_2$ of two signals f_1 and f_2 of different nature, for instance, the "harmonic" and the "spiky" component of an acoustic signal, or stars and filaments in an astronomical image. The task is to separate the two components f_1 and f_2 from the knowledge of f. Knowing that both f_1 and f_2 have sparse representations in dictionaries $\{\psi_j^j\}$ and $\{\psi_j^2\}$ of "different nature," one can indeed recover both f_1 and f_2 by similar algorithms as outlined above, for instance, by solving the ℓ_1-minimization problem

$$\min_{\mathbf{z}^1,\mathbf{z}^2} \|\mathbf{z}^1\| + \|\mathbf{z}^2\|_1 \;\text{ subject to }\; f = \sum_j z_j^1 \psi_j^1 + \sum_j z_j^2 \psi_j^2.$$

The solution $(\mathbf{x}^1, \mathbf{x}^2)$ defines the reconstructions $\widetilde{f}_1 = \sum_j x_j^1 \psi_j^1$ and $\widetilde{f}_2 = \sum_j x_j^2 \psi_j^2$. If, for instance, $\{\psi_j^1\}_{j=1}^N$ and $\{\psi_j^2\}_{j=1}^N$ are mutually incoherent bases in $\mathbb{C}^N$, that is, $\|\psi_j^1\|_2 = \|\psi_j^2\|_2 = 1$ for all j and the maximal inner product $\mu = |\langle \psi_j^1, \psi_j^2 \rangle|$ is small, then the above optimization problem recovers both f_1 and f_2 provided they have representations with altogether s terms where $s < 1/(2\mu)$ [15]. An example of two mutually incoherent bases are the Fourier basis and the canonical basis, where $\mu = 1/\sqrt{N}$ [17]. Under a probabilistic model, better estimates are possible [5, 35].

Compressed Sensing

The theory of compressed sensing [4, 21, 22, 26] builds on sparse representations. Assuming that a vector

$\mathbf{x} \in \mathbb{C}^N$ is s-sparse (or approximated well by a sparse vector), one would like to reconstruct it from only limited information, that is, from

$$\mathbf{y} = A\mathbf{x}, \quad \text{with } A \in \mathbb{C}^{m \times N}$$

where m is much smaller than N. Again the algorithms outlined above apply, for instance, basis pursuit (4) with A replacing Ψ. In this context, one would like to design matrices A with the minimal number m of rows (i.e., the minimal number of linear measurements), which are required to reconstruct $\mathbf{x}$ from $\mathbf{y}$. The recovery criterion based on coherence μ of A described above applies but is highly suboptimal. In fact, it can be shown that for certain random matrices $m \geq cs \log(eN/s)$ measurements suffice to (stably) reconstruct an s-sparse vector using ℓ_1-minimization with high probability, where c is a (small) universal constant. This bound is sufficiently better than the ones that can be deduced from coherence based bounds as described above. A particular case of interest arise when A consists of randomly selected rows of the discrete Fourier transform matrix. This setup corresponds to randomly sampling entries of the Fourier transform of a sparse vector. When $m \geq cs \log^4 N$, then ℓ_1-minimization succeeds to (stably) recover s-sparse vectors from m samples [26, 28].

This setup generalizes to the situation that one takes limited measurements of a vector $f \in \mathbb{C}^N$, which is sparse with respect to a basis or frame $\{\psi_j\}_{j=1}^M$. In fact, then $f = \Psi\mathbf{x}$ for a sparse $\mathbf{x} \in \mathbb{C}^M$ and with a measurement matrix $A \in \mathbb{C}^{m \times N}$, we have

$$\mathbf{y} = Af = A\Psi\mathbf{x},$$

so that we reduce to the initial situation with $A' = A\Psi$ replacing A. Once $\mathbf{x}$ is recovered, one forms $f = \Psi\mathbf{x}$.

Applications of compressed sensing can be found in various signal processing tasks, for instance, in medical imaging, analog-to-digital conversion, and radar.

References

1. Blumensath, T., Davies, M.: Iterative thresholding for sparse approximations. J. Fourier Anal. Appl. **14**, 629–654 (2008)
2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
3. Bruckstein, A., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
4. Candès, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid (2006)
5. Candès, E.J., Romberg, J.K.: Quantitative robust uncertainty principles and optimally sparse decompositions. Found. Comput. Math. **6**(2), 227–254 (2006)
6. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
7. Christensen, O.: An Introduction to Frames and Riesz Bases. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2003)
8. Cohen, A.: Numerical Analysis of Wavelet Methods. Studies in Mathematics and its Applications, vol. 32, xviii, 336p. EUR 95.00. North-Holland, Amsterdam (2003)
9. Daubechies, I.: Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics, vol 61. SIAM, Philadelphia (1992)
10. Daubechies, I., DeVore, R.A., Fornasier, M., Güntürk, C.: Iteratively re-weighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
11. Davis, G., Mallat, S., Avellaneda, M.: Adaptive greedy approximations. Constr. Approx. **13**(1), 57–98 (1997)
12. DeVore, R.A.: Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
13. Donoho, D.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory **41**(3), 613–627 (1995)
14. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. U.S.A. **100**(5), 2197–2202 (2003)
15. Donoho, D.L., Huo, X.: Uncertainty principles and ideal atomic decompositions. IEEE Trans. Inf. Theory **47**(7), 2845–2862 (2001)
16. Donoho, D.L., Johnstone, I.M.: Minimax estimation via wavelet shrinkage. Ann. Stat. **26**(3), 879–921 (1998)
17. Donoho, D.L., Stark, P.B.: Uncertainty principles and signal recovery. SIAM J. Appl. Math. **48**(3), 906–931 (1989)
18. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory **52**(1), 6–18 (2006)
19. Duffin, R.J., Schaeffer, A.C.: A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
20. Fornasier, M., Gröchenig, K.: Intrinsic localization of frames. Constr. Approx. **22**(3), 395–415 (2005)
21. Fornasier, M., Rauhut, H.: Compressive sensing. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, pp. 187–228. Springer, New York (2011)
22. Foucart, S., Rauhut, H.: A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhäuser, Basel
23. Gröchenig, K.: Foundations of Time-Frequency Analysis. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2001)
24. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
25. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM J. Comput. **24**, 227–234 (1995)
26. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations

and Numerical Methods for Sparse Recovery. Radon Series on Computational and Applied Mathematics, vol. 9, pp. 1–92. De Gruyter, Berlin/New York (2010)
27. Roberts, D.H.: Time series analysis with clean I. Derivation of a spectrum. Astronom. J. **93**, 968–989 (1987)
28. Rudelson, M., Vershynin, R.: On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**, 1025–1045 (2008)
29. Strohmer, T., Heath, R.W.: Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon Anal. **14**(3), 257–275 (2003)
30. Temlyakov, V.N.: Nonlinear methods of approximation. Found Comput. Math. **3**(1), 33–107 (2003)
31. Temlyakov, V.: Greedy Approximation. Cambridge Monographs on Applied and Computational Mathematics, No. 20. Cambridge University Press, Cambridge (2011)
32. Triebel, H.: Theory of Function Spaces. Monographs in Mathematics, vol. 78. Birkhäuser, Boston (1983)
33. Tropp, J.A.: Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**(10), 2231–2242 (2004)
34. Tropp, J.A.: Just relax: Convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory **51**(3), 1030–1051 (2006)
35. Tropp, J.A.: On the conditioning of random subdictionaries. Appl. Comput. Harmon Anal. **25**(1), 1–24 (2008)
36. Wojtaszczyk, P.: A Mathematical Introduction to Wavelets. Cambridge University Press, Cambridge (1997)

Special Functions: Computation

Amparo Gil[1], Javier Segura[2], and Nico M. Temme[3]
[1]Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, E.T.S. Caminos, Canales y Puertos, Santander, Spain
[2]Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Santander, Spain
[3]Centrum voor Wiskunde and Informatica (CWI), Amsterdam, The Netherlands

Mathematics Subject Classification

65D20; 41A60; 33C05; 33C10; 33C15

Overview

The special functions of mathematical physics [4] are those functions that play a key role in many problems in science and engineering. For example, Bessel, Legendre, or parabolic cylinder functions are well known for everyone involved in physics. This is not surprising because Bessel functions appear in the solution of partial differential equations in cylindrically symmetric domains (such as optical fibers) or in the Fourier transform of radially symmetric functions, to mention just a couple of applications. On the other hand, Legendre functions appear in the solution of electromagnetic problems involving spherical or spheroidal geometries. Finally, parabolic cylinder functions are involved, for example, in the analysis of the wave scattering by a parabolic cylinder, in the study of gravitational fields or quantum mechanical problems such as quantum tunneling or particle production.

But there are many more functions under the term "special functions" which, differently from the examples mentioned above, are not of hypergeometric type, such as some cumulative distribution functions [3, Chap. 10]. These functions also need to be evaluated in many problems in statistics, probability theory, communication theory, or econometrics.

Basic Methods

The methods used for the computation of special functions are varied, depending on the function under consideration as well as on the efficiency and the accuracy demanded. Usual tools for evaluating special functions are the evaluation of convergent and divergent series, the computation of continued fractions, the use of Chebyshev approximations, the computation of the function using integral representations (numerical quadrature), and the numerical integration of ODEs. Usually, several of these methods are needed in order to build an algorithm able to compute a given function for a large range of values of parameters and argument. Also, an important *bonus* in this kind of algorithms will be the possibility of evaluating scaled functions: if, for example, a function $f(z)$ increases exponentially for large $|z|$, the factorization of the exponential term and the computation of a scaled function (without the exponential term) can be used to avoid degradations in the accuracy of the functions and overflow problems as z increases. Therefore, the appropriate scaling of a special function could be useful for increasing both the range of computation and the accuracy of the computed expression.

Next, we briefly describe three important techniques for computing special functions which appear ubiquitously in algorithms for special function evaluation: convergent and divergent series, recurrence relations, and numerical quadrature.

Convergent and Divergent Series

Convergent series for special functions usually arise in the form of hypergeometric series:

$$ {}_pF_q\left(\begin{matrix} a_1,\cdots,a_p \\ ;z \\ b_1,\cdots,b_q \end{matrix}\right) = \sum_{n=0}^{\infty} \frac{(a_1)_n\cdots(a_p)_n}{(b_1)_n\cdots(b_q)_n}\frac{z^n}{n!}, \qquad (1)$$

where $p \leq q+1$ and $(a)_n$ is the Pochhammer symbol, also called the shifted factorial, defined by

$$(a)_0 = 1, \quad (a)_n = a(a+1)\cdots(a+n-1)\ (n \geq 1),$$
$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}. \qquad (2)$$

The series is easy to evaluate because of the recursion $(a)_{n+1} = (a+n)(a)_n$, $n \geq 0$, of the Pochhammer symbols. For example, for the modified Bessel function

$$I_\nu(z) = \left(\tfrac{1}{2}z\right)^\nu \sum_{n=0}^{\infty} \frac{(\frac{1}{4}z^2)^n}{\Gamma(\nu+n+1)\,n!}$$
$$= \left(\tfrac{1}{2}z\right)^\nu {}_0F_1\left(\begin{matrix} - \\ ;\tfrac{1}{4}z^2 \\ \nu+1 \end{matrix}\right), \qquad (3)$$

this is a stable representation when $z > 0$ and $\nu \geq 0$ and it is an efficient representation when z is not large compared with ν.

With *divergent expansion* we mean asymptotic expansions of the form

$$F(z) \sim \sum_{n=0}^{\infty} \frac{c_n}{z^n}, \quad z \to \infty. \qquad (4)$$

The series usually diverges, but it has the property

$$F(z) = \sum_{n=0}^{N-1} \frac{c_n}{z^n} + R_N(z), \quad R_N(z) = \mathcal{O}\left(z^{-N}\right),$$
$$z \to \infty, \qquad (5)$$

for $N = 0, 1, 2, \ldots$, and the order estimate holds for fixed N. This is the Poincaré-type expansion and for special functions like the gamma and Bessel functions they are crucial for evaluating these functions. Other variants of the expansion are also important, in particular expansions that hold for a certain range of additional parameters (this leads to the uniform asymptotic expansions in terms of other special functions like Airy functions, which are useful in turning point problems).

Recurrence Relations

In many important cases, there exist recurrence relations relating different values of the function for different values of its variables; in particular, one can usually find three-term recurrence relations [3, Chap. 4]. In these cases, the efficient computation of special functions uses at some stage the recurrence relations satisfied by such families of functions. In fact, it is difficult to find a computational task which does not rely on recursive techniques: the great advantage of having recursive relations is that they can be implemented with ease. However, the application of recurrence relations can be risky: each step of a recursive process generates not only its own rounding errors but also accumulates the errors of the previous steps. An important aspect is then the study of the numerical condition of the recurrence relations, depending on the initial values for starting recursion.

If we write the three-term recurrence satisfied by the function y_n as

$$y_{n+1} + b_n y_n + a_n y_{n-1} = 0, \qquad (6)$$

then, if a solution $y_n^{(m)}$ of (6) exists that satisfies $\lim_{n\to+\infty} \frac{y_n^{(m)}}{y_n^{(D)}} = 0$ for all solutions $y_n^{(D)}$ that are linearly independent of $y_n^{(m)}$, we will call $y_n^{(m)}$ the *minimal solution*. The solution $y_n^{(D)}$ is said to be a *dominant solution* of the three-term recurrence relation. From a computational point of view, the crucial point is the identification of the character of the function to be evaluated (either minimal or dominant) because the stable direction of application of the recurrence relation is different for evaluating the minimal or a dominant solution of (6): forward for dominant solutions and backward for minimal solutions.

For analyzing whether a special function is minimal or not, analytical information is needed regarding its behavior as $n \to +\infty$.

Assume that for large values of n the coefficients a_n, b_n behave as follows.

$$a_n \sim a n^{\alpha}, \quad b_n \sim b n^{\beta}, \quad ab \neq 0 \tag{7}$$

with α and β real; assume that t_1, t_2 are the zeros of the characteristic polynomial $\Phi(t) = t^2 + bt + a$ with $|t_1| \geq |t_2|$. Then it follows from Perron's theorem [3, p. 93] that we have the following results:

1. If $\beta > \frac{1}{2}\alpha$, then the difference equation (6) has two linearly independent solutions f_n and g_n, with the property

$$\frac{f_n}{f_{n-1}} \sim -\frac{a}{b} n^{\alpha-\beta}, \quad \frac{g_n}{g_{n-1}} \sim -b n^{\beta}, \quad n \to \infty. \tag{8}$$

   In this case, the solution f_n is minimal.
2. If $\beta = \frac{1}{2}\alpha$ and $|t_1| > |t_2|$, then the difference equation (6) has two linear independent solutions f_n and g_n, with the property

$$\frac{f_n}{f_{n-1}} \sim t_1 n^{\beta}, \quad \frac{g_n}{g_{n-1}} \sim t_2 n^{\beta}, \quad n \to \infty, \tag{9}$$

   In this case, the solution f_n is minimal.
3. If $\beta = \frac{1}{2}\alpha$ and $|t_1| = |t_2|$, or if $\beta < \frac{1}{2}\alpha$, then some information is still available, but the theorem is inconclusive with respect to the existence of minimal and dominant solutions.

Let's consider three-term recurrence relations satisfy by Bessel functions as examples. Ordinary Bessel functions satisfy the recurrence relation

$$y_{n+1} - \frac{2n}{z} y_n + y_{n-1} = 0, \quad z \neq 0, \tag{10}$$

with solutions $J_n(z)$ (the Bessel function of the first kind) and $Y_n(z)$ (the Bessel function of the second kind). This three-term recurrence relation corresponds to (8), with the values $a = 1$, $\alpha = 0$, $b = -\frac{2}{z}$, $\beta = 1$. Then, there exist two independent solutions f_n and g_n satisfying

$$\frac{f_{n+1}}{f_n} \sim \frac{z}{2n}, \quad \frac{g_{n+1}}{g_n} \sim \frac{2n}{z}. \tag{11}$$

As the known asymptotic behavior of the Bessel functions reads

$$J_n(z) \sim \frac{1}{n!}\left(\frac{z}{2}\right)^n, \quad Y_n(z) \sim -\frac{(n-1)!}{\pi}\left(\frac{2}{z}\right)^n, \quad n \to \infty, \tag{12}$$

it is easy to identify $J_n(z)$ and $Y_n(z)$ as the minimal (f_n) and a dominant (g_n) solutions, respectively, of the three-term recurrence relation (10).

Similar results hold for the modified Bessel functions, with recurrence relation

$$y_{n+1} + \frac{2n}{z} y_n - y_{n-1} = 0, \quad z \neq 0, \tag{13}$$

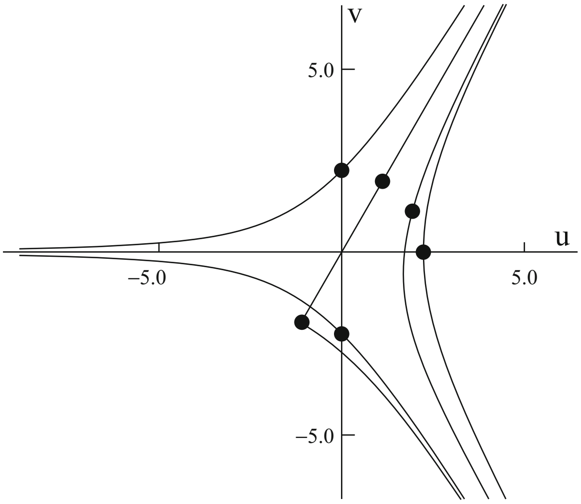
with solutions $I_n(z)$ (minimal) and $(-1)^n K_n(z)$ (dominant).

Numerical Quadrature

Another example where the study of numerical stability is of concern is the computation of special functions via integral representations. It is tempting, but usually wrong, to believe that once an integral representation is given, the computational problem is solved. One has to choose a stable quadrature rule and this choice depends on the integral under consideration. Particularly problematic is the integration of strongly oscillating integrals (Bessel and Airy functions, for instance); in these cases an alternative approach consists in finding non-oscillatory representations by properly deforming the integration path in the complex plane. Particularly useful is the *saddle point method* for obtaining integral representations which are suitable for applying the trapezoidal rule, which is optimal for computing certain integrals in **R**. Let's explain the saddle point method taking the Airy function $\mathrm{Ai}(z)$ as example. Quadrature methods for evaluating complex Airy functions can be found in, for example, [1, 2].

We start from the following integral representation in the complex plane:

$$\mathrm{Ai}(z) = \frac{1}{2\pi i} \int_{\mathcal{C}} e^{\frac{1}{3}w^3 - zw}\, dw, \tag{14}$$



Special Functions: Computation, Fig. 1 Saddle point contours for $\theta = 0, \frac{1}{3}\pi, \frac{2}{3}\pi, \pi$ and $r = 5$

where $z \in \mathbb{C}$ and $\mathcal{C}$ is a contour starting at $\infty e^{-i\pi/3}$ and terminating at $\infty e^{+i\pi/3}$ (in the valleys of the integrand). In the example we take $\operatorname{ph} z \in [0, \frac{2}{3}\pi]$.

Let $\phi(w) = \frac{1}{3}w^3 - zw$. The saddle points are $w_0 = \sqrt{z}$ and $-w_0$ and follow from solving $\phi'(w) = w^2 - z = 0$. The saddle point contour (the path of steepest descent) that runs through the saddle point w_0 is defined by $\Im[\phi(w)] = \Im[\phi(w_0)]$.

We write

$$z = x + iy = re^{i\theta}, \quad w = u + iv, \quad w_0 = u_0 + iv_0. \tag{15}$$

Then

$$u_0 = \sqrt{r}\cos\tfrac{1}{2}\theta, \quad v_0 = \sqrt{r}\sin\tfrac{1}{2}\theta, \quad x = u_0^2 - v_0^2,$$
$$y = 2u_0v_0. \tag{16}$$

The path of steepest descent through w_0 is given by the equation

$$u = u_0 + \frac{(v - v_0)(v + 2v_0)}{3\left[u_0 + \sqrt{\frac{1}{3}(v^2 + 2v_0v + 3u_0^2)}\right]},$$
$$-\infty < v < \infty. \tag{17}$$

Examples for $r = 5$ and a few $\theta-$values are shown in Fig. 1. The relevant saddle points are located on the circle with radius $\sqrt{r}$ and are indicated by small dots.

The saddle point on the positive real axis corresponds with the case $\theta = 0$ and the two saddles on the imaginary axis with the case $\theta = \pi$. It is interesting to see that the contour may split up and run through both saddle points $\pm w_0$. When $\theta = \frac{2}{3}\pi$ both saddle points are on one path, and the half-line in the $z-$plane corresponding with this θ is called a *Stokes line*.

Integrating with respect to $\tau = v - v_0$ (and writing $\sigma = u - u_0$), we obtain

$$\operatorname{Ai}(z) = \frac{e^{-\zeta}}{2\pi i}\int_{-\infty}^{\infty} e^{\psi_r(\sigma,\tau)}\left(\frac{d\sigma}{d\tau} + i\right)d\tau, \tag{18}$$

where $\zeta = \frac{2}{3}z^{\frac{3}{2}}$ and

$$\sigma = \frac{\tau(\tau + 3v_0)}{3\left[u_0 + \sqrt{\frac{1}{3}(\tau^2 + 4v_0\tau + 3r)}\right]}, \quad -\infty < \tau < \infty, \tag{19}$$

$$\psi_r(\sigma,\tau) = \Re[\phi(w) - \phi(w_0)] = u_0(\sigma^2 - \tau^2)$$
$$-2v_0\sigma\tau + \tfrac{1}{3}\sigma^3 - \sigma\tau^2. \tag{20}$$

The integral representation for the Airy function in (18) is now suitable for applying the trapezoidal rule. The resulting algorithm will be flexible and efficient.

References

1. Gautschi, W.: Computation of Bessel and Airy functions and of related Gaussian quadrature formulae. BIT **42**(1), 110–118 (2002)
2. Gil, A., Segura, J., Temme, N.M.: Algorithm 819: AIZ, BIZ: two Fortran 77 routines for the computation of complex Airy functions. ACM Trans. Math. Softw. **28**(3), 325–336 (2002)
3. Gil, A., Segura, J., Temme, N.M.: Numerical Methods for Special Functions. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2007)
4. Olver, F., Lozier, D., Boisvert, R., Clark, C.: NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge/New York (2010). http://dlmf.nist.gov

Splitting Methods

Sergio Blanes[1], Fernando Casas[2], and Ander Murua[3]
[1]Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, València, Spain
[2]Departament de Matemàtiques and IMAC, Universitat Jaume I, Castellón, Spain
[3]Konputazio Zientziak eta A.A. Saila, Informatika Fakultatea, UPV/EHU, Donostia/San Sebastián, Spain

Synonyms

Fractional step methods; Operator-splitting methods

Introduction

Splitting methods constitute a general class of numerical integration schemes for differential equations whose vector field can be decomposed in such a way that each subproblem is simpler to integrate than the original system. For ordinary differential equations (ODEs), this idea can be formulated as follows. Given the initial value problem

$$x' = f(x), \quad x_0 = x(0) \in \mathbb{R}^D \tag{1}$$

with $f : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ and solution $\varphi_t(x_0)$, assume that f can be expressed as $f = \sum_{i=1}^{m} f^{[i]}$ for certain functions $f^{[i]}$, such that the equations

$$x' = f^{[i]}(x), \quad x_0 = x(0) \in \mathbb{R}^D, \quad i = 1, \ldots, m \tag{2}$$

can be integrated exactly, with solutions $x(h) = \varphi_h^{[i]}(x_0)$ at $t = h$, the time step. The different parts of f may correspond to physically different contributions. Then, by combining these solutions as

$$\chi_h = \varphi_h^{[m]} \circ \cdots \circ \varphi_h^{[2]} \circ \varphi_h^{[1]} \tag{3}$$

and expanding into series in powers of h, one finds that $\chi_h(x_0) = \varphi_h(x_0) + \mathcal{O}(h^2)$, so that χ_h provides a first-order approximation to the exact solution. Higher-order approximations can be achieved by introducing more flows with additional coefficients, $\varphi_{a_{ij}h}^{[i]}$, in composition (3).

Splitting methods involve thus three steps: (i) choosing the set of functions $f^{[i]}$ such that $f = \sum_i f^{[i]}$, (ii) solving either exactly or approximately each equation $x' = f^{[i]}(x)$, and (iii) combining these solutions to construct an approximation for (1) up to the desired order.

The splitting idea can also be applied to partial differential equations (PDEs) involving time and one or more space dimensions. Thus, if the spatial differential operator contains parts of a different character (such as advection and diffusion), then different discretization techniques may be applied to each part, as well as for the time integration.

Splitting methods have a long history and have been applied (sometimes with different names) in many different fields, ranging from parabolic and reaction-diffusion PDEs to quantum statistical mechanics, chemical physics, and Hamiltonian dynamical systems [7].

Some of the advantages of splitting methods are the following: they are simple to implement, are explicit if each subproblem is solved with an explicit method, and often preserve qualitative properties the differential equation might possess.

Splitting Methods for ODEs

Increasing the Order

Very often in applications, the function f in the ODE (1) can be split in just two parts, $f(x) = f^{[a]}(x) + f^{[b]}(x)$. Then both $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$ and its adjoint, $\chi_h^* \equiv \chi_{-h}^{-1} = \varphi_h^{[a]} \circ \varphi_h^{[b]}$, are first-order integration schemes. These formulae are often called the Lie–Trotter splitting. On the other hand, the symmetric version

$$\mathcal{S}_h^{[2]} \equiv \varphi_{h/2}^{[a]} \circ \varphi_h^{[b]} \circ \varphi_{h/2}^{[a]} \tag{4}$$

provides a second-order integrator, known as the Strang–Marchuk splitting, the leapfrog, or the Störmer–Verlet method, depending on the context where it is used [2]. Notice that $\mathcal{S}_h^{[2]} = \chi_{h/2}^* \circ \chi_{h/2}$.

More generally, one may consider a composition of the form

$$\psi_h = \varphi_{a_{s+1}h}^{[a]} \circ \varphi_{b_s h}^{[b]} \circ \varphi_{a_s h}^{[a]} \circ \cdots \circ \varphi_{a_2 h}^{[a]} \circ \varphi_{b_1 h}^{[b]} \circ \varphi_{a_1 h}^{[a]} \tag{5}$$

and try to increase the order of approximation by suitably determining the parameters a_i, b_i. The number

s of $\varphi_h^{[b]}$ (or $\varphi_h^{[a]}$) evaluations in (5) is usually referred to as the number of stages of the integrator. This is called time-symmetric if $\psi_h = \psi_h^*$, in which case one has a left-right palindromic composition. Equivalently, in (5), one has

$$a_1 = a_{s+1}, \quad b_1 = b_s, \quad a_2 = a_s, \quad b_2 = b_{s-1}, \ldots \tag{6}$$

The order conditions the parameters a_i, b_i have to satisfy can be obtained by relating the previous integrator ψ_h with a formal series Ψ_h of differential operators [1]: it is known that the h-flow φ_h of the original system $x' = f^{[a]}(x) + f^{[b]}(x)$ satisfies, for each $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$, the identity $g(\varphi_h(x)) = \mathrm{e}^{h(F^{[a]}+F^{[b]})}[g](x)$, where $F^{[a]}$ and $F^{[b]}$ are the Lie derivatives corresponding to $f^{[a]}$ and $f^{[b]}$, respectively, acting as

$$F^{[a]}[g](x) = \sum_{j=1}^{D} f_j^{[a]}(x) \frac{\partial g}{\partial x_j}(x),$$
$$F^{[b]}[g](x) = \sum_{j=1}^{D} f_j^{[b]}(x) \frac{\partial g}{\partial x_j}(x). \tag{7}$$

Similarly, the approximation $\psi_h(x) \approx \varphi_h(x)$ given by the splitting method (5) satisfies the identity $g(\psi_h(x)) = \Psi(h)[g](x)$, where

$$\Psi(h) = \mathrm{e}^{a_1 h F^{[a]}}\, \mathrm{e}^{b_1 h F^{[b]}} \cdots \mathrm{e}^{a_s h F^{[a]}}\, \mathrm{e}^{b_s h F^{[b]}}\, \mathrm{e}^{a_{s+1} h F^{[a]}}. \tag{8}$$

Hence, the coefficients a_i, b_i must be chosen in such a way that the operator $\Psi(h)$ is a good approximation of $\mathrm{e}^{h(F^{[a]}+F^{[b]})}$, or equivalently, $h^{-1}\log(\Psi) \approx F^{[a]} + F^{[b]}$.

Applying repeatedly the Baker–Campbell–Hausdorff (BCH) formula [2], one arrives at

$$\begin{aligned}\frac{1}{h}\log(\Psi(h)) = {} & (v_a F^{[a]} + v_b F^{[b]}) + h v_{ab} F^{[ab]} \\ & + h^2 (v_{abb} F^{[abb]} + v_{aba} F^{[aba]}) \\ & + h^3 (v_{abbb} F^{[abbb]} + v_{abba} F^{[abba]} \\ & + v_{abaa} F^{[abaa]}) + \mathcal{O}(h^4),\end{aligned} \tag{9}$$

where

$$F^{[ab]} = [F^{[a]}, F^{[b]}], \quad F^{[abb]} = [F^{[ab]}, F^{[b]}],$$
$$F^{[aba]} = [F^{[ab]}, F^{[a]}],\ F^{[abbb]} = [F^{[abb]}, F^{[b]}],$$
$$F^{[abba]} = [F^{[abb]}, F^{[a]}],\ F^{[abaa]} = [F^{[aba]}, F^{[a]}],$$

the symbol $[\cdot,\cdot]$ stands for the Lie bracket, and $v_a, v_b, v_{ab}, v_{abb}, v_{aba}, v_{abbb}, \ldots$ are polynomials in the parameters a_i, b_i of the splitting scheme (5). In particular, one gets $v_a = \sum_{i=1}^{s+1} a_i$, $v_b = \sum_{i=1}^{s} b_i$, $v_{ab} = \frac{1}{2} - \sum_{i=1}^{s} b_i \sum_{j=1}^{i} a_j$. The order conditions then read $v_a = v_b = 1$ and $v_{ab} = v_{abb} = v_{aba} = \cdots = 0$ up to the order considered. To achieve order $r = 1, 2, 3, \ldots, 10$, the number of conditions to be fulfilled is $\sum_{j=1}^{r} n_j$, where $n_j = 2, 1, 2, 3, 6, 9, 18, 30, 56, 99$. This number is smaller for $r > 3$ when dealing with second-order ODEs of the form $y'' = g(y)$ when they are rewritten as (1) [1].

For time-symmetric methods, the order conditions at even orders are automatically satisfied, which leads to $n_1 + n_3 + \cdots + n_{2k-1}$ order conditions to achieve order $r = 2k$. For instance, $n_1 + n_3 = 4$ conditions need to be fulfilled for a symmetric method (5–6) to be of order 4.

Splitting and Composition Methods

When the original system (1) is split in $m > 2$ parts, higher-order schemes can be obtained by considering a composition of the basic first-order splitting method (3) and its adjoint $\chi_h^* = \varphi_h^{[1]} \circ \cdots \circ \varphi_h^{[m-1]} \circ \varphi_h^{[m]}$. More specifically, compositions of the general form

$$\psi_h = \chi_{\alpha_{2s} h}^* \circ \chi_{\alpha_{2s-1} h} \circ \cdots \circ \chi_{\alpha_2 h}^* \circ \chi_{\alpha_1 h}, \tag{10}$$

can be considered with appropriately chosen coefficients $(\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s}$ so as to achieve a prescribed order of approximation.

In the particular case when system (1) is split in $m = 2$ parts so that $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$, method (10) reduces to (5) with $a_1 = \alpha_1$ and

$$\begin{aligned} b_j = \alpha_{2j-1} + \alpha_{2j}, \quad & a_{j+1} = \alpha_{2j} + \alpha_{2j+1}, \\ \text{for} \quad & j = 1, \ldots, s, \end{aligned} \tag{11}$$

where $\alpha_{2s+1} = 0$. In that case, the coefficients a_i and b_i are such that

$$\sum_{i=1}^{s+1} a_i = \sum_{i=1}^{s} b_i. \tag{12}$$

Conversely, any splitting method (5) satisfying (12) can be written in the form (10) with $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$.

Moreover, compositions of the form (10) make sense for an arbitrary basic first-order integrator χ_h (and its adjoint χ_h^*) of the original system (1). Obviously, if the coefficients α_j of a composition method (10) are such that ψ_h is of order r for arbitrary basic integrators χ_h of (1), then the splitting method (5) with (11) is also of order r. Actually, as shown in [6], the integrator (5) is of order r for ODEs of the form (1) with $f = f^{[a]} + f^{[b]}$ if and only if the integrator (10) (with coefficients α_j obtained from (11)) is of order r for arbitrary first-order integrators χ_h.

This close relationship allows one to establish in an elementary way a defining feature of splitting methods (5) of order $r \geq 3$: at least one a_i and one b_i are necessarily negative [1]. In other words, splitting schemes of order $r \geq 3$ always involve backward fractional time steps.

Preserving Properties

Assume that the individual flows $\varphi_h^{[i]}$ share with the exact flow φ_h some defining property which is preserved by composition. Then it is clear that any composition of the form (5) and (10) with χ_h given by (3) also possesses this property. Examples of such features are symplecticity, unitarity, volume preservation, conservation of first integrals, etc. [7]. In this sense, splitting methods form an important class of geometric numerical integrators [2]. Repeated application of the BCH formula can be used (see (9)) to show that there exists a modified (formal) differential equation

$$\tilde{x}' = f_h(\tilde{x}) \equiv f(\tilde{x}) + h f_2(\tilde{x}) + h^2 f_3(\tilde{x}) + \cdots, \\ \tilde{x}(0) = x_0, \tag{13}$$

associated to any splitting method ψ_h such that the numerical solution $x_n = \psi_h(x_{n-1})$ $(n = 1, 2, \ldots)$ satisfies $x_n = \tilde{x}(nh)$ for the exact solution $\tilde{x}(t)$ of (13). An important observation is that the vector fields f_k in (13) belong to the Lie algebra generated by $f^{[1]}, \ldots, f^{[m]}$. In the particular case of autonomous Hamiltonian systems, if $f^{[i]}$ are Hamiltonian, then each f_k is also Hamiltonian. Then one may study the long-time behavior of the numerical integrator by analyzing the solutions of (13) viewed as a small perturbation of the original system (1) and obtain rigorous statements with techniques of backward error analysis [2].

Further Extensions

Several extensions can be considered to reduce the number of stages necessary to achieve a given order and get more efficient methods. One of them is the use of a processor or corrector. The idea consists in enhancing an integrator ψ_h (the kernel) with a map π_h (the processor) as $\hat{\psi}_h = \pi_h \circ \psi_h \circ \pi_h^{-1}$. Then, after n steps, one has $\hat{\psi}_h^n = \pi_h \circ \psi_h^n \circ \pi_h^{-1}$, and so only the cost of ψ_h is relevant. The simplest example of a processed integrator is provided by the Störmer–Verlet method (4). In that case, $\psi_h = \chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$ and $\pi_h = \varphi_{h/2}^{[a]}$. The use of processing allows one to get methods with fewer stages in the kernel and smaller error terms than standard compositions [1].

The second extension uses the flows corresponding to other vector fields in addition to $F^{[a]}$ and $F^{[b]}$. For instance, one could consider methods (5) such that, in addition to $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$, use the h-flow $\varphi_h^{[abb]}$ of the vector field $F^{[abb]}$ when its computation is straightforward. This happens, for instance, for second-order ODEs $y'' = g(y)$ [1, 7].

Splitting is particularly appropriate when $\|f^{[a]}\| \ll \|f^{[b]}\|$ in (1). Introducing a small parameter ε, we can write $x' = \varepsilon f^{[a]}(x) + f^{[b]}(x)$, so that the error of scheme (5) is $\mathcal{O}(\varepsilon)$. Moreover, since in many practical applications $\varepsilon < h$, one is mainly interested in eliminating error terms with small powers of ε instead of satisfying all the order conditions. In this way, it is possible to get more efficient schemes. In addition, the use of a processor allows one to eliminate the errors of order εh^k for all $1 < k < n$ and all n [7].

Although only autonomous differential equations have been considered here, several strategies exist for adapting splitting methods also to nonautonomous systems without deteriorating their overall efficiency [1].

S

Some Good Fourth-Order Splitting Methods

In the following table, we collect the coefficients of a few selected fourth-order symmetric methods of the form (5–6). Higher-order and more elaborated schemes can be found in [1, 2, 7] and references therein. They are denoted as X$_s$4, where s indicates the number of stages. S$_6$4 is a general splitting method, whereas SN$_6$4 refers to a method tailored for second-order ODEs of the form $y'' = g(y)$ when they are rewritten as a first-order system (1), and the coefficients a_i are associated to $g(y)$. Finally, SNI$_5$4 is a method especially designed for problems of the form $x' = \varepsilon f^{[a]}(x) + f^{[b]}(x)$. With $s = 3$ stages, there is only one solution, S$_3$4, given by $a_1 = b_1/2$, $b_1 = 2/(2 - 2^{1/3})$. In all cases, the remaining coefficients are fixed by symmetry and consistency ($\sum_i a_i = \sum_i b_i = 1$).

S$_6$4	$a_1 = 0.0792036964311\,9565$	$b_1 = 0.209515106613362$
	$a_2 = 0.353172906049774$	$b_2 = -0.143851773179818$
	$a_3 = -0.0420650803577\,1952$	
SN$_6$4	$a_1 = 0.0829844064174\,0515$	$b_1 = 0.245298957184271$
	$a_2 = 0.396309801498368$	$b_2 = 0.604872665711080$
	$a_3 = -0.3905630492234859$	
SNI$_5$4	$a_1 = 0.81186273854451628884$	$b_1 = -0.0075869131187744738$
	$a_2 = -0.67748039953216912289$	$b_2 = 0.31721827797316981388$

Numerical Example: A Perturbed Kepler Problem

To illustrate the performance of the previous splitting methods, we apply them to the time integration of the perturbed Kepler problem described by the Hamiltonian

$$H = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{r} - \frac{\varepsilon}{2r^5}\left(q_2^2 - 2q_1^2\right), \qquad (14)$$

where $r = \sqrt{q_1^2 + q_2^2}$. We take $\varepsilon = 0.001$ and integrate the equations of motion $q_i' = p_i$, $p_i' = -\partial H/\partial q_i$, $i = 1, 2$, with initial conditions $q_1 = 4/5$, $q_2 = p_1 = 0$, $p_2 = \sqrt{3/2}$. Splitting methods are used with the partition into kinetic and potential energy. We measure the two-norm error in the position at $t_f = 2{,}000$, $(q_1, q_2) = (0.318965403761932, 1.15731646810481)$, for different time steps and plot the corresponding error as a function of the number of evaluations for each method in Fig. 1. Notice that although the generic method S$_6$4 has three more stages than the minimum given by S$_3$4, this extra cost is greatly compensated by a higher accuracy. On the other hand, since this system corresponds to the second-order ODE $q'' = g(q)$, method SN$_6$4 leads to a higher accuracy with the same computational cost. Finally, SNI$_5$4 takes profit of the near-integrable character of the Hamiltonian (14) and the two extra stages to achieve an even higher efficiency. It requires solving the Kepler problem separately from the perturbation. This requires a more elaborated algorithm with a slightly increase in the computational cost (not reflected in the figure). Results provided by the leapfrog method S2 and the standard fourth-order Runge–Kutta integrator RK4 are also included for reference.

Splitting Methods for PDEs

In the numerical treatment of evolutionary PDEs of parabolic or mixed hyperbolic-parabolic type, splitting time-integration methods are also widely used. In this setting, the overall evolution operator is formally written as a sum of evolution operators, typically representing different aspects of a given model. Consider an evolutionary PDE formulated as an abstract Cauchy problem in a certain function space $\mathcal{U} \subset \{u : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}\}$,

$$u_t = L(u), \quad u(t_0) = u_0, \qquad (15)$$

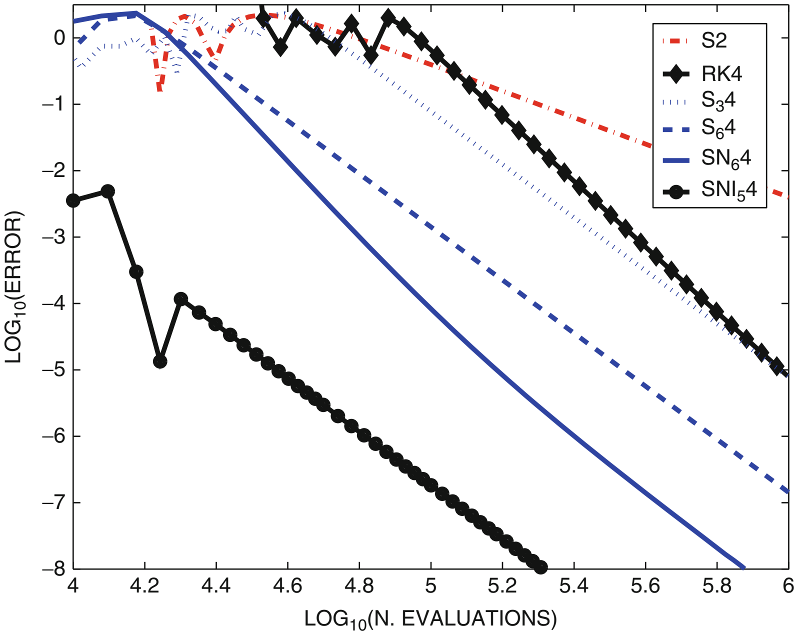
where L is a spatial partial differential operator. For instance,

$$\frac{\partial}{\partial t}u(x,t) = \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\left(\sum_{i=1}^{d} c_i(x)\frac{\partial}{\partial x_i}u(x,t)\right) + f(x, u(x,t)), \quad u(x,t_0) = u_0(x)$$

or in short, $L(x,u) = \nabla \cdot (c\nabla u) + f(u)$ corresponds to a diffusion-reaction problem. In that case, it makes sense to split the problem into two subequations, corresponding to the different physical contributions,

$$u_t = L_a(u) \equiv \nabla \cdot (c\nabla u), \quad u_t = L_b(u) \equiv f(u), \qquad (16)$$

solve numerically each equation in (16), thus giving $u^{[a]}(h) = \varphi_h^{[a]}(u_0)$, $u^{[b]}(h) = \varphi_h^{[b]}(u_0)$, respectively, for a time step h, and then compose the operators $\varphi_h^{[a]}$, $\varphi_h^{[b]}$ to construct an approximation to the solution of (15). Thus, $u(h) \approx \varphi_h^{[b]}(\varphi_h^{[a]}(u_0))$ provides a first-order approximation, whereas the Strang splitting $u(h) \approx \varphi_{h/2}^{[a]}(\varphi_h^{[b]}(\varphi_{h/2}^{[a]}(u_0)))$ is formally second-order accurate for sufficiently smooth solutions. In this way, especially adapted numerical methods can be used to integrate each subproblem, even in parallel [3, 4].



Splitting Methods, Fig. 1 Error in the solution $(q_1(t_f), q_2(t_f))$ vs. the number of evaluations for different fourth-order splitting methods (the extra cost in the method $\mathrm{SNI}_5 4$, designed for perturbed problems, is not taken into account)

Systems of hyperbolic conservation laws, such as

$$u_t + f(u)_x + g(u)_x = 0, \quad u(x, t_0) = u_0(x),$$

can also be treated with splitting methods, in this case, by fixing a step size h and applying a especially tailored numerical scheme to each scalar conservation law $u_t + f(u)_x = 0$ and $u_t + g(u)_x = 0$. This is a particular example of dimensional splitting where the original problem is approximated by solving one space direction at a time. Early examples of dimensional splitting are the so-called locally one-dimensional (LOD) methods (such as LOD-backward Euler and LOD Crank–Nicolson schemes) and alternating direction implicit (ADI) methods (e.g., the Peaceman–Rachford algorithm) [4].

Although the formal analysis of splitting methods in this setting can also be carried out by power series expansions, several fundamental difficulties arise, however. First, nonlinear PDEs in general possess solutions that exhibit complex behavior in small regions of space and time, such as sharp transitions and discontinuities. Second, even if the exact solution of the original problem is smooth, it might happen that the composition defining the splitting method provides nonsmooth approximations. Therefore, it is necessary to develop sophisticated tools to analyze whether the numerical solution constructed with a splitting method leads to the correct solution of the original problem or not [3].

On the other hand, even if the solution is sufficiently smooth, applying splitting methods of order higher than two is not possible for certain problems. This happens, in particular, when there is a diffusion term in the equation; since then the presence of negative coefficients in the method leads to an ill-posed problem. When $c = 1$ in (16), this order barrier has been circumvented, however, with the use of complex-valued coefficients with positive real parts: the operator $\varphi_{zh}^{[a]}$ corresponding to the Laplacian L_a is still well defined in a reasonable distribution set for $z \in \mathbb{C}$, provided that $\Re(z) \geq 0$.

There exist also relevant problems where high-order splitting methods can be safely used as is in the integration of the time-dependent Schrödinger equation $iu_t = -\frac{1}{2m}\Delta u + V(x)u$ split into kinetic $T = -(2m)^{-1}\Delta$ and potential V energy operators and with periodic boundary conditions. In this case, the combination of the Strang splitting in time and the Fourier collocation in space is quite popular in chemical physics (with the name of split-step Fourier method). These schemes have appealing structure-preserving properties, such as unitarity, symplecticity, and time-symmetry [5]. Moreover, it has been shown that for a method (5) of order r with the splitting into kinetic and potential energy and under relatively mild assumptions on T and V, one has an rth-order error bound $\|\psi_h^n u_0 - u(nh)\| \le Cnh^{r+1} \max_{0\le s\le nh} \|u(s)\|_r$ in terms of the rth-order Sobolev norm [5].

Cross-References

▶ Composition Methods
▶ One-Step Methods, Order, Convergence
▶ Symmetric Methods
▶ Symplectic Methods

References

1. Blanes, S., Casas, F., Murua, A.: Splitting and composition methods in the numerical integration of differential equations. Bol. Soc. Esp. Mat. Apl. **45**, 89–145 (2008)
2. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn, Springer, Berlin (2006)
3. Holden, H., Karlsen, K.H., Lie, K.A., Risebro, N.H.: Splitting Methods for Partial Differential Equations with Rough Solutions. European Mathematical Society, Zürich (2010)
4. Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Springer, Berlin (2003)
5. Lubich, C.: From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis. European Mathematical Society, Zürich (2008)
6. McLachlan, R.I.: On the numerical integration of ordinary differential equations by symmetric composition methods. SIAM J. Numer. Anal. **16**, 151–168 (1995)
7. McLachlan, R.I., Quispel, R.: Splitting methods. Acta Numer. **11**, 341–434 (2002)

Stability, Consistency, and Convergence of Numerical Discretizations

Douglas N. Arnold
School of Mathematics, University of Minnesota,
Minneapolis, MN, USA

Overview

A problem in differential equations can rarely be solved analytically, and so often is discretized, resulting in a discrete problem which can be solved in a finite sequence of algebraic operations, efficiently implementable on a computer. The *error* in a discretization is the difference between the solution of the original problem and the solution of the discrete problem, which must be defined so that the difference makes sense and can be quantified. *Consistency* of a discretization refers to a quantitative measure of the extent to which the exact solution satisfies the discrete problem. *Stability* of a discretization refers to a quantitative measure of the well-posedness of the discrete problem. A fundamental result in numerical analysis is that *the error of a discretization may be bounded in terms of its consistency and stability.*

A Framework for Assessing Discretizations

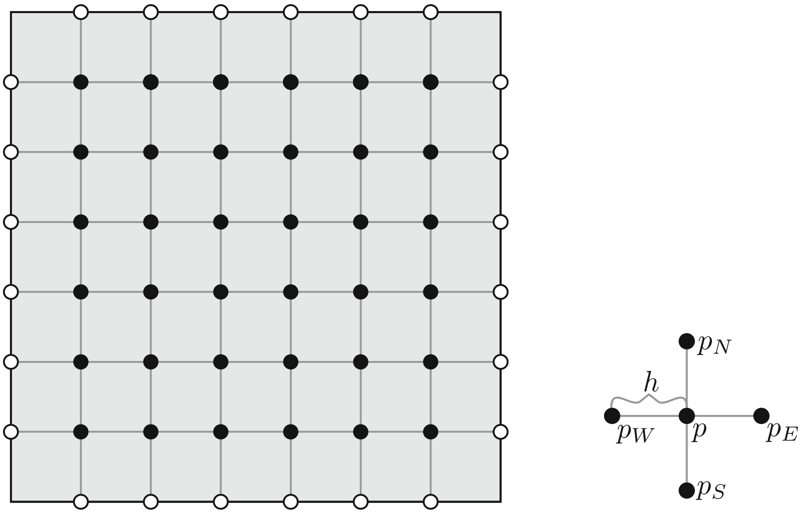
Many different approaches are used to discretize differential equations: finite differences, finite elements, spectral methods, integral equation approaches, etc. Despite the diversity of methods, fundamental concepts such as error, consistency, and stability are relevant to all of them. Here, we describe a framework general enough to encompass all these methods, although we do restrict to linear problems to avoid many complications. To understand the definitions, it is good to keep some concrete examples in mind, and so we start with two of these.

A Finite Difference Method

As a first example, consider the solution of the Poisson equation, $\Delta u = f$, on a domain $\Omega \subset \mathbb{R}^2$, subject to the Dirichlet boundary condition $u = 0$ on $\partial\Omega$. One possible discretization is a finite difference method, which we describe in the case $\Omega = (0,1)\times(0,1)$ is the unit square. Making reference to Fig. 1, let $h = 1/n$, $n > 1$ integer, be the grid size, and define the grid domain, $\Omega_h = \{(lh, mh) \,|\, 0 < l, m < n\}$, as the set of grid points in Ω. The nearest neighbors of a grid point $p = (p_1, p_2)$ are the four grid points $p_W = (p_1 - h, p_2)$, $p_E = (p_1 + h, p_2)$, $p_S = (p_1, p_2 - h)$, and $p_N = (p_1, p_2 + h)$. The grid points which do not themselves belong to Ω, but which have a nearest neighbor in Ω constitute the grid boundary, $\partial\Omega_h$, and we set $\bar{\Omega}_h = \Omega_h \cup \partial\Omega_h$. Now let $v : \bar{\Omega}_h \to \mathbb{R}$ be a grid function. Its five-point Laplacian $\Delta_h v$ is defined by

$$\Delta_h v(p) = \frac{v(p_E)+v(p_W)+v(p_S)+v(p_N)-4v(p)}{h^2}, \qquad p \in \Omega_h.$$

The finite difference discretization then seeks $u_h : \bar{\Omega}_h \to \mathbb{R}$ satisfying



Stability, Consistency, and Convergence of Numerical Discretizations, Fig. 1 The grid domain $\bar{\Omega}_h$ consists of the points in Ω_h, marked with *solid dots*, and in $\partial\Omega_h$, marked with *hollow dots*. On the *right* is the stencil of the five-point Laplacian, which consists of a grid point p and its four nearest neighbors

$$\Delta_h u_h(p) = f(p),\ p \in \Omega_h, \quad u_h(p) = 0,\ p \in \partial\Omega_h.$$

If we regard as unknowns, the $N = (n-1)^2$ values $u_h(p)$ for $p \in \Omega_h$, this gives us a systems of N linear equations in N unknowns which may be solved very efficiently.

A Finite Element Method

A second example of a discretization is provided by a finite element solution of the same problem. In this case we assume that Ω is a polygon furnished with a triangulation $\mathcal{T}_h$, such as pictured in Fig. 2. The finite element method seeks a function $u_h : \Omega \to \mathbb{R}$ which is continuous and piecewise linear with respect to the mesh and vanishing on $\partial\Omega$, and which satisfies

$$-\int_\Omega \nabla u_h \cdot \nabla v\, dx = \int_\Omega f v\, dx,$$

for all test functions v which are themselves continuous and piecewise linear with respect to the mesh and vanish on $\partial\Omega$. If we choose a basis for this set of space of test functions, then the computation of u_h may be reduced to an efficiently solvable system of N linear equations in N unknowns, where, in this case, N is the number of interior vertices in the triangulation.

Discretization

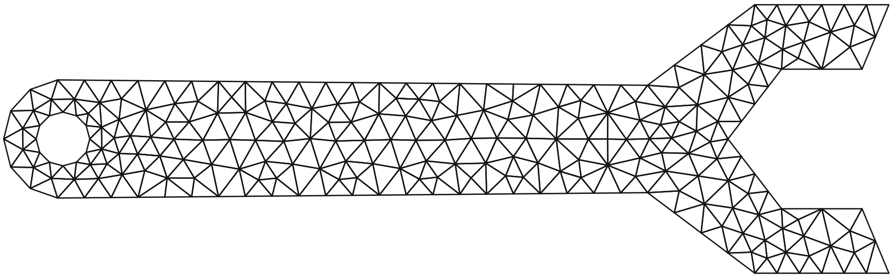
We may treat both these examples, and many other discretizations, in a common framework. We regard the discrete operator as a linear map L_h from a vector space V_h, called the discrete solution space, to a second vector space W_h, called the discrete data space. In the case of the finite difference operator, the discrete solution space is the space of mesh functions on $\bar{\Omega}_h$ which vanish on $\partial\Omega_h$, the discrete data space is the space of mesh functions on Ω_h, and the discrete operator $L_h = \Delta_h$, the five-point Laplacian. In the case of the finite element method, V_h is the space of continuous piecewise linear functions with respect to the given triangulation that vanish on $\partial\Omega$, and $W_h = V_h^*$, the dual space of V_h. The operator L_h is given by

$$(L_h w)(v) = -\int_\Omega \nabla w \cdot \nabla v\, dx, \quad w, v \in V_h.$$

For the finite difference method, we define the discrete data $f_h \in W_h$ by $f_h = f|_{\Omega_h}$, while for the finite element method $f_h \in W_h$ is given by $f_h(v) = \int f v\, dx$. In both cases, the discrete solution $u_h \in V_h$ is found by solving the discrete equation

$$L_h u_h = f_h. \tag{1}$$

Of course, a minimal requirement on the discretization is that the finite dimensional linear system (1) has a unique solution, i.e., that the associated matrix is invertible (so V_h and W_h must have the same dimension). Then, the discrete solution u_h is well-defined. The primary goal of numerical analysis is to ensure that the discrete solution is a good approximation of the true solution u in an appropriate sense.



Stability, Consistency, and Convergence of Numerical Discretizations, Fig. 2 A finite element mesh of the domain Ω. The solution is sought as a piecewise linear function with respect to the mesh

Representative and Error

Since we are interested in the difference between u and u_h, we must bring these into a common vector space, where the difference makes sense. To this end, we suppose that a *representative* $U_h \in V_h$ of u is given. The representative is taken to be an element of V_h which, though not practically computable, is a good approximation of u. For the finite difference method, a natural choice of representative is the grid function $U_h = u|_{\Omega_h}$. If we show that the difference $U_h - u_h$ is small, we know that the grid values $u_h(p)$ which determine the discrete solution are close to the exact values $u(p)$. For the finite element method, a good possibility for U_h is the piecewise linear interpolant of u, that is, U_h is the piecewise linear function that coincides with u at each vertex of the triangulation. Another popular possibility is to take U_h to be the best approximation of u in V_h in an appropriate norm. In any case, the quantity $U_h - u_h$, which is the difference between the representative of the true solution and the discrete solution, defines the *error* of the discretization.

At this point we have made our goal more concrete: we wish to ensure that the error, $U_h - u_h \in V_h$, is small. To render this quantitative, we need to select a norm on the finite dimensional vector space V_h with which to measure the error. The choice of norm is an important aspect of the problem presentation, and an appropriate choice must reflect the goal of the computation. For example, in some applications, a large error at a single point of the domain could be catastrophic, while in others only the average error over the domain is significant. In yet other cases, derivatives of u are the true quantities of interest. These cases would lead to different choices of norms. We shall denote the chosen norm of $v \in V_h$ by $\|v\|_h$. Thus, we now have a quantitative goal for our computation that the error $\|U_h - u_h\|_h$ be sufficiently small.

Consistency and Stability

Consistency Error

Having used the representative U_h of the solution to define the error, we also use it to define a second sort of error, the *consistency error*, also sometimes called the truncation error. The consistency error is defined to be $L_h U_h - f_h$, which is an element of W_h. Now U_h represents the true solution u, so the consistency error should be understood as a quantity measuring the extent to which the true solution satisfies the discrete equation (1). Since $Lu = f$, the consistency error should be small if L_h is a good representative of L and f_h a good representative of f. In order to relate the norm of the error to the consistency error, we need a norm on the discrete data space W_h as well. We denote this norm by $\|w\|'_h$ for $w \in W_h$ and so our measure of the consistency error is $\|L_h U_h - f_h\|'_h$.

Stability

If a problem in differential equations is well-posed, then, by definition, the solution u depends continuously on the data f. On the discrete level, this continuous dependence is called *stability*. Thus, stability refers to the continuity of the mapping $L_h^{-1} : W_h \to V_h$, which takes the discrete data f_h to the discrete solution u_h. Stability is a matter of degree, and an unstable discretization is one for which the modulus of continuity of L_h^{-1} is very large.

To illustrate the notion of instability, and to motivate the quantitative measure of stability we shall introduce below, we consider a simpler numerical problem than the discretization of a differential equation. Suppose we wish to compute the definite integral

$$\gamma_{n+1} = \int_0^1 x^n e^{x-1}\,dx, \tag{2}$$

for $n = 15$. Using integration by parts, we obtain a simple recipe to compute the integral in short sequence of arithmetic operations:

$$\begin{aligned}\gamma_{n+1} &= 1 - n\gamma_n,\ n = 1, \dots, 15,\\ \gamma_1 &= 1 - e^{-1} = 0.632121\dots.\end{aligned} \tag{3}$$

Now suppose we carry out this computation, beginning with $\gamma_1 = 0.632121$ (so truncated after six decimal places). We then find that $\gamma_{16} = -576,909$, which is truly a massive error, since the correct value is $\gamma_{16} = 0.0590175\dots$. If we think of (3) as a discrete solution operator (analogous to L_h^{-1} above) taking the data γ_1 to the solution γ_{16}, then it is a highly unstable scheme: a perturbation of the data of less than 10^{-6} leads to a change in the solution of nearly 6×10^5. In fact, it is easy to see that for (3), a perturbation ϵ in the data leads to an error of $15! \times \epsilon$ in solution – a huge instability. It is important to note that the numerical computation of the integral (2) is not a difficult numerical problem. It could be easily computed with Simpson's rule, for example. The crime here is solving the problem with the unstable algorithm (3).

Returning to the case of the discretization (1), imagine that we perturb the discrete data f_h to some $\tilde{f}_h = f_h + \epsilon_h$, resulting in a perturbation of the discrete solution to $\tilde{u}_h = L_h^{-1}\tilde{f}_h$. Using the norms in W_h and V_h to measure the perturbations and then computing the ratio, we obtain

$$\frac{\text{solution perturbation}}{\text{data perturbation}} = \frac{\|\tilde{u}_h - u_h\|_h}{\|\tilde{f}_h - f_h\|_h'} = \frac{\|L_h^{-1}\epsilon_h\|_h}{\|\epsilon_h\|_h'}.$$

We define the *stability constant* C_h^{stab}, which is our quantitative measure of stability, as the maximum value this ratio achieves for any perturbation ϵ_h of the data. In other words, the stability constant is the norm of the operator L_h^{-1}:

$$C_h^{\text{stab}} = \sup_{0 \neq \epsilon_h \in W_h} \frac{\|L_h^{-1}\epsilon_h\|_h}{\|\epsilon_h\|_h'} = \|L_h^{-1}\|_{\mathcal{L}(W_h,V_h)}.$$

Relating Consistency, Stability, and Error

The Fundamental Error Bound

Let us summarize the ingredients we have introduced in our framework to assess a discretization:

- The discrete solution space, V_h, a finite dimensional vector space, normed by $\|\cdot\|_h$
- The discrete data space, W_h, a finite dimensional vector space, normed by $\|\cdot\|_h'$
- The discrete operator, $L_h : V_h \to W_h$, an invertible linear operator
- The discrete data $f_h \in W_h$
- The discrete solution u_h determined by the equation $L_h u_h = f_h$
- The solution representative $U_h \in V_h$
- The error $U_h - u_h \in V_h$
- The consistency error $L_h U_h - f_h \in W_h$
- The stability constant $C_h^{\text{stab}} = \|L_h^{-1}\|_{\mathcal{L}(W_h,V_h)}$

With this framework in place, we may prove a rigorous error bound, stating that the error is bounded by the product of the stability constant and the consistency error:

$$\|U_h - u_h\|_h \le C_h^{\text{stab}} \|L_h U_h - f_h\|_h'. \tag{4}$$

The proof is straightforward. Since L_h is invertible,

$$\begin{aligned}U_h - u_h &= L_h^{-1}[L_h(U_h - u_h)] = L_h^{-1}(L_h U_h - L_h u_h)\\ &= L_h^{-1}(L_h U_h - f_h).\end{aligned}$$

Taking norms, gives

$$\|U_h - u_h\|_h \le \|L_h^{-1}\|_{\mathcal{L}(W_h,V_h)} \|L_h U_h - f_h\|_h',$$

as claimed.

The Fundamental Theorem

A discretization of a differential equation always entails a certain amount of error. If the error is not small enough for the needs of the application, one generally refines the discretization, for example, using a finer grid size in a finite difference method or a triangulation with smaller elements in a finite element method. Thus, we may consider a whole sequence or family of discretizations, corresponding to finer and finer grids or triangulations or whatever. It is conventional to parametrize these by a positive real number h called the discretization parameter. For example, in the finite difference method, we may use the same h as before, the grid size, and in the finite element method, we can take h to be the maximal triangle diameter or something related to it. We shall call such a family of discretizations a discretization scheme. The scheme is called *convergent* if the error norm $\|U_h - u_h\|_h$ tends to 0 as h tends to 0. Clearly convergence is a highly desirable property: it means that we can achieve whatever level of accuracy we need, as long as we do a fine enough computation. Two more definitions apply to a discretization scheme. The scheme is *consistent* if the consistency error norm $\|L_h U_h - f_h\|'_h$ tends to 0 with h. The scheme is *stable* if the stability constant C_h^{stab} is bounded uniformly in h: $C_h^{\text{stab}} \leq C^{\text{stab}}$ for some number C^{stab} and all h. From the fundamental error bound, we immediately obtain what may be called the fundamental theorem of numerical analysis: *a discretization scheme which is consistent and stable is convergent.*

Historical Perspective

Consistency essentially requires that the discrete equations defining the approximate solution are at least approximately satisfied by the true solution. This is an evident requirement and has implicitly guided the construction of virtually all discretization methods, from the earliest examples. Bounds on the consistency error are often not difficult to obtain. For finite difference methods, for example, they may be derived from Taylor's theorem, and, for finite element methods, from simple approximation theory. Stability is another matter. Its central role was not understood until the mid-twentieth century, and there are still many differential equations for which it is difficult to devise or to assess stable methods.

That consistency alone is insufficient for the convergence of a finite difference method was pointed out in a seminal paper of Courant, Friedrichs, and Lewy [2] in 1928. They considered the one-dimensional wave equation and used a finite difference method, analogous to the five-point Laplacian, with a space-time grid of points (jh, lk) with $0 \leq j \leq n$, $0 \leq l \leq m$ integers and $h, k > 0$ giving the spatial and temporal grid size, respectively. It is easy to bound the consistency error by $O(h^2 + k^2)$, so setting $k = \lambda h$ for some constant $\lambda > 0$ and letting h tend to 0, one obtains a consistent scheme. However, by comparing the domains of dependence of the true solution and of the discrete solution on the initial data, one sees that this method, though consistent, cannot be convergent if $\lambda > 1$.

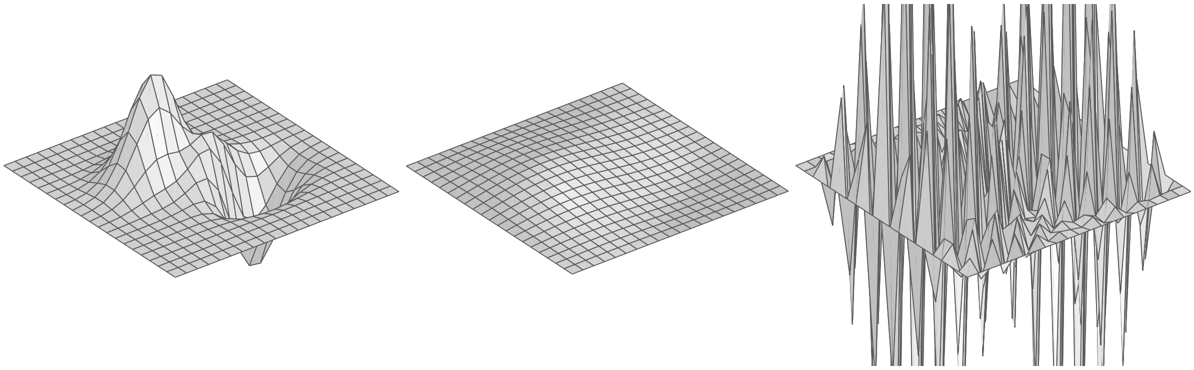
Twenty years later, the property of stability of discretizations began to emerge in the work of von Neumann and his collaborators. First, in von Neumann's work with Goldstine on solving systems of linear equations [5], they studied the magnification of round-off error by the repeated algebraic operations involved, somewhat like the simple example (3) of an unstable recursion considered above. A few years later, in a 1950 article with Charney and Fjørtoft [1] on numerical solution of a convection diffusion equation arising in atmospheric modeling, the authors clearly highlighted the importance of what they called computational stability of the finite difference equations, and they used Fourier analysis techniques to assess the stability of their method. This approach developed into von Neumann stability analysis, still one of the most widely used techniques for determining stability of finite difference methods for evolution equations.

During the 1950s, there was a great deal of study of the nature of stability of finite difference equations for initial value problems, achieving its capstone in the 1956 survey paper [3] of Lax and Richtmeyer. In that context, they formulated the definition of stability given above and proved that, for a consistent difference approximation, stability ensured convergence.

Techniques for Ensuring Stability

Finite Difference Methods

We first consider an initial value problem, for example, the heat equation or wave equation, discretized by a finite difference method using grid size h and time step k. The finite difference method advances the solution from some initial time t_0 to a terminal time T by a sequence of steps, with the lth step advancing the discrete solution from time $(l-1)k$ to time lk. At each time level lk, the discrete solution is a spatial grid function u_h^l, and so the finite difference method defines an operator $G(h, k)$ mapping u_h^{l-1} to u_h^l, called the *amplification matrix*. Since the amplification matrix is applied many times in the course of the calculation ($m = (T - t_0)/k$ times to be precise, a number which tends to infinity as k tends to 0), the solution at the final step u_h^m involves a high power of the amplification matrix, namely $G(h, k)^m$, applied to the



Stability, Consistency, and Convergence of Numerical Discretizations, Fig. 3 Finite difference solution of the heat equation using (5). *Left*: initial data. *Middle*: discrete solution at $t = 0.03$ computed with $h = 1/20$, $k = 1/2{,}000$ (stable). *Right*: same computation with $k = 1/1{,}000$ (unstable)

data u_h^0. Therefore, the stability constant will depend on a bound for $\|G(h,k)^m\|$. Usually this can only be obtained by showing that $\|G(h,k)\| \le 1$ or, at most, $\|G(h,k)\| \le 1 + O(k)$. As a simple example, we may consider an initial value problem for the heat equation with homogeneous boundary conditions on the unit square:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Delta u, \quad x \in \Omega,\ 0 < t \le T, \\ u(x,t) &= 0, \quad x \in \partial\Omega,\ 0 < t \le T, \\ u(x,0) &= u_0(x), \quad x \in \Omega, \end{aligned}$$

which we discretize with the five-point Laplacian and forward differences in time:

$$\frac{u^l(p) - u^{l-1}(p)}{k} = \Delta_h u^{l-1}(p),\ p \in \Omega_h, \quad 0 < l \le m, \tag{5}$$

$$\begin{aligned} u^l(p) &= 0, \quad p \in \partial\Omega_h,\ 0 < l \le m, \\ u^0(p) &= u_0(p), \quad p \in \Omega_h. \end{aligned} \tag{6}$$

In this case the norm condition on the amplification matrix $\|G(h,k)\| \le 1$ holds if $4k \le h^2$, but not otherwise, and, indeed, it can be shown that this discretization scheme is stable, if and only if that condition is satisfied. Figure 3 illustrates the tremendous difference between a stable and unstable choice of time step.

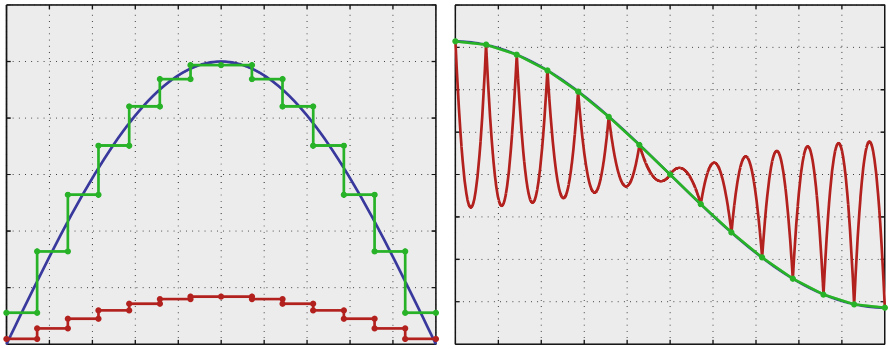
Several methods are used to bound the norm of the amplification matrix. If an L^∞ norm is chosen, one can often use a discrete maximum principle based on the structure of the matrix. If an L^2 norm is chosen, then Fourier analysis may be used if the problem has constant coefficients and simple enough boundary conditions. In other circumstances, more sophisticated matrix or eigenvalue analysis is used.

For time-independent PDEs, such as the Poisson equation, the requirement is to show that the inverse of the discretization operator is bounded uniformly in the grid size h. Similar techniques as for the time-dependent problems are applied.

Galerkin Methods

Galerkin methods, of which finite element methods are an important case, treat a problem which can be put into the form: find $u \in V$ such that $B(u,v) = F(v)$ for all $v \in V$. Here, V is a Hilbert space, $B : V \times V \to \mathbb{R}$ is a bounded bilinear form, and $F \in V^*$, the dual space of V. (Many generalizations are possible, e.g., to the case where B acts on two different Hilbert spaces or the case of Banach spaces.) This problem is equivalent to a problem in operator form, find u such $Lu = F$, where the operator $L : V \to V^*$ is defined by $Lu(v) = B(u,v)$. An example is the Dirichlet problem for the Poisson equation considered earlier. Then, $V = \overset{\circ}{H}{}^1(\Omega)$, $B(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$, and $F(v) = \int_\Omega f v \, dx$. The operator is $L = -\Delta : \overset{\circ}{H}{}^1(\Omega) \to \overset{\circ}{H}{}^1(\Omega)^*$.

A Galerkin method is a discretization which seeks u_h in a subspace V_h of V satisfying $B(u_h, v) = F(v)$ for all $v \in V_h$. The finite element method discussed



Stability, Consistency, and Convergence of Numerical Discretizations, Fig. 4 Approximation of the problem (7), with $u = \cos \pi x$ shown on *left* and $\sigma = u'$ on the *right*. The exact solution is shown in *blue*, and the stable finite element method, using piecewise linears for σ and piecewise constants for u, is shown in *green* (in the *right* plot, the *blue curve* essentially coincides with the *green curve*, and so is not visible). An unstable finite element method, using piecewise quadratics for σ, is shown in *red*

above took V_h to be the subspace of continuous piecewise linears. If the bilinear form B is coercive in the sense that there exists a constant $\gamma > 0$ for which

$$B(v, v) \geq \gamma \|v\|_V^2, \quad v \in V,$$

then stability of the Galerkin method with respect to the V norm is automatic. No matter how the subspace V_h is chosen, the stability constant is bounded by $1/\gamma$. If the bilinear form is not coercive (or if we consider a norm other than the norm in which the bilinear form is coercive), then finding stable subspaces for Galerkin's method may be quite difficult. As a very simple example, consider a problem on the unit interval $I = (0, 1)$, to find $(\sigma, u) \in H^1(I) \times L^2(I)$ such that

$$\int_0^1 \sigma\tau\,dx + \int_0^1 \tau' u\,dx + \int_0^1 \sigma' v\,dx = \int_0^1 f v\,dx, \quad (\tau, v) \in H^1(I) \times L^2(I). \tag{7}$$

This is a weak formulation of system $\sigma = u'$, $\sigma' = f$, with Dirichlet boundary conditions (which arise from this weak formulation as natural boundary conditions), so this is another form of the Dirichlet problem for Poisson's equation $u'' = f$ on I, $u(0) = u(1) = 0$. In higher dimensions, there are circumstances where such a first-order formulation is preferable to a standard second-order form. This problem can be discretized by a Galerkin method, based on subspaces $S_h \subset H^1(I)$ and $W_h \subset L^2(I)$. However, the choice of subspaces is delicate, even in this one-dimensional context. If we partition I into subintervals and choose S_h and W_h both to be the space of continuous piecewise linears, then the resulting matrix problem is *singular*, so the method is unusable. If we choose S_h to continuous piecewise linears, and W_h to be piecewise constants, we obtain a stable method. But if we choose S_h to contain all continuous piecewise quadratic functions and retain the space of piecewise constants for W_h, we obtain an unstable scheme. The stable and unstable methods can be compared in Fig. 4. For the same problem of the Poisson equation in first-order form, but in more than one dimension, the first stable elements were discovered in 1975 [4].

References

1. Charney, J.G., Fjørtoft, R., von Neumann, J.: Numerical integration of the barotropic vorticity equation. Tellus **2**, 237–254 (1950)
2. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. Math. Ann. **100**(1), 32–74 (1928)
3. Lax, P.D., Richtmyer, R.D.: Survey of the stability of linear finite difference equations. Commun. Pure Appl. Math. **9**(2), 267–293 (1956)
4. Raviart, P.A., Thomas, J.M.: A mixed finite element method for 2nd order elliptic problems. In: Mathematical Aspects of Finite Element Methods. Proceedings of the Conference, Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975. Volume 606 of Lecture Notes in Mathematics, pp. 292–315. Springer, Berlin (1977)
5. von Neumann, J., Goldstine, H.H.: Numerical inverting of matrices of high order. Bull. Am. Math. Soc. **53**, 1021–1099 (1947)

Statistical Methods for Uncertainty Quantification for Linear Inverse Problems

Luis Tenorio
Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO, USA

To solve an inverse problem means to recover an unknown object from indirect noisy observations. As an illustration, consider an idealized example of the blurring of a one-dimensional signal, $f(x)$, by a measuring instrument. Assume the function is parametrized so that $x \in [0, 1]$ and that the actual data can be modeled as noisy observations of a blurred version of f. We may model the blurring as a convolution with a kernel, $K(x)$, determined by the instrument. The forward operator maps f to the blurred function μ given by $\mu(x) = \int_0^1 K(x-t)\, f(t)\, dt$ (i.e., a *Fredholm integral equation* of the first kind.) The statistical model for the data is then $y(x) = \mu(x) + \varepsilon(x)$, where $\varepsilon(x)$ is measurement noise. The inverse problem consists of recovering f from finitely many measurements $y(x_1),\dots,y(x_n)$. However, inverse problems are usually ill-posed (e.g., the estimates may be very sensitive to small perturbations of the data) and deblurring is one such example. A regularization method is required to solve the problem. An introduction to regularization of inverse problems can be found in [11] and more general references are [10, 27].

Since the observations $y(x_i)$ are subject to systematic errors (e.g., discretization) as well as measurement errors that will be modeled as random variables, the solution of an inverse problem should include a summary of the statistical characteristics of the inversion estimate such as means, standard deviations, bias, mean squared errors, and confidence sets. However, the selection of proper statistical methods to assess estimators depends, of course, on the class of estimators, which in turn is determined by the type of inverse problem and chosen regularization method. Here we use a general framework that encompasses several different and widely used approaches.

We consider the problem of assessing the statistics of solutions of linear inverse problems whose data are modeled as $y_i = \mathcal{K}_i[f] + \varepsilon_i$ $(i = 1,\dots,n)$, where the functions f belong to a linear space H, each $\mathcal{K}_i$ is a continuous linear operator defined on H, and the errors ε_i are random variables. Since the errors are random, an estimate $\hat{f}$ of f is a random variable taking values in H. Given the finite amount of data, we can only hope to recover components of f that admit a finite-dimensional parametrization. Such parametrizations also help us avoid defining probability measures in function spaces. For example, we can discretize the operators and the function so that the estimate $\hat{\boldsymbol{f}}$ is a vector in $\mathbb{R}^m$. Alternatively, one may be able to use a finite-dimensional parametrization such as $f = \sum_{k=1}^m a_k \psi_k$, where ψ_k are fixed functions defined on H. This time the random variable is the estimate $\hat{\boldsymbol{a}}$ of the vector of coefficients $\boldsymbol{a} = (a_k)$. In either case the problem of finding an estimate of a function reduces to a finite-dimensional linear algebra problem.

Example 1 Consider the inverse problem for a Fredholm integral equation:

$$y_i = y(x_i) = \int_0^1 K(x_i - t)\, f(t)\, dt + \varepsilon_i, \quad (i = 1,\dots,n).$$

To discretize the integral, we can use m equally spaced points t_i in $[0, 1]$ and define $t_j' = (t_j + t_{j-1})/2$. Then,

$$\mu(x_i) = \int_0^1 K(x_i - y)\, f(y)\, dy \approx \frac{1}{m} \sum_{j=1}^{m-1} K(x_i - t_j')\; f(t_j').$$

Hence we have the approximation $\boldsymbol{\mu} = (\mu(x_1),\dots,\mu(x_n))^t \approx \boldsymbol{K}\boldsymbol{f}$ with $K_{ij} = K(x_i - t_j')$ and $f_i = f(t_i')$. Writing the discretization error as $\boldsymbol{\delta} = \boldsymbol{\mu} - \boldsymbol{K}\boldsymbol{f}$, we arrive at the following model for the data vector $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{f} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}. \tag{1}$$

As $m \to \infty$ the approximation of the integral improves but the matrix $\boldsymbol{K}$ becomes more ill-conditioned. To regularize the problem, we define an estimate $\hat{\boldsymbol{f}}$ of $\boldsymbol{f}$ using *penalized least squares*:

$$\begin{aligned}\hat{\boldsymbol{f}} &= \arg\min_{\boldsymbol{g}\in\mathbb{R}^m} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{g}\|^2 + \lambda^2 \|\boldsymbol{D}\boldsymbol{g}\|^2 \\ &= (\boldsymbol{K}^t\boldsymbol{K} + \lambda^2 \boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{K}^t\,\boldsymbol{y} \equiv \boldsymbol{L}\,\boldsymbol{y},\end{aligned}$$

where $\lambda > 0$ is a fixed *regularization parameter* and $\boldsymbol{D}$ is a chosen matrix (e.g., a matrix that computes discrete derivatives). This regularization addresses the ill-conditioning of the matrix $\boldsymbol{K}$; it is a way of adding the prior information that we expect $\|Df\|$ to be small. The case $\boldsymbol{D} = \boldsymbol{I}$ is known as (discrete) *Tikhonov-Phillips regularization*. Note that we may write $\hat{f}(x_i)$ as a linear function of $\boldsymbol{y}$: $\hat{f}(x_i) = \boldsymbol{e}_i^t \boldsymbol{L}\boldsymbol{y}$ where $\{\boldsymbol{e}_i\}$ is the standard orthonormal basis in $\mathbb{R}^m$. □

In the next two examples, we assume that H is a Hilbert space, and each $\mathcal{K}_i : H \to \mathbb{R}$ is a bounded linear operator (and thus continuous). We write $\mathcal{K}[f] = (\mathcal{K}_1[f], \ldots, \mathcal{K}_n[f])^t$.

Example 2 Since $\mathcal{K}(H) \subset \mathbb{R}^n$ is finite dimensional, it follows that $\mathcal{K}$ is compact as is its adjoint $\mathcal{K}^* : \mathbb{R}^n \to H$, and $\mathcal{K}^*\mathcal{K}$ is a self-adjoint compact operator on H. In addition, there is a collection of orthonormal functions $\{\phi_k\}$ in H, orthonormal vectors $\{\boldsymbol{v}_k\}$ in $\mathbb{R}^n$, and a positive, nonincreasing sequence (λ_k) such that [22]: (a) $\{\phi_k\}$ is an orthonormal basis for $\text{Null}(\mathcal{K})^\perp$; (b) $\{\boldsymbol{v}_k\}$ is an orthonormal basis for the closure of $\text{Range}(\mathcal{K})$ in $\mathbb{R}^n$; and (c) $\mathcal{K}[\phi_k] = \lambda_k \boldsymbol{v}_k$ and $\mathcal{K}^*[\boldsymbol{v}_k] = \lambda_k \phi_k$. Write $f = f_0 + f_1$, with $f_0 \in \text{Null}(\mathcal{K})$ and $f_1 \in \text{Null}(\mathcal{K})^\perp$. Then, there are constants a_k such that $f_1 = \sum_{k=1}^n a_k \phi_k$. The data do not provide any information about f_0 so without any other information we have no way of estimating such component of f. This introduces a systematic bias. The problem of estimating f is thus reduced to estimating f_1, that is, the coefficients a_k. In fact, we may transform the data to $\langle \boldsymbol{y}, \boldsymbol{v}_k \rangle = \lambda_k a_k + \langle \boldsymbol{\varepsilon}, \boldsymbol{v}_k \rangle$ and use them to estimate the vector of coefficients $\boldsymbol{a} = (a_k)$; the transformed data based on this sequence define a *sequence space model* [7, 18]. We may also rewrite the data as $\boldsymbol{y} = \boldsymbol{V}\boldsymbol{a} + \boldsymbol{\varepsilon}$, where $\boldsymbol{V} = (\lambda_1 \boldsymbol{v}_1 \cdots \lambda_n \boldsymbol{v}_n)$. An estimate of f is obtained using a penalized least-squares estimate of $\boldsymbol{a}$:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{V}\boldsymbol{b}\|^2 + \lambda^2 \|\boldsymbol{b}\|^2.$$

This leads again to an estimate that is linear in $\boldsymbol{y}$; write it as $\hat{\boldsymbol{a}} = \boldsymbol{L}\boldsymbol{y}$ for some matrix $\boldsymbol{L}$. The estimate of $f(x)$ is then similar to that in Example 1:

$$\hat{f}(x) = \boldsymbol{\phi}(x)^t \hat{\boldsymbol{a}} = \boldsymbol{\phi}(x)^t \boldsymbol{L}\boldsymbol{y}, \tag{2}$$

with $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_n(x))^t$. □

If the goal is to estimate pointwise values of $f \in H$, then the Hilbert space H needs to be defined appropriately. For example, if $H = L^2([0,1])$, then pointwise values of f are not well defined. The following example introduces spaces where evaluation at a point (i.e., $f \to f(x)$) is a continuous linear functional.

Example 3 Let $I = [0,1]$. Let $W_m(I)$ be the linear space of real-valued functions on I such that f has $m-1$ continuous derivatives on I, $f^{(m-1)}$ is absolutely continuous on I (so $f^{(m)}$ exists almost everywhere on I), and $f^{(m)} \in L^2(I)$. The space $W_m(I)$ is a Hilbert space with inner product

$$\langle f, g \rangle = \sum_{k=0}^{m-1} f^{(k)}(0)\, g^{(k)}(0) + \int_I f^{(m)}(x)\, g^{(m)}(x)\, dx$$

and has the following properties [2, 29]: (a) For every $x \in I$, there is a function $\rho_x \in W_m(I)$ such that the linear functional $f \to f(x)$ is continuous on $W_m(I)$ and given by $f \to \langle \rho_x, f \rangle$. The function $R : I \times I \to \mathbb{R}$, $R(x,y) = \langle \rho_x, \rho_y \rangle$ is called a *reproducing kernel* of the Hilbert space, and (b) $W_m(I) = \mathcal{N}_{m-1} \oplus H_m$, where $\mathcal{N}_{m-1}$ is the space of polynomials of degree at most $m-1$ and $H_m = \{ f \in W_m(I) : f^{(k)}(0) = 0 \text{ for } k = 0, \ldots, m-1 \}$. Since the space $W_m(I)$ satisfies (a), it is called a *reproducing kernel Hilbert space* (RKHS). To control the smoothness of the Tikhonov estimate, we put a penalty on the derivative of f_1, which is the projection $f_1 = P_H f$ onto H_m. To write the penalized sum of squares, we use the fact that each functional $K_i : W_m(I) \to \mathbb{R}$ is continuous and thus $K_i f = \langle \kappa_i, f \rangle$ for some function $\kappa_i \in W_m(I)$. We can then write

$$\begin{aligned} \| \boldsymbol{y} - Kf \|^2 + \lambda^2 \int_I (f_1^{(m)}(x))^2\, dx \\ = \sum_{j=1}^n (y_i - \langle \kappa_i, f \rangle)^2 + \lambda^2 \| P_H f \|^2. \end{aligned} \tag{3}$$

Define $\phi_k(x) = x^{k-1}$ for $k = 1, \ldots, m$ and $\phi_k = P_H \kappa_{k-m}$ for $k = m+1, \ldots, m+n$. Then $f = \sum_k a_k \phi_k + \delta$, where δ belongs to the orthogonal complement of the span of $\{\phi_k\}$. It can be shown that the minimizer $\hat{f}$ of (3) is again of the form (2) [29]. To estimate $\boldsymbol{a}$ we rewrite (3) as a function of $\boldsymbol{a}$ and use the following estimate:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{b}} \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \|^2 + \lambda^2 \boldsymbol{b}_H^t \boldsymbol{P} \boldsymbol{b}_H,$$

where $\boldsymbol{X}$ is the matrix of inner products $\langle \kappa_i, \phi_j \rangle$, $P_{ij} = \langle P_H \kappa_i, P_H \kappa_j \rangle$ and $\boldsymbol{a}_H = (a_{m+1}, \ldots, a_{m+n})^t$. □

These examples describe three different frameworks where the functional estimation is reduced to a finite-dimensional penalized least-squares problem. They serve to motivate the framework we will use in the statistical analysis. We will focus on frequentist statistical methods. Bayesian methods for inverse problems are discussed in [17]; [1,23] provide a tutorial comparison of frequentist and Bayesian procedures for inverse problems.

Consider first the simpler case of *general linear regression*: the $n \times 1$ data vector is modeled as $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{a} + \boldsymbol{\varepsilon}$, where $\boldsymbol{K}$ is an $n \times m$ matrix, $n > m$, $\boldsymbol{K}^t\boldsymbol{K}$ is non-singular and $\boldsymbol{\varepsilon}$ is a random vector with mean zero and covariance matrix $\sigma^2\boldsymbol{I}$. The least-squares estimate of $\boldsymbol{a}$ is

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{b}\|^2 = (\boldsymbol{K}^t\boldsymbol{K})^{-1}\boldsymbol{K}^t\boldsymbol{y}, \quad (4)$$

and it has the following properties: Its expected value is $\mathbb{E}(\hat{\boldsymbol{a}}) = \boldsymbol{a}$ regardless of the true value $\boldsymbol{a}$; that is, $\hat{\boldsymbol{a}}$ is an *unbiased estimator* of $\boldsymbol{a}$. The covariance matrix of $\hat{\boldsymbol{a}}$ is $\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{a}}) \equiv \sigma^2(\boldsymbol{K}^t\boldsymbol{K})^{-1}$. An unbiased estimator of σ^2 is $\hat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{K}\hat{\boldsymbol{a}}\|^2/(n-m)$. Note that the denominator is the difference between the number of observations; it is a kind of "effective number of observations." It can be shown that $m = \mathrm{tr}(\boldsymbol{H})$, where $\boldsymbol{H} = \boldsymbol{K}(\boldsymbol{K}^t\boldsymbol{K})^{-1}\boldsymbol{K}^t$ is the *hat matrix*; it is the matrix defined by $\boldsymbol{H}\boldsymbol{y} \equiv \hat{\boldsymbol{y}} = \boldsymbol{K}\hat{\boldsymbol{a}}$. The *degrees of freedom* (dof) of $\hat{\boldsymbol{y}}$ is defined as the sum of the covariances of $(\boldsymbol{K}\hat{\boldsymbol{a}})_i$ with y_i divided by σ^2 [24]. For linear regression we have $\mathrm{dof}(\boldsymbol{K}\hat{\boldsymbol{a}}) = m = \mathrm{tr}(\boldsymbol{H})$. Hence we may write $\hat{\sigma}^2$ as the residual sum of squares normalized by the effective number of observations:

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{n - \mathrm{dof}(\hat{\boldsymbol{y}})} = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H})}. \quad (5)$$

We now return to ill-posed inverse problems and define a general framework motivated by Examples 1–2 that is similar to general linear regression. We assume that the data vector has a representation of the form $\boldsymbol{y} = \mathcal{K}[f] + \boldsymbol{\varepsilon} = \boldsymbol{K}\boldsymbol{a} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$, where $\mathcal{K}$ is a linear operator $H \to \mathbb{R}^n$, $\boldsymbol{K}$ is an $n \times m$ matrix, $\boldsymbol{\delta}$ is a fixed unknown vector (e.g., discretization error), and $\boldsymbol{\varepsilon}$ is a random vector of mean zero and covariance matrix $\sigma^2\boldsymbol{I}$. We also assume that there is an $n \times 1$ vector $\boldsymbol{a}$ and a vector function $\boldsymbol{\phi}$ such that $f(x) = \boldsymbol{\phi}(x)^t\boldsymbol{a}$ for all x. The vector $\boldsymbol{a}$ is estimated using penalized least squares:

$$\begin{aligned}\hat{\boldsymbol{a}} &= \arg\min_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{b}\|^2 + \lambda^2\boldsymbol{b}^t\boldsymbol{S}\boldsymbol{b} \\ &= (\boldsymbol{K}^t\boldsymbol{K} + \lambda^2\boldsymbol{S})^{-1}\boldsymbol{K}^t\boldsymbol{y}, \quad (6)\end{aligned}$$

where $\boldsymbol{S}$ is a symmetric non-negative matrix and $\lambda > 0$ is a fixed regularization parameter. The estimate of f is defined as $\hat{f}(x) = \boldsymbol{\phi}(x)^t\hat{\boldsymbol{a}}$.

Bias, Variance, and MSE

For a fixed regularization parameter λ, the estimator $\hat{f}(x)$ is linear in $\boldsymbol{y}$, and therefore its mean, bias, variance, and mean squared error can be determined using only knowledge of the first two moments of the distribution of the noise vector $\boldsymbol{\varepsilon}$. Using (6) we find the mean, bias, and variance of $\hat{f}(x)$:

$$\begin{aligned}\mathbb{E}(\hat{f}(x)) &= \boldsymbol{\phi}(x)^t\boldsymbol{G_\lambda}^{-1}\boldsymbol{K}^t\mathcal{K}[f], \\ \mathrm{Bias}(\hat{f}(x)) &= \boldsymbol{\phi}(x)^t\boldsymbol{G_\lambda}^{-1}\boldsymbol{K}^t\mathcal{K}[f] - f(x) \\ \mathbb{V}\mathrm{ar}(\hat{f}(x)) &= \sigma^2\|\boldsymbol{K}\boldsymbol{G}_\lambda\boldsymbol{\phi}(x)\|^2,\end{aligned}$$

where $\boldsymbol{G}_\lambda = (\boldsymbol{K}^t\boldsymbol{K} + \lambda^2\boldsymbol{S})^{-1}$. Hence, unlike the least-squares estimate of $\boldsymbol{a}$ (4), the penalized least-squares estimate (6) is biased even when $\boldsymbol{\delta} = \boldsymbol{0}$. This bias introduces a bias in the estimates of f. In terms of $\boldsymbol{a}$ and $\boldsymbol{\delta}$, this bias is

$$\begin{aligned}\mathrm{Bias}(\hat{f}(x)) &= \mathbb{E}(\hat{f}(x)) - f(x) \\ &= \boldsymbol{\phi}(x)^t\boldsymbol{B}_\lambda\boldsymbol{a} + \boldsymbol{\phi}(x)^t\boldsymbol{G}_\lambda\boldsymbol{K}^t\boldsymbol{\delta}, \quad (7)\end{aligned}$$

where $\boldsymbol{B}_\lambda = -\lambda^2\boldsymbol{G}_\lambda\boldsymbol{S}$. Prior information about $\boldsymbol{a}$ should be used to choose the matrix $\boldsymbol{S}$ so that $\|\boldsymbol{B}_\lambda\boldsymbol{a}\|$ is small. Note that similar formulas can be derived for correlated noise provided the covariance matrix is known. Also, analogous closed formulas can be derived for estimates of linear functionals of f.

The *mean squared error* (MSE) can be used to include the bias and variance in the uncertainty evaluation of $\hat{f}(x)$; it is defined as the expected value of $(\hat{f}(x) - f(x))^2$, which is equivalent to the squared bias plus the variance:

$$\mathrm{MSE}(\hat{f}(x)) = \mathrm{Bias}(\hat{f}(x))^2 + \mathbb{V}\mathrm{ar}(\hat{f}(x)).$$

The *integrated mean squared error* of $\hat{f}$ is

$$\begin{aligned}\mathrm{IMSE}(\hat{f}) &= \mathbb{E}\int |\hat{f}(x) - f(x)|^2\,dx \\ &= \mathrm{Bias}(\hat{\boldsymbol{a}})^t\boldsymbol{F}\,\mathrm{Bias}(\hat{\boldsymbol{a}}) + \mathrm{tr}(\boldsymbol{F}\boldsymbol{G}_\lambda\boldsymbol{K}^t\boldsymbol{K}\boldsymbol{G}_\lambda),\end{aligned}$$

where $\boldsymbol{F} = \int \boldsymbol{\phi}(x)\boldsymbol{\phi}(x)^t\, dx$.

The bias component of the MSE is the most difficult to assess as it depends on the unknown f (or $\boldsymbol{a}$ and $\boldsymbol{\delta}$), but, depending on the available prior information, some inequalities can be derived [20, 26].

Example 4 If H is a Hilbert space and the functionals $\mathcal{K}_i$ are bounded, then there are function $\kappa_i \in H$ such that $\mathcal{K}_i[f] = \langle \kappa_i, f \rangle$, and we may write

$$\mathbb{E}[\,\hat{f}(x)\,] = \langle \sum_i a_i(x)\kappa_i, f \,\rangle = \langle A_x\,, f\,\rangle,$$

where the function $A_x(y) = \sum_i a_i(x)\kappa_i(y)$ is called the Backus-Gilbert averaging kernel for $\hat{f}$ at x [3]. In particular, since we would like to have $f(x) = \langle A_x\,, f\,\rangle$, we would like A_x to be as concentrated as possible around x; a plot of the function A_x may provide useful information about the mean of the estimate $\hat{f}(x)$. One may also summarize characteristics of $|A_x|$ such as its center and spread about the center (e.g., [20]). Heuristically, $|A_x|$ should be like a δ-function centered at x. This can be formalized in an RKHS H. In this case, there is a function $\rho_x \in H$ such that $f(x) = \langle \rho_x, f \rangle$ and the bias of $\hat{f}(x)$ can be written as $\text{Bias}(\,\hat{f}(x)\,) = \langle\, A_x - \rho_x, f\,\rangle$ and therefore

$$|\,\text{Bias}(\,\hat{f}(x)\,)\,| \le \|\, A_x - \rho_x \,\|\, \|\, f\,\|.$$

We can guarantee a small bias when A_x is close to ρ_x in H. In actual computations, averaging kernels can be approximated using splines. A discussion of this topic as well as information about available software for splines and reproducing kernels can be found in [14, 20]. □

Another bound for the bias follows from (7) via the Cauchy-Schwarz and triangle inequalities:

$$|\,\text{Bias}(\,\hat{f}(x)\,)\,| \le \|\boldsymbol{G}_\lambda \boldsymbol{\phi}(x)\|(\,\lambda^2\|\boldsymbol{S}\,\boldsymbol{a}\| + \|\boldsymbol{K}^t\boldsymbol{\delta}\|\,).$$

Plots of $\|\boldsymbol{G}_\lambda\boldsymbol{\phi}(x)\|$ (or $\|\, A_x - \rho_x \,\|$) as a function of x may provide geometric information (usually conservative) about the bias. Other measures such as the worst or an average bias can be obtained depending on the available prior information we have on f or its parametric representation. For example, if $\boldsymbol{a}$ and $\boldsymbol{\delta}$ are known to lie in convex sets S_1 and S_2, respectively, then we may determine the maximum of $|\,\text{Bias}(\,\hat{f}\,)\,|$ subject to $\boldsymbol{a} \in S_1$ and $\boldsymbol{\delta} \in S_2$. Or, if the prior information leads to the modeling of $\boldsymbol{a}$ and $\boldsymbol{\delta}$ as random variables with means and covariance matrices $\boldsymbol{\mu}_a = \mathbb{E}\boldsymbol{a}$, $\boldsymbol{\Sigma}_a = \mathbb{V}\text{ar}(\boldsymbol{a})$, $\boldsymbol{\mu}_\delta = \mathbb{E}\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_\delta = \mathbb{V}\text{ar}(\boldsymbol{\delta})$, then the average bias is

$$\mathbb{E}[\,\text{Bias}(\,\hat{f}(x)\,)\,] = \boldsymbol{\phi}(x)^t\boldsymbol{B}_\lambda\boldsymbol{\mu}_a + \boldsymbol{\phi}(x)^t\boldsymbol{G}_\lambda\boldsymbol{K}^t\boldsymbol{\mu}_\delta.$$

Similarly, we can easily derive a bound for the mean squared bias that can be used to put a bound on the average MSE.

Since the bias may play a significant factor in the inference (in some geophysical applications the bias is the dominant component of the MSE), it is important to study the residuals of the fit to determine if a significant bias is present. The mean and covariance matrix of the *residual* vector $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{K}\hat{\boldsymbol{a}}$ are

$$\mathbb{E}\boldsymbol{r} = -\boldsymbol{K}\,\text{Bias}(\hat{\boldsymbol{a}}) + \boldsymbol{\delta} = -\boldsymbol{K}\,\boldsymbol{B}_\lambda\boldsymbol{a} + (\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{\delta} \quad (8)$$

$$\mathbb{V}\text{ar}(\boldsymbol{r}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}_\lambda)^2, \quad (9)$$

where $\boldsymbol{H}_\lambda = \boldsymbol{K}\boldsymbol{G}_\lambda\boldsymbol{K}^t$ is the hat matrix. Equation (8) shows that if there is a significant bias, then we may see a trend in the residuals. From (8) we see that the residuals are correlated and heteroscedastic (i.e., $\mathbb{V}\text{ar}(r_i)$ depends on i) even if the bias is zero, which complicates the interpretation of the plots. To stabilize the variance, it is better to plot residuals that have been corrected for heteroscedasticity, for example, $r_i' = r_i/(1 - (H_\lambda)_{ii})$.

Confidence Intervals

In addition to the mean and variance of $\hat{f}(x)$, we may construct *confidence intervals* that are expected to contain $\mathbb{E}(\hat{f}(x))$ with some prescribed probability. We now assume that the noise is Gaussian, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. We use Φ to denote the cumulative distribution function of the standard Gaussian $N(0,1)$ and write $z_\alpha = \Phi^{-1}(1-\alpha)$.

Since $\hat{f}(x)$ is a biased estimate of $f(x)$, we can only construct confidence intervals for $\mathbb{E}\hat{f}(x)$. We should therefore interpret the intervals with caution as they may be incorrectly centered if the bias is significant.

Under the Gaussianity assumption, a confidence interval $I_\alpha(x,\sigma)$ for $\mathbb{E}\hat{f}(x)$ of coverage $1-\alpha$ is

$$I_\alpha(x,\sigma) = \hat{f}(x) \pm z_{\alpha/2}\,\sigma\,\|\boldsymbol{K}\boldsymbol{G}_\lambda\,\boldsymbol{\phi}(x)\|.$$

That is, for each x the probability that $I_\alpha(x,\sigma)$ contains $\mathbb{E}\hat{f}(x)$ is $1-\alpha$. If we have prior information to find an upper bound for the bias, $|\text{Bias}(\hat{f}(x))| \leq B(x)$, then a confidence interval for $f(x)$ with coverage at least $1-\alpha$ is $\hat{f}(x) \pm (z_{\alpha/2}\,\sigma \|\boldsymbol{K G}_\lambda\,\boldsymbol{\phi}(x)\| + B(x))$.

If the goal is to detect structure by studying the confidence intervals $I_\alpha(x,\sigma)$ for a range of values of x, then it is advisable to correct for the total number of intervals considered so as to control the rate of incorrect detections. One way to do this is by constructing $1-\alpha$ confidence intervals of the form $I_\alpha(x,\sigma\beta)$ with simultaneous coverage for all x in some closed set S. This requires finding a constant $\beta > 0$ such that

$$\mathbb{P}[\mathbb{E}\hat{f}(x) \in I_\alpha(x,\sigma\beta), \forall x \in S\,] \geq 1-\alpha,$$

which is equivalent to

$$\mathbb{P}\left[\sup_{x\in S} |\boldsymbol{Z}^t \boldsymbol{V}(x)| \geq \beta\right] \leq \alpha,$$

where $\boldsymbol{V}(x) = \boldsymbol{K G}_\lambda \boldsymbol{\phi}(x)/\|\boldsymbol{K G}_\lambda \boldsymbol{\phi}(x)\|$ and $Z_1, \ldots, Z_n$ are independent $N(0,1)$. We can use results regarding the tail behavior of maxima of Gaussian processes to find an appropriate value of β. For example, for the case when $x \in [a,b]$ [19] (see also [25]) shows that for β large

$$\mathbb{P}\left[\sup_{x\in S} |\boldsymbol{Z}^t \boldsymbol{V}(x)| \geq \beta\right] \approx \frac{v}{\pi}\, e^{-\beta^2/2} + 2(1-\Phi(\beta)),$$

where $v = \int_a^b \|\boldsymbol{V}'(x)\|\,dx$. It is not difficult to find a root of this nonlinear equation; the only potential problem may be computing v, but even an upper bound for it leads to intervals with simultaneous coverage at least $1-\alpha$. Similar results can be derived for the case when S is a subset of $\mathbb{R}^2$ or $\mathbb{R}^3$ [25].

An alternative approach is to use methods based on controlling the *false discovery rate* to correct for the interval coverage after the pointwise confidence intervals have been selected [4].

Estimating σ and λ

The formulas for the bias, variance, and confidence intervals described so far require knowledge of σ and a selection of λ that is independent of the data $\boldsymbol{y}$. If $\boldsymbol{y}$ is also used to estimate σ or choose λ, then $\hat{f}(x)$ is no longer linear in $\boldsymbol{y}$ and closed formulas for the moments, bias, or confidence intervals are not available. Still, the formulas derived above with "reasonable" estimates $\hat{\sigma}$ and $\hat{\lambda}$ in place of σ and λ are approximately valid. This depends of course on the class of possible functions f, the noise distribution, the signal-to-noise ratio, and the ill-posedness of the problem. We recommend conducting realistic simulation studies to understand the actual performance of the estimates for a particular problem.

Generalized cross-validation (GCV) methods to select λ have proved useful in applications and theoretical studies. A discussion of these methods can be found in [13, 14, 29, 30]. We now summarize a few methods for obtaining an estimate of σ.

The estimate of σ^2 given by (5) could be readily used for, once again, $\text{dof}(\boldsymbol{K\hat{a}}) = \text{tr}(\boldsymbol{H}_\lambda)$, with the corresponding hat matrix $\boldsymbol{H}_\lambda$ (provided $\boldsymbol{\delta} = \boldsymbol{0}$). However, because of the bias of $\boldsymbol{K\hat{a}}$ and the fixed error $\boldsymbol{\delta}$, it is sometimes better to estimate σ by considering the data as noisy observations of $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{y} = \mathcal{K}[f]$ – in which case we may assume $\boldsymbol{\delta} = \boldsymbol{0}$; that is, $y_i = \mu_i + \varepsilon_i$. This approach is natural as σ^2 is the variance of the errors, ε_i, in the observations of μ_i.

To estimate the variance of ε_i, we need to remove the trend μ_i. This trend may be seen as the values of a function $\mu(x)$: $\mu_i = \mu(x_i)$. A variety of nonparametric regression methods can be used to estimate the function μ so it can be removed, and an estimate of the noise variance can be obtained (e.g., [15, 16]). If the function can be assumed to be reasonably smooth, then we can use the framework described in 3 with $\mathcal{K}_i[\mu] = \mu(x_i)$ and a penalty in the second derivative of μ. In this case $\hat{\mu}(x) = \sum a_i \phi_i(x) = \boldsymbol{a}^t\boldsymbol{\phi}(x)$ is called a *spline smoothing* estimate because it is a finite linear combination of spline functions ϕ_i [15, 29]. The estimate of σ^2 defined by (5) with the corresponding hat matrix $\boldsymbol{H}_\lambda$ was proposed by [28]. Using (8) and (9) we find that the expected value of the residual sum of squares is

$$\mathbb{E}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \sigma^2\,\text{tr}[\,(\boldsymbol{I} - \boldsymbol{H}_\lambda)^2\,] + \boldsymbol{a}^t \boldsymbol{B}_\lambda \boldsymbol{K}^t \boldsymbol{K}\,\boldsymbol{B}_\lambda \boldsymbol{a}, \tag{10}$$

and thus $\hat{\sigma}^2$ is not an unbiased estimator of σ^2 even if the bias is zero (i.e., $\boldsymbol{B}_\lambda\boldsymbol{a} = \boldsymbol{0}$, which happens when μ is linear), but it has been shown to have good asymptotic properties when λ is selected using generalized cross-validation [13,28]. From (10) we see that a slight modification of $\hat{\sigma}^2$ leads to an estimate that is unbiased when $\boldsymbol{B}_\lambda\boldsymbol{a} = \boldsymbol{0}$ [5]

$$\hat{\sigma}_B^2 = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{\mathrm{tr}[\,(\boldsymbol{I} - \boldsymbol{H}_\lambda)^2\,]}.$$

Simulation studies seem to indicate that this estimate has a smaller bias for a wider set of values of λ [6]. This property is desirable as λ is usually chosen adaptively.

In some cases the effect of the trend can also be reduced using first- or second-order finite differences without having to choose a regularization parameter λ. For example, a first-order finite-difference estimate proposed by [21] is

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2.$$

The bias of $\hat{\sigma}_R^2$ is small if the local changes of $\mu(x)$ are small. In particular, the bias is zero for a linear trend. Other estimators of σ as well performance comparisons can be found in [5, 6].

Resampling Methods

We have assumed a Gaussian noise distribution for the construction of confidence intervals. In addition, we have only considered linear operators and linear estimators. Nonlinear estimators arise even when the operator $\mathcal{K}$ is linear. For example, if σ and λ are estimated using the same data or if the penalized least-squares estimate of $\boldsymbol{a}$ includes interval constraints (e.g., positivity), then the estimate $\hat{\boldsymbol{a}}$ is no longer linear in $\boldsymbol{y}$. In some cases the use of *bootstrap* (resampling) methods allows us to assess statistical properties while relaxing the distributional and linearity assumptions.

The idea is to simulate data $\boldsymbol{y}^*$ as follows: the function estimate $\hat{f}$ is used as a proxy for the unknown function f. Noise $\boldsymbol{\varepsilon}^*$ is simulated using a parametric or nonparametric method. In the *parametric bootstrap*, $\boldsymbol{\varepsilon}^*$ is sampled from the assumed distribution whose parameters are estimated from the data. For example, if the ε_i are independent $N(0, \sigma^2)$, then ε_i^* is sampled from $N(0, \hat{\sigma}^2)$. In the *nonparametric bootstrap*, ε_i^* is sampled with replacement from the vector of residuals of the fit. However, as Eqs. (8) and (9) show, even in the linear case, the residuals have to be corrected to behave approximately like the true errors. Of course, due to the bias and correlation of the residuals, these corrections are often difficult to derive and implement. Using ε_i^* and $\hat{f}$, one generates simulated data vectors $\boldsymbol{y}_j^* = \mathcal{K}[\hat{f}] + \boldsymbol{\varepsilon}_j^*$. For each such $\boldsymbol{y}_j^*$ one computes an estimate $\hat{f}_j$ of f following the same procedure used to obtain $\hat{f}$. The statistics of the sample of $\hat{f}_j$ are used as estimates of those of $\hat{f}$. One problem with this approach is that the bias of $\hat{f}$ may lead to a poor estimate of $\mathcal{K}[f]$ and thus to unrealistic simulated data.

An introduction to bootstrap methods can be found in [9, 12]. For an example of bootstrap methods to construct confidence intervals for estimates of a function based on smoothing splines, see [31].

References

1. Aguilar, O., Allmaras, M., Bangerth, W., Tenorio, L.: Statistics of parameter estimates: a concrete example. SIAM Rev. **57**, 131 (2015)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **89**, 337 (1950)
3. Backus, G., Gilbert, F.: Uniqueness in the inversion of inaccurate gross earth data. Philos. Trans. R. Soc. Lond. A **266**, 123 (1970)
4. Benjamini, Y., Yekutieli, D.: False discovery rate–adjusted multiple confidence intervals for selected parameters. J. Am. Stat. Assoc. **100**, 71 (2005)
5. Buckley, M.J., Eagleson, G.K., Silverman, B.W.: The estimation of residual variance in nonparametric regression. Biometrika **75**, 189 (1988)
6. Carter, C.K., Eagleson, G.K.: A comparison of variance estimators in nonparametric regression. J. R. Stat. Soc. B **54**, 773 (1992)
7. Cavalier, L.: Nonparametric statistical inverse problems. Inverse Probl. **24**, 034004 (2008)
8. Craven, P., Wahba, G.: Smoothing noisy data with splines. Numer. Math. **31**, 377 (1979)
9. Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press, Cambridge (1997)
10. Engl, H., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
11. Engl H his chapter
12. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)
13. Gu, C.: Smoothing Spline ANOVA Models. Springer, Berlin/Heidelberg/New York (2002)
14. Gu, C.: Smoothing noisy data via regularization: statistical perspectives. Inverse Probl. **24**, 034002 (2008)
15. Green, P.J., Silverman, B.W.: Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall, London (1993)
16. Härdle, W.K., Müller, M., Sperlich, S., Werwatz, A.: Nonparametric and Semiparametric Models. Springer, Berlin/Heidelberg/New York (2004)
17. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, Berlin/Heidelberg/New York (2004)

18. Mair, B., Ruymgaart, F.H.: Statistical estimation in Hilbert scale. SIAM J. Appl. Math. **56**, 1424 (1996)
19. Naiman, D.Q.: Conservative confidence bands in curvilinear regression. Ann. Stat. **14**, 896 (1986)
20. O'Sullivan, F.: A statistical perspective on ill-posed inverse problems. Stat. Sci. **1**, 502 (1986)
21. Rice, J.: Bandwidth choice for nonparametric regression. Ann. Stat. **12**, 1215 (1984)
22. Rudin, W.: Functional Analysis. McGraw-Hill, New York (1973)
23. Stark, P.B., Tenorio, L.: A primer of frequentist and Bayesian inference in inverse problems. In: Biegler, L., et al. (eds.) Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty, pp. 9–32. Wiley, Chichester (2011)
24. Stein, C.: Estimation of the mean of a multivariate normal distribution. Ann. Stat. **9**, 1135 (1981)
25. Sun, J., Loader, C.R.: Simultaneous confidence bands for liner regression and smoothing. Ann. Stat. **22**, 1328 (1994)
26. Tenorio, L., Andersson, F., de Hoop, M., Ma, P.: Data analysis tools for uncertainty quantification of inverse problems. Inverse Probl. **29**, 045001 (2011)
27. Vogel, C.R.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)
28. Wahba, G.: Bayesian "confidence intervals" for the cross-validated smoothing spline. J. R. Stat. Soc. B **45**, 133 (1983)
29. Wahba, G.: Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 50. SIAM, Philadelphia (1990)
30. Wahba G her chapter
31. Wang, Y., Wahba, G.: Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. J. Stat. Comput. Simul. **51**, 263 (1995)

Step Size Control

Gustaf Söderlind
Centre for Mathematical Sciences, Numerical
Analysis, Lund University, Lund, Sweden

Introduction

Step size control is used to make a numerical method that proceeds in a step-by-step fashion *adaptive*. This includes time stepping methods for solving initial value problems, nonlinear optimization methods, and continuation methods for solving nonlinear equations. The objective is to increase efficiency, but also includes managing the stability of the computation.

This entry focuses exclusively on time stepping adaptivity in initial value problems. Special control algorithms continually adjust the step size in accordance with the local variation of the solution, attempting to compute a numerical solution to within a given error tolerance at minimal cost. As a typical integration may run over thousands of steps, the task is ideally suited to proven methods from automatic control.

Assume that the problem to be solved is a dynamical system,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y); \qquad y(0) = y_0, \tag{1}$$

with $y(t) \in \mathbb{R}^m$. Without loss of generality, we may assume that the problem is solved numerically using a one-step integration procedure, explicit or implicit, written formally as

$$y_{n+1} = \Phi_h(y_n); \qquad y_0 = y(0), \tag{2}$$

where the map Φ_h advances the solution one step of size h, from time t_n to $t_{n+1} = t_n + h$. Here the sequence y_n is the numerical approximations to the exact solution, $y(t_n)$. The difference $e_n = y_n - y(t_n)$ is the *global error* of the numerical solution. If the method is of *convergence order* p, and the vector field f in (1) is sufficiently differentiable, then $\|e_n\| = \mathrm{O}(h^p)$ as $h \to 0$.

The accuracy of the numerical solution can also be evaluated locally. The *local error* l_n is defined by

$$y(t_{n+1}) + l_n = \Phi_h\left(y(t_n)\right). \tag{3}$$

Thus, if the method would take a step of size h, starting on the exact solution $y(t_n)$, it will deviate from the exact solution at t_{n+1} by a small amount, l_n. If the method is of order p, the local error will satisfy

$$\|l_n\| = \varphi_n h^{p+1} + \mathrm{O}(h^{p+2}); \quad h \to 0. \tag{4}$$

Here the *principal error function* φ_n varies along the solution, and depends on the problem (in terms of derivatives of f) as well as on the method.

Using differential inequalities, it can be shown that the global and local errors are related by the a priori *error bound*

$$\|e_n\| \lessapprox \max_{m \le n} \frac{\|l_m\|}{h} \cdot \frac{\mathrm{e}^{M[f]t_n} - 1}{M[f]}, \tag{5}$$

where $M[f]$ is the logarithmic Lipschitz constant of f. Thus, the global error is bounded in terms of the *local error per unit step*, l_n/h. For this reason, one can manage the global error by choosing the step size h so as to keep $\|l_n\|/h = \text{TOL}$ during the integration, where TOL is a user-prescribed *local error tolerance*. The *global error is then proportional to* TOL, by a factor that reflects the intrinsic growth or decay of solutions to (1). Good initial value problem solvers usually produce numerical results that reflect this *tolerance proportionality*. By reducing TOL, one reduces the local error as well as the global error, while computational cost increases, as $h \sim \text{TOL}^{1/p}$.

Although it is possible to compute a posteriori global error estimates, such estimates are often costly. All widely used solvers therefore control the local error, relying on the relation (5), and the possibility of comparing several different numerical solutions computed for different values of TOL. There is no claim that the step size sequences are "optimal," but in all problems where the principal error function varies by several orders of magnitude, as is the case in stiff differential equations, local error control is an inexpensive tool that offers vastly increased performance. It is a necessity for efficient computations.

A time stepping method is made adaptive by providing a separate procedure for updating the step size as a function of the numerical solution. Thus, an adaptive method can be written formally as the interactive recursion

$$y_{n+1} = \Phi_{h_n}(y_n) \tag{6}$$

$$h_{n+1} = \Psi_{y_{n+1}}(h_n), \tag{7}$$

where the first equation represents the numerical method and the second the step size control. If $\Psi_y \equiv I$ (the identity map), the scheme reduces to a constant step size method. Otherwise, the interaction between the two dynamical systems implies that *step size control interferes with the stability of the numerical method*. For this reason, it is important that step size control algorithms are designed to increase efficiency *without compromising stability*.

Basic Multiplicative Control

Modern time stepping methods provide a *local error estimate*. By using two methods of different orders, computing two results, y_{n+1} and $\hat{y}_{n+1}$ from y_n, the solver estimates the local error by $r_n = \|y_{n+1} - \hat{y}_{n+1}\|$. To control the error, the relation between step size and error is modeled by

$$r_n = \hat{\varphi}_n h_n^k. \tag{8}$$

Here k is the order of the local error estimator. Depending on the estimator's construction, k may or may not equal $p+1$, where p is the order of the method used to advance the solution. For control purposes, however, it is sufficient that k is known, and that the method operates in the *asymptotic regime*, meaning that (8) is an *accurate model of the error* for the step sizes in use.

The common approach to varying the step size is *multiplicative control*

$$h_{n+1} = \theta_n \cdot h_n, \tag{9}$$

where the factor θ_n needs to be determined so that the error estimate r_n is kept near the target value TOL for all n.

A simple control heuristic is derived by requiring that the next step size h_{n+1} solves the equation $\text{TOL} = \hat{\varphi}_n h_{n+1}^k$; this assumes that φ_n varies slowly. Thus, dividing this equation by (8), one obtains

$$h_{n+1} = \left(\frac{\text{TOL}}{r_n}\right)^{1/k} h_n. \tag{10}$$

This multiplicative control is found in many solvers. It is usually complemented by a range of safety measures, such as limiting the maximum step size increase, preventing "too small" step size changes, and special schemes for recomputing a step, should the estimated error be much larger than TOL.

Although it often works well, the control law (10) and its safety measures have several disadvantages that call for more advanced feedback control schemes. *Control theory* and *digital signal processing*, both based on linear difference equations, offer a wide range of proven tools that are suitable for controlling the step size. Taking logarithms, (10) can be written as the linear difference equation

$$\log h_{n+1} = \log h_n - \frac{1}{k} \log \hat{r}_n, \tag{11}$$

where $\hat{r}_n = r_n/\text{TOL}$. This recursion continually changes the step size, unless $\log\hat{r}_n$ is zero (i.e., $r_n = \text{TOL}$). If $\hat{r}_n > 1$ the step size decreases, and if $\hat{r}_n < 1$ it increases. Thus, the error r_n is kept near the *set point* TOL. As (11) is a summation process, the controller is referred to as an *integrating controller*, or *I control*. This integral action is necessary in order to eliminate a persistent error, and to find the step size that makes $r_n = \text{TOL}$.

The difference equation (11) may be viewed as using the explicit Euler method for integrating a differential equation that represents a continuous control. Just as there are many different methods for solving differential equations, however, there are many different discrete-time controllers that can potentially be optimized for different numerical methods or problem types, and offering different stability properties.

General Multiplicative Control

In place of (11), a general controller takes the error sequence $\log\hat{r} = \{\log\hat{r}_n\}$ as input, and produces a step size sequence $\log h = \{\log h_n\}$ via a linear difference equation,

$$(E-1)Q(E)\log h = -P(E)\log\hat{r}. \qquad (12)$$

Here E is the forward shift operator, and P and Q are two polynomials of equal degree, making the recursion explicit. The special case (11) has $Q(E) \equiv 1$ and $P(E) \equiv 1/k$, and is a one-step controller, while (12) in general is a multistep controller. Finally, the factor $E - 1$ in (12) is akin to the consistency condition in linear multistep methods. Thus, if $\log\hat{r} \equiv 0$, a solution to (12) is $\log h \equiv \text{const}$.

If, for example, $P(z) = \beta_1 z + \beta_0$ and $Q(z) = z + \alpha_0$, then the recursion (12) is equivalent to the two-step multiplicative control,

$$h_{n+1} = \left(\frac{\text{TOL}}{r_n}\right)^{\beta_1}\left(\frac{\text{TOL}}{r_{n-1}}\right)^{\beta_0}\left(\frac{h_n}{h_{n-1}}\right)^{-\alpha_0} h_n. \qquad (13)$$

By taking logarithms, it is easily seen to correspond to (12). One could include more factors following the same pattern, but in general, it rarely pays off to use a longer step size – error history than two to three steps. Because of the simple structure of (13), it is relatively straightforward to include more advanced controllers in existing codes, keeping in mind that a multistep controller is started either by using (10), or by merely putting all factors representing nonexistent starting data equal to one. Examples of how to choose the parameters in (13) are found in Table 1.

A causal *digital filter* is a linear difference equation of the form (12), converting the input signal $\log\hat{r}$ to an output $\log h$. This implies that digital control and filtering are intimately related. There are several important filter structures that fit the purpose of step size control, all covered by the general controller (13). Among them are *finite impulse response* (FIR) filters; *proportional–integral* (PI) controllers; *autoregressive* (AR) filters; and *moving average* (MA) filters. These filter classes are not mutually exclusive but can be combined.

The elementary controller (10) is a FIR filter, also known as a *deadbeat controller*. Such controllers have the quickest dynamic response to variations in $\log\hat{\varphi}$, but also tend to produce nonsmooth step size sequences, and sometimes display ringing or stability problems. These problems can be eliminated by using PI controllers and MA filters that improve stability and suppress step size oscillations. Filter design is a matter of determining the filter coefficients with respect to *order conditions* and *stability criteria*, and is reminiscent of the construction of linear multistep methods [11].

Stability and Frequency Response

Controllers are analyzed and designed by investigating the *closed loop transfer function*. In terms of the z transform of (12), the control action is

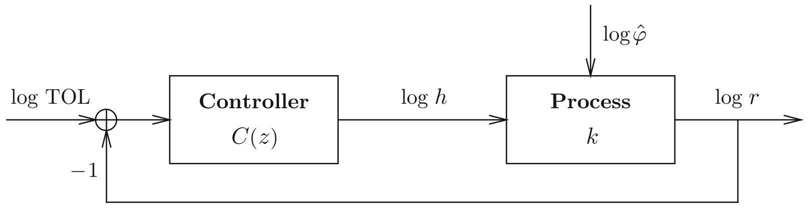
$$\log h = -C(z)\log\hat{r} \qquad (14)$$

where the control transfer function is given by

$$C(z) = \frac{P(z)}{(z-1)Q(z)}. \qquad (15)$$

Similarly, the error model (8) can be written

$$\log\hat{r} = k\cdot\log h + \log\hat{\varphi} - \log\text{TOL}. \qquad (16)$$



Step Size Control, Fig. 1 *Time step adaptivity viewed as a feedback control system.* The computational process takes a stepsize $\log h$ as input and produces an error estimate $\log r = k \log h + \log\hat{\varphi}$. Representing the ODE, the principal error function $\log\hat{\varphi}$ enters as an additive disturbance, to be compensated by the controller. The error estimate $\log r$ is fed back and compared to $\log \text{TOL}$. The controller constructs the next stepsize through $\log h = C(z)\cdot(\log \text{TOL} - \log r)$ (From [11])

These relations and their interaction are usually illustrated in a block diagram, see Fig. 1.

Overall stability depends on the interaction between the controller $C(z)$ and the computational process. Inserting (16) into (14) and solving for $\log h$ yields

$$\log h = \frac{-C(z)}{1+kC(z)}\log\hat{\varphi} + \frac{C(z)}{1+kC(z)}\log \text{TOL}. \tag{17}$$

Here the *closed loop transfer function* $H(z) : \log\hat{\varphi} \mapsto \log h$ is defined by

$$H(z) = \frac{-C(z)}{1+kC(z)} = \frac{-P(z)}{(z-1)Q(z)+kP(z)}. \tag{18}$$

It determines the performance of the combined system of step size controller and computational process, and, in particular, how successful the controller will be in adjusting $\log h$ to $\log\hat{\varphi}$ so that $\log r \approx \log \text{TOL}$.

For a controller or a filter to be useful, the closed loop must be *stable*. This is determined by the poles of $H(z)$, which are the roots of the characteristic equation $(z-1)Q(z) + kP(z) = 0$. These must be located well inside the unit circle, and preferably have positive real parts, so that homogeneous solutions decay quickly without oscillations.

To asses *frequency response*, one takes $\log\hat{\varphi} = \{\mathrm{e}^{\mathrm{i}\omega n}\}$ with $\omega \in [0,\pi]$ to investigate the output $\log h = H(\mathrm{e}^{\mathrm{i}\omega})\{\mathrm{e}^{\mathrm{i}\omega n}\}$. The amplitude $|H(\mathrm{e}^{\mathrm{i}\omega})|$ measures the attenuation of the frequency ω. By choosing P such that $P(\mathrm{e}^{\mathrm{i}\omega}) = 0$ for some ω^*, it follows that $H(\mathrm{e}^{\mathrm{i}\omega^*}) = 0$. Thus, zeros of $P(z)$ *block signal transmission*. The natural choice is $\omega^* = \pi$ so that $P(-1) = 0$, as this will annihilate $(-1)^n$ oscillations, and produce a smooth step size sequence. This is achieved by the two $H211$ controllers in Table 1. A smooth step size sequence is of importance, for example, to avoid higher order BDF methods to suffer stability problems.

Step Size Control, Table 1 Some recommended two-step controllers. The $H211$ controllers produce smooth step size sequences, using a moving average low-pass filter. In $H211b$, the filter can be adjusted. Starting at $b = 2$ it is a deadbeat (FIR) filter; as the parameter b increases, dynamic response slows and high frequency suppression (smoothing) increases. Note that the β_j coefficients are given in terms of the product $k\beta_j$ for use with error estimators of different orders k. The α coefficient is however independent of k (From [11])

$k\beta_1$	$k\beta_0$	α_0	Type	Name	Usage
$3/5$	$-1/5$	–	PI	PI.4.2	Nonstiff solvers
$1/b$	$1/b$	$1/b$	MA	$H211b$	Stiff solvers; $b \in [2,6]$
$1/6$	$1/6$	–	MA+PI	$H211$ PI	Stiff problems, smooth solutions

Implementation and Modes of Operation

Carefully implemented adaptivity algorithms are central for the code to operate efficiently and reliably for broad classes of problems. Apart from the accuracy requirements, which may be formulated in many different ways, there are several other factors of importance in connection with step size control.

EPS Versus EPUS

For a code that emphasizes asymptotically correct error estimates, controlling the local *error per unit step* $\|l_n\|/h_n$ is necessary in order to accumulate a targeted global error, over a fixed integration range, regardless of the number of steps needed to complete the integration. Abbreviated EPUS, this approach is viable for nonstiff problems, but tends to be costly for stiff problems, where strong dissipation usually means that the global error is dominated by the most recent local errors. There, controlling the local *error per step* $\|l_n\|$, referred to as EPS, is often a far more efficient option, if less well aligned with theory. In modern codes, the trend is generally to put less emphasis on asymptotically correct estimates, and control $\|l_n\|$ directly. This has few practical drawbacks, but it makes it less straightforward to compare the performance of two different codes.

Computational Stability

Just as a well-conditioned problem depends continuously on the data, the computational procedure should depend continuously on the various parameters that control the computation. In particular, for *tolerance proportionality*, there should be constants c and C such that the global error e can be bounded above and below,

$$c \cdot \mathrm{TOL}^{\gamma} \le \|e\| \le C \cdot \mathrm{TOL}^{\gamma}, \qquad (19)$$

where the method is tolerance proportional if $\gamma = 1$. The smaller the ratio C/c, the better is the *computational stability*, but C/c can only be made small with carefully implemented tools for adaptivity. Thus, with the elementary controller (10), prevented from making small step size changes, C/c is typically large, whereas if the controller is based on a digital filter (here $H211b$ with $b = 4$, cf. Table 1) allowing a continual change of the step size, the global error becomes a smooth function of TOL, see Fig. 2. This also shows that the behavior of an implementation is significantly affected by the control algorithms, and how they are implemented.

Absolute and Relative Errors

All modern codes provide options for controlling both absolute and relative errors. If at any given time, the estimated error is $\hat{l}$ and the computed solution is y, then a weighted error vector d with components

$$d_i = \frac{\hat{l}_i}{\eta_i + |y_i|} \qquad (20)$$

is constructed, where η_i is a scaling factor, determining a gradual switchover from relative to absolute error as $y_i \to 0$. The error (and the step) is accepted if $\|d\| \le \mathrm{TOL}$, and the expression $\mathrm{TOL}/\|d\|$ corresponds to the factors TOL/r in (10) and (13). By (20),

$$\frac{d_i}{\mathrm{TOL}} = \frac{\hat{l}_i}{\mathrm{TOL} \cdot \eta_i + \mathrm{TOL} \cdot |y_i|}. \qquad (21)$$

Most codes employ *two different tolerance parameters*, ATOL and RTOL, defined by $\mathrm{ATOL}_i = \mathrm{TOL} \cdot \eta_i$ and $\mathrm{RTOL} = \mathrm{TOL}$, respectively, replacing the denominator in (21) by $\mathrm{ATOL}_i + \mathrm{RTOL} \cdot |y_i|$. Thus, the user controls the accuracy by the vector ATOL and the scalar RTOL. For scaling purposes, it is also important to note that TOL and RTOL are dimensionless, whereas ATOL is not. The actual computational setup will make the step size control operate differently as the user-selected tolerance parameters affect both the set point and the control objective.

Interfering with the Controller

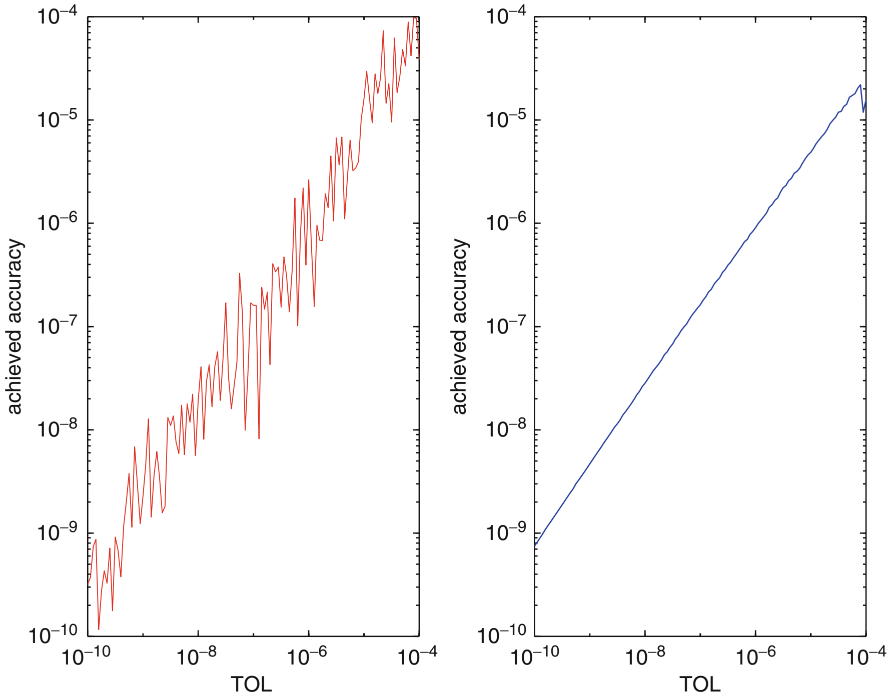
In most codes, a step is rejected if it exceeds TOL by a small amount, say 20 %, calling for the step to be recomputed. As a correctly implemented controller is expectation-value correct, a too large error is almost invariably compensated by other errors being too small. It is, therefore, in general harmless to accept steps that exceed TOL by as much as a factor of 2, and indeed often preferable to minimize interference with the controller's dynamics.

Other types of interference may come from conditions that prevent "small" step size changes, as this might call for a refactorization of the Jacobian. However, such concerns are not warranted with smooth controllers, which usually make small enough changes not to disturb the Newton process beyond what can be managed. On the contrary, a smoothly changing step size is beneficial for avoiding instability in multistep methods such as the BDF methods.

It is however necessary to interfere with the controller's action when there is a change of method order, or when a too large step size change is suggested. This is equivalent to encountering an error that is much larger or smaller than TOL. In the first case, the step needs to be rejected, and in the second, the step size increase must be held back by a limiter.

Special Problems

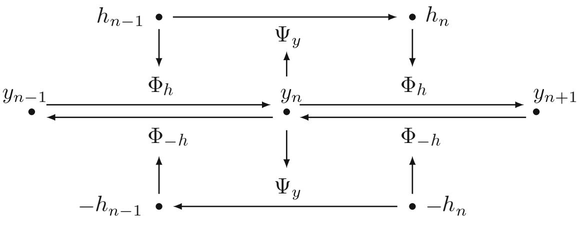
Conventional multiplicative control is not useful in connection with *geometric integration*, where it fails to preserve structure. The interaction (6, 7) shows that



Step Size Control, Fig. 2 Global error vs. TOL for a linear multistep code applied to a stiff nonlinear test problem. Left panel shows results when the controller is based on (10). In the right panel, it has been replaced by the digital filter $H211b$. Although computational effort remains unchanged, stability is much enhanced. The graphs also reveal that the code is not tolerance proportional

adaptive step size selection adds dynamics, interfering with structure preserving integrators.

A one-step method $\Phi_h : y_n \mapsto y_{n+1}$ is called *symmetric* if $\Phi_h^{-1} = \Phi_{-h}$. This is a minimal requirement for the numerical integration of, for example, *reversible Hamiltonian systems*, in order to nearly preserve action variables in integrable problems. To make such a method adaptive, *symmetric step size control* is also needed. An invertible step size map $\Psi_y : \mathbb{R} \to \mathbb{R}$ is called *symmetric* if $-\Psi_y$ is an involution, see Fig. 3. A symmetric Ψ_y then maps h_{n-1} to h_n and $-h_n$ to $-h_{n-1}$, and only depends on y_n; with these conditions satisfied, the adaptive integration can be run in reverse time and retrace the numerical trajectory that was generated in forward time, [8]. However, this cannot be achieved by multiplicative controllers (9), and a special, nonlinear controller is therefore necessary.


$$\begin{tikzcd}
 & h_{n-1} \arrow[rr] \arrow[d] & & h_n \arrow[d] & \\
 & \Phi_h & \Psi_y & \Phi_h & \\
y_{n-1} \arrow[rr, shift left] & & y_n \arrow[ll, shift left] \arrow[rr, shift left] \arrow[u] \arrow[d] & & y_{n+1} \arrow[ll, shift left] \\
 & \Phi_{-h} & \Psi_y & \Phi_{-h} & \\
 & -h_{n-1} \arrow[u] & & -h_n \arrow[ll] \arrow[u] &
\end{tikzcd}$$

Step Size Control, Fig. 3 Symmetric adaptive integration in forward time (*upper part*), and reverse time (*lower part*) illustrate the interaction (6, 7). The symmetric step size map Ψ_y governs both h and $-h$ (From [8])

An explicit control recursion satisfying the requirements is either *additive* or *inverse-additive*, with the latter being preferable. Thus, a controller of the form

$$\frac{1}{h_n} - \frac{1}{h_{n-1}} = G(y_n) \tag{22}$$

can be used, where the function G needs to be chosen with respect to the symmetry and geometric properties of the differential equation to be solved. This approach corresponds to constructing a *Hamiltonian continuous control system*, which is converted to the *discrete controller* (22) by geometric integration of the control system. This leaves the long-term behavior of the geometric integrator intact, even in the presence of step size variation. It is also worth noting that (22) generates a smooth step size sequence, as $h_n - h_{n-1} = \mathrm{O}(h_n h_{n-1})$.

This type of control does not work with an error estimate, but rather tracks a prescribed target function; it corresponds to keeping $hQ(y) = \text{const.}$, where Q is a given functional reflecting the geometric structure of (1). One can then take $G(y) = \operatorname{grad} Q(y) \cdot f(y)/Q(y)$. For example, in celestial mechanics, $Q(y)$ could be selected as total centripetal acceleration; then the step size is small when centripetal acceleration is large and vice versa, concentrating the computational effort to those intervals where the solution of the problem changes rapidly and is more sensitive to perturbations.

Literature

Step size control has a long history, starting with the first initial value problem solvers around 1960, often using a simple step doubling/halving strategy. The controller (10) was soon introduced, and further developments quickly followed. Although the schemes were largely heuristic, performance tests and practical experience developed working standards. Monographs such as [1, 2, 6, 7, 10] all offer detailed descriptions.

The first full control theoretic analysis is found in [3, 4], explaining and overcoming some previously noted difficulties, developing proportional-integral (PI) and autoregressive (AR) controllers. Synchronization with Newton iteration is discussed in [5]. A complete framework for using digital filters and signal processing is developed in [11], focusing on moving average (MA) controllers. Further developments on how to obtain improved computational stability are discussed in [12].

The special needs of geometric integration are discussed in [8], although the symmetric controllers are not based on error control. Error control in implicit, symmetric methods is analyzed in [13].

References

1. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. Wiley, Chichester (2008)
2. Gear, C.W.: Numerical Initial Value Problems in Ordinary Differential Equations. Prentice Hall, Englewood Cliffs (1971)
3. Gustafsson, K.: Control theoretic techniques for stepsize selection in explicit Runge–Kutta methods. ACM TOMS **17**, 533–554 (1991)
4. Gustafsson, K.: Control theoretic techniques for stepsize selection in implicit Runge–Kutta methods. ACM TOMS **20**, 496–517 (1994)
5. Gustafsson, K., Söderlind, G.: Control strategies for the iterative solution of nonlinear equations in ODE solvers. SIAM J. Sci. Comp. **18**, 23–40 (1997)
6. Hairer, E., Nϕrsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems, 2nd edn. Springer, Berlin (1993)
7. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, 2nd edn. Springer, Berlin (1996)
8. Hairer, E., Söderlind, G.: Explicit, time reversible, adaptive step size control. SIAM J. Sci. Comp. **26**, 1838–1851 (2005)
9. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer, Berlin (2006)
10. Shampine, L., Gordon, M.: Computer Solution of Ordinary Differential Equations: The Initial Value Problem. Freeman, San Francisco (1975)
11. Söderlind, G.: Digital filters in adaptive time-stepping. ACM Trans. Math. Softw. **29**, 1–26 (2003)
12. Söderlind, G., Wang, L.: Adaptive time-stepping and computational stability. J. Comp. Methods Sci. Eng. **185**, 225–243 (2006)
13. Stoffer, D.: Variable steps for reversible integration methods. Computing **55**, 1–22 (1995)

Stochastic and Statistical Methods in Climate, Atmosphere, and Ocean Science

Daan Crommelin[1,2] and Boualem Khouider[3]
[1] Scientific Computing Group, Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands
[2] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands
[3] Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada

Introduction

The behavior of the atmosphere, oceans, and climate is intrinsically uncertain. The basic physical principles that govern atmospheric and oceanic flows are well known, for example, the Navier-Stokes equations for fluid flow, thermodynamic properties of moist air, and the effects of density stratification and Coriolis force.

Notwithstanding, there are major sources of randomness and uncertainty that prevent perfect prediction and complete understanding of these flows.

The climate system involves a wide spectrum of space and time scales due to processes occurring on the order of microns and milliseconds such as the formation of cloud and rain droplets to global phenomena involving annual and decadal oscillations such as the EL Nio-Southern Oscillation (ENSO) and the Pacific Decadal Oscillation (PDO) [5]. Moreover, climate records display a spectral variability ranging from 1 cycle per month to 1 cycle per 100,000 years [23]. The complexity of the climate system stems in large part from the inherent nonlinearities of fluid mechanics and the phase changes of water substances. The atmosphere and oceans are turbulent, nonlinear systems that display chaotic behavior (e.g., [39]). The time evolutions of the same chaotic system starting from two slightly different initial states diverge exponentially fast, so that chaotic systems are marked by limited predictability. Beyond the so-called predictability horizon (on the order of 10 days for the atmosphere), initial state uncertainties (e.g., due to imperfect observations) have grown to the point that straightforward forecasts are no longer useful.

Another major source of uncertainty stems from the fact that numerical models for atmospheric and oceanic flows cannot describe all relevant physical processes at once. These models are in essence discretized partial differential equations (PDEs), and the derivation of suitable PDEs (e.g., the so-called primitive equations) from more general ones that are less convenient for computation (e.g., the full Navier-Stokes equations) involves approximations and simplifications that introduce errors in the equations. Furthermore, as a result of spatial discretization of the PDEs, numerical models have finite resolution so that small-scale processes with length scales below the model grid scale are not resolved. These limitations are unavoidable, leading to model error and uncertainty.

The uncertainties due to chaotic behavior and unresolved processes motivate the use of stochastic and statistical methods for modeling and understanding climate, atmosphere, and oceans. Models can be augmented with random elements in order to represent time-evolving uncertainties, leading to stochastic models. Weather forecasts and climate predictions are increasingly expressed in probabilistic terms, making explicit the margins of uncertainty inherent to any prediction.

Statistical Methods

For assessment and validation of models, a comparison of individual model trajectories is typically not suitable, because of the uncertainties described earlier. Rather, the statistical properties of models are used to summarize model behavior and to compare against other models and against observations. Examples are the mean and variance of spatial patterns of rainfall or sea surface temperature, the time evolution of global mean temperature, and the statistics of extreme events (e.g., hurricanes or heat waves). Part of the statistical methods used in this context is fairly general, not specifically tied to climate-atmosphere-ocean science (CAOS). However, other methods are rather specific for CAOS applications, and we will highlight some of these here. General references on statistical methods in CAOS are [61, 62].

EOFs

A technique that is used widely in CAOS is Principal Component Analysis (PCA), also known as Empirical Orthogonal Function (EOF) analysis in CAOS. Consider a multivariate dataset $\Phi \in \mathbf{R}^{M \times N}$. In CAOS this will typically be a time series $\phi(t_1), \phi(t_2), \ldots, \phi(t_N)$ where each $\phi(t_n) \in \mathbf{R}^M$ is a spatial field (of, e.g., temperature or pressure). For simplicity we assume that the time mean has been subtracted from the dataset, so $\sum_{n=1}^{N} \Phi_{mn} = 0 \;\; \forall m$. Let C be the $M \times M$ (sample) covariance matrix for this dataset:

$$C = \frac{1}{N-1} \, \Phi \, \Phi^T \, .$$

We denote by (λ_m, v^m), $m = 1, \ldots, M$ the ordered eigenpairs of C:

$$C \, v^m = \lambda_m \, v^m \, , \quad \lambda_m \geq \lambda_{m+1} \;\; \forall m .$$

The ordering of the (positive) eigenvalues implies that the projection of the dataset onto the leading eigenvector v^1 gives the maximum variance among all projections. The next eigenvector v^2 gives the maximum variance among all projections orthogonal to v^1, v^3 gives maximum variance among all projections orthogonal to v^1 and v^2, etc. The fraction $\lambda_m / \sum_l \lambda_l$ equals the fraction of the total variance of the data captured by projection onto the m-th eigenvector v^m.

The eigenvectors v^m are called the Empirical Orthogonal Functions (EOFs) or Principal Components (PCs). Projecting the original dataset Φ onto the leading EOFs, i.e., the projection/reduction

$$\phi^r(t_n) = \sum_{m=1}^{M'} \alpha_m(t_n) v^m \,, \qquad M' \ll M,$$

can result in a substantial data reduction while retaining most of the variance of the original data.

PCA is discussed in great detail in [27] and [59]. Over the years, various generalizations and alternatives for PCA have been formulated, for example, Principal Interaction and Oscillation Patterns [24], Nonlinear Principal Component Analysis (NLPCA) [49], and Nonlinear Laplacian Spectral Analysis (NLSA) [22]. These more advanced methods are designed to overcome limitations of PCA relating to the nonlinear or dynamical structure of datasets.

In CAOS, the EOFs v^m often correspond to spatial patterns. The shape of the patterns of leading EOFs can give insight in the physical-dynamical processes underlying the dataset Φ. However, this must be done with caution, as the EOFs are statistical constructions and cannot always be interpreted as having physical or dynamical meaning in themselves (see [50] for a discussion).

The temporal properties of the (time-dependent) coefficients $\alpha_m(t)$ can be analyzed by calculating, e.g., autocorrelation functions. Also, models for these coefficients can be formulated (in terms of ordinary differential equations (ODEs), stochastic differential equations (SDEs), etc.) that aim to capture the main dynamical properties of the original dataset or model variables $\phi(t)$. For such reduced models, the emphasis is usually on the dynamics on large spatial scales and long time scales. These are embodied by the leading EOFs v^m, $m = 1, \ldots, M'$, and their corresponding coefficients $\alpha_m(t)$, so that a reduced model ($M' \ll M$) can be well capable of capturing the main large-scale dynamical properties of the original dataset.

Inverse Modeling

One way of arriving at reduced models is inverse modeling, i.e., the dynamical model is obtained through statistical inference from time series data. The data can be the result of, e.g., projecting the dataset Φ onto the EOFs (in which case the data are time series of $\alpha(t)$). These models are often cast as SDEs whose parameters must be estimated from the available time series. If the SDEs are restricted to have linear drift and additive noise (i.e., restricted to be those of a multivariate Ornstein-Uhlenbeck (OU) process), the estimation can be carried out for high-dimensional SDEs rather easily. That is, assume the SDEs have the form

$$d\alpha(t) = B\,\alpha(t)\,dt + \sigma\,dW(t)\,, \tag{1}$$

in which B and σ are both a constant real $M' \times M'$ matrix and $W(t)$ is an M'-dimensional vector of independent Wiener processes (for simplicity we assume that α has zero mean). The parameters of this model are the matrix elements of B an σ. They can be estimated from two (lagged) covariance matrices of the time series. If we define

$$R^0_{ij} = \mathbf{E}\alpha_i(t)\alpha_j(t)\,, \qquad R^\tau_{ij} = \mathbf{E}\alpha_i(t)\alpha_j(t+\tau)\,,$$

with $\mathbf{E}$ denoting expectation, then for the OU process (1), we have the relations

$$R^\tau = \exp(B\,\tau)\,R^0$$

and

$$BR^0 + R^0B^T + \sigma\sigma^T = 0$$

The latter of these is the fluctuation-dissipation relation for the OU process. By estimating R^0 and R^τ (with some $\tau > 0$) from time series of α, estimates for B and $A := \sigma\sigma^T$ can be easily computed using these relations. This procedure is sometimes referred to as linear inverse modeling (LIM) in CAOS [55]. The matrix σ cannot be uniquely determined from A; however, any σ for which $A = \sigma\sigma^T$ (e.g., obtained by Cholesky decomposition of A) will result in an OU process with the desired covariances R^0 and R^τ.

As mentioned, LIM can be carried out rather easily for multivariate processes. This is a major advantage of LIM. A drawback is that the OU process (1) cannot capture non-Gaussian properties, so that LIM can only be used for data with Gaussian distributions. Also, the estimated B and A are sensitive to the choice of τ, unless the available time series is a an exact sampling of (1).

Estimating diffusion processes with non-Gaussian properties is much more complicated. There are various estimation procedures available for SDEs with nonlinear drift and/or multiplicative noise; see, e.g., [30, 58] for an overview. However, the practical use of these procedures is often limited to SDEs with very low dimensions, due to curse of dimension or to computational feasibility. For an example application in CAOS, see, e.g., [4].

The dynamics of given time series can also be captured by reduced models that have discrete state spaces, rather than continuous ones as in the case of SDEs. There are a number of studies in CAOS that employ finite-state Markov chains for this purpose (e.g., [8,48,53]). It usually requires discretization of the state space; this can be achieved with, e.g., clustering methods. A more advanced methodology, building on the concept of Markov chains yet resulting in continuous state spaces, is that of hidden Markov models. These have been used, e.g., to model rainfall data (e.g., [3, 63]) and to study regime behavior in large-scale atmospheric dynamics [41]. Yet a more sophisticated methodology that combines the clustering and Markov chain concepts, specifically designed for nonstationary processes, can be found in [25].

Extreme Events

The occurrence of extreme meteorological events, such as hurricanes, extreme rainfall, and heat waves, is of great importance because of their societal impact. Statistical methods to study extreme events are therefore used extensively in CAOS. The key question for studying extremes with statistical methods is to be able to assess the probability of certain events, having only a dataset available that is too short to contain more than a few of these events (and occasionally, too short to contain even a single event of interest). For example, how can one assess the probability of sea water level at some coastal location being more than 5 m above average if only 100 years of observational data for that location is available, with a maximum of 4 m above average? Such questions can be made accessible using extreme value theory. General introductions to extreme value theory are, e.g., [7] and [11]. For recent research on extremes in the context of climate science, see, e.g., [29] and the collection [1].

The classical theory deals with sequences or observations of N independent and identically distributed (iid) random variables, denoted here by $r_1, \dots, r_N$. Let M_N be the maximum of this sequence, $M_N = \max\{r_1, \dots, r_N\}$. If the probability distribution for M_N can be rescaled so that it converges in the limit of increasingly long sequences (i.e., $N \to \infty$), it converges to a generalized extreme value (GEV) distribution. More precisely, if there are sequences $a_N (> 0)$ and b_N such that $\text{Prob}((M_N - b_N)/a_N \leq z) \to G(z)$ as $N \to \infty$, then

$$G(z) = \exp\left(-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right]\right).$$

$G(z)$ is a GEV distribution, with parameters μ (location), $\sigma > 0$ (scale), and ξ (shape). It combines the Fréchet ($\xi > 0$), Weibull ($\xi < 0$), and Gumbel ($\xi \to 0$) families of extreme value distributions. Note that this result is independent of the precise distribution of the random variables r_n. The parameters μ, σ, ξ can be inferred by dividing the observations $r_1, r_2, \dots$ in blocks of equal length and considering the maxima on these blocks (the so-called block maxima approach).

An alternative method for characterizing extremes, making more efficient use of available data than the block maxima approach, is known as the peaks-over-threshold (POT) approach. The idea is to set a threshold, say r^*, and study the distribution of all observations r_n that exceed this threshold. Thus, the object of interest is the conditional probability distribution $\text{Prob}(r_n - r^* > z \,|\, r_n > r^*)$, with $z > 0$. Under fairly general conditions, this distribution converges to $1 - H(z)$ for high thresholds r^*, where $H(z)$ is the generalized Pareto distribution (GPD):

$$H(z) = 1 - \left(1 + \frac{\xi z}{\tilde{\sigma}}\right)^{-1/\xi}.$$

The parameters of the GPD family of distributions are directly related to those of the GEV distribution: the shape parameter ξ is the same in both, whereas the threshold-dependent scale parameter is $\tilde{\sigma} = \sigma + \xi(r^* - \mu)$ with μ and σ as in the GEV distribution.

By inferring the parameters of the GPD or GEV distributions from a given dataset, one can calculate probabilities of extremes that are not present themselves in that dataset (but have the same underlying distribution as the available data). In principle, this makes it possible to assess risks of events that have not been observed, provided the conditions on convergence to GPD or GEV distributions are met.

As mentioned, classical results on extreme value theory apply to iid random variables. These results have been generalized to time-correlated random variables, both stationary and nonstationary [7]. This is important for weather and climate applications, where datasets considered in the context of extremes are often time series. Another relevant topic is the development of multivariate extreme value theory [11].

Stochastic Methods

Given the sheer complexity of climate-atmosphere-ocean (CAO) dynamics, when studying the global climate system or some parts of global oscillation patterns such ENSO or PDO, it is natural to try to separate the global dynamics occurring on longer time scales from local processes which occur on much shorter scales. Moreover, as mentioned before, climate and weather prediction models are based on a numerical discretization of the equations of motion, and due to limitations in computing resources, it is simply impossible to represent the wide range of space and time scales involved in CAO. Instead, general circulation models (GCMs) rely on parameterization schemes to represent the effect of the small/unresolved scales on the large/resolved scales. Below, we briefly illustrate how stochastic models are used in CAO both to build theoretical models that separate small-scale (noise) and large-scale dynamics and to "parameterize" the effect of small scales on large scales. A good snapshot on the state of the art, during the last two decades or so, in stochastic climate modeling research can be found in [26, 52].

Model Reduction for Noise-Driven Large-Scale Dynamics

In an attempt to explain the observed low-frequency variability of CAO, Hasselmann [23] splits the system into slow climate components (e.g., oceans, biosphere, cryosphere), denoted by the vector x, and fast components representing the weather, i.e., atmospheric variability, denoted by a vector y. The full climate system takes the form

$$\frac{dx}{dt} = u(x, y) \tag{2}$$
$$\frac{dy}{dt} = v(x, y),$$

where t is time and $u(x, v)$ and $v(x, y)$ contain the external forcing and internal dynamics that couple the slow and fast variables.

Hasselmann assumes a large scale-separation between the slow and fast time scales: $\tau_y = O\left(y_j \left(\frac{dy_j}{dt}\right)^{-1}\right) \ll \tau_x = O\left(x_i \left(\frac{dx_i}{dt}\right)^{-1}\right)$, for all components i and j. The time scale separation was used earlier to justify statistical dynamical models (SDM) used then to track the dynamics of the climate system alone under the influence of external forcing. Without the variability due to the internal interactions of CAO, the SDMs failed badly to explain the observed "red" spectrum which characterizes low-frequency variability of CAO.

Hasselmann made the analogy with the Brownian motion (BM), modeling the erratic movements of a few large particles immersed in a fluid that are subject to bombardments by the rapidly moving fluid molecules as a "natural" extension of the SDM models. Moreover, Hasselmann [23] assumes that the variability of x can be divided into a mean tendency $\langle dx/dt \rangle = \langle u(x, y) \rangle$ (Here $\langle . \rangle$ denotes average with respect to the joint distribution of the fast variables.) and a fluctuation tendency $dx'/dt = u(x, y) - \langle u(x, y) \rangle = u'(x, y)$ which, according to the Brownian motion problem, is assumed to be a pure diffusion process or white noise. However, unlike BM, Hasselmann argued that for the weather and climate system, the statistics of y are not in equilibrium but depend on the slowly evolving large-scale dynamics and thus can only be obtained empirically. To avoid linear growth of the covariance matrix $\langle x' \otimes x' \rangle$, Hasselmann assumes a damping term proportional to the divergence of the background frequency $F(0)$ of $\langle x' \otimes x' \rangle$, where $\delta(\omega - \omega')F_{ij}(\omega) = \langle V_i(\omega) V_j(\omega') \rangle$ with $V(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} u'(t) e^{-i\omega t} dt$. This leads to the Fokker-Plank equation: [23]

$$\frac{\partial p(x, t)}{\partial t} + \nabla_x \cdot (\hat{u}(x) p(x, t)) = \nabla_x \cdot (D \nabla_x p(x, t)) \tag{3}$$

for the distribution $p(x, t)$ of $x(t)$ as a stochastic process given that $x(0) = x_0$, where D is the normalized covariance matrix $D = \langle x' \otimes x' \rangle / 2t$ and $\hat{u} = \langle u \rangle - \pi \nabla_x \cdot F(0)$. Given the knowledge of the mean statistical forcing $\langle u \rangle$, the evolution equation for p can be determined from the time series of x obtained either from a climate model simulation of from observations. Notice also that for a large number of slow variables x_i, the PDE in (3) is impractical; instead, one can always resort to Monte Carlo simulations using the associated Langevin equation:

$$dx = \hat{u}(x)dt + \Sigma(x)dW_t \tag{4}$$

where $\Sigma(x)\Sigma(x)^T = D(x)$. However, the functional dependence of $\hat{u}$ and D remains ambiguous, and relying on rather empirical methods to define such terms is unsatisfactory. Nonetheless, Hasselmann introduced a "linear feedback" version of his model where the drift or propagation term is a negative definite linear operator: $\hat{u}(x) = Ux$ and D is constant, independent of x as an approximation for short time excursions of the climate variables. In this case, $p(x,t)$ is simply a Gaussian distribution whose time-dependent mean and variance are determined by the matrices D and U as noted in the inverse modeling section above.

Due to its simplicity, the linear feedback model is widely used to study the low-frequency variability of various climate processes. It is, for instance, used in [17] to reproduce the observed red spectrum of the sea surface temperature in midlatitudes using simulation data from a simplified coupled ocean-atmosphere model. However, this linear model has severe limitations of, for example, not being able to represent deviations from Gaussian distribution of some climate phenomena [13, 14, 17, 36, 51, 54]. It is thus natural to try to reincorporate a nonlinearity of some kind into the model. The most popular idea consisted in making the matrix D or equivalently Σ dependent on x (quadratically for D or linearly for Σ as a next order Taylor correction) to which is tied the notion of multiplicative versus additive (when D is constant) noise [37, 60]. Beside the crude approximation, the apparent advantage of this approach is the maintaining of the stabilizing linear operator U in place although it is not universally justified.

A mathematical justification for Hasselmann's framework is provided by Arnold and his collaborators (see [2] and references therein). It is based on the well-known technique of averaging (the law of large numbers) and the central limit theorem. However, as in Hasselmann's original work, it assumes the existence and knowledge of the invariant measure of the fast variables. Nonetheless, a rigorous mathematical derivation of such Langevin-type models for the slow climate dynamics, using the equations of motion in discrete form, is possible as illustrated by the MTV theory presented next.

The Systematic Mode Reduction MTV Methodology

A systematic mathematical methodology to derive Langevin-type equations (4) à la Hasselmann, for the slow climate dynamics from the coupled atmosphere-ocean-land equations of motion, which yields the propagation (or drift) and diffusion terms $\hat{u}(x)$ and $D(x)$ in closed form, is presented in [44, 45] by Majda, Timofeyev, and Vanden-Eijnden (MTV).

Starting from the generalized form of the discretized equations of motion

$$\frac{dz}{dt} = Lz + B(z,z) + f(t)$$

where L and B are a linear and a bilinear operators while $f(t)$ represent external forcing, MTV operate the same dichotomy as Hasselmann did of splitting the vector z into slow and fast variables x and y, respectively. However, they introduced a nondimensional parameter $\epsilon = \tau_y/\tau_x$ which measures the degree of time scale separation between the two sets of variables. This leads to the slow-fast coupled system

$$\begin{aligned} dx =& \epsilon^{-1}(L_{11}x + L_{12}y)\,dt + B^1_{11}(x,x)dt \\ &+ \epsilon^{-1}\left(B^1_{12}(x,y) + B^1_{22}(y,y)\right)dt \\ &+ Dxdt + F_1(t)dt + \epsilon^{-1} f_1(\epsilon^{-1}t) \\ dy =& \epsilon^{-1}\left(L_{21}x + L_{22}y + B^2_{12}(x,y) + B^2_{22}(y,y)\right)dt \\ &- \epsilon^{-2}\Gamma y dt + \epsilon^{-1}\sigma dW_t + \epsilon^{-1} f_2(\epsilon^{-1}t) \end{aligned} \tag{5}$$

under a few key assumptions, including (1) the nonlinear self interaction term of the fast variables is "parameterized" by an Ornstein-Uhlenbeck process: $B^2_{22}(y,y)dt := -\epsilon^{-1}\Gamma y dt + \sqrt{\epsilon}^{-1} dW_t$ and (2) a small dissipation term $\epsilon D x dt$ is added to the slow dynamics while (3) the slow variable forcing term assumes slow and fast contributions $f_1(t) = \epsilon F_1(\epsilon t) + f_1(t)$. Moreover, the system in (5) is written in terms of the slow time $t \longrightarrow \epsilon t$.

MTV used the theory of asymptotic expansion applied to the backward Fokker-Plank equation associated with the stochastic differential system in (5) to obtain an effective reduced Langevin equation (4) for the slow variables x in the limit of large separation of time scales $\epsilon \longrightarrow 0$ [44, 45]. The main advantage of the MTV theory is that unlike Hasselmann's ad hoc formulation, the functional form of the drift and diffusion coefficients, in terms of the slow variables, are obtained and new physical phenomena can emerge from the large-scale feedback besides the assumed stabilization effect. It turns out that the drift term is not always stabilizing, but there are dynamical regimes where growing modes can be excited and, depending on the dynamical configuration, the Langevin equation (4) can support either additive or multiplicative noise.

Even though MTV assumes strict separation of scales, $\epsilon \ll 1$, it is successfully used for a wide range of examples including cases where $\epsilon = O(1)$ [46]. Also in [47], MTV is successfully extended to fully deterministic systems where the requirement that the fast-fast interaction term $B_{22}(y, y)$ in (5) is parameterized by an Ornstein-Uhlenbeck process is relaxed. Furthermore, MTV is applied to a wide range of climate problems. It is used, for instance, in [19] for a realistic barotropic model and extended in [18] to a three-layer quasi-geostrophic model. The example of midlatitude teleconnection patterns where multiplicative noise plays a crucial role is studied in [42]. MTV is also applied to the triad and dyad normal mode (EOF) interactions for arbitrary time series [40].

Stochastic Parametrization

In a typical GCM, the parametrization of unresolved processes is based on theoretical and/or empirical deterministic equations. Perhaps the area where deterministic parameterizations have failed the most is moist convection. GCMs fail very badly in simulating the planetary and intra-seasonal variability of winds and rainfall in the tropics due to the inadequate representation of the unresolved variability of convection and the associated cross-scale interactions behind the multiscale organization of tropical convection [35]. To overcome this problem, some climate scientists introduced random variables to mimic the variability of such unresolved processes. Unfortunately, as illustrated below, many of the existing stochastic parametrizations were based on the assumptions of statistical equilibrium and/or of a stationary distribution for the unresolved variability, which are only valid to some extent when there is scale separation.

The first use of random variables in CGMs appeared in Buizza et al. [6] as means for improving the skill of the ECMWF ensemble prediction system (EPS). Buizza et al. [6] used uniformly distributed random scalars to rescale the parameterized tendencies in the governing equations. Similarly, Lin and Neelin [38] introduced a random perturbation in the tendency of convective available potential energy (CAPE). In [38], the random noise is assumed to be a Markov process of the form $\xi_{t+\Delta t} = \epsilon_t \xi_t + z_t$ where z_t is a white noise with a fixed standard deviation and ϵ_t is a parameter. Plant and Craig [57] used extensive cloud-permitting numerical simulations to empirically derive the parameters for the PDF of the cloud base mass flux itself whose Poisson shape is determined according to arguments drawn from equilibrium statistical mechanics. Careful simulations conducted by Davoudi et al. [10] revealed that while the Poisson PDF is more or less accurate for isolated deep convective clouds, it fails to extend to cloud clusters where a variety of cloud types interact with each other: a crucial feature of organized tropical convection.

Majda and Khouider [43] borrowed an idea from material science [28] of using the Ising model of ferromagnetization to represent convective inhibition (CIN). An order parameter σ, defined on a rectangular lattice, embedded within each horizontal grid box of the climate model, takes values 1 or 0 at a given site, according to whether there is CIN or there is potential for deep convection (PAC). The lattice model makes transitions at a given site according to intuitive probability rules depending both on the large-scale climate model variables and on local interactions between lattice sites based on a Hamiltonian energy principle. The Hamiltonian is given by

$$H(\sigma, U) = -\frac{1}{2} \sum_{x,y} J(|x-y|)\sigma(x)\sigma(y) + h(U) \sum_{x} \sigma_x$$

where $J(r)$ is the local interaction potential and $h(U)$ is the external potential which depends on the climate variables U and where the summations are taken over all lattice sites x, y. A transition (spin-flip by analogy to the Ising model of magnetization) occurs at a site y if for a small time τ, we have $\sigma_{t+\tau}(y) = 1 - \sigma_t(y)$ and $\sigma_{t+\tau}(x) = \sigma_t(x)$ if $x \neq y$. Transitions occur at a rate $C(y, \sigma, U)$ set by Arrhenius dynamics: $C(x, \sigma, U) = \frac{1}{\tau_I} \exp(-\Delta_x H(\sigma, U))$ if $\sigma_x = 0$ and $C(x, \sigma, U) = \frac{1}{\tau_I}$ if $\sigma_x = 1$ so that the resulting Markov process satisfies detailed balance with respect to the Gibbs distribution $\mu(\sigma, U) \propto \exp(-H(\sigma, U)$. Here

$\Delta_x H(\sigma, U)) = H(\sigma + [1-\sigma(x)]e_x), U) - H(\sigma, U) = -\sum_z J(|x-z|)\sigma(z) + h(U)$ with $e_x(y) = 1$ if $y = x$ and 0 otherwise.

For computational efficiency, a coarse graining of the stochastic CIN model is used in [34] to derive a stochastic birth-death process for the mesoscopic area coverage $\eta_X = \sum_{x \in X} \sigma(x)$ where X represents a generic site of a mesoscopic lattice, which in practice can be considered to be the GCM grid. The stochastic CIN model is coupled to a toy GCM where it is successfully demonstrated how the addition of such a stochastic model could improve the climatology and waves dynamics in a deficient GCM [34, 42].

This Ising-type modeling framework is extended in [33] to represent the variability of organized tropical convection (OTC). A multi-type order parameter is introduced to mimic the multimodal nature of OTC. Based on observations, tropical convective systems (TCS) are characterized by three cloud types, cumulus congestus whose height does not exceed the freezing level develop when the atmosphere is dry, and there is convective instability, positive CAPE. In return congestus clouds moisten the environment for deep convective towers. Stratiform clouds that develop in the upper troposphere lag deep convection as a natural freezing phase in the upper troposphere. Accordingly, the new order parameter σ takes the multiple values 0,1,2,3, on a given lattice site, according to whether the given site is, respectively, clear sky or occupied by a congestus, deep, or stratiform cloud.

Similar Arrhenius-type dynamics are used to build transition rates resulting in an ergodic Markov process with a well-defined equilibrium measure. Unphysical transitions of congestus to stratiform, stratiform to deep, stratiform to congestus, clear to stratiform, and deep to congestus were eliminated by setting the associated rates to zero. When local interactions are ignored, the equilibrium measure and the transition rates depend only on the large-scale climate variables U where CAPE and midlevel moisture are used as triggers and the coarse-graining process is carried with exact statistics. It leads to a multidimensional birth-death process with immigration for the area fractions of the associated three cloud types. The stochastic multicloud model (SMCM) is used very successfully in [20, 21] to capture the unresolved variability of organized convection in a toy GCM. The simulation of convectively coupled gravity waves and mean climatology were improved drastically when compared to their deterministic counterparts. The realistic statistical behavior of the SMCM is successfully assessed against observations in [56]. Local interaction effects are reintroduced in [32] where a coarse-graining approximation based on conditional expectation is used to recover the multidimensional birth-death process dynamics with local interactions. A Bayesian methodology for inferring key parameters for the SMCM is developed and validated in [12]. A review of the basic methodology of the CIN and SMCM models, which is suitable for undergraduates, is found in [31].

A systematic data-based methodology for inferring a suitable stochastic process for unresolved processes conditional on resolved model variables was proposed in [9]. The local feedback from unresolved processes on resolved ones is represented by a small Markov chain whose transition probability matrix is made dependent on the resolved-scale state. The matrix is estimated from time series data that is obtained from highly resolved numerical simulations or observations. This approach was developed and successfully tested on the Lorenz '96 system [39] in [9]. [16] applied it to parameterize shallow cumulus convection, using data from large eddy simulation (LES) of moist atmospheric convection. A two-dimensional lattice, with at each lattice node a Markov chain, was used to mimic (or emulate) the convection as simulated by the high-resolution LES model, at a fraction of the computational cost.

Subsequently, [15] combined the conditional Markov chain methodology with elements from the SMCM [33]. They applied it to deep convection but without making use of the Arrhenius functional forms of the transition rates in terms of the large-scale variables (as was done in [33]). Similar to [16], LES data was used for estimation of the Markov chain transition probabilities. The inferred stochastic model in [15] was well capable of generating cloud fractions very similar to those observed in the LES data. While the main cloud types of the original SMCM were preserved, an important improvement in [15] resides in the addition of a fifth state for shallow cumulus clouds. As an experiment, direct spatial coupling of the Markov chains on the lattice was also considered in [15]. Such coupling amounts to the structure of a stochastic cellular automaton (SCA). Without this direct coupling, the Markov chains are still coupled, but indirectly, through their interaction with the large-scale variables (see, e.g., [9]).

References

1. AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., Sorooshian, S. (eds.): Extremes in a Changing Climate, p. 423. Springer, Dordrecht (2013)
2. Arnold, L.: Hasselmann's program revisited: the analysis of stochasticity in deterministic climate models. In: Imkeller, P., von Storch, J.-S. (eds) Stochastic Climate Models. Birkhäuser, Basel (2001)
3. Bellone, E., Hughes, J.P., Guttorp, P.: A Hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Clim. Res. **15**, 1–12 (2000)
4. Berner, J.: Linking nonlinearity and non-Gaussianity of planetary wave behavior by the Fokker-Planck equation. J. Atmos. Sci. **62**, 2098–2117 (2005)
5. Bond, N.A., Harrison, D.E.: The pacific decadal oscillation, air-sea interaction and central north pacific winter atmospheric regimes. Geophys. Res. Lett. **27**, 731–734 (2000)
6. Buizza, R., Milleer, M., Palmer, T.N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Quart. J. R. Meteorol. Soc. **125**, 2887–2908 (1999)
7. Coles, S.: An Introduction to Statistical Modeling of Extreme Values, p. 208. Springer, London (2001); Statistical Methods in the Atmospheric Sciences, 3rd edn., p. 704. Academic, Oxford
8. Crommelin, D.T.: Observed non-diffusive dynamics in large-scale atmospheric flow. J. Atmos. Sci. **61**, 2384–239 (2004)
9. Crommelin, D.T., Vanden-Eijnden, E.: Subgrid-scale parameterization with conditional Markov chains. J. Atmos. Sci. **65**, 2661–2675 (2008)
10. Davoudi, J., McFarlane, N.A., Birner, T.: Fluctuation of mass flux in a cloud-resolving simulation with interactive radiation. J. Atmos. Sci. **67**, 400–418 (2010)
11. de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction, p. 417. Springer, New York (2006)
12. de la Chevrotière, M., Khouider, B., Majda, A.: Calibration of the stochastic multicloud model using Bayesian inference. SIAM J. Sci. Comput. **36**(3), B538–B560 (2014)
13. DelSole, T.: Stochastic models of quasi-geostrophic turbulence. Surv. Geophys. **25**, 107–149 (2004)
14. DelSole, T., Farrel, B.F.: Quasi-linear equilibration of a thermally maintained, stochastically excited jet in a quasi-geostrophic model. J. Atmos. Sci. **53**, 1781–1797 (1996)
15. Dorrestijn, J., Crommelin, D.T., Biello, J.A., Böing, S.J.: A data-driven multicloud model for stochastic parameterization of deep convection. Phil. Trans. R. Soc. A **371**, 20120374 (2013)
16. Dorrestijn, J., Crommelin, D.T., Siebesma, A.P., Jonker, H.J.J.: Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. Theor. Comput. Fluid Dyn. **27**, 133–148 (2013)
17. Frankignoul, C., Hasselmann, K.: Stochastic climate models, Part II application to sea-surface temperature anomalies and thermocline variability. Tellus **29**, 289–305 (1977)
18. Franzke, C., Majda, A.: Low order stochastic mode reduction for a prototype atmospheric GCM. J. Atmos. Sci. **63**, 457–479 (2006)
19. Franzke, C., Majda, A., Vanden-Eijnden, E.: Low-order stochastic mode reduction for a realistic barotropic model climate. J. Atmos. Sci. **62**, 1722–1745 (2005)
20. Frenkel, Y., Majda, J., Khouider, B.: Using the stochastic multicloud model to improve tropical convective parameterization: a paradigm example. J. Atmos. Sci. **69**, 1080–1105 (2012)
21. Frenkel, Y., Majda, J., Khouider, B.: Stochastic and deterministic multicloud parameterizations for tropical convection. Clim. Dyn. **41**, 1527–1551 (2013)
22. Giannakis, D., Majda, A.J.: Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. Proc. Natl. Acad. Sci. USA **109**, 2222–2227 (2012)
23. Hasselmann, K.: Stochastic climate models. Part I, theory. Tellus **28**, 473–485 (1976)
24. Hasselmann, K.: PIPs and POPs: the reduction of complex dynamical systems using principal interaction and oscillation patterns. J. Geophys. Res. **93**(D9), 11015–11021 (1988)
25. Horenko, I.: Nonstationarity in multifactor models of discrete jump processes, memory, and application to cloud modeling. J. Atmos. Sci. **68**, 1493–1506 (2011)
26. Imkeller P., von Storch J.-S. (eds.): Stochastic Climate Models. Birkhäuser, Basel (2001)
27. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
28. Katsoulakis, M., Majda, A.J., Vlachos, D.: Coarse-Grained stochastic processes and Monte-Carlo simulations in lattice systems. J. Comp. Phys. **186**, 250–278 (2003)
29. Katz, R.W., Naveau P.: Editorial: special issue on statistics of extremes in weather and climate. Extremes **13**, 107–108 (2010)
30. Kessler, M., Lindner, A., Sørensen, M. (eds.): Statistical Methods for Stochastic Differential Equations. CRC, Boca Raton (2012)
31. Khouider, B.: Markov-jump stochastic models for organized tropical convection. In: Yang, X.-S. (ed.) Mathematical Modeling With Multidisciplinary Applications. Wiley, (2013). ISBN: 978-1-1182-9441-3
32. Khouider, B.: A stochastic coarse grained multi-type particle interacting model for tropical convection. Commun. Math. Sci. **12**, 1379–1407 (2014)
33. Khouider, B., Biello, J., Majda, A.J.: A stochastic multicloud model for tropical convection. Commun. Math. Sci. **8**, 187–216 (2010)
34. Khouider, B., Majda, A.J., Katsoulakis, M.: Coarse grained stochastic models for tropical convection. Proc. Natl. Acad. Sci. USA **100**, 11941–11946 (2003)
35. Khouider, B., Majda, A.J., Stechmann, S.: Climate science in the tropics: waves, vortices, and PDEs. Nonlinearity **26**, R1-R68 (2013)
36. Kleeman, R., Moore, A.: A theory for the limitation of ENSO predictability due to stochastic atmospheric transients. J. Atmos. Sci. **54**, 753–767 (1997)
37. Kondrasov, D., Krastov, S., Gill, M.: Empirical mode reduction in a model of extra-tropical low-frequency variability. J. Atmos. Sci. **63**, 1859–1877 (2006)

38. Lin, J.B.W., Neelin, D.: Toward stochastic deep convective parameterization in general circulation models. Geophy. Res. Lett. **27**, 3691–3694 (2000)
39. Lorenz, E.N.: Predictability: a problem partly solved. In: Proceedings, Seminar on Predictability ECMWF, Reading, vol. 1, pp. 1–18 (1996)
40. Majda, A., Franzke, C., Crommelin, D.: Normal forms for reduced stochastic climate models. Proc. Natl. Acad. Sci. USA **16**, 3649–3653 (2009)
41. Majda, A.J., Franzke, C.L., Fischer, A., Crommelin, D.T.: Distinct metastable atmospheric regimes despite nearly Gaussian statistics: a paradigm model. Proc. Natl. Acad. Sci. USA **103**, 8309–8314 (2006)
42. Majda, A.J., Franzke, C.L., Khouider, B.: An applied mathematics perspective on stochastic modelling for climate. Phil. Trans. R. Soc. **366**, 2427–2453 (2008)
43. Majda, A.J., Khouider, B.: Stochastic and mesoscopic models for tropical convection. Proc. Natl. Acad. Sci. **99**, 1123–1128 (2002)
44. Majda, A.J., Timofeyev, I., Vanden-Eijnden, E.: Models for stochastic climate prediction. Proc. Natl. Acad. Sci. **96**, 14687–14691 (1999)
45. Majda, A.J., Timofeyev, I., Vanden-Eijnden, E.: A mathematical framework for stochastic climate models. Commun. Pure Appl. Math. **LIV**, 891–974 (2001)
46. Majda, A.J., Timofeyev, I., Vanden-Eijnden, E.: A priori tests of a stochastic mode reduction strategy. Physica D **170**, 206–252 (2002)
47. Majda A., Timofeyev, I., Vanden-Eijnden, E.: Stochastic models for selected slow variables in large deterministic systems. Nonlinearity **19**, 769–794 (2006)
48. Mo, K.C., Ghil, M.: Statistics and dynamics of persistent anomalies. J. Atmos. Sci. **44**, 877–901 (1987)
49. Monahan, A.H.: Nonlinear principal component analysis by neural networks: theory and application to the Lorenz system. J. Clim. **13**, 821–835 (2000)
50. Monahan, A.H., Fyfe, J.C., Ambaum, M.H.P., Stephenson, D.B., North, G.R.: Empirical orthogonal functions: the medium is the message. J. Clim. **22**, 6501–6514 (2009)
51. Newman, M., Sardeshmukh, P., Penland, C.: Stochastic forcing of the wintertime extratropical flow. J. Atmos. Sci. **54**, 435–455 (1997)
52. Palmer, T., Williams, P. (eds.): Stochastic Physics and Climate Modelling. Cambridge University Press, Cambridge (2009)
53. Pasmanter, R.A., Timmermann, A.: Cyclic Markov chains with an application to an intermediate ENSO model. Nonlinear Process. Geophys. **10**, 197–210 (2003)
54. Penland, C., Ghil, M.: Forecasting northern hemisphere 700-mb geopotential height anomalies using empirical normal modes. Mon. Weather Rev. **121**, 2355–2372 (1993)
55. Penland, C., Sardeshmukh, P.D.: The optimal growth of tropical sea surface temperature anomalies. J. Clim. **8**, 1999–2024 (1995)
56. Peters, K., Jakob, C., Davies, L., Khouider, B., Majda, A.: Stochastic behaviour of tropical convection in observations and a multicloud model. J. Atmos. Sci. **70**, 3556–3575 (2013)
57. Plant, R.S., Craig, G.C.: A stochastic parameterization for deep convection based on equilibrium statistics. J. Atmos. Sci. **65**, 87–105 (2008)
58. Prakasa Rao, B.L.S.: Statistical Inference for Diffusion Type Processes. Arnold Publishers, London (1999)
59. Preisendorfer, R.W.: Principal Component Analysis in Meteorology and Oceanography. Elsevier, Amsterdam (1988)
60. Sura, P., Newman, M., Penland, C., Sardeshmukh, P.: Multiplicative noise and non-Gaussianity: a paradigm for atmospheric regimes. J. Atmos. Sci. **62**, 1391–1409 (2005)
61. Von Storch, H., Zwiers, F.W.: Statistical Analysis in Climate Research. Cambridge University Press, Cambridge (1999)
62. Wilks, D.S.: Statistical Methods in the Atmospheric Sciences, 3rd edn., p. 704. Academic, Oxford (2011)
63. Zucchini, W., Guttorp, P.: A Hidden Markov model for space-time precipitation. Water Resour. Res. **27**, 1917–1923 (1991)

Stochastic Eulerian-Lagrangian Methods

Paul J. Atzberger
Department of Mathematics, University of California
Santa Barbara (UCSB), Santa Barbara, CA, USA

Synonyms

Fluid-structure interaction; Fluid dynamics; Fluctuating hydrodynamics; Immersed Boundary Method; SELM; Statistical mechanics; Stochastic Eulerian Lagrangian method; Thermal fluctuations

Abstract

We present approaches for the study of fluid-structure interactions subject to thermal fluctuations. A mechanical description is utilized combining Eulerian and Lagrangian reference frames. We establish general conditions for the derivation of operators coupling these descriptions and for the derivation of stochastic driving fields consistent with statistical mechanics. We present stochastic numerical methods for the fluid-structure dynamics and methods to generate efficiently the required stochastic driving fields. To help establish the validity of the proposed approach, we perform analysis of the invariant probability distribution of the stochastic dynamics and relate our results to statistical mechanics. Overall, the presented approaches are expected to be applicable to a wide variety of systems involving fluid-structure interactions subject to thermal fluctuations.

Introduction

Recent scientific and technological advances motivate the study of fluid-structure interactions in physical regimes often involving very small length and time scales [26, 30, 35, 36]. This includes the study of microstructure in soft materials and complex fluids, the study of biological systems such as cell motility and microorganism swimming, and the study of processes within microfluidic and nanofluidic devices. At such scales thermal fluctuations play an important role and pose significant challenges in the study of such fluid-structure systems. Significant past work has been done on the formulation of descriptions for fluid-structure interactions subject to thermal fluctuations. To obtain descriptions tractable for analysis and numerical simulation, these approaches typically place an emphasis on approximations which retain only the structure degrees of freedom (eliminating the fluid dynamics). This often results in simplifications in the descriptions having substantial analytic and computational advantages. In particular, this eliminates the many degrees of freedom associated with the fluid and avoids having to resolve the potentially intricate and stiff stochastic dynamics of the fluid. These approaches have worked especially well for the study of bulk phenomena in free solution and the study of many types of complex fluids and soft materials [3, 3, 9, 13, 17, 23].

Recent applications arising in the sciences and in technological fields present situations in which resolving the dynamics of the fluid may be important and even advantageous both for modeling and computation. This includes modeling the spectroscopic responses of biological materials [19, 25, 37], studying transport in microfluidic and nanofluidic devices [16, 30], and investigating dynamics in biological systems [2, 11]. There are also other motivations for representing the fluid explicitly and resolving its stochastic dynamics. This includes the development of hybrid fluid-particle models in which thermal fluctuations mediate important effects when coupling continuum and particle descriptions [12, 14], the study of hydrodynamic coupling and diffusion in the vicinity of surfaces having complicated geometries [30], and the study of systems in which there are many interacting mechanical structures [8, 27, 28]. To facilitate the development of methods for studying such phenomena in fluid-structure systems, we present a rather general formalism which captures essential features of the coupled stochastic dynamics of the fluid and structures.

To model the fluid-structure system, a mechanical description is utilized involving both Eulerian and Lagrangian reference frames. Such mixed descriptions arise rather naturally, since it is often convenient to describe the structure configurations in a Lagrangian reference frame while it is convenient to describe the fluid in an Eulerian reference frame. In practice, this presents a number of challenges for analysis and numerical studies. A central issue concerns how to couple the descriptions to represent accurately the fluid-structure interactions, while obtaining a coupled description which can be treated efficiently by numerical methods. Another important issue concerns how to account properly for thermal fluctuations in such approximate descriptions. This must be done carefully to be consistent with statistical mechanics. A third issue concerns the development of efficient computational methods. This requires discretizations of the stochastic differential equations and the development of efficient methods for numerical integration and stochastic field generation.

We present a set of approaches to address these issues. The formalism and general conditions for the operators which couple the Eulerian and Lagrangian descriptions are presented in section "Stochastic Eulerian Lagrangian Method." We discuss a convenient description of the fluid-structure system useful for working with the formalism in practice in section "Derivations for the Stochastic Eulerian Lagrangian Method." A derivation of the stochastic driving fields used to represent the thermal fluctuations is also presented in section "Derivations for the Stochastic Eulerian Lagrangian Method." Stochastic numerical methods are discussed for the approximation of the stochastic dynamics and generation of stochastic fields in sections "Computational Methodology." To validate the methodology, we perform analysis of the invariant probability distribution of the stochastic dynamics of the fluid-structure formalism. We compare this analysis with results from statistical mechanics in section "Equilibrium Statistical Mechanics of SELM Dynamics." A more detailed and comprehensive discussion of the approaches presented here can be found in our paper [6].

Stochastic Eulerian Lagrangian Method

To study the dynamics of fluid-structure interactions subject to thermal fluctuations, we utilize a mechanical description involving Eulerian and Lagrangian reference frames. Such mixed descriptions arise rather naturally, since it is often convenient to describe the structure configurations in a Lagrangian reference frame while it is convenient to describe the fluid in an Eulerian reference frame. In principle more general descriptions using other reference frames could also be considered. Descriptions for fluid-structure systems having these features can be described rather generally by the following dynamic equations

$$\rho \frac{d\mathbf{u}}{dt} = \mathcal{L}\mathbf{u} + \Lambda[\Upsilon(\mathbf{v} - \Gamma\mathbf{u})] + \lambda + \mathbf{f}_{\text{thm}} \tag{1}$$

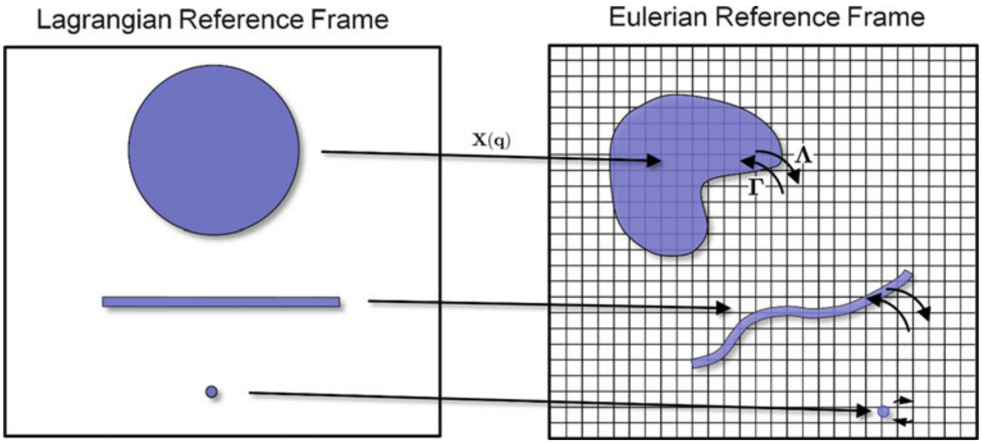
$$m \frac{d\mathbf{v}}{dt} = -\Upsilon\,(\mathbf{v} - \Gamma\mathbf{u}) - \nabla_{\mathbf{X}}\Phi[\mathbf{X}] + \zeta + \mathbf{F}_{\text{thm}} \tag{2}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{v}. \tag{3}$$

The $\mathbf{u}$ denotes the velocity of the fluid, and ρ the uniform fluid density. The $\mathbf{X}$ denotes the configuration of the structure, and $\mathbf{v}$ the velocity of the structure. The mass of the structure is denoted by m. To simplify the presentation, we treat here only the case when ρ and m are constant, but with some modifications these could also be treated as variable. The λ, ζ are Lagrange multipliers for imposed constraints, such as incompressibility of the fluid or a rigid body constraint of a structure. The operator $\mathcal{L}$ is used to account for dissipation in the fluid, such as associated with Newtonian fluid stresses [1]. To account for how the fluid and structures are coupled, a few general operators are introduced, $\Gamma, \Upsilon, \Lambda$.

The linear operators $\Gamma, \Lambda, \Upsilon$ are used to model the fluid-structure coupling. The Γ operator describes how a structure depends on the fluid flow, while $-\Upsilon$ is a negative definite dissipative operator describing the viscous interactions coupling the structure to the fluid. We assume throughout that this dissipative operator is symmetric, $\Upsilon = \Upsilon^T$. The linear operator Λ is used to attribute a spatial location for the viscous interactions between the structure and fluid. The linear operators are assumed to have dependence only on the configuration degrees of freedom $\Gamma = \Gamma[\mathbf{X}]$, $\Lambda = \Lambda[\mathbf{X}]$. We assume further that Υ does not have any dependence on $\mathbf{X}$. For an illustration of the role these coupling operators play, see Fig. 1.

To account for the mechanics of structures, $\Phi[\mathbf{X}]$ denotes the potential energy of the configuration $\mathbf{X}$. The total energy associated with this fluid-structure system is given by



Stochastic Eulerian-Lagrangian Methods, Fig. 1 The description of the fluid-structure system utilizes both Eulerian and Lagrangian reference frames. The structure mechanics are often most naturally described using a Lagrangian reference frame. The fluid mechanics are often most naturally described using an Eulerian reference frame. The mapping **X(q)** relates the Lagrangian reference frame to the Eulerian reference frame. The operator Γ prescribes how structures are to be coupled to the fluid. The operator Λ prescribes how the fluid is to be coupled to the structures. A variety of fluid-structure interactions can be represented in this way. This includes rigid and deformable bodies, membrane structures, polymeric structures, or point particles

$$E[\mathbf{u},\mathbf{v},\mathbf{X}] = \int_{\Omega} \frac{1}{2}\rho|\mathbf{u}(\mathbf{y})|^2 d\mathbf{y} + \frac{1}{2}m\mathbf{v}^2 + \Phi[\mathbf{X}]. \tag{4}$$

The first two terms give the kinetic energy of the fluid and structures. The last term gives the potential energy of the structures.

As we shall discuss, it is natural to consider coupling operators Λ and Γ which are adjoint in the sense

$$\int_{\mathcal{S}} (\Gamma\mathbf{u})(\mathbf{q})\cdot\mathbf{v}(\mathbf{q})d\mathbf{q} = \int_{\Omega} \mathbf{u}(\mathbf{x})\cdot(\Lambda\mathbf{v})(\mathbf{x})d\mathbf{x} \tag{5}$$

for any $\mathbf{u}$ and $\mathbf{v}$. The $\mathcal{S}$ and Ω denote the spaces used to parameterize respectively the structures and the fluid. We denote such an adjoint by $\Lambda = \Gamma^{\dagger}$ or $\Gamma = \Lambda^{\dagger}$. This adjoint condition can be shown to have the important consequence that the fluid-structure coupling conserves energy when $\Upsilon \rightarrow \infty$ in the inviscid and zero temperature limit.

To account for thermal fluctuations, a random force density $\mathbf{f}_{\text{thm}}$ is introduced in the fluid equations and $\mathbf{F}_{\text{thm}}$ in the structure equations. These account for spontaneous changes in the system momentum which occurs as a result of the influence of unresolved microscopic degrees of freedom and unresolved events occurring in the fluid and in the fluid-structure interactions.

The thermal fluctuations consistent with the form of the total energy and relaxation dynamics of the system are taken into account by the introduction of stochastic driving fields in the momentum equations of the fluid and structures. The stochastic driving fields are taken to be Gaussian processes with mean zero and with δ-correlation in time [29]. By the fluctuation-dissipation principle [29], these have covariances given by

$$\langle \mathbf{f}_{\text{thm}}(s)\mathbf{f}_{\text{thm}}^{T}(t)\rangle = -(2k_B T)(\mathcal{L} - \Lambda\Upsilon\Gamma)\,\delta(t-s) \tag{6}$$

$$\langle \mathbf{F}_{\text{thm}}(s)\mathbf{F}_{\text{thm}}^{T}(t)\rangle = (2k_B T)\,\Upsilon\,\delta(t-s) \tag{7}$$

$$\langle \mathbf{f}_{\text{thm}}(s)\mathbf{F}_{\text{thm}}^{T}(t)\rangle = -(2k_B T)\,\Lambda\Upsilon\,\delta(t-s). \tag{8}$$

We have used that $\Gamma = \Lambda^{\dagger}$ and $\Upsilon = \Upsilon^{T}$. We remark that the notation $\mathbf{g}\mathbf{h}^{T}$ which is used for the covariance operators should be interpreted as the tensor product. This notation is meant to suggest the analogue to the outer-product operation which holds in the discrete setting [5]. A more detailed discussion and derivation of the thermal fluctuations is given in section "Derivations for the Stochastic Eulerian Lagrangian Method."

It is important to mention that some care must be taken when using the above formalism in practice and when choosing operators. An important issue concerns the treatment of the material derivative of the fluid, $d\mathbf{u}/dt = \partial\mathbf{u}/\partial t + \mathbf{u}\cdot\nabla\mathbf{u}$. For stochastic systems the field $\mathbf{u}$ is often highly irregular and not defined in a point-wise sense, but rather only in the sense of a generalized function (distribution) [10, 24]. To avoid these issues, we shall treat $d\mathbf{u}/dt = \partial\mathbf{u}/\partial t$ in this initial presentation of the approach [6]. The SELM provides a rather general framework for the study of fluid-structure interactions subject to thermal fluctuations. To use the approach for specific applications requires the formulation of appropriate coupling operators Λ and Γ to model the fluid-structure interaction. We provide some concrete examples of such operators in the paper [6].

Formulation in Terms of Total Momentum Field

When working with the formalism in practice, it turns out to be convenient to reformulate the description in terms of a field describing the total momentum of the fluid-structure system at a given spatial location. As we shall discuss, this description results in simplifications in the stochastic driving fields. For this purpose, we define

$$\mathbf{p}(\mathbf{x},t) = \rho\mathbf{u}(\mathbf{x},t) + \Lambda[m\mathbf{v}(t)](\mathbf{x}). \tag{9}$$

The operator Λ is used to give the distribution in space of the momentum associated with the structures for given configuration $\mathbf{X}(t)$. Using this approach, the fluid-structure dynamics are described by

$$\frac{d\mathbf{p}}{dt} = \mathcal{L}\mathbf{u} + \Lambda[-\nabla_{\mathbf{X}}\Phi(\mathbf{X})] + (\nabla_{\mathbf{X}}\Lambda[m\mathbf{v}])\cdot\mathbf{v} + \lambda + \mathbf{g}_{\text{thm}} \tag{10}$$

$$m\frac{d\mathbf{v}}{dt} = -\Upsilon(\mathbf{v} - \Gamma\mathbf{u}) - \nabla_{\mathbf{X}}\Phi(\mathbf{X}) + \zeta + \mathbf{F}_{\text{thm}} \tag{11}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{v} \tag{12}$$

where $\mathbf{u} = \rho^{-1}(\mathbf{p} - \Lambda[m\mathbf{v}])$ and $\mathbf{g}_{\text{thm}} = \mathbf{f}_{\text{thm}} + \Lambda[\mathbf{F}_{\text{thm}}]$. The third term in the first equation arises from the dependence of Λ on the configuration of the structures, $\Lambda[m\mathbf{v}] = (\Lambda[X])[m\mathbf{v}]$. The Lagrange

S

multipliers for imposed constraints are denoted by λ, ζ. For the constraints, we use rather liberally the notation with the Lagrange multipliers denoted here not necessarily assumed to be equal to the previous definition. The stochastic driving fields are again Gaussian with mean zero and δ-correlation in time [29]. The stochastic driving fields have the covariance structure given by

$$\langle \mathbf{g}_{\text{thm}}(s)\mathbf{g}_{\text{thm}}^T(t)\rangle = -(2k_B T)\,\mathcal{L}\,\delta(t-s) \tag{13}$$

$$\langle \mathbf{F}_{\text{thm}}(s)\mathbf{F}_{\text{thm}}^T(t)\rangle = (2k_B T)\,\Upsilon\,\delta(t-s) \tag{14}$$

$$\langle \mathbf{g}_{\text{thm}}(s)\mathbf{F}_{\text{thm}}^T(t)\rangle = 0. \tag{15}$$

This formulation has the convenient feature that the stochastic driving fields become independent. This is a consequence of using the field for the total momentum for which the dissipative exchange of momentum between the fluid and structure no longer arises. In the equations for the total momentum, the only source of dissipation remaining occurs from the stresses of the fluid. This approach simplifies the effort required to generate numerically the stochastic driving fields and will be used throughout.

Derivations for the Stochastic Eulerian Lagrangian Method

We now discuss formal derivations to motivate the stochastic differential equations used in each of the physical regimes. For this purpose, we do not present the most general derivation of the equations. For brevity, we make simplifying assumptions when convenient.

In the initial formulation of SELM, the fluid-structure system is described by

$$\rho\frac{d\mathbf{u}}{dt} = \mathcal{L}\mathbf{u} + \Lambda[\Upsilon(\mathbf{v} - \Gamma\mathbf{u})] + \lambda + \mathbf{f}_{\text{thm}} \tag{16}$$

$$m\frac{d\mathbf{v}}{dt} = -\Upsilon(\mathbf{v} - \Gamma\mathbf{u}) - \nabla_{\mathbf{X}}\Phi(\mathbf{X}) + \zeta$$
$$+\,\mathbf{F}_{\text{thm}} \tag{17}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{v}. \tag{18}$$

The notation and operators appearing in these equations have been discussed in detail in section "Stochastic Eulerian Lagrangian Method." For these equations, we focus primarily on the motivation for the stochastic driving fields used for the fluid-structure system.

For the thermal fluctuations of the system, we assume Gaussian random fields with mean zero and δ-correlated in time. For such stochastic fields, the central challenge is to determine an appropriate covariance structure. For this purpose, we use the fluctuation-dissipation principle of statistical mechanics [22, 29]. For linear stochastic differential equations of the form

$$d\mathbf{Z}_t = L\mathbf{Z}_t\,dt + Q\,d\mathbf{B}_t \tag{19}$$

the fluctuation-dissipation principle can be expressed as

$$G = QQ^T = -(LC) - (LC)^T. \tag{20}$$

This relates the equilibrium covariance structure C of the system to the covariance structure G of the stochastic driving field. The operator L accounts for the dissipative dynamics of the system. For the Eqs. 16–18, the dissipative operators only appear in the momentum equations. This can be shown to have the consequence that there is no thermal forcing in the equation for $\mathbf{X}(t)$; this will also be confirmed in section "Formulation in Terms of Total Momentum Field." To simplify the presentation, we do not represent explicitly the stochastic dynamics of the structure configuration $\mathbf{X}$.

For the fluid-structure system, it is convenient to work with the stochastic driving fields by defining

$$\mathbf{q} = [\rho^{-1}\mathbf{f}_{\text{thm}}, m^{-1}\mathbf{F}_{\text{thm}}]^T. \tag{21}$$

The field $\mathbf{q}$ formally is given by $\mathbf{q} = Q\,d\mathbf{B}_t/dt$ and determined by the covariance structure $G = QQ^T$. This covariance structure is determined by the fluctuation-dissipation principle expressed in Eq. 20 with

$$L = \begin{bmatrix} \rho^{-1}(\mathcal{L} - \Lambda\Upsilon\Gamma) & \rho^{-1}\Lambda\Upsilon \\ m^{-1}\Upsilon\Gamma & -m^{-1}\Upsilon \end{bmatrix} \tag{22}$$

$$C = \begin{bmatrix} \rho^{-1}k_B T\mathcal{I} & 0 \\ 0 & m^{-1}k_B T\mathcal{I} \end{bmatrix}. \tag{23}$$

The $\mathcal{I}$ denotes the identity operator. The covariance C was obtained by considering the fluctuations at equilibrium. The covariance C is easily found since the Gibbs-Boltzmann distribution is a Gaussian with formal density $\Psi(\mathbf{u},\mathbf{v}) = \frac{1}{Z_0}\exp\left[-E/k_BT\right]$. The Z_0 is the normalization constant for Ψ. The energy is given by Eq. 4. For this purpose, we need only consider the energy E in the case when $\Phi = 0$. This gives the covariance structure

$$G = (2k_BT)\begin{bmatrix} -\rho^{-2}\left(\mathcal{L} - \Lambda\Upsilon\Gamma\right) & -m^{-1}\rho^{-1}\Lambda\Upsilon \\ -m^{-1}\rho^{-1}\Upsilon\Gamma & m^{-2}\Upsilon \end{bmatrix}. \tag{24}$$

To obtain this result, we use that $\Gamma = \Lambda^{\dagger}$ and $\Upsilon = \Upsilon^{\dagger}$. From the definition of $\mathbf{q}$, it is found that the covariance of the stochastic driving fields of SELM is given by Eqs. 6–8. This provides a description of the thermal fluctuations in the fluid-structure system.

Formulation in Terms of Total Momentum Field

It is convenient to reformulate the description of the fluid-structure system in terms of a field for the total momentum of the system associated with spatial location $\mathbf{x}$. For this purpose we define

$$\mathbf{p}(\mathbf{x},t) = \rho\mathbf{u}(\mathbf{x},t) + \Lambda[m\mathbf{v}(t)](\mathbf{x}). \tag{25}$$

The operator Λ is used to give the distribution in space of the momentum associated with the structures. Using this approach, the fluid-structure dynamics are described by

$$\begin{aligned} \frac{d\mathbf{p}}{dt} =& \mathcal{L}\mathbf{u} + \Lambda[-\nabla_{\mathbf{X}}\Phi(\mathbf{X})] \\ & + (\nabla_{\mathbf{X}}\Lambda[m\mathbf{v}])\cdot\mathbf{v} + \lambda + \mathbf{g}_{\mathrm{thm}} \end{aligned} \tag{26}$$

$$\begin{aligned} m\frac{d\mathbf{v}}{dt} =& -\Upsilon\left(\mathbf{v} - \Gamma\mathbf{u}\right) - \nabla_{\mathbf{X}}\Phi(\mathbf{X}) + \zeta \\ & + \mathbf{F}_{\mathrm{thm}} \end{aligned} \tag{27}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{v} \tag{28}$$

where $\mathbf{u} = \rho^{-1}\left(\mathbf{p} - \Lambda[m\mathbf{v}]\right)$ and $\mathbf{g}_{\mathrm{thm}} = \mathbf{f}_{\mathrm{thm}} + \Lambda[\mathbf{F}_{\mathrm{thm}}]$. The third term in the first equation arises from the dependence of Λ on the configuration of the structures, $\Lambda[m\mathbf{v}(t)] = (\Lambda[X])[m\mathbf{v}(t)]$.

The thermal fluctuations are taken into account by two stochastic fields $\mathbf{g}_{\mathrm{thm}}$ and $\mathbf{F}_{\mathrm{thm}}$. The covariance of $\mathbf{g}_{\mathrm{thm}}$ is obtained from

$$\begin{aligned} \langle \mathbf{g}_{\mathrm{thm}}\mathbf{g}_{\mathrm{thm}}^T\rangle =& \langle \mathbf{f}_{\mathrm{thm}}\mathbf{f}_{\mathrm{thm}}^T\rangle + \langle \mathbf{f}_{\mathrm{thm}}\mathbf{F}_{\mathrm{thm}}^T\Lambda^T\rangle \\ & + \langle \Lambda\mathbf{F}_{\mathrm{thm}}\mathbf{f}_{\mathrm{thm}}^T\rangle + \langle \Lambda\mathbf{F}_{\mathrm{thm}}\mathbf{F}_{\mathrm{thm}}^T\Lambda^T\rangle \\ =& (2k_BT)\big(-\mathcal{L} + \Lambda\Upsilon\Gamma \\ & - \Lambda\Upsilon\Lambda^T - \Lambda\Upsilon\Lambda^T + \Lambda\Upsilon\Lambda^T\big) \\ =& -(2k_BT)\,\mathcal{L}. \end{aligned} \tag{29}$$

This makes use of the adjoint property of the coupling operators $\Lambda^{\dagger} = \Gamma$.

One particularly convenient feature of this reformulation is that the stochastic driving fields $\mathbf{F}_{\mathrm{thm}}$ and $\mathbf{g}_{\mathrm{thm}}$ become independent. This can be seen as follows:

$$\begin{aligned} \langle \mathbf{g}_{\mathrm{thm}}\mathbf{F}_{\mathrm{thm}}^T\rangle &= \langle \mathbf{f}_{\mathrm{thm}}\mathbf{F}_{\mathrm{thm}}^T\rangle + \langle \Lambda\mathbf{F}_{\mathrm{thm}}\mathbf{F}_{\mathrm{thm}}^T\rangle \\ &= (2k_BT)(-\Lambda\Upsilon + \Lambda\Upsilon) = 0. \end{aligned} \tag{30}$$

This decoupling of the stochastic driving fields greatly reduces the computational effort to generate the fields with the required covariance structure. This shows that the covariance structure of the stochastic driving fields of SELM is given by Eqs. 13–15.

Computational Methodology

We now discuss briefly numerical methods for the SELM formalism. For concreteness we consider the specific case in which the fluid is Newtonian and incompressible. For now, the other operators of the SELM formalism will be treated rather generally. This case corresponds to the dissipative operator for the fluid

$$\mathcal{L}\mathbf{u} = \mu\Delta\mathbf{u}. \tag{31}$$

The Δ denotes the Laplacian $\Delta\mathbf{u} = \partial_{xx}\mathbf{u} + \partial_{yy}\mathbf{u} + \partial_{zz}\mathbf{u}$. The incompressibility of the fluid corresponds to the constraint

$$\nabla\cdot\mathbf{u} = 0. \tag{32}$$

This is imposed by the Lagrange multiplier λ. By the Hodge decomposition, λ is given by the gradient of a function p with $\lambda = -\nabla p$. The p can be interpreted as the local pressure of the fluid.

A variety of methods could be used in practice to discretize the SELM formalism, such as finite difference methods, spectral methods, and finite element methods [20, 32, 33]. We present here discretizations based on finite difference methods.

Numerical Semi-discretizations for Incompressible Newtonian Fluid

The Laplacian will be approximated by central differences on a uniform periodic lattice by

$$[L\mathbf{u}]_{\mathbf{m}} = \sum_{j=1}^{3} \frac{\mathbf{u}_{\mathbf{m}+\mathbf{e}_j} - 2\mathbf{u}_{\mathbf{m}} + \mathbf{u}_{\mathbf{m}-\mathbf{e}_j}}{\Delta x^2}. \tag{33}$$

The $\mathbf{m} = (m_1, m_2, m_3)$ denotes the index of the lattice site. The $\mathbf{e}_j$ denotes the standard basis vector in three dimensions. The incompressibility of the fluid will be approximated by imposing the constraint

$$[D \cdot \mathbf{u}]_{\mathbf{m}} = \sum_{j=1}^{3} \frac{\mathbf{u}^{j}_{\mathbf{m}+\mathbf{e}_j} - \mathbf{u}^{j}_{\mathbf{m}-\mathbf{e}_j}}{2\Delta x}. \tag{34}$$

The superscripts denote the vector component. In practice, this will be imposed by computing the projection of a vector $\mathbf{u}^*$ to the subspace $\{\mathbf{u} \in \mathbb{R}^{3N} \mid D \cdot \mathbf{u} = 0\}$, where N is the total number of lattice sites. We denote this projection operation by

$$\mathbf{u} = \wp \mathbf{u}^*. \tag{35}$$

The semi-discretized equations for SELM to be used in practice are

$$\frac{d\mathbf{p}}{dt} = L\mathbf{u} + \Lambda[-\nabla_{\mathbf{X}}\Phi] + (\nabla_{\mathbf{X}}\Lambda[m\mathbf{v}]) \cdot \mathbf{v} + \lambda + \mathbf{g}_{\text{thm}} \tag{36}$$

$$\frac{d\mathbf{v}}{dt} = -\Upsilon[\mathbf{v} - \Gamma\mathbf{u}] + \mathbf{F}_{\text{thm}} \tag{37}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{v}. \tag{38}$$

The component $\mathbf{u}_{\mathbf{m}} = \rho^{-1}(\mathbf{p}_{\mathbf{m}} - \Lambda[m\mathbf{v}]_{\mathbf{m}})$. Each of the operators now appearing is understood to be discretized. We discuss specific discretizations for Γ and Λ in paper [6]. To obtain the Lagrange multiplier λ which imposes incompressibility, we use the projection operator and

$$\lambda = -(\mathcal{I} - \wp)\,(L\mathbf{u} + \Upsilon[\mathbf{v} - \Gamma\mathbf{u}] + \mathbf{f}_{\text{thm}}) \tag{39}$$

In this expression, we let $\mathbf{f}_{\text{thm}} = \mathbf{g}_{\text{thm}} - \Lambda[\mathbf{F}_{\text{thm}}]$ for the particular realized values of the fields $\mathbf{g}_{\text{thm}}$ and $\mathbf{F}_{\text{thm}}$.

We remark that in fact the semi-discretized equations of the SELM formalism in this regime can also be given in terms of $\mathbf{u}$ directly, which may provide a simpler approach in practice. The identity $\mathbf{f}_{\text{thm}} = \mathbf{g}_{\text{thm}} - \Lambda[\mathbf{F}_{\text{thm}}]$ could be used to efficiently generate the required stochastic driving fields in the equations for $\mathbf{u}$. We present the reformulation here, since it more directly suggests the semi-discretized equations to be used for the reduced stochastic equations.

For this semi-discretization, we consider a total energy for the system given by

$$E[\mathbf{u}, \mathbf{v}, \mathbf{X}] = \frac{\rho}{2} \sum_{\mathbf{m}} |\mathbf{u}(\mathbf{x}_{\mathbf{m}})|^2 \Delta \mathbf{x}^3_{\mathbf{m}} + \frac{m}{2}|\mathbf{v}|^2 + \Phi[\mathbf{X}]. \tag{40}$$

This is useful in formulating an adjoint condition 5 for the semi-discretized system. This can be derived by considering the requirements on the coupling operators Γ and Λ which ensure the energy is conserved when $\Upsilon \to \infty$ in the inviscid and zero temperature limit.

To obtain appropriate behaviors for the thermal fluctuations, it is important to develop stochastic driving fields which are tailored to the specific semi-discretizations used in the numerical methods. Once the stochastic driving fields are determined, which is the subject of the next section, the equations can be integrated in time using traditional methods for SDEs, such as the Euler-Maruyama method or a stochastic Runge-Kutta method [21]. More sophisticated integrators in time can also be developed to cope with sources of stiffness but are beyond the scope of this entry [7]. For each of the reduced equations, similar semi-discretizations can be developed as the one presented above.

Stochastic Driving Fields for Semi-discretizations

To obtain behaviors consistent with statistical mechanics, it is important stochastic driving fields be used which are tailored to the specific numerical discretization employed [5–7, 15]. To ensure consistency with statistical mechanics, we will again use the fluctuation-dissipation principle but now apply it to the semi-discretized equations. For each regime, we then discuss the important issues arising in practice concerning the efficient generation of these stochastic driving fields.

Formulation in Terms of Total Momentum Field

To obtain the covariance structure for this regime, we apply the fluctuation-dissipation principle as expressed in Eq. 20 to the semi-discretized Eqs. 36–38. This gives the covariance

$$G = -2LC = (2k_B T)\begin{bmatrix} -\rho^{-2}\Delta x^{-3}L & 0 & 0 \\ 0 & m^{-2}\Upsilon & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{41}$$

The factor of Δx^{-3} arises from the form of the energy for the discretized system which gives covariance for the equilibrium fluctuations of the total momentum $\rho^{-1}\Delta x^{-3}k_B T$; see Eq. 40. In practice, achieving the covariance associated with the dissipative operator of the fluid L is typically the most challenging to generate efficiently. This arises from the large number N of lattice sites in the discretization.

One approach is to determine a factor Q such that the block $G_{\mathbf{p},\mathbf{p}} = QQ^T$; subscripts indicate block entry of the matrix. The required random field with covariance $G_{\mathbf{p},\mathbf{p}}$ is then given by $\mathbf{g} = Q\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is the uncorrelated Gaussian field with the covariance structure $\mathcal{I}$. For the discretization used on the uniform periodic mesh, the matrices L and C are cyclic [31]. This has the important consequence that they are both diagonalizable in the discrete Fourier basis of the lattice. As a result, the field $\mathbf{f}_{\text{thm}}$ can be generated using the fast Fourier transform (FFT) with at most $O(N\log(N))$ computational steps. In fact, in this special case of the discretization, "random fluxes" at the cell faces can be used to generate the field in $O(N)$ computational steps [5]. Other approaches can be used to generate the random fields on nonperiodic meshes and on multilevel meshes; see [4, 5].

Equilibrium Statistical Mechanics of SELM Dynamics

We now discuss how the SELM formalism and the presented numerical methods capture the equilibrium statistical mechanics of the fluid-structure system. This is done through an analysis of the invariant probability distribution of the stochastic dynamics. For the fluid-structure systems considered, the appropriate probability distribution is given by the Gibbs-Boltzmann distribution

$$\Psi_{\text{GB}}(\mathbf{z}) = \frac{1}{Z}\exp\left[-E(\mathbf{z})/k_B T\right]. \tag{42}$$

The $\mathbf{z}$ is the state of the system, E is the energy, k_B is Boltzmann's constant, T is the system temperature, and Z is a normalization constant for the distribution [29]. We show that this Gibbs-Boltzmann distribution is the equilibrium distribution of both the full stochastic dynamics and the reduced stochastic dynamics in each physical regime.

We present here both a verification of the invariance of the Gibbs-Boltzmann distribution for the general formalism and for numerical discretizations of the formalism. The verification is rather formal for the undiscretized formalism given technical issues which would need to be addressed for such an infinite dimensional dynamical system. However, the verification is rigorous for the semi-discretization of the formalism, which yields a finite dimensional dynamical system. The latter is likely the most relevant case in practice. Given the nearly identical calculations involved in the verification for the general formalism and its semi-discretizations, we use a notation in which the key differences between the two cases primarily arise in the definition of the energy. In particular, the energy is understood to be given by Eq. 4 when considering the general SELM formalism and Eq. 40 when considering semi-discretizations.

Formulation in Terms of Total Momentum Field

The stochastic dynamics given by Eqs. 10–12 is a change of variable of the full stochastic dynamics of the SELM formalism given by Eqs. 1–3. Thus verifying the invariance using the reformulated description is also applicable to Eqs. 1–3 and vice versa. To verify the invariance in the other regimes, it is convenient to work with the reformulated description. The energy associated with the reformulated description is given by

$$E[\mathbf{p},\mathbf{v},\mathbf{X}] = \frac{1}{2\rho}\int_{\Omega}|\mathbf{p}(\mathbf{y}) - \Lambda[m\mathbf{v}](\mathbf{y})|^2 d\mathbf{y} + \frac{m}{2}|\mathbf{v}|^2 + \Phi[\mathbf{X}]. \tag{43}$$

The energy associated with the semi-discretization is

$$E[\mathbf{p},\mathbf{v},\mathbf{X}] = \frac{1}{2\rho}\sum_{\mathbf{m}}|\mathbf{p}(\mathbf{x}_{\mathbf{m}}) - \Lambda[m\mathbf{v}]_{\mathbf{m}}|^2\Delta\mathbf{x}_{\mathbf{m}}^3 + \frac{m}{2}|\mathbf{v}|^2 + \Phi[\mathbf{X}]. \tag{44}$$

The probability density $\Psi(\mathbf{p}, \mathbf{v}, \mathbf{X}, t)$ for the current state of the system under the SELM dynamics is governed by the Fokker-Planck equation

$$\frac{\partial \Psi}{\partial t} = -\nabla \cdot \mathbf{J} \tag{45}$$

with probability flux

$$\mathbf{J} = \begin{bmatrix} \mathcal{L} + \Lambda + \nabla_{\mathbf{X}} \Lambda \cdot \mathbf{v} + \lambda \\ -\Upsilon - \nabla_{\mathbf{X}} \Phi + \zeta \\ \mathbf{v} \end{bmatrix} \Psi - \frac{1}{2}(\nabla \cdot G)\Psi - \frac{1}{2} G \nabla \Psi. \tag{46}$$

The covariance operator G is associated with the Gaussian field $\mathbf{g} = [\mathbf{g}_{\text{thm}}, \mathbf{F}_{\text{thm}}, 0]^T$ by $\langle \mathbf{g}(s) \mathbf{g}^T(t) \rangle = G\delta(t - s)$. In this regime, G is given by Eq. 13 or 41. In the notation $[\nabla \cdot G(\mathbf{z})]_i = \partial_{z_j} G_{ij}(\mathbf{z})$ with the summation convention for repeated indices. To simplify the notation, we have suppressed denoting the specific functions on which each of the operators acts; see Eqs. 10–12 for these details.

The requirement that the Gibbs-Boltzmann distribution Ψ_{GB} given by Eq. 42 be invariant under the stochastic dynamics is equivalent to the distribution yielding $\nabla \cdot \mathbf{J} = 0$. We find it convenient to group terms and express this condition as

$$\nabla \cdot \mathbf{J} = A_1 + A_2 + \nabla \cdot \mathbf{A}_3 + \nabla \cdot \mathbf{A}_4 = 0 \tag{47}$$

where

$$\begin{aligned}
A_1 &= \left[(\Lambda + \nabla_{\mathbf{X}} \Lambda \cdot \mathbf{v} + \lambda_1) \cdot \nabla_{\mathbf{p}} E + (-\nabla_{\mathbf{X}} \Phi + \zeta_1) \cdot \nabla_{\mathbf{v}} E + (\mathbf{v}) \cdot \nabla_{\mathbf{X}} E\right] (-k_B T)^{-1} \Psi_{\text{GB}} \\
A_2 &= \left[\nabla_{\mathbf{p}} \cdot (\Lambda + \nabla_{\mathbf{X}} \Lambda \cdot \mathbf{v} + \lambda_1) + \nabla_{\mathbf{v}} \cdot (-\nabla_{\mathbf{X}} \Phi + \zeta_2) + \nabla_{\mathbf{X}} \cdot (\mathbf{v})\right] \Psi_{\text{GB}} \\
\mathbf{A}_3 &= -\frac{1}{2}(\nabla \cdot G) \Psi_{\text{GB}} \\
\mathbf{A}_4 &= \begin{bmatrix} \mathcal{L}\mathbf{u} + \lambda_2 + \left[G_{\mathbf{pp}} \nabla_{\mathbf{p}} E + G_{\mathbf{pv}} \nabla_{\mathbf{v}} E + G_{\mathbf{pX}} \nabla_{\mathbf{X}} E\right] (2k_B T)^{-1} \\ -\Upsilon + \zeta_2 + \left[G_{\mathbf{vp}} \nabla_{\mathbf{p}} E + G_{\mathbf{vv}} \nabla_{\mathbf{v}} E + G_{\mathbf{vX}} \nabla_{\mathbf{X}} E\right] (2k_B T)^{-1} \\ \left[G_{\mathbf{Xp}} \nabla_{\mathbf{p}} E + G_{\mathbf{Xv}} \nabla_{\mathbf{v}} E + G_{\mathbf{XX}} \nabla_{\mathbf{X}} E\right] (2k_B T)^{-1} \end{bmatrix} \Psi_{\text{GB}}.
\end{aligned} \tag{48}$$

We assume here that the Lagrange multipliers can be split $\lambda = \lambda_1 + \lambda_2$ and $\zeta = \zeta_1 + \zeta_2$ to impose the constraints by considering in isolation different terms contributing to the dynamics; see Eq. 48. This is always possible for linear constraints. The block entries of the covariance operator G are denoted by $G_{i,j}$ with $i, j \in \{\mathbf{p}, \mathbf{v}, \mathbf{X}\}$. For the energy of the discretized system given by Eq. 4, we have

$$\nabla_{\mathbf{p}_{\mathbf{n}}} E = \mathbf{u}(\mathbf{x}_{\mathbf{n}}) \Delta x_{\mathbf{n}}^3 \tag{49}$$

$$\nabla_{\mathbf{v}_q} E = \sum_{\mathbf{m}} \mathbf{u}(\mathbf{x}_{\mathbf{m}}) \cdot \left(-\nabla_{\mathbf{v}_q} \Lambda[m\mathbf{v}]_{\mathbf{m}}\right) \Delta x_{\mathbf{m}}^3 + m\mathbf{v}_q \tag{50}$$

$$\nabla_{\mathbf{X}_q} E = \sum_{\mathbf{m}} \mathbf{u}(\mathbf{x}_{\mathbf{m}}) \cdot \left(-\nabla_{\mathbf{X}_q} \Lambda[m\mathbf{v}]_{\mathbf{m}}\right) \Delta x_{\mathbf{m}}^3 + \nabla_{\mathbf{X}_q} \Phi. \tag{51}$$

where $\mathbf{u} = \rho^{-1}(\mathbf{p} - \Lambda[m\mathbf{v}])$. Similar expressions for the energy of the undiscretized formalism can be obtained by using the calculus of variations [18].

We now consider $\nabla \cdot \mathbf{J}$ and each term $A_1, A_2, A_3, \mathbf{A}_4$. The term A_1 can be shown to be the time derivative of the energy $A_1 = dE/dt$ when considering only a subset of the contributions to the dynamics. Thus, conservation of the energy under this restricted dynamics would result in A_1 being zero. For the SELM formalism, we find by direct substitution of the gradients of E given by Eqs. 49–51 into Eq. 48 that $A_1 = 0$. When there are constraints, it is important to consider only admissible states $(\mathbf{p}, \mathbf{v}, \mathbf{X})$. This shows in the inviscid and zero temperature limit of SELM, the resulting dynamics are nondissipative. This property imposes constraints on the coupling operators and can be viewed as a further motivation for the adjoint conditions imposed in Eq. 5.

The term A_2 gives the compressibility of the phase-space flow generated by the nondissipative dynamics of the SELM formalism. The flow is generated by the vector field $(\Lambda + \nabla_{\mathbf{X}}\Lambda \cdot \mathbf{v} + \lambda_1,\ -\nabla_{\mathbf{X}}\Phi + \zeta_1,\ \mathbf{v})$ on the phase-space $(\mathbf{p}, \mathbf{v}, \mathbf{X})$. When this term is nonzero, there are important implications for the Liouville theorem and statistical mechanics of the system [34]. For the current regime, we have $A_2 = 0$ since in the divergence each component of the vector field is seen to be independent of the variable on which the derivative is computed. This shows in the inviscid and zero temperature limit of SELM, the phase-space flow is incompressible. For the reduced SELM descriptions, we shall see this is not always the case.

The term $\mathbf{A}_3$ corresponds to fluxes arising from multiplicative features of the stochastic driving fields. When the covariance G has a dependence on the current state of the system, this can result in possible changes in the amplitude and correlations in the fluctuations. These changes can yield asymmetries in the stochastic dynamics which manifest as a net probability flux. In the SELM formalism, it is found that in the divergence of G, each contributing entry is independent of the variable on which the derivative is being computed. This shows for the SELM dynamics there is no such probability fluxes, $\mathbf{A}_3 = 0$.

The last term $\mathbf{A}_4$ accounts for the fluxes arising from the primarily dissipative dynamics and the stochastic driving fields. This term is calculated by substituting the gradients of the energy given by Eqs. 49–51 and using the choice of covariance structure given by Eq. 13 or 41. By direct substitution this term is found to be zero, $\mathbf{A}_4 = 0$.

This shows the invariance of the Gibbs-Boltzmann distribution under the SELM dynamics. This provides a rather strong validation of the stochastic driving fields introduced for the SELM formalism. This shows the SELM stochastic dynamics are consistent with equilibrium statistical mechanics [29].

Conclusions

An approach for fluid-structure interactions subject to thermal fluctuations was presented based on a mechanical description utilizing both Eulerian and Lagrangian reference frames. General conditions were established for operators coupling these descriptions. A reformulated description was presented for the stochastic dynamics of the fluid-structure system having convenient features for analysis and for computational methods. Analysis was presented establishing for the SELM stochastic dynamics that the Gibbs-Boltzmann distribution is invariant. The SELM formalism provides a general framework for the development of computational methods for applications requiring a consistent treatment of structure elastic mechanics, hydrodynamic coupling, and thermal fluctuations. A more detailed and comprehensive discussion of SELM can be found in our paper [6].

Acknowledgements The author P. J. A. acknowledges support from research grant NSF CAREER DMS - 0956210.

References

1. Acheson, D.J.: Elementary Fluid Dynamics. Oxford Applied Mathematics and Computing Science Series. Oxford University Press, Oxford/Clarendon Press/New York (1990)
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walker, P.: Molecular Biology of the Cell. Garland Publishing, New York (2002)
3. Armstrong, R.C., Byron Bird, R., Hassager, O.:Dynamic Polymeric Liquids, Vols. I, II. Wiley, New York (1987)
4. Atzberger, P.: Spatially adaptive stochastic multigrid methods for fluid-structure systems with thermal fluctuations, technical report (2010). arXiv:1003.2680
5. Atzberger, P.J.: Spatially adaptive stochastic numerical methods for intrinsic fluctuations in reaction-diffusion systems. J. Comput. Phys. **229**, 3474–3501 (2010)
6. Atzberger, P.J.: Stochastic eulerian lagrangian methods for fluid-structure interactions with thermal fluctuations. J. Comput. Phys. **230**, 2821–2837 (2011)
7. Atzberger, P.J., Kramer, P.R., Peskin, C.S.: A stochastic immersed boundary method for fluid-structure dynamics at microscopic length scales. J. Comput. Phys. **224**, 1255–1292 (2007)
8. Banchio, A.J., Brady, J.F.: Accelerated stokesian dynamics: Brownian motion. J. Chem. Phys. **118**, 10323–10332 (2003)
9. Brady, J.F. Bossis, G.: Stokesian dynamics. Annu. Rev. Fluid Mech. **20**, 111–157 (1988)
10. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Cambridge University Press, Cambridge/New York (1992)
11. Danuser, G., Waterman-Storer, C.M: Quantitative fluorescent speckle microscopy of cytoskeleton dynamics. Annu. Rev. Biophys. Biomol. Struct. **35**, 361–387 (2006)
12. De Fabritiis, G., Serrano, M., Delgado-Buscalioni, R., Coveney, P.V.: Fluctuating hydrodynamic modeling of fluids at the nanoscale. Phys. Rev. E **75**, 026307 (2007)
13. Doi, M., Edwards, S.F.: The Theory of Polymer Dynamics. Oxford University Press, New York (1986)
14. Donev, A., Bell, J.B., Garcia, A.L., Alder, B.J.: A hybrid particle-continuum method for hydrodynamics of complex fluids. SIAM J. Multiscale Model. Simul. **8** 871–911 (2010)

15. Donev, A., Vanden-Eijnden, E., Garcia, A.L., Bell, J.B.: On the accuracy of finite-volume schemes for fluctuating hydrodynamics, Commun. Appl. Math. Comput. Sci. **5**(2), 149–197 (2010)
16. Eijkel, J.C.T., Napoli, M., Pennathur, S.: Nanofluidic technology for biomolecule applications: a critical review. Lab on a Chip **10**, 957–985 (2010)
17. Ermak, D.L. McCammon, J.A.: Brownian dynamics with hydrodynamic interactions. J. Chem. Phys. **69**, 1352–1360 (1978)
18. Gelfand, I.M., Fomin, S.V.: Calculus of Variations. Dover, Mineola (2000)
19. Gotter, R., Kroy, K., Frey, E., Barmann, M., Sackmann, E.: Dynamic light scattering from semidilute actin solutions: a study of hydrodynamic screening, filament bending stiffness, and the effect of tropomyosin/troponin-binding. Macromolecules **29** 30–36 (1996)
20. Gottlieb, D., Orszag, S.A.: Numerical Analysis of Spectral Methods Theory and Applications. SIAM, Philadelphia (1993)
21. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin/New York (1992)
22. Landau, L.D., Lifshitz, E.M.: Course of Theoretical Physics. Statistical Physics, vol. 9. Pergamon Press, Oxford (1980)
23. Larson, R.G.: The Structure and Rheology of Complex Fluids. Oxford University Press, New York (1999)
24. Lieb, E.H., Loss, M.: Analysis. American Mathematical Society, Providence (2001)
25. Mezei, F., Pappas, C., Gutberlet, T.: Neutron Spin Echo Spectroscopy: Basics, Trends, and Applications. Spinger, Berlin/New York (2003)
26. Moffitt, J.R., Chemla, Y.R., Smith, S.B., Bustamante, C.: Recent advances in optical tweezers. Annu. Rev. Biochem. **77**, 205–228 (2008)
27. Peskin, C.S.: Numerical analysis of blood flow in the heart. J. Comput. Phys. **25**, 220–252 (1977)
28. Peskin, C.S.: The immersed boundary method. Acta Numerica **11**, 479–517 (2002)
29. Reichl, L.E.: A Modern Course in Statistical Physics. Wiley, New York (1998)
30. Squires, T.M., Quake, S.R.: Microfluidics: fluid physics at the nanoliter scale. Rev. Mod. Phys. **77**, 977–1026 (2005)
31. Strang, G.: Linear Algebra and Its Applications. Harcourt Brace Jovanovich College Publishers, San Diego (1988)
32. Strang, G., Fix, G.: An Analysis of the Finite Element Method. Wellesley-Cambridge Press, Wellesley (2008)
33. Strikwerda, J.C.: Finite Difference Schemes and Partial Differential Equations. SIAM, Philadelphia (2004)
34. Tuckerman, M.E., Mundy, C.J., Martyna, G.J.: On the classical statistical mechanics of non-hamiltonian systems. EPL (Europhys. Lett.) **45**, 149–155 (1999)
35. Valentine, M.T., Weeks, E.R., Gisler, T., Kaplan, P.D., Yodh, A.G., Crocker, J.C., Weitz, D.A.: Two-point microrheology of inhomogeneous soft materials. Phys. Rev. Lett. **85**, 888–91 (2000)
36. Watari, N., Doi, M., Larson, R.G.: Fluidic trapping of deformable polymers in microflows. Phys. Rev. E **78**, 011801 (2008)
37. Watson, M.C., Brown, F.L.H.: Interpreting membrane scattering experiments at the mesoscale: the contribution of dissipation within the bilayer. Biophys. J. **98**, L9–L11 (2010)

Stochastic Filtering

M.V. Tretyakov
School of Mathematical Sciences, University of Nottingham, Nottingham, UK

Mathematics Subject Classification

60G35; 93E11; 93E10; 62M20; 60H15; 65C30; 65C35; 60H35

Synonyms

Filtering problem for hidden Markov models; Nonlinear filtering

Short Definition

Let $\mathbb{T}_1$ and $\mathbb{T}_2$ be either a time interval $[0, T]$ or a discrete set. The stochastic filtering problem consists in estimating an unobservable signal $X_t,\ t \in \mathbb{T}_1$, based on an observation $\{y_s,\ s \leq t,\ s \in \mathbb{T}_2\}$, where the process y_t is related to X_t via a stochastic model.

Description

We restrict ourselves to the case when an unobservable signal and observation are continuous time processes with $\mathbb{T}_1 = \mathbb{T}_2 = [0, T]$ (see discrete filtering in, e.g., [1, 3, 7]). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $\mathcal{F}_t,\ 0 \leq t \leq T$, be a filtration satisfying the usual hypotheses, and $(w_t, \mathcal{F}_t)$ and $(v_t, \mathcal{F}_t)$ be d_1-dimensional and r-dimensional independent standard Wiener processes, respectively. We consider the classical filtering scheme in which the unobservable signal process ("hidden" state) $X_t \in \mathbb{R}^d$ and the observation process $y_t \in \mathbb{R}^r$ satisfy the system of Îto stochastic differential equations (SDE):

$$\begin{aligned} dX &= \alpha(X)ds + \sigma(X)dw_s + \gamma(X)dv_s, \\ X_0 &= x, \end{aligned} \tag{1}$$

$$dy = \beta(X)ds + dv_s,\ y_0 = 0, \tag{2}$$

where $\alpha(x)$ and $\beta(x)$ are d-dimensional and r-dimensional vector functions, respectively, and $\sigma(x)$ and $\gamma(x)$ are $d \times d_1$-dimensional and $d \times r$-dimensional matrix functions, respectively. The vector $X_0 = x$ in the initial condition for (1) is usually random (i.e., uncertain), it is assumed to be independent of both w and v, and its density $\varphi(\cdot)$ is assumed to be known.

Let $f(x)$ be a function on $\mathbb{R}^d$. We assume that the coefficients in (1), (2) and the function f are bounded and have bounded derivatives up to some order. The stochastic filtering problem consists in constructing the estimate $\hat{f}(X_t)$ based on the observation y_s, $0 \le s \le t$, which is the best in the mean-square sense, i.e., the problem amounts to computing the conditional expectation:

$$\begin{aligned} \pi_t[f] &= \hat{f}(X_t) = \mathbb{E}\left(f(X_t) \mid y_s,\ 0 \le s \le t\right) \\ &=: \mathbb{E}^y f(X_t). \qquad (3) \end{aligned}$$

Applications of nonlinear filtering include tracking, navigation systems, cryptography, image processing, weather forecasting, financial engineering, speech recognition, and many others (see, e.g., [2, 11] and references therein). For a historical account, see, e.g., [1].

Optimal Filter Equations

In this section we give a number of expressions for the optimal filter which involve solving some stochastic evolution equations. Proofs of the results presented in this section and their detailed exposition and extensions are available, e.g., in [1, 4, 6–11].

The solution $\pi_t[f]$ to the filtering problem (3), (1)–(2) satisfies the nonlinear stochastic evolution equation:

$$\begin{aligned} d\pi_t[f] = \pi_t[\mathcal{L}f]dt + \left(\pi_t[\mathcal{M}^\top f] - \pi_t[f]\pi_t[\beta^\top]\right) \\ (dy - \pi_t[\beta]dt), \qquad (4) \end{aligned}$$

where $\mathcal{L}$ is the generator for the diffusion process X_t:

$$\mathcal{L}f := \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_{i=1}^{d} \alpha_i \frac{\partial f}{\partial x_i}$$

with $a = \{a_{ij}\}$ being a $d \times d$-dimensional matrix defined by $a(x) = \sigma(x)\sigma^\top(x) + \gamma(x)\gamma^\top(x)$ and $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_r)^\top$ is a vector of the operators $\mathcal{M}_j f := \sum_{i=1}^d \gamma_{ij}\frac{\partial f}{\partial x_i} + \beta_j f$. The equation of optimal nonlinear filtering (4) is usually called the Kushner-Stratonovich equation or the Fujisaki-Kallianpur-Kunita equation. If the conditional measure $\mathbb{E}(I(X(t) \in dx) \mid y_s,\ 0 \le s \le t)$ has a smooth density $\pi(t,x)$ with respect to the Lebesgue measure, then it solves the following nonlinear stochastic equation:

$$\begin{aligned} d\pi(t,x) = \mathcal{L}^*\pi(t,x)dt + (\mathcal{M}^* - \pi_t[\beta])^\top \pi(t,x) \\ (dy - \pi_t[\beta]dt),\ \pi(0,x) = \varphi(x), \qquad (5) \end{aligned}$$

where $\mathcal{L}^*$ is an adjoint operator to $\mathcal{L}$: $\mathcal{L}^* f := \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x_j}\left(a_{ij} f\right) - \sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\alpha_i f\right)$ and $\mathcal{M}^*$ is an adjoint operator to $\mathcal{M}$: $\mathcal{M}^* = (\mathcal{M}_1^*, \ldots, \mathcal{M}_r^*)^\top$ with $\mathcal{M}_j^* f = -\sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\gamma_{ij} f\right) + \beta_j f$. We note that $\pi_t[f] = \int_{\mathbb{R}^d} f(x)\pi(t,x)dx$. We also remark that the process $\bar{v}_t := y_t - \int_0^t \pi_s[\beta]ds$ is called the innovation process.

Now we will give another expression for the optimal filter. Let

$$\begin{aligned} \eta_t &:= \exp\left\{\int_0^t \beta^\top(X_s)dv_s + \frac{1}{2}\int_0^t \beta^2(X_s)ds\right\} \\ &= \exp\left\{\int_0^t \beta^\top(X_s)dy_s - \frac{1}{2}\int_0^t \beta^2(X_s)ds\right\}. \end{aligned}$$

According to our assumptions, we have $\mathbb{E}\eta_t^{-1} = 1$, $0 \le t \le T$. We introduce the new probability measure $\tilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$: $\tilde{\mathbb{P}}(\Gamma) = \int_\Gamma \eta_T^{-1} d\mathbb{P}(\omega)$. The measures $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are mutually absolutely continuous. Due to the Girsanov theorem, y_s is a Wiener process on $(\Omega, \mathcal{F}, \mathcal{F}_t, \tilde{\mathbb{P}})$, the processes X_s and y_s are independent on $(\Omega, \mathcal{F}, \mathcal{F}_s, \tilde{\mathbb{P}})$, and the process X_s satisfies the Îto SDE

$$\begin{aligned} dX = (\alpha(X) - \gamma(X)\beta(X))\,ds + \sigma(X)dw_s \\ +\gamma(X)dy_s, X_0 = x. \qquad (6) \end{aligned}$$

Due to the Kallianpur-Striebel formula (a particular case of the general Bayes formula [7, 9]) for the conditional mean (3), we have

$$\pi_t[f] = \frac{\tilde{\mathbb{E}}\left(f(X_t)\eta_t \mid y_s,\ 0 \le s \le t\right)}{\tilde{\mathbb{E}}\left(\eta_t \mid y_s,\ 0 \le s \le t\right)} = \frac{\tilde{\mathbb{E}}^y\left(f(X_t)\eta_t\right)}{\tilde{\mathbb{E}}^y \eta_t}, \qquad (7)$$

where X_t is from (6), $\tilde{\mathbb{E}}$ means expectation according to the measure $\tilde{\mathbb{P}}$, and $\tilde{\mathbb{E}}^y(\cdot) := \tilde{\mathbb{E}}(\cdot \mid y_s,\ 0 \le s \le t)$. Let

$$\rho_t[g] := \tilde{\mathbb{E}}^y\left(g(X_t)\eta_t\right),$$

where g is a scalar function on $\mathbb{R}^d$. The process ρ_t is often called the unnormalized optimal filter. It satisfies the linear evolution equation

$$d\rho_t[g] = \rho_t[\mathcal{L}g]dt + \rho_t[\mathcal{M}^\top g]dy_t, \quad \rho_0[g] = \pi_0[g], \tag{8}$$

which is known as the Zakai equation. Assuming that there is the corresponding smooth unnormalized filtering density $\rho(t,x)$, it solves the linear stochastic partial differential equation (SPDE) of parabolic type:

$$\begin{aligned} d\rho(t,x) &= \mathcal{L}^*\rho(t,x)dt + \left(\mathcal{M}^*\rho(t,x)\right)^\top dy_t, \\ \rho(0,x) &= \varphi(x). \end{aligned} \tag{9}$$

We note that $\rho_t[g] = \int_{\mathbb{R}^d} g(x)\rho(t,x)dx$.

The unnormalized optimal filter can also be found as a solution of a backward SPDE. To this end, let us fix a time moment t and introduce the function

$$u_g(s,x;t) = \tilde{\mathbb{E}}^y \left(g(X_t^{s,x})\eta_t^{s,x,1} \right), \tag{10}$$

where $x \in \mathbb{R}^d$ is deterministic and $X_{s'}^{s,x}, \eta_{s'}^{s,x,1}, s' \geq s$, is the solution of the Îto SDE

$$\begin{aligned} dX &= (\alpha(X) - \gamma(X)\beta(X))\,ds' + \sigma(X)dw_{s'} \\ &\quad + \gamma(X)dy_{s'}, \quad X_s = x, \\ d\eta &= \beta^\top(X)\eta dy_{s'}, \eta_s = 1. \end{aligned}$$

The function $u_g(s,x;t)$, $s \leq t$, is the solution of the Cauchy problem for the backward linear SPDE:

$$-du = \mathcal{L}uds + \mathcal{M}^\top u * dy_s, \quad u(t,x) = g(x). \tag{11}$$

The notation "$*dy$" means backward Îto integral [5,9]. If $X_0 = x = \xi$ is a random variable with the density $\varphi(\cdot)$, we can write

$$\pi_t[f] = \frac{u_{f,\varphi}(0,t)}{u_{\mathbf{1},\varphi}(0,t)}, \tag{12}$$

where $u_{g,\varphi}(0,t) := \int_{\mathbb{R}^d} u_g(0,x;t)\varphi(x)dx = \tilde{\mathbb{E}}^y \left(g(X_t^{0,\xi})\eta_t^{0,\xi,1} \right) = \rho_t[g]$.

Generally, numerical methods are required to solve optimal filtering equations. For an overview of various numerical approximations for the nonlinear filtering problem, see [1, Chap. 8] together with references therein and for a number of recent developments see [2].

Linear Filtering and Kalman-Bucy Filter

There are a very few cases when explicit formulas for optimal filters are available [1, 7]. The most notable case is when the filtering problem is linear. Consider the system of linear SDE:

$$\begin{aligned} dX &= (a_s + A_s X)\,ds + Q_s dw_s + G_s dv_s, \\ X_0 &= x, \end{aligned} \tag{13}$$

$$dy = (b_s + B_s X)\,ds + dv_s, \; y_0 = 0, \tag{14}$$

where A_s, B_s, Q_s, and G_s are deterministic matrix functions of time s having the appropriate dimensions; a_s and b_s are deterministic vector functions of time s having the appropriate dimensions; the initial condition $X_0 = x$ is a Gaussian random vector with mean $M_0 \in \mathbb{R}^d$ and covariance matrix $C_0 \in \mathbb{R}^d \times \mathbb{R}^d$ and it is independent of both w and v; the other notation is as in (1) and (2).

We note that the solution X_t, y_t of the SDE (13) and (14) is a Gaussian process. The conditional distribution of X_t given $\{y_s, \; 0 \leq s \leq t\}$ is Gaussian with mean $\hat{X}_t$ and covariance P_t, which satisfy the following system of differential equations [1,7,10,11]:

$$\begin{aligned} d\hat{X} &= \left(a_s + A_s\hat{X}\right)ds + \left(G_s + PB_s^\top\right) \\ &\quad (dy_s - (b_s + B_s\hat{X})ds), \; \hat{X}_0 = M_0, \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{d}{dt}P &= PA_s^\top + A_sP - \left(G_s + PB_s^\top\right)\left(G_s + PB_s^\top\right)^\top \\ &\quad + Q_sQ_s^\top + G_sG_s^\top, \; P_0 = C_0. \end{aligned} \tag{16}$$

The solution $\hat{X}_t$, P_t is called the Kalman-Bucy filter (or linear quadratic estimation). We remark that (15) for the conditional mean $\hat{X}_t = \mathbb{E}(X_t \mid y_s, \; 0 \leq s \leq t)$ is a linear SDE, while the solution P_t of the matrix Riccati equation (16) is deterministic and it can be pre-computed off-line. Online updating of $\hat{X}_t$ with arrival of new data y_t from observations is computationally very cheap, and the Kalman-Bucy filter and its various modifications are widely used in practical applications.

References

1. Bain, A., Crisan, D.: Fundamentals of Stochastic Filtering. Springer, New York/London (2008)
2. Crisan, D., Rozovskii, B. (eds.): Handbook of Nonlinear Filtering. Oxford University Press, Oxford (2011)
3. Fristedt, B., Jain, N., Krylov, N.: Filtering and Prediction: A Primer. AMS, Providence (2007)
4. Kallianpur, G.: Stochastic Filtering Theory. Springer, New York (1980)
5. Kunita, H.: Stochastic Flows and Stochastic Differential Equations. Cambridge University Press, Cambridge/New York (1990)
6. Kushner, H.J.: Probability Methods for Approximations in Stochastic Control and for Elliptic Equations. Academic, New York (1977)
7. Liptser, R.S., Shiryaev, A.N.: Statistics of Random Processes. Springer, New York (1977)
8. Pardoux, E.: Équations du filtrage non linéaire de la prédiction et du lissage. Stochastics **6**, 193–231 (1982)
9. Rozovskii, B.L.: Stochastic Evolution Systems, Linear Theory and Application to Nonlinear Filtering. Kluwer Academic, Dordrecht/Boston/London (1991)
10. Stratonovich, R.L.: Conditional Markov Processes and Their Applications to Optimal Control Theory. Elsevier, New York (1968)
11. Xiong, J.: An Introduction to Stochastic Filtering Theory. Oxford University Press, Oxford (2008)

Stochastic ODEs

Peter Kloeden
FB Mathematik, J.W. Goethe-Universität,
Frankfurt am Main, Germany

A scalar stochastic ordinary differential equation (SODE)

$$dX_t = f(t, X_t)\,dt + g(t, X_t)\,dW_t \qquad (1)$$

involves a Wiener process W_t, $t \geq 0$, which is one of the most fundamental stochastic processes and is often called a Brownian motion. A Wiener process is a Gaussian process with $W_0 = 0$ with probability 1 and normally distributed increments $W_t - W_s$ for $0 \leq s < t$ with

$$\mathbb{E}\,(W_t - W_s) = 0, \qquad \mathbb{E}\,(W_t - W_s)^2 = t - s,$$

where the increments $W_{t_2} - W_{t_1}$ and $W_{t_4} - W_{t_3}$ on non-overlapping intervals (i.e., with $0 \leq t_1 < t_2 \leq t_3 < t_4$) are independent random variables. The sample paths of a Wiener process are continuous, but they are nowhere differentiable.

Consequently, an SODE is not a differential equation at all, but just a symbolic representation for the stochastic integral equation

$$X_t = X_{t_0} + \int_{t_0}^{t} f(s, X_s)\,ds + \int_{t_0}^{t} g(s, X_s)\,dW_s,$$

where the first integral is a deterministic Riemann integral for each sample path. The second integral cannot be defined pathwise as a Riemann-Stieltjes integral because the sample paths of the Wiener process do not have even bounded variation on any bounded time interval. Thus, a new type of stochastic integral is required. An Itô stochastic integral $\int_{t_0}^{T} f(t)dW_t$ is defined as the mean-square limit of sums of products of an integrand f evaluated at the left end point of each partition subinterval times $[t_n, t_{n+1}]$, the increment of the Wiener process, i.e.,

$$\int_{t_0}^{T} f(t)dW_t := \text{m.s.} - \lim_{N_\Delta \to \infty} \sum_{j=0}^{N_\Delta - 1} f(t_n)$$

$$\left(W_{t_{n+1}} - W_{t_n}\right),$$

where $t_{n+1} - t_n = \Delta / N_\Delta$ for $n = 0, 1, \ldots, N_\Delta - 1$. The integrand function f may be random or even depend on the path of the Wiener process, but $f(t)$ should be independent of future increments of the Wiener process, i.e., $W_{t+h} - W_t$ for $h > 0$.

The Itô stochastic integral has the important properties (the second is called the Itô isometry) that

$$\mathbb{E}\left[\int_{t_0}^{T} f(t)dW_t\right] = 0, \quad \mathbb{E}\left[\left(\int_{t_0}^{T} f(t)dW_t\right)^2\right]$$

$$= \int_{t_0}^{T} \mathbb{E}\left[f(t)^2\right] dt.$$

However, the solutions of Itô SODE satisfy a different chain rule to that in deterministic calculus, called the Itô formula, i.e.,

$$U(t, X_t) = U(t_0, X_{t_0}) + \int_{t_0}^{t} L^0 U(s, X_s)\, ds + \int_{t_0}^{t} L^1(s, X_s)\, dW_s,$$

where

$$L^0 U = \frac{\partial U}{\partial t} + f \frac{\partial U}{\partial x} + \frac{1}{2} g^2 \frac{\partial^2 U}{\partial x^2}, \qquad L^1 U = g \frac{\partial U}{\partial x}.$$

An immediate consequence is that the integration rules and tricks from deterministic calculus do not hold and different expressions result, e.g.,

$$\int_0^T W_s\, dW_s = \frac{1}{2} W_T^2 - \frac{1}{2} T.$$

The situation for vector-valued SODE and vector-valued Wiener processes is similar. Details can be found in Refs. [3,4,6].

Stratonovich SODEs

There is another stochastic integral called the Stratonovich integral, for which the integrand function is evaluated at the midpoint of each partition subinterval rather than at the left end point. It is written with $\circ dW_t$ to distinguish it from the Itô integral. A Stratonovich SODE is thus written

$$dX_t = f(t, X_t)\, dt + g(t, X_t) \circ dW_t.$$

Stratonovich stochastic calculus has the same chain rule as deterministic calculus, which means that Stratonovich SODE can be solved with the same integration tricks as for ordinary differential equations. However, Stratonovich stochastic integrals do not satisfy the nice properties above for Itô stochastic integrals, which are very convenient for estimates in proofs. Nor does the Stratonovich SODE have same direct connection with diffusion process theory as the Itô SODE, e.g., the coefficient of the Fokker-Planck equation correspond to those of the Itô SODE (1), i.e.,

$$\frac{\partial p}{\partial t} + f \frac{\partial}{\partial x} + \frac{1}{2} g^2 \frac{\partial^2 p}{\partial x^2} = 0.$$

The Itô and Stratonovich stochastic calculi are both mathematically correct. Which one should be used is really a modeling issue, but once one has been chosen, the advantages of the other can be used through a modification of the drift term to obtain the corresponding SODE of the other type that has the same solutions.

Numerical Solution of SODEs

The simplest numerical method for the above SODE (1) is the *Euler-Maruyama scheme* given by

$$Y_{n+1} = Y_n + f(t_n, Y_n)\, \Delta_n + g(t_n, Y_n)\, \Delta W_n,$$

where $\Delta_n = t_{n+1} - t_n$ and $\Delta W_n = W_{t_{n+1}} - W_{t_n}$. This is intuitively consistent with the definition of the Itô integral. Here Y_n is a random variable, which is supposed to be an approximation on X_{t_n}. The stochastic increments ΔW_n, which are $\mathcal{N}(0, \Delta_n)$ distributed, can be generated using, for example, the Box-Muller method. In practice, however, only individual realizations can be computed.

Depending on whether the realizations of the solutions or only their probability distributions are required to be close, one distinguishes between strong and weak convergence of numerical schemes, respectively, on a given interval $[t_0, T]$. Let $\Delta = \max_n \Delta_n$ be the maximum step size. Then a numerical scheme is said to converge with *strong order* γ if, for sufficiently small Δ,

$$\mathbb{E}\left(\left|X_T - Y_{N_T}^{(\Delta)}\right|\right) \le K_T\, \Delta^{\gamma}$$

and with *weak order* β if

$$\left|\mathbb{E}\left(p(X_T)\right) - \mathbb{E}\left(p(Y_{N_T}^{(\Delta)})\right)\right| \le K_{p,T}\, \Delta^{\beta}$$

for each polynomial p. These are global discretization errors, and the largest possible values of γ and β give the corresponding strong and weak orders, respectively, of the scheme for a whole class of stochastic differential equations, e.g., with sufficiently often continuously differentiable coefficient functions. For example, the Euler-Maruyama scheme has strong order $\gamma = \frac{1}{2}$ and weak order $\beta = 1$, while the *Milstein scheme*

$$Y_{n+1} = Y_n + f(t_n, Y_n)\,\Delta_n + g(t_n, Y_n)\,\Delta W_n + \frac{1}{2}\,g(t_n, Y_n)\frac{\partial g}{\partial x}(t_n, Y_n)\,\left\{(\Delta W_n)^2 - \Delta_n\right\}$$

has strong order $\gamma = 1$ and weak order $\beta = 1$; see [2,3,5].

Note that these convergence orders may be better for specific SODE within the given class, e.g., the Euler-Maruyama scheme has strong order $\gamma = 1$ for SODE with additive noise, i.e., for which g does not depend on x, since it then coincides with the Milstein scheme.

The Milstein scheme is derived by expanding the integrand of the stochastic integral with the Itô formula, the stochastic chain rule. The additional term involves the double stochastic integral $\int_{t_n}^{t_{n+1}}\int_{t_n}^{s} dW_u\,dW_s$, which provides more information about the non-smooth Wiener process inside the discretization subinterval and is equal to $\frac{1}{2}\left\{(\Delta W_n)^2 - \Delta_n\right\}$. Numerical schemes of even higher order can be obtained in a similar way.

In general, different schemes are used for strong and weak convergence. The strong stochastic Taylor schemes have strong order $\gamma = \frac{1}{2}, 1, \frac{3}{2}, 2, \ldots$, whereas weak stochastic Taylor schemes have weak order $\beta = 1, 2, 3, \ldots$. See [3] for more details. In particular, one should not use heuristic adaptations of numerical schemes for ordinary differential equations such as Runge-Kutta schemes, since these may not converge to the right solution or even converge at all.

The proofs of convergence rates in the literature assume that the coefficient functions in the above stochastic Taylor schemes are uniformly bounded, i.e., the partial derivatives of appropriately high order of the SODE coefficient functions f and g exist and are uniformly bounded. This assumption, however, is not satisfied in many basic and important applications, for example, with polynomial coefficients such as

$$dX_t = -(1 + X_t)(1 - X_t^2)\,dt + (1 - X_t^2)\,dW_t$$

or with square-root coefficients such as in the Cox-Ingersoll-Ross volatility model

$$dV_t = \kappa\,(\vartheta - V_t)\;dt + \mu\sqrt{V_t}\,dW_t,$$

which requires $V_t \geq 0$. The second is more difficult because there is a small probability that numerical iterations may become negative, and various ad hoc methods have been suggested to prevent this. The paper [1] provides a systematic method to handle both of these problems by using pathwise convergence, i.e.,

$$\sup_{n=0,\ldots,N_T}\left|X_{t_n}(\omega) - Y_n^{(\Delta)}(\omega)\right| \longrightarrow 0 \;\text{ as }\; \Delta \to 0, \quad \omega \in \Omega.$$

It is quite natural to consider pathwise convergence since numerical calculations are actually carried out path by path. Moreover, the solutions of some SODE do not have bounded moments, so pathwise convergence may be the only option.

Iterated Stochastic Integrals

Vector-valued SODE with vector-valued Wiener processes can be handled similarly. The main new difficulty is how to simulate the multiple stochastic integrals since these cannot be written as simple formulas of the basic increments as in the double integral above when they involve different Wiener processes. In general, such multiple stochastic integrals cannot be avoided, so they must be approximated somehow. One possibility is to use random Fourier series for Brownian bridge processes based on the given Wiener processes; see [3,5].

Another way is to simulate the integrals themselves by a simpler numerical scheme. For example, double integral

$$I_{(2,1),n} = \int_{t_n}^{t_{n+1}}\int_{t_n}^{t} dW_s^2\,dW_t^1$$

for two independent Wiener processes W_t^1 and W_t^2 can be approximated by applying the (vector-valued) Euler-Maruyama scheme to the 2-dimensional Itô SODE (with superscripts labeling components)

$$dX_t^1 = X_t^2\,dW_t^1, \qquad dX_t^2 = dW_t^2, \tag{2}$$

over the discretization subinterval $[t_n, t_{n+1}]$ with a suitable step size $\delta = (t_{n+1} - t_n)/K$. The solution of the SODE (2) with the initial condition $X_{t_n}^1 = 0$, $X_{t_n}^2 = W_{t_n}^2$ at time $t = t_{n+1}$ is given by

$$X_{t_{n+1}}^1 = I_{(2,1),n}, \qquad X_{t_{n+1}}^2 = \Delta W_n^2.$$

Writing $\tau_k = t_n + k\delta$ and $\delta W^j_{n,k} = W^j_{\tau_{k+1}} - W^j_{\tau_k}$, the stochastic Euler scheme for the SDE (2) reads

$$Y^1_{k+1} = Y^1_k + Y^2_k\,\delta W^1_{n,k}, \quad Y^2_{k+1} = Y^2_k + \delta W^2_{n,k}, \qquad \text{for } 0 \le k \le K-1, \tag{3}$$

with the initial value $Y^1_0 = 0$, $Y^2_0 = W^2_{t_n}$. The strong order of convergence of $\gamma = \frac{1}{2}$ of the Euler-Maruyama scheme ensures that

$$\mathbb{E}\left(\left|Y^1_K - I_{(2,1),n}\right|\right) \le C\sqrt{\delta},$$

so $I_{(2,1),n}$ can be approximated in the Milstein scheme by Y^1_K with $\delta \approx \Delta^2_n$, i.e., $K \approx \Delta^{-1}_n$, without affecting the overall order of convergence.

Commutative Noise

Identities such as

$$\int_{t_n}^{t_{n+1}}\int_{t_n}^{t} dW^{j_1}_s\,dW^{j_2}_t + \int_{t_n}^{t_{n+1}}\int_{t_n}^{t} dW^{j_2}_s\,dW^{j_1}_t = \Delta W^{j_1}_n\,\Delta W^{j_2}_n$$

allow one to avoid calculating the multiple integrals if the corresponding coefficients in the numerical scheme are identical, in this case if $L^1 g_2(t,x) \equiv L^2 g_1(t,x)$ (where L^2 is defined analogously to L^1) for an SODE of the form

$$dX_t = f(t,X_t)\,dt + g_1(t,X_t)\,dW^1_t + g_2(t,X_t)\,dW^2_t. \tag{4}$$

Then the SODE (4) is said to have *commutative noise*.

Concluding Remarks

The need to approximate multiple stochastic integrals places a practical restriction on the order of strong schemes that can be implemented for a general SODE. Wherever possible special structural properties like commutative noise of the SODE under investigation should be exploited to simplify strong schemes as much as possible. For weak schemes the situation is easier as the multiple integrals do not need to be approximated so accurately. Moreover, extrapolation of weak schemes is possible.

The important thing is to decide first what kind of approximation one wants, strong or weak, as this will determine the type of scheme that should be used, and then to exploit the structural properties of the SODE under consideration to simplify the scheme that has been chosen to be implemented.

References

1. Jentzen, A., Kloeden, P.E., Neuenkirch, A.: Convergence of numerical approximations of stochastic differential equations on domains: higher order convergence rates without global Lipschitz coefficients. Numer. Math. **112**, 41–64 (2009)
2. Kloeden, P.E.: The systematic deviation of higher order numerical methods for stochastic differential equations. Milan J. Math. **70**, 187–207 (2002)
3. Kloeden, P.E., Platen, E.: The Numerical Solution of Stochastic Differential Equations, 3rd rev. printing. Springer, Berlin, (1999)
4. Mao, X.: Stochastic Differential Equations and Applications, 2nd edn. Horwood, Chichester (2008)
5. Milstein, G.N.: Numerical Integration of Stochastic Differential Equations. Kluwer, Dordrecht (1995)
6. Øksendal, B.: Stochastic Differential Equations. An Introduction with Applications, 6th edn. 2003, Corr. 4th printing. Springer, Berlin (2007)

Stochastic Simulation

Dieter W. Heermann
Institute for Theoretical Physics, Heidelberg
University, Heidelberg, Germany

Synonyms

Brownian Dynamics Simulation; Langevin Simulation; Monte Carlo Simulations

Modelling a system or data one is often faced with the following:

- Exact data is unavailable or expensive to obtain
- Data is uncertain and/or specified by a probability distribution

or decisions have to be made with respect to the degrees of freedom that are taken explicitly into account. This can be seen by looking at a system with two components. One of the components could be water molecules and the other component large molecules. The decision is to take the water molecules explicitly into account or to treat them implicitly. Since the water molecules move much faster than the large molecules, we can eliminate the water by subsuming their action on the larger molecules by a stochastic force, i.e., as a random variable with a specific distribution. Thus we have eliminated some details in favor of a probabilistic description where perhaps some elements of the model description are given by deterministic rules and other contributes stochastically. Overall a model derived along the outlined path can be viewed as if an individual state has a probability that may depend on model parameters.

In most of the interesting cases, the number of available states the model has will be so large that they simply cannot be enumerated. A sampling of the states is necessary such that the most relevant states will be sampled with the correct probability. Assume that the model has some deterministic part. In the above example, the motion of larger molecules is governed by Newton's equation of motion. These are augmented by stochastic forces mimicking the water molecules. Depending on how exactly this is implemented results in Langevin equations

$$m\ddot{x} = -\nabla U(x) - \gamma m\dot{x} + \xi(t)\sqrt{2\gamma k_B T m}, \tag{1}$$

where x denotes the state (here the position), U the potential, m the mass of the large molecule, k_B the Boltzmann constant, T the temperature, and ξ the stochastic force with the properties:

$$\langle \xi(t) \rangle = 0 \tag{2}$$

$$\left\langle \xi(t)\xi(t') \right\rangle = \delta(t - t'). \tag{3}$$

Neglecting the acceleration in the Langevin equation yields the Brownian dynamics equation of motion

$$\dot{x}(t) = -\nabla U(x)/\zeta + \xi(t)\sqrt{2D} \tag{4}$$

with $\zeta = \gamma m$ and $D = k_B T/\zeta$.

Hence the sampling is obtained using the equations of motion to transition from one state to the next. If enough of the available states (here x) are sampled, quantities of interest that depend on the states can be calculated as averages over the generated states:

$$\bar{A} = \sum_x A(x)P(x,\alpha), \tag{5}$$

where $P(x,\alpha)$ is the probability of the state and α a set of parameters (e.g., the temperature T, mass m, etc.).

A point of view that can be taken is that what the Eqs. (1) and (4) accomplish is the generation of a stochastic trajectory through the available states. This can equally be well established by other means. As long as we satisfy the condition that the right probability distribution is generated, we could generate the trajectory by a Monte Carlo method [1].

In a Monte Carlo formulation, a transition probability from a state x to another state x' is specified

$$W(x'|x). \tag{6}$$

Together with the proposition probability for the state x', a decision is made to accept or reject the state (Metropolis-Hastings Monte Carlo Method). An advantage of this formulation is that it allows freedom in the choice of proposition of states and the efficient sampling of the states (importance sampling). In more general terms, what one does is to set up a biased random walk that explores the target distribution (Markov Chain Monte Carlo).

A special case of the sampling that yields a Markov chain is the Gibbs sampler. Assume $x = (x^1, x^2)$ with target $P(x,\alpha)$

Algorithm 1 Gibbs Sampler Algorithm:

```
initialize $x_0 = (x_0^1, x_0^2)$
while $i \leq$ max number of samples do
    sample $x_i^1 \sim P(x^1|x_{i-1}^2, \alpha)$
    sample $x_i^2 \sim P(x^2|x_i^1, \alpha)$
end while
```

then $\{x^1, x^2\}$ is a Markov chain. Thus we obtain a sequence of states such that we can again apply (5) to compute the quantities of interest.

Common to all of the above stochastic simulation methods is the use of random numbers. They are either used for the implementation of the random contribution to the force or the decision whether a state is accepted or rejected. Thus the quality of the result depends on the quality of the random number generator.

Reference

1. Binder, K., Heermann, D.W.: Monte Carlo Simulation in Statistical Physics: An Introduction. Graduate Texts in Physics, 5th edn. Springer, Heidelberg (2010)

Stochastic Systems

Guang Lin[1,2,3] and George Em Karniadakis[4]
[1]Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA
[2]School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA
[3]Department of Mathematics, Purdue University, West Lafayette, IN, USA
[4]Division of Applied Mathematics, Brown University, Providence, RI, USA

Mathematics Subject Classification

93E03

Synonyms

Noisy Systems; Random Systems; Stochastic Systems

Short Definition

A stochastic system may contain one or more elements of random, i.e., nondeterministic behavior. Compared to a deterministic system, a stochastic system does not always generate the same output for a given input. The elements of systems that can be stochastic in nature may include noisy initial conditions, random boundary conditions, random forcing, etc.

Description

Stochastic systems (SS) are encountered in many application domains in science, engineering, and business. They include logistics, transportation, communication networks, financial markets, supply chains, social systems, robust engineering design, statistical physics, systems biology, etc. Stochasticity or randomness is perhaps associated with a bad outcome, but harnessing stochasticity has been pursued in arts to create beauty, e.g., in the paintings of Jackson Pollock or in the music of Iannis Xenakis. Similarly, it can be exploited in science and engineering to design new devices (e.g., stochastic resonances in the Bang and Olufsen speakers) or to design robust and cost-effective products under the framework of uncertainty-based design and real options [17].

Stochasticity is often associated with uncertainty, either intrinsic or extrinsic, and specifically with the lack of knowledge of the properties of the system; hence quantifying uncertain outcomes and system responses is of great interest in applied probability and scientific computing. This uncertainty can be further classified as aleatory, i.e., statistical, and epistemic which can be reduced by further measurements or computations of higher resolution. Mathematically, stochasticity can be described by either deterministic or stochastic differential equations. For example, the molecular dynamics of a simple fluid is described by the classical deterministic Newton's law of motion or by the deterministic Navier-Stokes equations whose outputs, in both cases, however may be stochastic. On the other hand, stochastic elliptic equations can be used to predict the random diffusion of water in porous media, and similarly a stochastic differential equation may be used to model neuronal activity in the brain [9, 10].

Here we will consider systems that are governed by stochastic ordinary and partial differential equations (SODEs and SPDEs), and we will present some effective methods for obtaining stochastic solutions in the next section. In the classical stochastic analysis, these terms refer to differential equations subject to *white noise* either additive or multiplicative, but in more recent years, the same terminology has been adopted for differential equations with color noise, i.e., processes that are correlated in space or time. The color of noise, which can also be pink or violet, may dictate the numerical method used to predict efficiently the response of a stochastic system, and hence it is important to consider this carefully at the modeling stage. Specifically, the correlation length or time scale is the most important parameter of a stochastic process as it determines the effective dimension of the process; the smaller the correlation scale, the larger the dimensionality of the stochastic system.

Example: To be more specific in the following, we present a tumor cell growth model that involves stochastic inputs that need to be represented according to the correlation structure of available empirical data [27]. The evolution equation is

$$\begin{aligned} \dot{x}(t;\omega) =& G(x) + g(x) f_1(t;\omega) + f_2(t;\omega), \\ x(0;\omega) =& x_0(\omega), \end{aligned} \qquad (1)$$

where $x(t;\omega)$ denotes the concentration of tumor cell at time t,

$$G(x) = x(1-\theta x) - \beta \frac{x}{x+1}, \qquad g(x) = -\frac{x}{x+1},$$

β is the immune rate, and θ is related to the rate of growth of cytotoxic cells. The random process $f_1(t;\omega)$ represents the strength of the treatment (i.e., the dosage of the medicine in chemotherapy or the intensity of the ray in radiotherapy), while the process $f_2(t;\omega)$ is related to other factors, such as drugs and radiotherapy, that restrain the number of tumor cells. The parameters β, θ and the covariance structure of random processes f_1 and f_2 are usually estimated based on empirical data.

If the processes $f_1(t,\omega)$ and $f_2(t,\omega)$ are independent, they can be represented using the Karhunen-Loeve (K-L) expansion by a zero mean, second-order random process $f(t,\omega)$ defined on a probability space $(\Omega, \mathsf{F}, \mathbf{P})$ and indexed over $t \in [a, b]$. Let us denote the continuous covariance function of $f(t;\omega)$ as $C(s, t)$. Then the process $f(t;\omega)$ can be represented as

$$f(t;\omega) = \sum_{k=1}^{N_d} \sqrt{\lambda_k} e_k(t) \xi_k(\omega),$$

where $\xi_k(\omega)$ are uncorrelated random variables with zero mean and unitary variance, while λ_k and $e_k(t)$ are, respectively, eigenvalues and (normalized) eigenfunctions of the integral operator with kernel $C(s, t)$, i.e.,

$$\int_a^b C(s, t) e_k(s) ds = \lambda_k e_k(t).$$

The dimension N_d depends strongly on the correlation scale of the kernel $C(s,t)$. If we rearrange the eigenvalues λ_k in a descending order, then any truncation of the expansion $f(t;\omega)$ is optimal in the sense that it minimizes the mean square error [3, 18, 20]. The K-L expansion has been employed to represent random input processes in many stochastic simulations (see, e.g., [7, 20]).

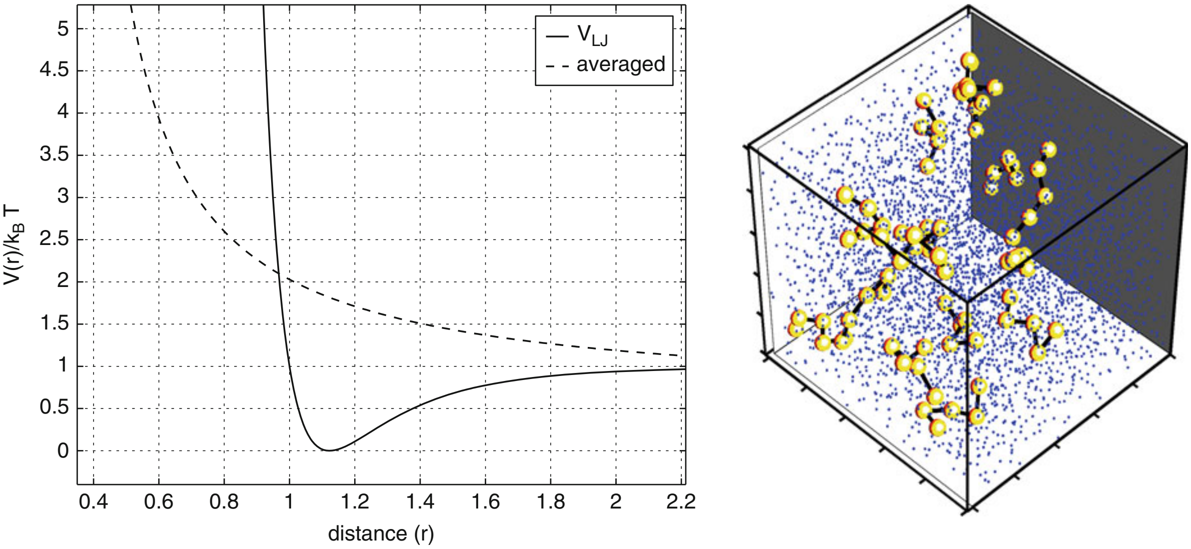
Stochastic Modeling and Computational Methods

We present two examples of two fundamentally different descriptions of SS in order to show some of the complexity but also rich stochastic response that can be obtained: the first is based on a discrete particle model and is governed by SODEs, and the second is a continuum system and is governed by SPDEs.

A Stochastic Particle System: We first describe a stochastic model for a mesoscopic system governed by a modified version of Newton's law of motion, the so-called dissipative particle dynamics (DPD) equations [19]. It consists of particles which correspond to *coarse-grained* entities, thus representing molecular clusters rather than individual atoms. The particles move off-lattice interacting with each other through a set of prescribed (conservative and stochastic) and velocity-dependent forces. Specifically, there are three types of forces acting on each dissipative particle: (a) a purely repulsive conservative force, (b) a dissipative force that reduces velocity differences between the particles, and (c) a stochastic force directed along the line connecting the center of the particles. The last two forces effectively implement a thermostat so that thermal equilibrium is achieved. Correspondingly, the amplitude of these forces is dictated by the fluctuation-dissipation theorem that ensures that in thermodynamic equilibrium the system will have a *canonical* distribution. All three forces are modulated by a weight function which specifies the range of interaction or cutoff radius r_c between the particles and renders the interaction local.

The DPD equations for a system consisting of N particles have equal mass (for simplicity in the presentation) m, position $\mathbf{r}_i$, and velocities $\mathbf{u}_i$, which are stochastic in nature. The aforementioned three types of forces exerted on a particle i by particle j are given by

$$\mathbf{F}_{ij}^c = F^{(c)}(r_{ij})\mathbf{e}_{ij}, \qquad \mathbf{F}_{ij}^d = -\gamma \omega^d(r_{ij})(\mathbf{v}_{ij} \cdot \mathbf{e}_{ij})\mathbf{e}_{ij},$$
$$\mathbf{F}_{ij}^r = \sigma \omega^r(r_{ij}) \xi_{ij} \mathbf{e}_{ij},$$



Stochastic Systems, Fig. 1 *Left*: Lennard-Jones potential and its averaged soft repulsive-only potential. *Right*: Polymer chains flowing in a sea of solvent in DPD. For more details see [19]

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$, $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$, $r_{ij} = |\mathbf{r}_{ij}|$ and the unit vector $\mathbf{e}_{ij} = \frac{\mathbf{r}_{ij}}{r_{ij}}$. The variables γ and σ determine the strength of the dissipative and random forces, respectively, ξ_{ij} are symmetric Gaussian random variables with zero mean and unit variance, and ω^d and ω^r are weight functions.

The time evolution of DPD particles is described by Newton's law

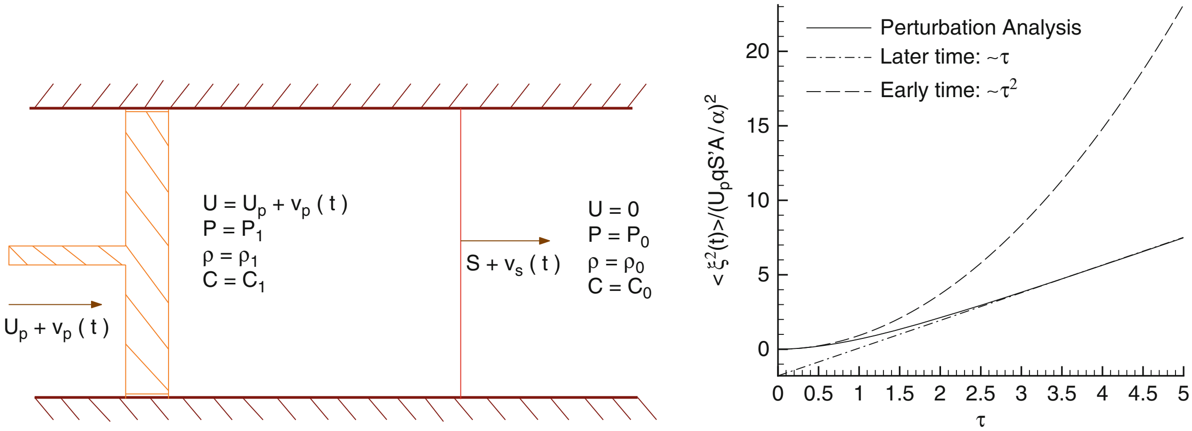
$$d\mathbf{r}_i = \mathbf{v}_i \delta t; \qquad d\mathbf{v}_i = \frac{\mathbf{F}_i^c \delta t + \mathbf{F}_i^d \delta t + \mathbf{F}_i^r \sqrt{\delta t}}{m_i},$$

where $\mathbf{F}_i^c = \sum_{i \neq j} \mathbf{F}_{ij}^c$ is the total conservative force acting on particle i; $\mathbf{F}_i^d$ and $\mathbf{F}_i^r$ are defined similarly. The velocity increment due to the random force has a factor $\sqrt{\delta t}$ since it represents Brownian motion, which is described by a standard Wiener process with a covariance kernel given by $C_{FF}(t_i, t_j) = e^{-\frac{|t_1 - t_2|}{A}}$, where A is the correlation time for this stochastic process. The conservative force $\mathbf{F}^c$ is typically given in terms of a soft potential in contrast to the Lennard-Jones potential used in molecular dynamics studies (see Fig. 1(left)). The dissipative and random forces are characterized by strengths $\omega^d(r_{ij})$ and $\omega^r(r_{ij})$ coupled due to the *fluctuation-dissipation* theorem.

Several complex fluid systems in industrial and biological applications (DNA chains, polymer gels, lubrication) involve multiscale processes and can be modeled using modifications of the above stochastic DPD equations [19]. Dilute polymer solutions are a typical example, since individual polymer chains form a group of large molecules by atomic standards but still governed by forces similar to intermolecular ones. Therefore, they form large repeated units exhibiting slow dynamics with possible nonlinear interactions (see Fig. 1(right)).

A Stochastic Continuum System: Here we present an example from classical aerodynamics on shock dynamics by reformulating the one-dimensional piston problem within the stochastic framework, i.e., we allow for random piston motions which may be changing in time [11]. In particular, we superimpose *small* random velocity fluctuations to the piston velocity and aim to obtain both analytical and numerical solutions of the stochastic flow response. We consider a piston having a constant velocity, U_p, moving into a straight tube filled with a homogeneous gas at rest. A shock wave will be generated ahead of the piston. A sketch of the piston-driven shock tube with random piston motion superimposed is shown in Fig. 2(left).

As shown in Fig. 2 (left), U_p and S are the deterministic speed of the piston and deterministic speed of the shock, respectively, and ρ_0, P_0, C_0, ρ_1, P_1, and C_1 are the deterministic density, pressure, and local sound



Stochastic Systems, Fig. 2 *Left*: Sketch of piston-driven shock tube with random piston motion. *Right*: Normalized variance of perturbed shock paths. *Solid line*: perturbation analysis results. *Dashed line*: early-time asymptotic results, $\langle \xi^2(\tau)\rangle \sim \tau^2$. *Dash-dotted line*: late-time asymptotic results, $\langle \xi^2(\tau)\rangle \sim \tau$

speed ahead and after of the shock, respectively. We now define the stochastic motion of the piston by superimposing a small stochastic component to the steady speed of the piston, i.e., $u_p(t) = U_p[1 + \epsilon V(t,\omega)]$, where ϵ is the amplitude of the random perturbation. Here $V(t,\omega)$ is modeled as a random process with zero mean and covariance $\langle V(t_1,\omega), V(t_2,\omega)\rangle = e^{-\frac{|t_1-t_2|}{A}}$, where A is the correlation time; it can be represented by a truncated K-L expansion as explained earlier. Our objective is to quantify the deviation of the perturbed shock paths due to the random piston motion from the unperturbed ones, which are given by $X(t) = S \cdot t$. If the amplitude ϵ is small, $0 < \epsilon \ll 1$, the analytical solutions for the perturbed shock paths can be expressed as follows:

$$\langle \xi^2(\tau)\rangle = (U_p q S' A/\alpha)^2 \left[2\sum_{n=1}^{\infty}\sum_{m=0}^{n-1}(-r)^{n+m} I_{n,m}(\tau) + \sum_{n=0}^{\infty} r^{2n} I_{n,n}(\tau)\right] \qquad (2)$$

where $\tau = \alpha t/A$, and

$$I_{n,m}(\tau) = \frac{2\tau}{\beta^m} + \frac{1}{\beta^{n+m}}\left[e^{-\beta^m \tau} + e^{-\beta^n \tau} - 1 - e^{-(\beta^m-\beta^n)\tau}\right],$$

where $S' = \frac{dS}{dU_p}$, $m < n$, $\alpha = \frac{C_1+U_p-S}{C_1}$, $\beta = \frac{C_1+U_p-S}{C_1+S-U_p}$, $q = \frac{2}{1+k}$ and $r = \frac{1-k}{1+k}$. Here $k = C\,\frac{S+S'U_p}{1+\gamma S U_p}$ and $\gamma = c_p/c_v$ is the ratio of specific heats.

In Fig. 2 (right), the variance of the perturbed shock path, $\langle \xi^2(\tau)\rangle/(U_p q S' A/\alpha)^2$, is plotted as a function of τ with $U_p = 1.25$, i.e., corresponding to Mach number of the shock $M = 2$. The asymptotic formulas for small and large τ are also included in the plot. In Fig. 2 (right), we observe that the variance of the shock location grows *quadratically* with time for early times and switches to *linear* growth for longer times.

The stochastic solutions for shock paths, for either small or large piston motions, can also be obtained numerically by solving the full nonlinear Euler equations with an unsteady stochastic boundary, namely, the piston position to model the stochastic piston problem. Classic Monte Carlo simulations [4] or quasi-Monte Carlo simulations [2] can be performed for these stochastic simulations. However, due to the slow convergence rate of Monte Carlo methods, it may take thousands of equivalent deterministic simulations to achieve acceptable accuracy. Recently, methods based on generalized polynomial chaos (gPC) expansions have become popular for such SPDEs due to their fast convergence rate for SS with color noise. The term *polynomial chaos* was first coined by Norbert Wiener in 1938 in his pioneering work on representing Gaussian stochastic processes [22] as generalized Fourier series. In Wiener's work, Hermite polynomials serve as an orthogonal basis. The gPC method for solving SPDEs is an extension of the polynomial chaos method developed in [7], inspired by the theory of Wiener-Hermite polynomial chaos. The use of Hermite polynomials may not be optimum in applications involving non-Gaussian processes, and hence gPC was proposed in [25] to alleviate the difficulty. In gPC, different kinds of orthogonal polynomials are chosen as a basis depending on the probability distribution of the random inputs. The Pth-order, gPC approximations of the solution $u(x,\xi)$ can be obtained by projecting u onto the space W_N^P, i.e.,

$$\mathbb{P}_N^P u = u_N^P(x,\xi) = \sum_{m=1}^{M} \hat{u}_m(x)\phi_m(\xi), \qquad (3)$$

where $\mathbb{P}_N^P u$ denotes the orthogonal projection operator from $L_\rho^2(\tau)$ onto W_N^P, $M+1 = \frac{(N+P)!}{N!P!}$, and $\hat{u}_m$ are the coefficients, and ρ the probability measure.

Although gPC was shown to exhibit exponential convergence in approximating stochastic solutions at finite times, gPC may converge slowly or fail to converge even in short-time integration due to a discontinuity of the approximated solution in random space. To this end, the Wiener-Haar method [14, 15] based on wavelets, random domain decomposition [12], multielement-gPC (ME-gPC) [21], and multielement probabilistic collocation method (ME-PCM) [5] were developed to address problems related to the aforementioned discontinuities in random space. Additionally, a more realistic representation of stochastic inputs associated with various sources of uncertainty in the stochastic systems may lead to high-dimensional representations, and hence exponential computational complexity, running into the so-called curse of dimensionality. Sparse grid stochastic collocation method [24] and various versions ANOVA (ANalysis Of VAriance) method [1, 6, 8, 13, 23, 26] have been employed as effective dimension-reduction techniques for quantifying the uncertainty in stochastic systems with dimensions up to 100.

Conclusion

Aristotle's logic has ruled our scientific thinking in the past two millennia. Most scientific models and theories have been constructed from exact models and logic reasoning. It is argued in [16] that SS models and statistical reasoning are more relevant "i) to the world, ii) to science and many parts of mathematics and iii) particularly to understanding the computations in our own minds, than exact models and logical reasoning." Indeed, many real-world problems can be viewed or modeled as SS with great potential benefits across disciplines from physical sciences and engineering to social sciences. Stochastic modeling can bring in more realism and flexibility and account for uncertain inputs and parametric uncertainty albeit at the expense of mathematical and computational complexity. However, the rapid mathematical and algorithmic advances already realized at the beginning of the twenty-first century along with the simultaneous advances in computer speeds and capacity will help alleviate such difficulties and will make stochastic modeling the standard norm rather than the exception in the years ahead. The three examples we presented in this chapter illustrate the diverse applications of stochastic modeling in biomedicine, materials processing, and fluid mechanics. The same methods or proper extensions can also be applied to quantifying uncertainties in climate modeling; in decision making under uncertainty, e.g., in robust engineering design and in financial markets; but also for modeling the plethora of emerging social networks. Further work on the mathematical and algorithmic formulations of such more complex and high-dimensional systems is required as current approaches cannot yet deal satisfactorily with white noise, system discontinuities, high dimensions, and long-time integration.

Acknowledgements The authors would like to acknowledge support by DOE/PNNL Collaboratory on Mathematics for Mesoscopic Modeling of Materials (CM4).

References

1. Bieri, M., Schwab, C.: Sparse high order fem for elliptic sPDEs. Comput. Methods Appl. Mech. Eng. **198**, 1149–1170 (2009)
2. Caflisch, R.: Monte Carlo and quais-Monte Carlo methods. Acta Numer. **7**, 1–49 (1998)
3. Chien, Y., Fu, K.S.: On the generalized Karhunen-Loève expansion. IEEE Trans. Inf. Theory **13**, 518–520 (1967)
4. Fishman, G.: Monte Carlo: Concepts, Algorithms, and Applications. Springer Series in Operations Research and Financial Engineering. Springer, New York (2003)
5. Foo, J., Wan, X., Karniadakis, G.E.: A multi-element probabilistic collocation method for PDEs with parametric un-

certainty: error analysis and applications. J. Comput. Phys. **227**, 9572–9595 (2008)

6. Foo, J.Y., Karniadakis, G.E.: Multi-element probabilistic collocation in high dimensions. J. Comput. Phys. **229**, 1536–1557 (2009)
7. Ghanem, R.G., Spanos, P.: Stochastic Finite Eelements: A Spectral Approach. Springer, New York (1991)
8. Griebel, M.: Sparse grids and related approximation schemes for higher-dimensional problems. In: Proceedings of the Conference on Foundations of Computational Mathematics, Santander (2005)
9. van Kampen, N.: Stochastic Processes in Physics and Chemistry, 3rd edn. Elsevier, Amsterdam (2008)
10. Laing, C., Lord, G.J.: Stochastic Methods in Neuroscience. Oxford University Press, Oxford (2009)
11. Lin, G., Su, C.H., Karniadakis, G.E.: The stochastic piston problem. Proc. Natl. Acad. Sci. U. S. A. **101**(45), 15,840–15,845 (2004)
12. Lin, G., Tartakovsky, A.M., Tartakovsky, D.M.: Uncertainty quantification via random domain decomposition and probabilistic collocation on sparse grids. J. Comput. Phys. **229**, 6995–7012 (2010)
13. Ma, X., Zabaras, N.: An adaptive hierarchical sparse grid collocation method for the solution of stochastic differential equations. J. Comput. Phys. **228**, 3084–3113 (2009)
14. Maitre, O.P.L., Njam, H.N., Ghanem, R.G., Knio, O.M.: Multi-resolution analysis of Wiener-type uncertainty propagation schemes. J. Comput. Phys. **197**, 502–531 (2004)
15. Maitre, O.P.L., Njam, H.N., Ghanem, R.G., Knio, O.M.: Uncertainty propagation using Wiener-Haar expansions. J. Comput. Phys. **197**, 28–57 (2004)
16. Mumford, D.: The dawning of the age of stochasticity. In: Mathematics Towards the Third Millennium, Rome (1999)
17. de Neufville, R.: Uncertainty Management for Engineering Systems Planning and Design. MIT, Cambridge (2004)
18. Papoulis, A.: Probability, Random Variables and Stochastic Processes, 3rd edn. McGraw-Hill, Europe (1991)
19. Symeonidis, V., Karniadakis, G., Caswell, McGraw-Hill Europe B.: A seamless approach to multiscale complex fluid simulation. Comput. Sci. Eng. **7**, 39–46 (2005)
20. Venturi, D.: On proper orthogonal decomposition of randomly perturbed fields with applications to flow past a cylinder and natural convection over a horizontal plate. J. Fluid Mech. **559**, 215–254 (2006)
21. Wan, X., Karniadakis, G.E.: An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. J. Comput. Phys. **209**, 617–642 (2005)
22. Wiener, N.: The homogeneous chaos. Am. J. Math. **60**, 897–936 (1938)
23. Winter, C., Guadagnini, A., Nychka, D., Tartakovsky, D.: Multivariate sensitivity analysis of saturated flow through simulated highly heterogeneous groundwater aquifers. J. Comput. Phys. **217**, 166–175 (2009)
24. Xiu, D., Hesthaven, J.: High order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**(3), 1118–1139 (2005)
25. Xiu, D., Karniadakis, G.E.: The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)
26. Yang, X., Choi, M., Lin, G., Karniadakis, G.E.: Adaptive anova decomposition of stochastic incompressible and compressible flows. J. Comput. Phys. **231**, 1587–1614 (2012)
27. Zeng, C., Wang, H.: Colored noise enhanced stability in a tumor cell growth system under immune response. J. Stat. Phys. **141**, 889–908 (2010)

Stokes or Navier-Stokes Flows

Vivette Girault and Frédéric Hecht
Laboratoire Jacques-Louis Lions, UPMC University of Paris 06 and CNRS, Paris, France

Mathematics Subject Classification

76D05; 76D07; 35Q30; 65N30; 65F10

Definition Terms/Glossary

Boundary layer It refers to the layer of fluid in the immediate vicinity of a bounding surface where the effects of viscosity are significant.

GMRES Abbreviation for the generalized minimal residual algorithm. It refers to an iterative method for the numerical solution of a nonsymmetric system of linear equations.

Precondition It consists in multiplying both sides of a system of linear equations by a suitable matrix, called the preconditioner, so as to reduce the condition number of the system.

The Incompressible Navier-Stokes Model

The incompressible Navier-Stokes system of equations is a widely accepted model for viscous Newtonian incompressible flows. It is extensively used in meteorology, oceanography, canal flows, pipeline flows, automotive industry, high-speed trains, wind turbines, etc. Computing accurately its solutions is a difficult and important challenge.

A Newtonian fluid is a model whose Cauchy stress tensor depends linearly on the strain tensor, in contrast to non-Newtonian fluids for which this relation is nonlinear and possibly implicit. For a Navier-Stokes fluid model, the constitutive equation defining the Cauchy stress tensor $\boldsymbol{T}$ is:

$$\boldsymbol{T} = -\pi \boldsymbol{I} + 2\mu \boldsymbol{D}(\boldsymbol{u}), \tag{1}$$

where $\mu > 0$ is the constant viscosity coefficient, representing friction between molecules, π is the pressure, $\boldsymbol{u}$ is the velocity, $\boldsymbol{D}(\boldsymbol{u}) = \frac{1}{2}\left(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\right)$ is the strain tensor, and $(\nabla \boldsymbol{u})_{ij} = \frac{\partial u_i}{\partial x_j}$ is the gradient tensor. When substituted into the balance of linear momentum

$$\rho \frac{d\boldsymbol{u}}{dt} = \operatorname{div} \boldsymbol{T} + \rho \boldsymbol{f}, \tag{2}$$

where $\rho > 0$ is the fluid's density, $\boldsymbol{f}$ is an external body force (e.g., gravity), and $\frac{d\boldsymbol{u}}{dt}$ is the material time derivative

$$\frac{d\boldsymbol{u}}{dt} = \frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla \boldsymbol{u}, \text{ where } \boldsymbol{u}\cdot\nabla \boldsymbol{u} = [\nabla \boldsymbol{u}]\boldsymbol{u} = \sum_i u_i \frac{\partial \boldsymbol{u}}{\partial x_i}, \tag{3}$$

(1) gives, after division by ρ,

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\frac{1}{\rho} \nabla \pi + 2\frac{\mu}{\rho} \operatorname{div} \boldsymbol{D}(\boldsymbol{u}) + \boldsymbol{f}.$$

But the density ρ is constant, since the fluid is incompressible. Therefore renaming the quantities $p = \frac{\pi}{\rho}$ and the kinematic viscosity $\nu = \frac{\mu}{\rho}$, the momentum equation reads:

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\nabla p + 2\nu \operatorname{div} \boldsymbol{D}(\boldsymbol{u}) + \boldsymbol{f}. \tag{4}$$

As the fluid is incompressible, the conservation of mass

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \boldsymbol{u}) = 0,$$

reduces to the incompressibility condition

$$\operatorname{div} \boldsymbol{u} = 0. \tag{5}$$

From now on, we assume that Ω is a *bounded, connected, open set* in $\mathbb{R}^3$, *with a suitably piecewise smooth boundary* $\partial\Omega$ (essentially, *without cusps or multiple points*). The relations (4) and (5) are the incompressible Navier-Stokes equations in Ω. They are complemented with boundary conditions, such as the no-slip condition

$$\boldsymbol{u} = \boldsymbol{0}, \text{ on } \partial\Omega, \tag{6}$$

and an initial condition

$$\begin{aligned} &\boldsymbol{u}(\cdot, 0) = \boldsymbol{u}_0(\cdot) \text{ in } \Omega, \\ &\text{satisfying } \operatorname{div}\boldsymbol{u}_0 = 0, \text{ and } \boldsymbol{u}_0 = \boldsymbol{0}, \text{ on } \partial\Omega. \end{aligned} \tag{7}$$

In practical situations, other boundary conditions may be prescribed. One of the most important occurs in flows past a moving obstacle, in which case (6) is replaced by

$$\boldsymbol{u} = \boldsymbol{g}, \text{ on } \partial\Omega \quad \text{where} \int_{\partial\Omega} \boldsymbol{g} \cdot \boldsymbol{n} = 0, \tag{8}$$

where $\boldsymbol{g}$ is the velocity of the moving body and $\boldsymbol{n}$ denotes the unit exterior normal vector to $\partial\Omega$. To simplify, we shall only discuss (6), but we shall present numerical experiments where (8) is prescribed.

If the boundary conditions do not involve the boundary traction vector $\boldsymbol{T}\boldsymbol{n}$, (4) can be substantially simplified by using the fluid's incompressibility. Indeed, (5) implies $2\operatorname{div} \boldsymbol{D}(\boldsymbol{u}) = \Delta \boldsymbol{u}$, and (4) becomes

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} - \nu \Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}. \tag{9}$$

When written in dimensionless variables (denoted by the same symbols), (9) reads

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} - \frac{1}{\mathrm{Re}} \Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \tag{10}$$

where Re is the Reynolds number, Re$= \frac{LU}{\nu}$, L is a characteristic length, and U a characteristic velocity. When $1 \leq \mathrm{Re} \leq 10^5$, the flow is said to be laminar. Finally, when the Reynolds number is small and the force $\boldsymbol{f}$ does not depend on time, the material time derivative can be neglected. Then reverting to the original variables, this yields the Stokes system:

$$-\nu \Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \tag{11}$$

complemented with (5) and (6).

Some Theoretical Results

Let us start with Stokes problems (11), (5), and (6). In view of both theory and numerics, it is useful to write it in variational form. Let

$$H^1(\Omega) = \{v \in L^2(\Omega)\,;\, \nabla\, v \in L^2(\Omega)^3\}\,,$$
$$H_0^1(\Omega) = \{v \in H^1(\Omega)\,;\, v = 0 \text{ on } \partial\Omega\},$$

$$L_\circ^2(\Omega) = \{v \in L^2(\Omega)\,;\, (v, 1) = 0\},$$

where $(\cdot,\cdot)$ denotes the scalar product of $L^2(\Omega)$:

$$(f, g) = \int_\Omega f\, g.$$

Let $H^{-1}(\Omega)$ denote the dual space of $H_0^1(\Omega)$ and $\langle\cdot,\cdot\rangle$ the duality pairing between them. The space H_0^1 takes into account the no-slip boundary condition on the velocity, and the space $L_\circ^2$ is introduced to lift the undetermined constant in the pressure; note that it is only defined by its gradient and hence up to one additive constant in a connected region. Assume that $\boldsymbol{f}$ belongs to $H^{-1}(\Omega)^3$. For our purpose, a suitable variational form is: Find a pair $(\boldsymbol{u}, p) \in H_0^1(\Omega)^3 \times L_\circ^2(\Omega)$ solution of

$$\forall(\boldsymbol{v}, q) \in H_0^1(\Omega)^3 \times L_\circ^2(\Omega)\,,$$
$$\nu(\nabla\,\boldsymbol{u}, \nabla\,\boldsymbol{v}) - (p, \operatorname{div}\boldsymbol{v}) - (q, \operatorname{div}\boldsymbol{u}) = \langle \boldsymbol{f}, \boldsymbol{v}\rangle. \tag{12}$$

Albeit linear, this problem is difficult both from theoretical and numerical standpoints. The pressure can be eliminated from (12) by working with the space V of divergence-free velocities:

$$V = \{\boldsymbol{v} \in H_0^1(\Omega)^3\,;\, \operatorname{div}\boldsymbol{v} = 0\},$$

but the difficulty lies in recovering the pressure. Existence and continuity of the pressure stem from the following deep result: *The divergence operator is an isomorphism from* $V^\perp$ *onto* $L_\circ^2(\Omega)$, where $V^\perp$ is the orthogonal of V in $H_0^1(\Omega)^3$. In other words, for every $q \in L_\circ^2(\Omega)$, there exists one and only one $\boldsymbol{v} \in V^\perp$ solution of $\operatorname{div}\boldsymbol{v} = q$. Moreover $\boldsymbol{v}$ depends continuously on q:

$$\|\nabla\,\boldsymbol{v}\|_{L^2(\Omega)} \le \frac{1}{\beta}\|q\|_{L^2(\Omega)}, \tag{13}$$

where $\beta > 0$, *only depends* on Ω. This inequality is equivalent to the following "inf-sup condition":

$$\inf_{q\in L_\circ^2(\Omega)}\ \sup_{\boldsymbol{v}\in H_0^1(\Omega)^3} \frac{(\operatorname{div}\boldsymbol{v}, q)}{\|\nabla\,\boldsymbol{v}\|_{L^2(\Omega)}\|q\|_{L^2(\Omega)}} \ge \beta. \tag{14}$$

Interestingly, (14) is not true when $\partial\Omega$ has an outward cusp, a situation that occurs, for instance, in a flow exterior to two colliding spheres. There is no simple proof of (14). Its difficulty lies in the no-slip boundary condition prescribed on $\boldsymbol{v}$: The proof is much simpler when it is replaced by the weaker condition $\boldsymbol{v}\cdot\boldsymbol{n} = 0$. The above isomorphism easily leads to the following result:

Theorem 1 *For any* $\boldsymbol{f}$ *in* $H^{-1}(\Omega)^3$ *and any* $\nu > 0$, *Problem* (12) *has exactly one solution and this solution depends continuously on the data:*

$$\|\nabla\,\boldsymbol{u}\|_{L^2(\Omega)} \le \frac{1}{\nu}\|\boldsymbol{f}\|_{H^{-1}(\Omega)},\ \|p\|_{L^2(\Omega)} \le \frac{1}{\beta}\|\boldsymbol{f}\|_{H^{-1}(\Omega)}. \tag{15}$$

Next we consider the steady Navier-Stokes system. The natural extension of (12) is: Find a pair $(\boldsymbol{u}, p) \in H_0^1(\Omega)^3 \times L_\circ^2(\Omega)$ solution of

$$\forall(\boldsymbol{v}, q) \in H_0^1(\Omega)^3 \times L_\circ^2(\Omega)\,,$$
$$\nu(\nabla\,\boldsymbol{u}, \nabla\,\boldsymbol{v}) + (\boldsymbol{u}\cdot\nabla\,\boldsymbol{u}, \boldsymbol{v})$$
$$- (p, \operatorname{div}\boldsymbol{v}) - (q, \operatorname{div}\boldsymbol{u}) = \langle \boldsymbol{f}, \boldsymbol{v}\rangle. \tag{16}$$

Its analysis is fairly simple because on one hand the nonlinear convection term $\boldsymbol{u}\cdot\nabla\,\boldsymbol{u}$ has the following antisymmetry:

$$\forall \boldsymbol{u} \in V, \forall \boldsymbol{v} \in H^1(\Omega)^3\,, (\boldsymbol{u}\cdot\nabla\,\boldsymbol{u}, \boldsymbol{v}) = -(\boldsymbol{u}\cdot\nabla\,\boldsymbol{v}, \boldsymbol{u}), \tag{17}$$

and on the other hand, it belongs to $L^{3/2}(\Omega)^3$, which, roughly speaking, is significantly smoother than the data $\boldsymbol{f}$ in $H^{-1}(\Omega)^3$. This enables to prove existence of solutions, but uniqueness is only guaranteed for small force or large viscosity. More precisely, let

$$\mathcal{N} = \sup_{\boldsymbol{w},\boldsymbol{u},\boldsymbol{v}\in V, \boldsymbol{w},\boldsymbol{u},\boldsymbol{v}\neq \boldsymbol{0}} \frac{(\boldsymbol{w}\cdot\nabla\,\boldsymbol{u}, \boldsymbol{v})}{\|\nabla\,\boldsymbol{w}\|_{L^2(\Omega)}\|\nabla\,\boldsymbol{u}\|_{L^2(\Omega)}\|\nabla\,\boldsymbol{v}\|_{L^2(\Omega)}}.$$

Then we have the next result.

Theorem 2 *For any* $\boldsymbol{f}$ *in* $H^{-1}(\Omega)^3$ *and any* $\nu > 0$*, Problem* (16) *has at least one solution. A sufficient condition for uniqueness is*

$$\frac{\mathcal{N}}{\nu^2}\|\boldsymbol{f}\|_{H^{-1}(\Omega)} < 1. \tag{18}$$

Now we turn to the time-dependent Navier-Stokes system. Its analysis is much more complex because in $\mathbb{R}^3$ the dependence of the pressure on time holds in a weaker space. To simplify, we do not treat the most general situation. For a given time interval $[0, T]$, Banach space X, and number $r \geq 1$, the relevant spaces are of the form $L^r(0, T; X)$, which is the space of functions defined and measurable in $]0, T[$, such that

$$\int_0^T \|v\|_X^r dt < \infty,$$

and

$$W^{1,r}(0, T; X) = \{v \in L^r(0, T; X); \frac{dv}{dt} \in L^r(0, T; X)\},$$

$$W_0^{1,r}(0, T; X) = \{v \in W^{1,r}(0, T; X); v(0) = v(T) = 0\},$$

with dual space $W^{-1,r'}(0, T; X)$, $\frac{1}{r} + \frac{1}{r'} = 1$. There are several weak formulations expressing (4)–(7). For numerical purposes, we shall use the following one: Find $\boldsymbol{u} \in L^2(0, T; V) \cap L^\infty(0, T; L^2(\Omega)^3)$, with $\frac{d\boldsymbol{u}}{dt}$ in $L^{3/2}(0, T; V')$, and $p \in W^{-1,\infty}(0, T; L^2_\circ(\Omega))$ satisfying a.e. in $]0, T[$

$$\forall(\boldsymbol{v}, q) \in H_0^1(\Omega)^3 \times L^2_\circ(\Omega),$$

$$\frac{d}{dt}(\boldsymbol{u}(t), \boldsymbol{v}) + \nu(\nabla \boldsymbol{u}(t), \nabla \boldsymbol{v}) + (\boldsymbol{u}(t) \cdot \nabla \boldsymbol{u}(t), \boldsymbol{v}) - (p(t), \operatorname{div} \boldsymbol{v}) - (q, \operatorname{div} \boldsymbol{u}(t)) = \langle \boldsymbol{f}(t), \boldsymbol{v}\rangle, \tag{19}$$

with the initial condition (7). This problem always has at least one solution.

Theorem 3 *For any* $\boldsymbol{f}$ *in* $L^2(0, T; H^{-1}(\Omega)^3)$*, any* $\nu > 0$*, and any initial data* $\boldsymbol{u}_0 \in V$*, Problem* (19)*,* (7) *has at least one solution.*

Unfortunately, unconditional uniqueness (which is true in $\mathbb{R}^2$) is to this date an open problem in $\mathbb{R}^3$. In fact, it is one of the Millennium Prize Problems.

Discretization

Solving numerically a steady Stokes system is costly because the theoretical difficulty brought by the pressure is inherited both by its discretization, whatever the scheme, and by the computer implementation of its scheme. This computational difficulty is aggravated by the need of fine meshes for capturing complex flows produced by the Navier-Stokes system. In comparison, when the flow is laminar, at reasonable Reynolds numbers, the additional cost of the nonlinear convection term is minor. There are some satisfactory schemes and algorithms but so far no "miracle" method.

Three important methods are used for discretizing flow problems: Finite-element, finite-difference, or finite-volume methods. For the sake of simplicity, we shall mainly consider discretization by finite-element methods. Usually, they consist in using polynomial functions on cells: triangles or quadrilaterals in $\mathbb{R}^2$ or tetrahedra or hexahedra in $\mathbb{R}^3$. Most finite-difference schemes can be derived from finite-element methods on rectangles in $\mathbb{R}^2$ or rectangular boxes in $\mathbb{R}^3$, coupled with quadrature formulas, in which case the mesh may not fit the boundary and a particular treatment may be required near the boundary. Finite volumes are closely related to finite differences but are more complex because they can be defined on very general cells and do not involve functions. All three methods require meshing of the domain, and the success of these methods depends not only on their accuracy but also on how well the mesh is adapted to the problem under consideration. For example, boundary layers may appear at large Reynolds numbers and require locally refined meshes. Constructing a "good" mesh can be difficult and costly, but these important meshing issues are outside the scope of this work. Last, but not least, in many practical applications where the Stokes system is coupled with other equations, it is important that the scheme be locally mass conservative, i.e., the integral mean value of the velocity's divergence be zero in each cell.

Discretization of the Stokes Problem

Let $\mathcal{T}_h$ be a triangulation of $\overline{\Omega}$ made of tetrahedra (also called elements) in $\mathbb{R}^3$, the discretization parameter h being the maximum diameter of the elements. For approximation purposes, $\mathcal{T}_h$ is not completely arbitrary:

it is assumed to be shape regular in the sense that its dihedral angles are uniformly bounded away from 0 and π, and it has no hanging node in the sense that the intersection of two cells is either empty, or a vertex, or a complete edge, or a complete face. For a given integer $k \geq 0$, let $\mathbb{P}_k$ denote the space of polynomials in three variables of *total* degree k and $\mathbb{Q}_k$ that of degree k in *each* variable. The accuracy of a finite-element space depends on the degree of the polynomials used in each cell; however, we shall concentrate on low-degree elements, as these are most frequently used.

Let us start with locally mass conservative methods and consider first conforming finite-element methods, i.e., where the finite-element space of discrete velocities, say X_h, is contained in $H_0^1(\Omega)^3$. Strictly speaking, the space of discrete pressures should be contained in $L_\circ^2(\Omega)$. However, the zero mean-value constraint destroys the band structure of the matrix, and therefore, this constraint is prescribed weakly by means of a small, consistent perturbation. Thus the space of discrete pressures, say Q_h, is simply a discrete subspace of $L^2(\Omega)$, and problem (12) is discretized by: Find $(\boldsymbol{u}_h, p_h) \in X_h \times Q_h$ solution of

$$\forall (\boldsymbol{v}_h, q_h) \in X_h \times Q_h\,,$$
$$\nu(\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h) - (p_h, \operatorname{div} \boldsymbol{v}_h) - (q_h, \operatorname{div} \boldsymbol{u}_h) - \varepsilon(p_h, q_h)$$
$$= \langle \boldsymbol{f}, \boldsymbol{v}_h \rangle, \tag{20}$$

where $\varepsilon > 0$ is a small parameter. Let $M_h = Q_h \cap L_\circ^2(\Omega)$. Regardless of their individual accuracy, X_h and M_h cannot be chosen independently of each other because they must satisfy a uniform discrete analogue of (14), namely,

$$\inf_{q_h \in M_h} \sup_{\boldsymbol{v}_h \in X_h} \frac{(\operatorname{div} \boldsymbol{v}_h, q_h)}{\|\nabla \boldsymbol{v}_h\|_{L^2(\Omega)} \|q_h\|_{L^2(\Omega)}} \geq \beta^*, \tag{21}$$

for some real number $\beta^* > 0$, independent of h. Elements that satisfy (21) are called inf-sup stable. For such elements, the accuracy of (20) depends directly on the individual approximation properties of X_h and Q_h.

Roughly speaking, (21) holds when a discrete velocity space is sufficiently rich compared to a given discrete pressure space. Observe also that the discrete velocity's degree in each cell must be at least one in order to guarantee continuity at the interfaces of elements.

We begin with constant pressures with one degree of freedom at the center of each tetrahedron. It can be checked that, except on some very particular meshes, a conforming $\mathbb{P}_1$ velocity space is not sufficiently rich to satisfy (21). This can be remedied by adding one degree of freedom (a vector in the normal direction) at the center of each face, and it is achieved by enriching the velocity space with one polynomial of $\mathbb{P}_3$ per face. This element, introduced by Bernardi and Raugel, is inf-sup stable and is of order one. It is frugal in number of degrees of freedom and is locally mass conservative but complex in its implementation because the velocity components are not independent. Of course, three polynomials of $\mathbb{P}_3$ (one per component) can be used on each face, and thus each component of the discrete velocity is the sum of a polynomial of $\mathbb{P}_1$, which guarantees accuracy, and a polynomial of $\mathbb{P}_3$, which guarantees inf-sup stability, but the element is more expensive.

The idea of degrees of freedom on faces motivates a nonconforming method where X_h is contained in $L^2(\Omega)^3$ and problem (12) is discretized by the following: Find $(\boldsymbol{u}_h, p_h) \in X_h \times Q_h$ solution of

$$\forall (\boldsymbol{v}_h, q_h) \in X_h \times Q_h\,,$$
$$\nu \sum_{T \in \mathcal{T}_h} (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h)_T - \sum_{T \in \mathcal{T}_h} (p_h, \operatorname{div} \boldsymbol{v}_h)_T$$
$$- \sum_{T \in \mathcal{T}_h} (q_h, \operatorname{div} \boldsymbol{u}_h)_T - \varepsilon(p_h, q_h) = \langle \boldsymbol{f}, \boldsymbol{v}_h \rangle. \tag{22}$$

The inf-sup condition (21) is replaced by:

$$\inf_{q_h \in M_h} \sup_{\boldsymbol{v}_h \in X_h} \frac{\sum_{T \in \mathcal{T}_h} (\operatorname{div} \boldsymbol{v}_h, q_h)_T}{\|\nabla \boldsymbol{v}_h\|_h \|q_h\|_{L^2(\Omega)}} \geq \beta^* \quad \text{where}$$
$$\|\cdot\|_h = \Big(\sum_{T \in \mathcal{T}_h} \|\cdot\|_{L^2(\Omega)}^2 \Big)^{1/2}. \tag{23}$$

The simplest example, introduced by Crouzeix and Raviart, is that of a constant pressure and each velocity's component $\mathbb{P}_1$ per tetrahedron and each velocity's component having one degree of freedom at the center of each face. The functions of X_h must be continuous at the center of each interior face and vanish at the center of each boundary face. Then it is not hard to prove that (23) is satisfied. Thus this element has order one, it is fairly economical and mass conservative, and its implementation is fairly straightforward.

The above methods easily extend to hexahedral triangulations with Cartesian structure (i.e., eight hexahedra meeting at any interior vertex) provided the polynomial space $\mathbb{P}_k$ is replaced by the inverse image of $\mathbb{Q}_k$ on the reference cube. Furthermore, such hexahedral triangulations offer more possibilities. For instance, a conforming, inf-sup stable, locally mass conservative scheme of order two can be obtained by taking, in each cell, a $\mathbb{P}_1$ pressure and each component of the velocity in $\mathbb{Q}_2$.

Now we turn to conforming methods that use continuous discrete pressures; thus the pressure must be at least $\mathbb{P}_1$ in each cell and continuous at the interfaces. Therefore the resulting schemes are not locally mass conservative. It can be checked that velocities with $\mathbb{P}_1$ components are not sufficiently rich. The simplest alternative, called "mini-element" or "$\mathbb{P}_1$–bubble," enriches each velocity component in each cell with a polynomial of $\mathbb{P}_4$ that vanishes on the cell's boundary, whence the name bubble. This element is inf-sup stable and has order one. Its extension to order two, introduced in $\mathbb{R}^2$ by Hood and Taylor, associates with the same pressure, velocities with components in $\mathbb{P}_2$. It is inf-sup stable and has order two.

Discretization of the Navier-Stokes System

Here we present straightforward discretizations of (19). The simplest one consists in using a linearized backward Euler finite-difference scheme in time. Let $N > 1$ be an integer, $\delta t = T/N$ the corresponding time step, and $t_n = n\delta t$ the discrete times. Starting from a finite-element approximation or interpolation, say $\boldsymbol{u}_h^0$ of $\boldsymbol{u}_0$ satisfying the discrete divergence constraint of (20), we construct a sequence $(\boldsymbol{u}_h^n, p_h^n) \in X_h \times Q_h$ such that for $1 \le n \le N$:

$$\forall(\boldsymbol{v}_h, q_h) \in X_h \times Q_h,$$
$$\frac{1}{\delta t}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) + \nu(\nabla \boldsymbol{u}_h^n, \nabla \boldsymbol{v}_h) + c(\boldsymbol{u}_h^{n-1}; \boldsymbol{u}_h^n, \boldsymbol{v}_h)$$
$$- (p_h^n, \operatorname{div} \boldsymbol{v}_h) - (q_h, \operatorname{div} \boldsymbol{u}_h^n) - \varepsilon(p_h^n, q_h) = \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle, \tag{24}$$

where $\boldsymbol{f}^n$ is an approximation of $\boldsymbol{f}(t_n, \cdot)$ and $c(\boldsymbol{w}_h; \boldsymbol{u}_h, \boldsymbol{v}_h)$ a suitable approximation of the convection term $(\boldsymbol{w} \cdot \nabla \boldsymbol{u}, \boldsymbol{v})$. As (17) does not necessarily extend to the discrete spaces, the preferred choice, from the standpoint of theory, is

$$c(\boldsymbol{w}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) = \frac{1}{2}\Big[(\boldsymbol{w}_h \cdot \nabla \boldsymbol{u}_h, \boldsymbol{v}_h) - (\boldsymbol{w}_h \cdot \nabla \boldsymbol{v}_h, \boldsymbol{u}_h)\Big], \tag{25}$$

because it is both consistent and antisymmetric, which makes the analysis easier. But from the standpoint of numerics, the choice

$$c(\boldsymbol{w}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) = (\boldsymbol{w}_h \cdot \nabla \boldsymbol{u}_h, \boldsymbol{v}_h) \tag{26}$$

is simpler and seems to maintain the same accuracy. Observe that at each step n, (24) is a discrete Stokes system with two or three additional linear terms, according to the choice of form c. In both cases, the matrix of the system is not symmetric, which is a strong disadvantage. This can be remedied by completely time lagging the form c, i.e., replacing it by $(\boldsymbol{u}_h^{n-1} \cdot \nabla \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h)$.

There are cases when none of the above linearizations are satisfactory, and the convection term is approximated by $c(\boldsymbol{u}_h^n; \boldsymbol{u}_h^n, \boldsymbol{v}_h)$ with c defined by (25) or (26). The resulting scheme is nonlinear and must be linearized, for instance, by an inner loop of Newton's iterations. Recall Newton's method for solving the equation $f(x) = 0$ in $\mathbb{R}$: Starting from an initial guess x_0, compute the sequence (x_k) for $k \ge 0$ by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Its generalization is straightforward and at step n, starting from $\boldsymbol{u}_{h,0} = \boldsymbol{u}_h^{n-1}$, the inner loop reads: Find $(\boldsymbol{u}_{h,k+1}, p_{h,k+1}) \in X_h \times Q_h$ solution of

$$\forall(\boldsymbol{v}_h, q_h) \in X_h \times Q_h,\ \frac{1}{\delta t}(\boldsymbol{u}_{h,k+1}, \boldsymbol{v}_h) + \nu(\nabla \boldsymbol{u}_{h,k+1}, \nabla \boldsymbol{v}_h)$$
$$+ c(\boldsymbol{u}_{h,k+1}; \boldsymbol{u}_{h,k}, \boldsymbol{v}_h) + c(\boldsymbol{u}_{h,k}; \boldsymbol{u}_{h,k+1}, \boldsymbol{v}_h)$$
$$- (p_{h,k+1}, \operatorname{div} \boldsymbol{v}_h) - (q_h, \operatorname{div} \boldsymbol{u}_{h,k+1}) - \varepsilon(p_{h,k+1}, q_h)$$
$$= c(\boldsymbol{u}_{h,k}; \boldsymbol{u}_{h,k}, \boldsymbol{v}_h) + \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle + \frac{1}{\delta t}(\boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h). \tag{27}$$

Experience shows that only a few iterations are sufficient to match the discretization error. Once this inner loop converges, we set $\boldsymbol{u}_h^n := \boldsymbol{u}_{h,k+1}$, $p_h^n := p_{h,k+1}$.

An interesting alternative to the above schemes is the characteristics method that uses a discretization of the material time derivative (see (3)):

$$\frac{d}{dt}\boldsymbol{u}(t_n,\boldsymbol{x}) \simeq \frac{1}{\delta t}\big(\boldsymbol{u}_h^n(\boldsymbol{x}) - \boldsymbol{u}_h^{n-1}(\chi^{n-1}(\boldsymbol{x}))\big),$$

where $\chi^{n-1}(\boldsymbol{x})$ gives the position at time t_{n-1} of a particle located at $\boldsymbol{x}$ at time t_n. Its first-order approximation is

$$\chi^{n-1}(\boldsymbol{x}) = \boldsymbol{x} - (\delta t)\boldsymbol{u}_h^{n-1}(\boldsymbol{x}).$$

Thus (24) is replaced by

$$\forall(\boldsymbol{v}_h, q_h) \in X_h \times Q_h,$$
$$\frac{1}{\delta t}\big(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\circ\chi^{n-1}, \boldsymbol{v}_h\big) + \nu(\nabla \boldsymbol{u}_h^n, \nabla \boldsymbol{v}_h)$$
$$-(p_h^n, \operatorname{div}\boldsymbol{v}_h) - (q_h, \operatorname{div}\boldsymbol{u}_h^n) - \varepsilon(p_h^n, q_h) = \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle, \tag{28}$$

whose matrix is symmetric and constant in time and requires no linearization. On the other hand, computing the right-hand side is more complex.

Algorithms

In this section we assume that the discrete spaces are inf-sup stable, and to simplify, we restrict the discussion to conforming discretizations. Any discretization of the time-dependent Navier-Stokes system requires the solution of at least one Stokes problem per time step, whence the importance of an efficient Stokes solver. But since the matrix of the discrete Stokes system is large and indefinite, in $\mathbb{R}^3$ the system is rarely solved simultaneously for $\boldsymbol{u}_h$ and p_h. Instead the computation of p_h is decoupled from that of $\boldsymbol{u}_h$.

Decoupling the Pressure and Velocity

Let $\boldsymbol{U}$ be the vector of velocity unknowns, $\boldsymbol{P}$ that of pressure unknowns, and $\boldsymbol{F}$ the vector of data represented by $(\boldsymbol{f}, \boldsymbol{v}_h)$. Let $\boldsymbol{A}$ be the (symmetric positive definite) matrix of the discrete Laplace operator represented by $\nu(\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h)$, $\boldsymbol{B}$ the matrix of the discrete divergence operator represented by $(q_h, \operatorname{div}\boldsymbol{u}_h)$, and $\boldsymbol{C}$ the matrix of the operator represented by $\varepsilon(p_h, q_h)$. Owing to (21), the matrix $\boldsymbol{B}$ has maximal rank. With this notation, (20) has the form:

$$\boldsymbol{A}\,\boldsymbol{U} - \boldsymbol{B}^T\boldsymbol{P} = \boldsymbol{F} \quad , \quad -\boldsymbol{B}\,\boldsymbol{U} - \boldsymbol{C}\,\boldsymbol{P} = \boldsymbol{0}, \tag{29}$$

whose matrix is symmetric but indeed indefinite. Since $\boldsymbol{A}$ is nonsingular, a partial solution of (29) is

$$(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T + \boldsymbol{C})\boldsymbol{P} = -\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{F},$$
$$\boldsymbol{U} = \boldsymbol{A}^{-1}(\boldsymbol{F} + \boldsymbol{B}^T\boldsymbol{P}). \tag{30}$$

As $\boldsymbol{B}$ has maximal rank, the Schur complement $\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T + \boldsymbol{C}$ is symmetric positive definite, and an iterative gradient algorithm is a good candidate for solving (30). Indeed, (30) is equivalent to minimizing with respect to $\boldsymbol{Q}$ the quadratic functional

$$K(\boldsymbol{Q}) = \frac{1}{2}(\boldsymbol{A}\boldsymbol{v}_Q, \boldsymbol{v}_Q) + \frac{1}{2}(\boldsymbol{C}\,\boldsymbol{Q}, \boldsymbol{Q}), \quad \text{with}$$
$$\boldsymbol{A}\boldsymbol{v}_Q = \boldsymbol{F} + \boldsymbol{B}^T\boldsymbol{Q}. \tag{31}$$

A variety of gradient algorithms for approximating the minimum are obtained by choosing a sequence of direction vectors $\boldsymbol{W}^k$ and an initial vector $\boldsymbol{P}_0$ and computing a sequence of vectors $\boldsymbol{P}_k$ defined for each $k \geq 1$ by:

$$\boldsymbol{P}_k = \boldsymbol{P}_{k-1} - \rho_{k-1}\boldsymbol{W}_{k-1}, \quad \text{where}$$
$$K(\boldsymbol{P}_{k-1} - \rho_{k-1}\boldsymbol{W}_{k-1}) = \inf_{\rho\in\mathbb{R}} K(\boldsymbol{P}_{k-1} - \rho\boldsymbol{W}_{k-1}). \tag{32}$$

Usually the direction vectors $\boldsymbol{W}_k$ are related to the gradient of K, whence the name of gradient algorithms. It can be shown that each step of these gradient algorithms requires the solution of a linear system with matrix $\boldsymbol{A}$, which is equivalent to solving a Laplace equation per step. This explains why solving the Stokes system is expensive.

The above strategy can be applied to (28) but not to (24) because its matrix $\boldsymbol{A}$ is no longer symmetric. In this case, a GMRES algorithm can be used, but this algorithm is expensive. For this reason, linearization by fully time lagging c may be preferable because the matrix $\boldsymbol{A}$ becomes symmetric. Of course, when Newton's iterations are performed, as in (27), this option is not available because $\boldsymbol{A}$ is not symmetric. In this case, a splitting strategy may be useful.

Splitting Algorithms

There is a wide variety of algorithms for splitting the nonlinearity from the divergence constraint. Here is an example where the divergence condition is enforced once every other step. At step n,

1. Knowing $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}) \in X_h \times Q_h$, compute an intermediate velocity $(\boldsymbol{w}_h^n, p_h^n) \in X_h \times Q_h$ solution of

$$\begin{aligned}&\forall(\boldsymbol{v}_h, q_h) \in X_h \times Q_h\,,\\ &\quad\frac{1}{\delta t}(\boldsymbol{w}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) - (p_h^n, \operatorname{div} \boldsymbol{v}_h) - (q_h, \operatorname{div} \boldsymbol{w}_h^n)\\ &\quad - \varepsilon(p_h^n, q_h) = \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle - \nu(\nabla \boldsymbol{u}_h^{n-1}, \nabla \boldsymbol{v}_h)\\ &\quad - c(\boldsymbol{u}_h^{n-1}; \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h). \qquad (33)\end{aligned}$$

2. Compute $\boldsymbol{u}_h^n \in X_h$ solution of

$$\begin{aligned}&\forall \boldsymbol{v}_h \in X_h\,,\ \frac{1}{\delta t}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) + \nu(\nabla \boldsymbol{u}_h^n, \nabla \boldsymbol{v}_h)\\ &\quad + c(\boldsymbol{w}_h^n; \boldsymbol{u}_h^n, \boldsymbol{v}_h) = \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle + (p_h^n, \operatorname{div} \boldsymbol{v}_h). \quad (34)\end{aligned}$$

The first step is fairly easy because it reduces to a "Laplace" operator with unknown boundary conditions and therefore can be preconditioned by a Laplace operator, while the second step is an implicit linearized system without constraint.

Numerical Experiments

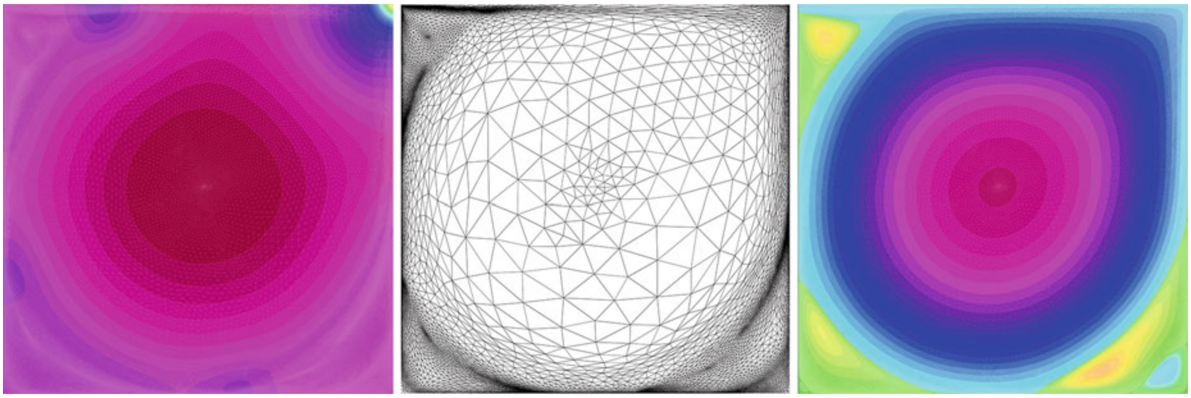
We present here two numerical experiments of benchmarks programmed with the software FreeFem++. More details including scripts and plots can be found online at http://www.ljll.math.upmc.fr/~hecht/ftp/ECM-2013.

The Driven Cavity in a Square

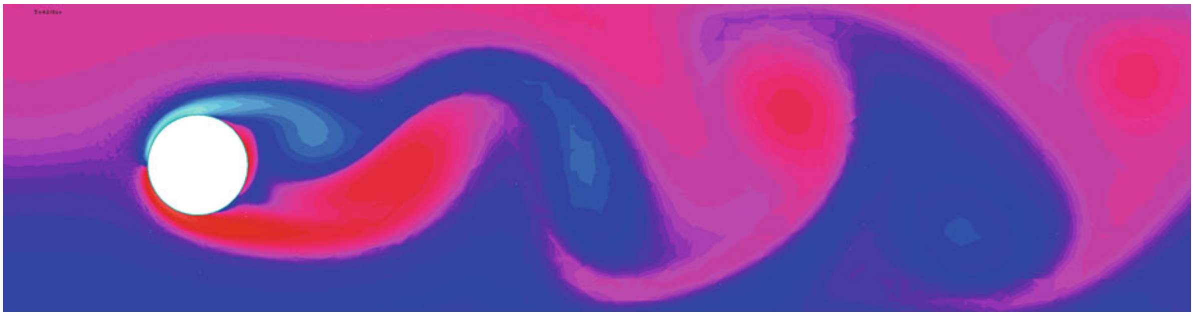
We use the Taylor-Hood $\mathbb{P}_2 - \mathbb{P}_1$ scheme to solve the steady Navier-Stokes equations in the square cavity $\Omega =]0,1[\times]0,1[$ with upper boundary $\Gamma_1 =]0,1[\times\{1\}$:

$$-\frac{1}{\mathrm{Re}}\Delta \boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} + \nabla p = \boldsymbol{0} \quad , \quad \operatorname{div}\boldsymbol{u} = 0,$$

$$\boldsymbol{u}|_{\Gamma_1} = (1,0) \quad , \quad \boldsymbol{u}|_{\partial\Omega\setminus\Gamma_1} = (0,0),$$



Stokes or Navier-Stokes Flows, Fig. 1 From left to right: pressure at Re 9,000, adapted mesh at Re 8,000, stream function at Re 9,000. Observe the cascade of corner eddies



Stokes or Navier-Stokes Flows, Fig. 2 Von Kármán's vortex street with different values of Re ranging from 1 to 9,000. The discontinuity of the boundary values at the two upper corners of the cavity produces a singularity of the pressure. The nonlinearity is solved by Newton's method, the initial guess being obtained by continuation on the Reynolds number, i.e., from the solution computed with the previous Reynolds number. The address of the script is cavityNewton.edp (Fig. 1).

Flow Past an Obstacle: Von Kármán's Vortex Street in a Rectangle

The Taylor-Hood $\mathbb{P}_2 - \mathbb{P}_1$ scheme in space and characteristics method in time are used to solve the time-dependent Navier-Stokes equations in a rectangle $2.2 \times 0.41\,\mathrm{m}$ with a circular hole of diameter $0.1\,\mathrm{m}$ located near the inlet. The density $\rho = 1.0\,\frac{\mathrm{Kg}}{\mathrm{m}^3}$ and the kinematic viscosity $\nu = 10^{-3}\,\frac{\mathrm{m}^2}{\mathrm{s}}$. All the relevant data are taken from the benchmark case 2D-2 that can be found online at http://www.mathematik.tu-dortmund.de/lsiii/cms/papers/SchaeferTurek1996.pdf. The address of the script is at http://www.ljll.math.upmc.fr/~hecht/ftp/ECM-2013 is NSCaraCyl-100-mpi.edp or NSCaraCyl-100-seq.edp and func-max.idp (Fig. 2).

Bibilographical Notes

The bibliography on Stokes and Navier-Stokes equations, theory and approximation, is very extensive and we have only selected a few references.

A mechanical derivation of the Navier-Stokes equations can be found in the book by L.D. Landau and E.M. Lifshitz:

Fluid Mechanics, Second Edition, Vol. **6** (Course of Theoretical Physics), Pergamon Press, 1959.

The reader can also refer to the book by C. Truesdell and K.R. Rajagopal:

An Introduction to the Mechanics of Fluids, Modeling and Simulation in Science, Engineering and Technology, Birkhauser, Basel, 2000.

A thorough theory and description of finite element methods can be found in the book by P.G. Ciarlet:

Basic error estimates for elliptic problems - Finite Element Methods, Part 1, in *Handbook of Numerical Analysis, II*, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1991.

The reader can also refer to the book by T. Oden and J.N. Reddy:

An introduction to the mathematical theory of finite elements, Wiley, New-York, 1976.

More computational aspects can be found in the book by A. Ern and J.L. Guermond:

Theory and Practice of Finite Elements, AMS **159**, Springer-Verlag, Berlin, 2004.

The reader will find an introduction to the theory and approximation of the Stokes and steady Navier-Stokes equations, including a thorough discussion on the inf-sup condition, in the book by V. Girault and P.A. Raviart:

Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms, SCM **5**, Springer-Verlag, Berlin, 1986.

An introduction to the theory and approximation of the time-dependent Navier-Stokes problem is treated in the Lecture Notes by V. Girault and P.A. Raviart:

Finite Element Approximation of the Navier-Stokes Equations, Lect. Notes in Math. **749**, Springer-Verlag, Berlin, 1979.

Non-conforming finite elements can be found in the reference by M. Crouzeix and P.A. Raviart:

Conforming and non-conforming finite element methods for solving the stationary Stokes problem, RAIRO Anal. Numér. **8** (1973), pp. 33–76.

We also refer to the book by R. Temam:

Navier-Stokes Equations, Theory and Numerical Analysis, North-Holland, Amsterdam, 1979.

The famous Millennium Prize Problem is described at the URL:

http://www.claymath.org/millennium/Navier-Stokes Equations.

The reader will find a wide range of numerical methods for fluids in the book by O. Pironneau:

Finite Element Methods for Fluids, Wiley, 1989. See also http://www.ljll.math.upmc.fr/~pironneau.

We also refer to the course by R. Rannacher available online at the URL:

http://numerik.iwr.uni-heidelberg.de/Oberwolfach-Seminar/CFD-Course.pdf.

The book by R. Glowinski proposes a vey extensive collection of numerical methods, algorithms, and experiments for Stokes and Navier-Stokes equations:

Finite Element Methods for Incompressible Viscous Flow, in *Handbook of numerical analysis, IX*, P.G. Ciarlet and J.L .Lions, eds., North-Holland, Amsterdam, 2003.

Stratosphere and Its Coupling to the Troposphere and Beyond

Edwin P. Gerber
Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Synonyms

Middle atmosphere

Glossary

Mesosphere an atmospheric layer between approximately 50 and 100 km height

Middle atmosphere a region of the atmosphere including the stratosphere and mesosphere

Stratosphere an atmospheric layer between approximate 12 and 50 km

Stratospheric polar vortex a strong, circumpolar jet that forms in the extratropical stratosphere during the winter season in each respective hemisphere

Sudden stratospheric warming a rapid break down of the stratospheric polar vortex, accompanied by a sharp warming of the polar stratosphere

Tropopause boundary between the troposphere and stratosphere, generally between 10 to 18 km.

Troposphere lowermost layer of the atmosphere, extending from the surface to between 10 and 18 km.

Climate engineering the deliberate modification of the Earth's climate system, primarily aimed at reducing the impact of global warming caused by anthropogenic greenhouse gas emissions

Geoengineering see climate engineering

Quasi-biennial oscillation an oscillating pattern of easterly and westerly jets which propagates downward in the tropical stratosphere with a slightly varying period around 28 months

Solar radiation management a form of climate engineering where the net incoming solar radiation to the surface is reduced to offset warming caused by greenhouse gases

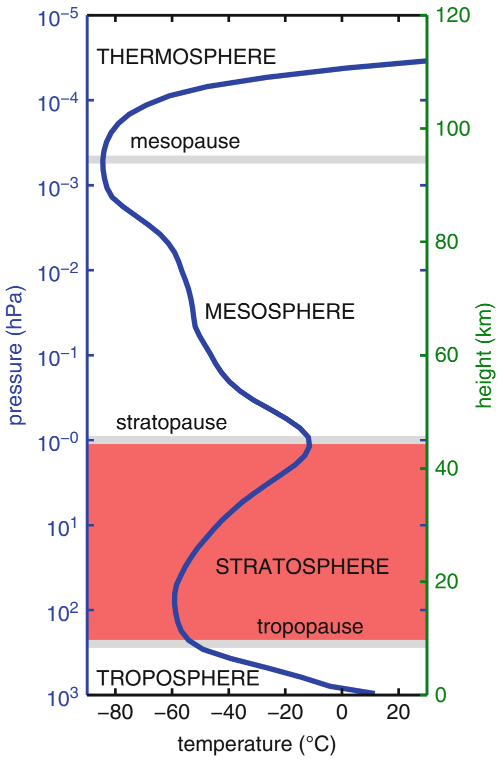
Definition

As illustrated in Fig. 1, the Earth's atmosphere can be separated into distinct regions, or "spheres," based on its vertical temperature structure. In the lowermost part of the atmosphere, the *troposphere*, the temperature declines steeply with height at an average rate of approximately 7 °C per kilometer. At a distinct level, generally between 10–12 km in the extratropics and 16–18 km in the tropics (The separation between these regimes is rather abrupt and can be used as a dynamical indicator delineating the tropics and extratropics.) this steep descent of temperature abruptly shallows, transitioning to a layer of the atmosphere where temperature is initially constant with height, and then begins to rise. This abrupt change in the vertical temperature gradient, denoted the *tropopause*, marks the lower boundary of the *stratosphere*, which extends to approximately 50 km in height, at which point the temperature begins to fall with height again. The region above is denoted the *mesosphere*, extending to a second temperature minimum between 85 and 100 km. Together, the stratosphere and mesosphere constitute the *middle atmosphere*.

The stratosphere was discovered at the dawn of the twentieth century. The vertical temperature gradient, or *lapse rate*, of the troposphere was established in the eighteenth century from temperature and pressure measurements taken on alpine treks, leading to speculation that the air temperature would approach absolute zero somewhere between 30 and 40 km: presumably the top of the atmosphere. Daring hot air balloon ascents in the late nineteenth century provided hints at a shallowing of the lapse rate – early evidence of the tropopause – but also led to the deaths of aspiring upper-atmosphere meteorologists. Teisserenc de Bort [15] and Assmann [2], working outside of Paris and Berlin, respectively, pioneered the first systematic, unmanned balloon observations of the upper atmosphere, establishing the distinct changes in the temperature structure that mark the stratosphere.

Overview

The lapse rate of the atmosphere reflects the stability of the atmosphere to vertical motion. In the troposphere, the steep decline in temperature reflects near-neutral stability to moist convection. This is the turbulent



Stratosphere and Its Coupling to the Troposphere and Beyond, Fig. 1 The vertical temperature structure of the atmosphere. This sample profile shows the January zonal mean temperature at 40Å N from the Committee on Space Research (COSPAR) International Reference Atmosphere 1986 model (CIRA-86). The changes in temperature gradients and hence stratification of the atmosphere reflect a difference in the dynamical and radiative processes active in each layer. The heights of the separation points (tropopause, stratopause, and mesopause) vary with latitude and season – and even on daily time scales due to dynamical variability – but are generally sharply defined in any given temperature profile

weather layer of the atmosphere where air is in close contact with the surface, with a turnover time scale on the order of days. In the stratosphere, the near-zero or positive lapse rates strongly stratify the flow. Here the air is comparatively isolated from the surface of the Earth, with a typical turnover time scale on the order of a year or more. This distinction in stratification and resulting impact on the circulation are reflected in the nomenclature: the “troposphere” and “stratosphere” were coined by Teisserenc de Bort, the former the “sphere of change” from the Greek *tropos*, to turn or whirl while the latter the “sphere of layers” from the Latin *stratus*, to spread out.

In this sense, the troposphere can be thought of as a boundary layer in the atmosphere that is well connected to the surface. This said, it is important to note that the mass of the atmosphere is proportional to the pressure: the tropospheric “boundary” layer constitutes roughly 85 % of the mass of the atmosphere and contains all of our weather. The stratosphere contains the vast majority of the remaining atmospheric mass and the mesosphere and layers above just 0.1 %.

Why Is There a Stratosphere?

The existence of the stratosphere depends on the radiative forcing of the atmosphere by the Sun. As the atmosphere is largely transparent to incoming solar radiation, the bulk of the energy is absorbed at the surface. The presence of greenhouse gases, which absorb infrared light, allows the atmosphere to interact with radiation emitted by the surface. If the atmosphere were “fixed,” and so unable to convect (described as a *radiative equilibrium*), this would lead to an unstable situation, where the air near the surface is much warmer – and so more buoyant – than that above it. At height, however, temperature eventually becomes isothermal, given a fairly uniform distribution of the infrared absorber throughout the atmosphere (The simplest model for this is the so-called gray radiation scheme, where one assumes that all solar radiation is absorbed at the surface and a single infrared band from the Earth interacts with a uniformly distributed greenhouse gas.).

If we allow the atmosphere to turn over in the vertical or convect in the nomenclature of atmospheric science, the circulation will produce to a well-mixed layer at the bottom with near-neutral stability: the troposphere. The energy available to the air at the surface is finite, however, only allowing it to penetrate so high into the atmosphere. Above the convection will sit the stratified isothermal layer that is closer to the radiative equilibrium: the stratosphere. This simplified view of a *radiative-convective equilibrium* obscures the role of dynamics in setting the stratification in both the troposphere and stratosphere but conveys the essential distinction between the layers. In this respect, “stratospheres” are found on other planets as well, marking the region where the atmosphere becomes more isolated from the surface.

The increase in temperature seen in the Earth’s stratosphere (as seen in Fig. 1) is due to the fact that the atmosphere does interact with the incoming solar radiation through ozone. Ozone is produced by the interaction between molecular oxygen and ultraviolet radiation in the stratosphere [4] and takes over as the dominant absorber of radiation in this band. The decrease in density with height leads to an optimal level for net ultraviolet warming and hence the temperature maximum near the *stratopause*, which provides the demarcation for the mesosphere above.

Absorption of ultraviolet radiation by stratospheric ozone protects the surface from high-energy radiation. The destruction of ozone over Antarctica by halogenated compounds has had significant health impacts, in addition to damaging all life in the biosphere. As described below, it has also had significant impacts on the tropospheric circulation in the Southern Hemisphere.

Compositional Differences

The separation in the turnover time scale between the troposphere and stratosphere leads to distinct chemical or compositional properties of air in these two regions. Indeed, given a sample of air randomly taken from some point in the atmosphere, one can easily tell whether it came from the troposphere or the stratosphere. The troposphere is rich in water vapor and reactive organic molecules, such as carbon monoxide, which are generated by the biosphere and anthropogenic activity. Stratospheric air is extremely dry, with an average water vapor concentration of approximate 3–5 parts per billion, and comparatively rich in ozone. Ozone is a highly reactive molecule (causing lung damage when it is formed in smog at the surface) and does not exist for long in the troposphere.

Scope and Limitations of this Entry

Stratospheric research, albeit only a small part of the Earth system science, is a fairly mature field covering a wide range of topics. The remaining goal of this brief entry is to highlight the dynamical interaction between the stratosphere and the troposphere, with particular emphasis on the impact of the stratosphere on surface climate. In the interest of brevity, references have been kept to a minimum, focusing primarily on seminal historical papers and reviews. More detailed references can be found in the review articles listed in further readings.

The stratosphere also interacts with the troposphere through the exchange of mass and trace chemical species, such as ozone. This exchange is critical for understanding the atmospheric chemistry in both the troposphere and stratosphere and has significant implications for tropospheric air quality, but will not be discussed. For further information, please see two review articles, [8] and [11]. The primary entry point for air into the stratosphere is through the tropics, where the boundary between the troposphere and stratosphere is less well defined. This region is known as the *tropical tropopause layer* and a review by [6] will provide the reader an introduction to research on this topic.

Dynamical Coupling Between the Stratosphere and Troposphere

The term "coupling" suggests interactions between independent components and so begs the question as to whether the convenient separation of the atmosphere into layers is merited in the first place. The key dynamical distinction between the troposphere and stratosphere lies in the differences in their stratification and the fact that moist processes (i.e., moist convection and latent heat transport) are restricted to the troposphere. The separation between the layers is partly historical, however, evolving in response to the development of weather forecasting and the availability of computational resources.

Midlatitude weather systems are associated with large-scale Rossby waves, which owe their existence to gradients in the effective rotation, or vertical component of vorticity, of the atmosphere due to variations in the angle between the surface plain and the axis of rotation with latitude. Pioneering work by [5] and [10] showed that the dominant energy containing waves in the troposphere, wavenumber roughly 4–8, the so-called *synoptic scales*, cannot effectively propagate into the stratosphere due the presence of easterly winds in the summer hemisphere and strong westerly winds in the winter hemisphere. For the purposes of weather prediction, then, the stratosphere could largely be viewed as an upper-boundary condition. Models thus resolved the stratosphere as parsimoniously as possible in order to focus numerical resources on the troposphere. The strong winds in the winter stratosphere also impose a stricter Courant-Friedrichs-Lewy condition on the time step of the model, although more advanced numerical techniques have alleviated this problem.

Despite the dynamical separation for weather system-scale waves, larger-scale Rossby waves (wavenumber 1–3, referred to as planetary scales) can penetrate into the winter stratosphere, allowing for momentum exchange between the layers. In addition, smaller-scale (on the order of 10–1,000 km) gravity waves (Gravity waves are generated in stratified fluids, where the restoring force is the gravitational acceleration of fluid parcels or buoyancy. They are completely distinct from relativistic gravity waves.) also transport momentum between the layers. As computation power increased, leading to a more accurate representation of tropospheric dynamics, it became increasingly clear that a better representation of the stratosphere was necessary to fully understand and simulate surface weather and climate.

Coupling on Daily to Intraseasonal Time Scales

Weather prediction centers have found that the increased representation of the stratosphere improves tropospheric forecasts. On short time scales, however, much of the gain comes from improvements to the tropospheric initial condition. This stems from better assimilation of satellite temperature measurements which project onto both the troposphere and stratosphere.

The stratosphere itself has a more prominent impact on intraseasonal time scales, due to the intrinsically longer time scales of variability in this region of the atmosphere. The gain in predictability, however, is conditional, depending on the state of the stratosphere. Under normal conditions, the winter stratosphere is very cold in the polar regions, associated with a strong westerly jet, or *stratospheric polar vortex*. As first observed in the 1950s [12], this strong vortex is sometimes disturbed by the planetary wave activity propagating below, leading to massive changes in temperature (up to 70 °C in a matter of days) and a reversal of the westerly jet, a phenomenon known as a *sudden stratospheric warming*, or SSW. While the predictability of SSWs are limited by the chaotic nature of tropospheric dynamics, after an SSW the stratosphere remains in an altered state for up to 2–3 months as the polar vortex slowly recovers from the top down.

Baldwin and Dunkerton [3] demonstrated the impact of these changes on the troposphere, showing that an abrupt warming of the stratosphere is followed by an equatorward shift in the tropospheric jet stream and associated storm track. An abnormally cold stratosphere is conversely associated with a poleward shift in the jet stream, although the onset of cold vortex events is not as abrupt. More significantly, the changes in the troposphere extend for up to 2–3 months on the slow time scale of the stratospheric recovery, while under normal conditions the chaotic nature of tropospheric flow restricts the time scale of jet variations to approximately 10 days. The associated changes in the stratospheric jet stream and tropospheric jet shift are conveniently described by the *Northern Annular Mode* (The NAM is also known as the *Arctic Oscillation*, although the annular mode nomenclature has become more prominent.) (NAM) pattern of variability.

The mechanism behind this interaction is still an active area of research. It has become clear, however, that key lies in the fact that the lower stratosphere influences the formation and dissipation of synoptic-scale Rossby waves, despite the fact that these waves do not penetrate far into the stratosphere.

A shift in the jet stream is associated with a large-scale rearrangement of tropospheric weather patterns. In the Northern Hemisphere, where the stratosphere is more variable due to the stronger planetary wave activity (in short, because there are more continents), an equatorward shift in the jet stream following an SSW leads to colder, stormier weather over much of northern Europe and eastern North America. Forecast skill of temperature, precipitation, and wind anomalies at the surface increases in seasonal forecasts following an SSW. SSWs can be further differentiated into "vortex displacements" and "vortex splits," depending on the dominant wavenumber (1 or 2, respectively) involved in the breakdown of the jet, and recent work has suggested this has an effect on the tropospheric impact of the warming.

SSWs occur approximately every other year in the Northern Hemisphere, although there is strong intermittency: few events were observed in the 1990s, while they have been occurring in most years in the first decades of the twenty-first century. In the Southern Hemisphere, the winter westerlies are stronger and less variable – only one SSW has ever been observed, in 2002 – but predictability may be gained around the time of the "final warming," when the stratosphere transitions to it's summer state with easterly winds. Some years, this transition is accelerated by planetary wave dynamics, as in an SSW, while in other years it is gradual, associated with a slow radiative relaxation to the summer state.

Coupling on Interannual Time Scales

On longer time scales, the impact of the stratosphere is often felt through a modulation of the intraseasonal coupling between the stratospheric and tropospheric jet streams. Stratospheric dynamics play an important role in internal modes of variability to the atmosphere-ocean system, such as El Niño and the Southern Oscillation (ENSO), and in the response of the climate system to "natural" forcing by the solar cycle and volcanic eruptions.

The quasi-biennial oscillation (QBO) is a nearly periodic oscillation of downward propagating easterly and westerly tropical jets in the tropical stratosphere, with a period of approximately 28 months. It is perhaps the most long-lived mode of variability intrinsic to the atmosphere alone. The QBO influences the surface by modulating the wave coupling between the troposphere and stratosphere in the Northern Hemisphere winter, altering the frequency and intensity of SSWs depending on the phase of the oscillation.

Isolating the impact of the QBO has been complicated by the possible overlap with the ENSO, a coupled mode of atmosphere-ocean variability with a time scale of approximately 3–7 years. The relatively short observational record makes it difficult to untangle the signals from measurements alone, and models have only recently been able to simulate these phenomenon with reasonable accuracy. ENSO is driven by interaction between the tropical Pacific Ocean and the zonal circulation of the tropical atmosphere (the Walker circulation). Its impact on the extratropical circulation in the Northern Hemisphere, however, is in part effected through its influence on the stratospheric polar vortex. A warm phase of ENSO is associated with stronger planetary wave propagation into the stratosphere, hence a weaker polar vortex and equatorward shift in the tropospheric jet stream.

Further complicating the statistical separation between the impacts of ENSO and the QBO is the influence of the 11-year solar cycle, associated with changes in the number of sunspots. While the overall intensity of solar radiation varies less than 0.1 % of its mean value over the cycle, the variation is stronger in the ultraviolet part of the spectrum. Ultraviolet radiation is primarily absorbed by ozone in the stratosphere, and it has been suggested that the associated changes in temperature structure alter the planetary wave propagation, along the lines of the influence of ENSO and QBO.

The role of the stratosphere in the climate response to volcanic eruptions is comparatively better understood. While volcanic aerosols are washed out of the troposphere on fairly short time scales by the hydrological cycle, sulfate particles in the stratosphere can last for 1–2 years. These particles reflect the incoming solar radiation, leading to a global cooling of the surface; following Pinatubo, the global surface cooled to 0.1–0.2 K. The overturning circulation of the stratosphere lifts mass up into the tropical stratosphere, transporting it poleward where it descends in the extratropics. Thus, only tropical eruptions have a persistent, global impact.

Sulfate aerosols warm the stratosphere, therefore modifying the planetary wave coupling. There is some evidence that the net result is a strengthening of the polar vortex which in turn drives a poleward shift in the tropospheric jets. Hence, eastern North America and northern Europe may experience warmer winters following eruptions, despite the overall cooling impact of the volcano.

The Stratosphere and Climate Change

Anthropogenic forcing has changed the stratosphere, with resulting impacts on the surface. While greenhouse gases warm the troposphere, they increase the radiative efficiency of the stratosphere, leading to a net cooling in this part of the atmosphere. The combination of a warming troposphere and cooling stratosphere leads to a rise in the tropopause and may be one of the most identifiable signatures of global warming on the atmospheric circulation.

While greenhouse gases will have a dominant long-term impact on the climate system, anthropogenic emissions of halogenated compounds, such as chlorofluorocarbons (CFCs), have had the strongest impact on the stratosphere in recent decades. Halogens have caused some destruction of ozone throughout the stratosphere, but the extremely cold temperatures of the Antarctic stratosphere in winter permit the formation of *polar stratospheric clouds*, which greatly accelerate the production of Cl and Br atoms that catalyze ozone destruction (e.g., [14]). This led to the ozone hole, the effective destruction of all ozone throughout the middle and lower stratosphere over Antarctica. The effect of ozone loss on ultraviolet radiation was quickly appreciated, and the use of halogenated compounds regulated and phased out under the Montreal Protocol (which came into force in 1989) and subsequent agreements. Chemistry climate models suggest that the ozone hole should recover by the end of this century, assuming the ban on halogenated compounds is observed.

It was not appreciated until the first decade of the twenty-first century, however, that the ozone hole also has impacted the circulation of the Southern Hemisphere. The loss of ozone leads to a cooling of the austral polar vortex in springtime and a subsequent poleward shift in the tropospheric jet stream. Note that this poleward shift in the tropospheric jet in response to a stronger stratospheric vortex mirrors the coupling associated with natural variability in the Northern Hemisphere. As reviewed by [16], this shift in the jet stream has had significant impacts on precipitation across much of the Southern Hemisphere.

Stratospheric trends in water vapor also have the potential to affect the surface climate. Despite the minuscule concentration of water vapor in the stratosphere (just 3–5 parts per billion), the radiative impact of a greenhouse gases scales logarithmically, so relatively large changes in small concentrations can have a strong impact. Decadal variations in stratospheric water vapor can have an influence on surface climate comparable to decadal changes in greenhouse gas forcing, and there is evidence of a positive feedback of stratospheric water vapor on greenhouse gas forcing.

The stratosphere has also been featured prominently in the discussion of *climate engineering* (or *geoengineering*), the deliberate alteration of the Earth system to offset the consequences of greenhouse-induced warming. Inspired by the natural cooling impact of volcanic aerosols, the idea is to inject hydrogen sulfide or sulfur dioxide into the stratosphere, where it will form sulfate aerosols. To date, this strategy of the so-called *solar radiation management* appears to be among the most feasible and cost-effective means of cooling the Earth's surface, but it comes with many dangers. In particular, it does not alleviate ocean acidification, and the effect is short-lived – a maximum of two years – and so would require continual action ad infinitum or until greenhouse gas concentrations were returned to safer levels. (In saying this, it is important to note that the natural time scale for carbon dioxide removal is 100,000s of years, and there are no known strategies for accelerating CO2 removal that appear feasible, given current technology.) In addition, the impact of sulfate aerosols on stratospheric ozone and the potential regional effects due to changes in the planetary wave coupling with the troposphere are not well understood.

Further Reading

There are a number of review papers on stratosphere-tropospheric coupling in the literature. In particular, [13] provides a comprehensive discussion of stratosphere-troposphere coupling, while [7] highlights developments in the last decade. Andrews et al. [1] provide a classic text on the dynamics of the stratosphere, and [9] provides a wider perspective on the stratosphere, including the history of field.

References

1. Andrews, D.G., Holton, J.R., Leovy, C.B.: Middle Atmosphere Dynamics. Academic Press, Waltham, MA (1987)
2. Assmann, R.A.: über die existenz eines wärmeren Lufttromes in der Höhe von 10 bis 15 km. Sitzungsber K. Preuss. Akad. Wiss. **24**, 495–504 (1902)
3. Baldwin, M.P., Dunkerton, T.J.: Stratospheric harbingers of anomalous weather regimes. Science **294**, 581–584 (2001)
4. Chapman, S.: A theory of upper-atmosphere ozone. Mem. R. Meteor. Soc. **3**, 103–125 (1930)
5. Charney, J.G., Drazin, P.G.: Propagation of planetary-scale disturbances from the lower into the upper atmosphere. J. Geophys. Res. **66**, 83–109 (1961)
6. Fuegistaler, S., Dessler, A.E., Dunkerton, T.J., Folkins, I., Fu, Q., Mote, P.W.: Tropical tropopause layer. Rev. Geophys. **47**, RG1004 (2009)
7. Gerber, E.P., Butler, A., Calvo, N., Charlton-Perez, A., Giorgetta, M., Manzini, E., Perlwitz, J., Polvani, L.M., Sassi, F., Scaife, A.A., Shaw, T.A., Son, S.W., Watanabe, S.: Assessing and understanding the impact of stratospheric dynamics and variability on the Earth system. Bull. Am. Meteor. Soc. **93**, 845–859 (2012)
8. Holton, J.R., Haynes, P.H., McIntyre, M.E., Douglass, A.R., Rood, R.B., Pfister, L.: Stratosphere-troposphere exchange. Rev. Geophys. **33**, 403–439 (1995)
9. Labitzke, K.G., Loon, H.V.: The Stratosphere: Phenomena, History, and Relevance. Springer, Berlin/New York (1999)
10. Matsuno, T.: Vertical propagation of stationary planetary waves in the winter Northern Hemisphere. J. Atmos. Sci. **27**, 871–883 (1970)
11. Plumb, R.A.: Stratospheric transport. J. Meteor. Soc. Jpn. **80**, 793–809 (2002)
12. Scherhag, R.: Die explosionsartige stratosphärenerwarmung des spätwinters 1951/52. Ber Dtsch Wetterd. US Zone **6**, 51–63 (1952)

13. Shepherd, T.G.: Issues in stratosphere-troposphere coupling. J. Meteor. Soc. Jpn, **80**, 769–792 (2002)
14. Solomon, S.: Stratospheric ozone depletion: a review of concepts and history. Rev. Geophys. **37**(3), 275–316 (1999). doi:10.1029/1999RG900008
15. Teisserence de Bort, L.: Variations de la temperature d l'air libre dans la zone comprise entre 8 km et 13 km d'altitude. C. R. Acad. Sci. Paris **138**, 42–45 (1902)
16. Thompson, D.W.J., Solomon, S., Kushner, P.J., England, M.H., Grise, K.M., Karoly, D.J.: Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change. Nature Geoscience **4**, 741–749 (2011). doi:10.1038/NGEO1296

Structural Dynamics

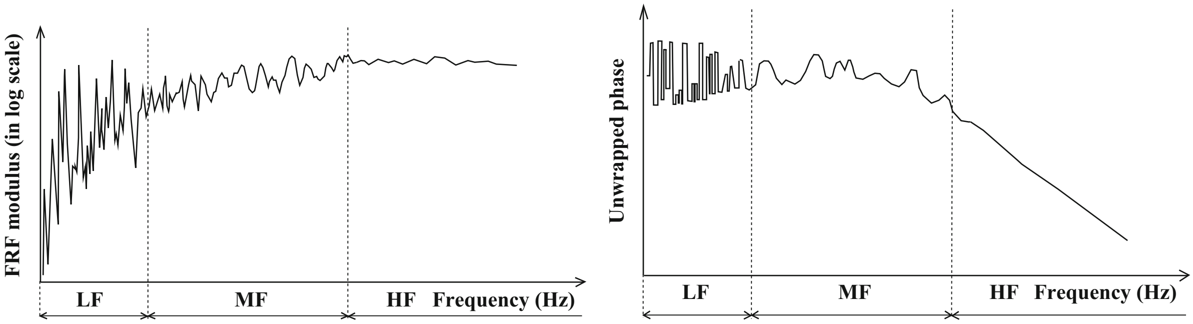
Roger Ohayon[1] and Christian Soize[2]
[1]Structural Mechanics and Coupled Systems Laboratory, LMSSC, Conservatoire National des Arts et Métiers (CNAM), Paris, France
[2]Laboratoire Modélisation et Simulation Multi-Echelle, MSME UMR 8208 CNRS, Universite Paris-Est, Marne-la-Vallée, France

Description

The computational structural dynamics is devoted to the computation of the dynamical responses in time or in frequency domains of complex structures, submitted to prescribed excitations. The complex structure is constituted of a deformable medium constituted of metallic materials, heterogeneous composite materials, and more generally, of metamaterials.

This chapter presents the linear dynamic analysis for complex structures, which is the most frequent case encountered in practice. For this situation, one of the most efficient modeling strategy is based on a formulation in the frequency domain (structural vibrations). There are many advantages to use a frequency domain formulation instead of a time domain formulation because the modeling can be adapted to the nature of the physical responses which are observed. This is the reason why the low-, the medium-, and the high-frequency ranges are introduced. The different types of vibration responses of a linear dissipative complex structure lead us to define the frequency ranges of analysis. Let $u_j(\mathbf{x}, \omega)$ be the Frequency Response Function (FRF) of a component j of the displacement $\mathbf{u}(\mathbf{x}, \omega)$, at a fixed point $\mathbf{x}$ of the structure and at a fixed circular frequency ω (in rad/s). Figure 1 represents the modulus $|u_j(\mathbf{x}, \omega)|$ in log scale and the unwrapped phase $\varphi_j(\mathbf{x}, \omega)$ of the FRF such that $u_j(\mathbf{x}, \omega) = |u_j(\mathbf{x}, \omega)| \exp\{-i\varphi_j(\mathbf{x}, \omega)\}$. The unwrapped phase is defined as a continuous function of ω obtained in adding multiples of $\pm 2\pi$ for jumps of the phase angle. The three frequency ranges can then be characterized as follows:

1. The *low-frequency range* (LF) is defined as the modal domain for which the modulus of the FRF exhibits isolated resonances due to a low modal density of elastic structural modes. The amplitudes of the resonances are driven by the damping and the phase rotates of π at the crossing of each isolated resonance (see Fig. 1). For the LF range, the strategy used consists in computing the elastic structural modes of the associated conservative dynamical system and then to construct a reduced-order model by the Ritz-Galerkin projection. The resulting matrix equation is solved in the time domain or in the frequency domain. It should be noted that substructuring techniques can also be introduced for complex structural systems. Those techniques consist in decomposing the structure into substructures and then in constructing a reduced-order model for each substructure for which the physical degrees of freedom on the coupling interfaces are kept.
2. The *high-frequency range* (HF) is defined as the range for which there is a high modal density which is constant on the considered frequency range. In this HF range the modulus of the FRF varies slowly as the function of the frequency and the phase is approximatively linear (see Fig. 1). Presently, this frequency range is relevant of various approaches such as *Statistical Energy Analysis* (SEA), diffusion of energy equation, and transport equation. However, due to the constant increase of computer power and advances in modeling of complex mechanical systems, this frequency domain becomes more and more accessible to the computational methods.
3. For complex structures (complex geometry, heterogeneous materials, complex junctions, complex boundary conditions, several attached equipments or mechanical subsystems, etc.), an intermediate frequency range, called the *medium-frequency range* (MF), appears. This MF range does not exist for a simple structure (e.g., a simply supported homogeneous straight beam). This MF range is defined as the intermediate frequency range



Structural Dynamics, Fig. 1 Modulus (*left*) and unwrapped phase (*right*) of the FRF as a function of the frequency. Definition of the LF, MF, and HF ranges

for which the modal density exhibits large variations over the frequency band. Due to the presence of the damping which yields an overlapping of elastic structural modes, the frequency response functions do not exhibit isolated resonances, and the phase slowly varies as a function of the frequency (see Fig. 1). In this MF range, the responses are sensitive to damping modeling (for weakly dissipative structure), which is frequency dependent, and sensitive to uncertainties. For this MF range, the computational model is constructed as follows: The reduced-order computational model of the LF range can be used in (i) adapting the finite element discretization to the MF range, (ii) introducing appropriate damping models (due to dissipation in the structure and to transfer of mechanical energy from the structure to mechanical subsystems which are not taken into account in the computational model), and (iii) introducing uncertainty quantification for both the system-parameter uncertainties and the model uncertainties induced by the modeling errors.

For sake of brevity, the case of nonlinear dynamical responses of structures (involving nonlinear constitutive equations, nonlinear geometrical effects, plays, etc.) is not considered in this chapter (see Bibliographical comments).

Formulation in the Low-Frequency Range for Complex Structures

We consider linear dynamics of a structure around a position of static equilibrium taken as the reference configuration, Ω, which is a three-dimensional bounded connected domain of $\mathbb{R}^3$, with a smooth boundary $\partial\Omega$ for which the external unit normal is denoted as $\mathbf{n}$. The generic point of Ω is $\mathbf{x} = (x_1, x_2, x_3)$. Let $\mathbf{u}(\mathbf{x}, t) = (u_1(\mathbf{x}, t), u_2(\mathbf{x}, t), u_3(\mathbf{x}, t))$ be the displacement of a particle located at point $\mathbf{x}$ in Ω and at a time t. The structure is assumed to be free ($\Gamma_0 = \emptyset$), a given surface force field $\mathbf{G}(\mathbf{x}, t) = (G_1(\mathbf{x}, t), G_2(\mathbf{x}, t), G_3(\mathbf{x}, t))$ is applied to the total boundary $\Gamma = \partial\Omega$, and a given body force field $\mathbf{g}(\mathbf{x}, t) = (g_1(\mathbf{x}, t), g_2(\mathbf{x}, t), g_3(\mathbf{x}, t))$ is applied in Ω. It is assumed that these external forces are in equilibrium. Below, if w is any quantity depending on $\mathbf{x}$, then $w_{,j}$ denotes the partial derivative of w with respect to x_j. The classical convention for summations over repeated Latin indices is also used.

The elastodynamic boundary value problem is written, in terms of $\mathbf{u}$ and at time t, as

$$\rho\, \partial_t^2 u_i(\mathbf{x}, t) - \sigma_{ij,j}(\mathbf{x}, t) = g_i(\mathbf{x}, t) \quad \text{in} \quad \Omega, \quad (1)$$

$$\sigma_{ij}(\mathbf{x}, t)\, n_j(\mathbf{x}) = G_i(\mathbf{x}, t) \quad \text{on} \quad \Gamma, \quad (2)$$

$$\sigma_{ij,j}(\mathbf{x},t) = a_{ijkh}(\mathbf{x})\, \varepsilon_{kh}(\mathbf{u}) + b_{ijkh}(\mathbf{x})\varepsilon_{kh}(\partial_t \mathbf{u}),$$

$$\varepsilon_{kh}(\mathbf{u}) = (u_{k,h} + u_{h,k})/2. \quad (3)$$

In (1), $\rho(\mathbf{x})$ is the mass density field, σ_{ij} is the Cauchy stress tensor. The constitutive equation is defined by (3) exhibiting an elastic part defined by the tensor $a_{ijkh}(\mathbf{x})$ and a dissipative part defined by the tensor $b_{ijkh}(\mathbf{x})$, independent of t because the model is developed for the low-frequency range, and $\varepsilon(\partial_t \mathbf{u})$ is the linearized strain tensor.

Let $\mathcal{C} = (H^1(\Omega))^3$ be the real Hilbert space of the admissible displacement fields, $\mathbf{x} \mapsto \mathbf{v}(\mathbf{x})$, on Ω. Considering t as a parameter, the variational formulation of the boundary value problem defined by

(1)–(3) consists, for fixed t, in finding $\mathbf{u}(.,t)$ in $\mathcal{C}$, such that

$$m(\partial_t^2\mathbf{u},\mathbf{v}) + d(\partial_t\mathbf{u},\mathbf{v}) + k(\mathbf{u},\mathbf{v}) = f(t;\mathbf{v}), \quad \forall\mathbf{v}\in\mathcal{C}, \tag{4}$$

in which the bilinear form m is symmetric and positive definite, the bilinear forms d and k are symmetric, positive semi-definite, and are such that

$$m(\mathbf{u},\mathbf{v}) = \int_\Omega \rho\, u_j\, v_j\, d\mathbf{x},$$

$$k(\mathbf{u},\mathbf{v}) = \int_\Omega a_{ijkh}\, \varepsilon_{kh}(\mathbf{u})\, \varepsilon_{ij}(\mathbf{v}) d\mathbf{x}, \tag{5}$$

$$d(\mathbf{u},\mathbf{v}) = \int_\Omega b_{ijkh}\, \varepsilon_{kh}(\mathbf{u})\, \varepsilon_{ij}(\mathbf{v}) d\mathbf{x},$$

$$f(t;\mathbf{v}) = \int_\Omega g_j(t)\, v_j\, d\mathbf{x} + \int_\Gamma G_j(t)\, v_j\, ds(\mathbf{x}). \tag{6}$$

The kernel of the bilinear forms k and d is the set of the rigid body displacements, $\mathcal{C}_{\text{rig}} \subset \mathcal{C}$ of dimension 6. Any displacement field $\mathbf{u}_{\text{rig}}$ in $\mathcal{C}_{\text{rig}}$ is such that, for all $\mathbf{x}$ in Ω, $\mathbf{u}_{\text{rig}}(\mathbf{x}) = \mathbf{t} + \boldsymbol{\theta} \times \mathbf{x}$ in which $\mathbf{t}$ and $\boldsymbol{\theta}$ are two arbitrary constant vectors in $\mathbb{R}^3$.

For the evolution problem with given Cauchy initial conditions $\mathbf{u}(.,0) = \mathbf{u}_0$ and $\partial_t\mathbf{u}(.,0) = \mathbf{v}_0$, the analysis of the existence and uniqueness of a solution requires the introduction of the following hypotheses: ρ is a positive bounded function on Ω; for all $\mathbf{x}$ in Ω, the fourth-order tensor $a_{ijkh}(\mathbf{x})$ (resp. $b_{ijkh}(\mathbf{x})$) is symmetric, $a_{ijkh}(\mathbf{x}) = a_{jikh}(\mathbf{x}) = a_{ijhk}(\mathbf{x}) = a_{khij}(\mathbf{x})$, and such that, for all second-order real symmetric tensor η_{ij}, there is a positive constant c independent of $\mathbf{x}$, such that $a_{ijkh}(\mathbf{x})\, \eta_{kh}\, \eta_{ij} \geq c\, \eta_{ij}\, \eta_{ij}$; the functions a_{ijkh} and b_{ijkh} are bounded on Ω; finally, $\mathbf{g}$ and $\mathbf{G}$ are such that the linear form $\mathbf{v} \mapsto f(t;\mathbf{v})$ is continuous on $\mathcal{C}$. Assuming that for all $\mathbf{v}$ in $\mathcal{C}$, $t \mapsto f(t;\mathbf{v})$ is a square integrable function on $\mathbb{R}$. Let $\mathcal{C}^c$ be the complexified vector space of $\mathcal{C}$ and let $\overline{v}$ be the complex conjugate of $\mathbf{v}$. Then, introducing the Fourier transforms $\mathbf{u}(\mathbf{x},\omega) = \int_\mathbb{R} e^{-i\omega t}\mathbf{u}(\mathbf{x},t)\, dt$ and $f(\omega;\mathbf{v}) = \int_\mathbb{R} e^{-i\omega t} f(t;\mathbf{v})\, dt$, the variational formulation defined by (4) can be rewritten as follows: For all fixed real $\omega \neq 0$, find $\mathbf{u}(.,\omega)$ with values in $\mathcal{C}^c$ such that

$$-\omega^2\, m(\mathbf{u},\overline{\mathbf{v}}) \;+\; i\omega\, d(\mathbf{u},\overline{\mathbf{v}}) \;+\; k(\mathbf{u},\overline{\mathbf{v}}) = f(\omega;\overline{\mathbf{v}}), \quad \forall\mathbf{v}\in\mathcal{C}^c. \tag{7}$$

The finite element discretization with n degrees of freedom of (4) yields the following second-order differential equation on $\mathbb{R}^n$:

$$[M]\,\ddot{\mathbf{U}}(t) + [D]\,\dot{\mathbf{U}}(t) + [K]\,\mathbf{U}(t) = \mathbf{F}(t), \tag{8}$$

and its Fourier transform, which corresponds to the finite element discretization of (7), yields the complex matrix equation which is written as

$$(-\omega^2\,[M] + i\omega\,[D] + [K])\,\mathbf{U}(\omega) = \mathbf{F}(\omega), \tag{9}$$

in which $[M]$ is the mass matrix which is a symmetric positive definite $(n \times n)$ real matrix and where $[D]$ and $[K]$ are the damping and stiffness matrices which are symmetric positive semi-definite $(n \times n)$ real matrices. *Case of a fixed structure.* If the structure is fixed on a part Γ_0 of boundary $\partial\Omega$ (Dirichlet condition $\mathbf{u} = \mathbf{0}$ on Γ_0), then the given surface force field $\mathbf{G}(\mathbf{x},t)$ is applied to the part $\Gamma = \partial\Omega\backslash\Gamma_0$. The space $\mathcal{C}$ of the admissible displacement fields must be replaced by

$$\mathcal{C}_0 = \{\mathbf{v}\in\mathcal{C}, \mathbf{v} = 0 \text{ on } \Gamma_0\}. \tag{10}$$

The complex vector space $\mathcal{C}^c$ must be replaced by the complex vector space $\mathcal{C}_0^c$ which is the complexified vector space of $\mathcal{C}_0$. The real matrices $[D]$ and $[K]$ are positive definite.

Associated Spectral Problem and Structural Modes

Setting $\lambda = \omega^2$, the spectral problem, associated with the variational formulation defined by (4) or (7), is stated as the following generalized eigenvalue problem. Find real $\lambda \geq 0$ and $\mathbf{u} \neq \mathbf{0}$ in $\mathcal{C}$ such that

$$k(\mathbf{u},\mathbf{v}) = \lambda\, m(\mathbf{u},\mathbf{v}), \quad \forall\mathbf{v}\in\mathcal{C}. \tag{11}$$

Rigid body modes (solutions for $\lambda = 0$). Since the dimension of $\mathcal{C}_{\text{rig}}$ is 6, then $\lambda = 0$ can be considered as a "zero eigenvalue" of multiplicity 6, denoted as $\lambda_{-5},\ldots,\lambda_0$. Let $\mathbf{u}_{-5},\ldots,\mathbf{u}_0$ be the corresponding eigenfunctions which are constructed such that the following orthogonality conditions are satisfied: for α and β in $\{-5,\ldots,0\}$, $m(\mathbf{u}_\alpha,\mathbf{u}_\beta) = \mu_\alpha\, \delta_{\alpha\beta}$ and $k(\mathbf{u}_\alpha,\mathbf{u}_\beta) = 0$. These eigenfunctions, called the *rigid body modes*, form a basis of $\mathcal{C}_{\text{rig}} \subset \mathcal{C}$ and any rigid body displacement $\mathbf{u}_{\text{rig}}$ in $\mathcal{C}_{\text{rig}}$ can then be expanded as $\mathbf{u}_{\text{rig}} = \sum_{\alpha=-5}^{0} q_\alpha\, \mathbf{u}_\alpha$.

Elastic structural modes (solutions for $\lambda \neq 0$). We introduce the subset $\mathcal{C}_{\text{elas}} = \mathcal{C} \setminus \mathcal{C}_{\text{rig}}$. It can be shown that $\mathcal{C} = \mathcal{C}_{\text{rig}} \oplus \mathcal{C}_{\text{elas}}$ which means that any displacement field $\mathbf{u}$ in $\mathcal{C}$ has the following unique decomposition $\mathbf{u} = \mathbf{u}_{\text{rig}} + \mathbf{u}_{\text{elas}}$ with $\mathbf{u}_{\text{rig}}$ in $\mathcal{C}_{\text{rig}}$ and $\mathbf{u}_{\text{elas}}$ in $\mathcal{C}_{\text{elas}}$. Consequently, $k(\mathbf{u}_{\text{elas}}, \mathbf{v}_{\text{elas}})$ defined on $\mathcal{C}_{\text{elas}} \times \mathcal{C}_{\text{elas}}$ is positive definite and we then have $k(\mathbf{u}_{\text{elas}}, \mathbf{u}_{\text{elas}}) > 0$ for all $\mathbf{u}_{\text{elas}} \neq \mathbf{0} \in \mathcal{C}_{\text{elas}}$.

Eigenvalue problem restricted to $\mathcal{C}_{\text{elas}}$. The eigenvalue problem restricted to $\mathcal{C}_{\text{elas}}$ is written as follows: Find $\lambda \neq 0$ and $\mathbf{u}_{\text{elas}} \neq \mathbf{0}$ in $\mathcal{C}_{\text{elas}}$ such that

$$k(\mathbf{u}_{\text{elas}}, \mathbf{v}_{\text{elas}}) = \lambda\, m(\mathbf{u}_{\text{elas}}, \mathbf{v}_{\text{elas}}), \quad \forall \mathbf{v}_{\text{elas}} \in \mathcal{C}_{\text{elas}}. \tag{12}$$

Countable number of positive eigenvalues. It can be proven that the eigenvalue problem, defined by (12), admits an increasing sequence of positive eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_\alpha \leq \ldots$. In addition, any multiple positive eigenvalue has a finite multiplicity (which means that a multiple positive eigenvalue is repeated a finite number of times).

Orthogonality of the eigenfunctions corresponding to the positive eigenvalues. The sequence of eigenfunctions $\{\mathbf{u}_\alpha\}_\alpha$ in $\mathcal{C}_{\text{elas}}$ corresponding to the positive eigenvalues satisfies the following orthogonality conditions:

$$m(\mathbf{u}_\alpha, \mathbf{u}_\beta) = \mu_\alpha\, \delta_{\alpha\beta}, \quad k(\mathbf{u}_\alpha, \mathbf{u}_\beta) = \mu_\alpha\, \omega_\alpha^2\, \delta_{\alpha\beta}, \tag{13}$$

in which $\omega_\alpha = \sqrt{\lambda_\alpha}$ and where μ_α is a positive real number depending on the normalization of eigenfunction $\mathbf{u}_\alpha$.

Completeness of the eigenfunctions corresponding to the positive eigenvalues. Let $\mathbf{u}_\alpha$ be the eigenfunction associated with eigenvalue $\lambda_\alpha > 0$. It can be shown that eigenfunctions $\{\mathbf{u}_\alpha\}_{\alpha \geq 1}$ form a complete family in $\mathcal{C}_{\text{elas}}$ and consequently, an arbitrary function $\mathbf{u}_{\text{elas}}$ belonging to $\mathcal{C}_{\text{elas}}$ can be expanded as $\mathbf{u}_{\text{elas}} = \sum_{\alpha=1}^{+\infty} q_\alpha\, \mathbf{u}_\alpha$ in which $\{q_\alpha\}_\alpha$ is a sequence of real numbers. These eigenfunctions are called the *elastic structural modes*.

Orthogonality between the elastic structural modes and the rigid body modes. We have $k(\mathbf{u}_\alpha, \mathbf{u}_{\text{rig}}) = 0$ and $m(\mathbf{u}_\alpha, \mathbf{u}_{\text{rig}}) = 0$. Substituting $\mathbf{u}_{\text{rig}}(\mathbf{x}) = \mathbf{t} + \boldsymbol{\theta} \times \mathbf{x}$ into (13) yields

$$\int_\Omega \mathbf{u}_\alpha(\mathbf{x})\, \rho(\mathbf{x})\, d\mathbf{x} = \mathbf{0}, \quad \int_\Omega \mathbf{x} \times \mathbf{u}_\alpha(\mathbf{x})\, \rho(\mathbf{x})\, d\mathbf{x} = \mathbf{0}, \tag{14}$$

which shows that the inertial center of the structure deformed under the elastic structural mode $\mathbf{u}_\alpha$, coincides with the inertial center of the undeformed structure.

Expansion of the displacement field using the rigid body modes and the elastic structural modes. Any displacement field $\mathbf{u}$ in $\mathcal{C}$ can then be written as $\mathbf{u} = \sum_{\alpha=-5}^{0} q_\alpha\, \mathbf{u}_\alpha + \sum_{\alpha=1}^{+\infty} q_\alpha\, \mathbf{u}_\alpha$.

Terminology. In structural vibrations, $\omega_\alpha > 0$ is called the *eigenfrequency* of elastic structural mode $\mathbf{u}_\alpha$ (or the eigenmode or mode shape of vibration) whose normalization is defined by the generalized mass μ_α. An elastic structural mode α is defined by the three quantities $\{\omega_\alpha, \mathbf{u}_\alpha, \mu_\alpha\}$.

Finite element discretization. The matrix equation of the generalized symmetric eigenvalue problem corresponding to the finite element discretization of (11) is written as

$$[\,K\,]\,\mathbf{U} = \lambda\; [\,M\,]\,\mathbf{U}. \tag{15}$$

For large computational model, this generalized eigenvalue problem is solved using iteration algorithms such as the Krylov sequence, the Lanczos method, and the subspace iteration method, which allows a prescribed number of eigenvalues and associated eigenvectors to be computed.

Case of a fixed structure. If the structure is fixed on Γ_0, $\mathcal{C}_{\text{rig}}$ is reduced to the empty set and admissible space $\mathcal{C}_{\text{elas}}$ must be replaced by $\mathcal{C}_0$. In this case the eigenvalues are strictly positive. In addition, the property given for the free structure concerning the inertial center of the structure does not hold.

Reduced-Order Computational Model in the Frequency Domain

In the frequency range, the reduced-order computational model is carried out using the Ritz-Galerkin projection. Let $\mathcal{C}_S$ be the admissible function space such that $\mathcal{C}_S = \mathcal{C}_{\text{elas}}$ for a free structure and $\mathcal{C}_S = \mathcal{C}_0$ for a structure fixed on Γ_0. Let $\mathcal{C}_{S,N}$ be the subspace of $\mathcal{C}_S$, of dimension $N \geq 1$, spanned by the finite family $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ of elastic structural modes $\mathbf{u}_\alpha$.

For all fixed ω, the projection $\mathbf{u}^N(\omega)$ of $\mathbf{u}(\omega)$ on the complexified vector space of $\mathcal{C}_{S,N}$ can be written as

$$\mathbf{u}^N(\mathbf{x},\omega) = \sum_{\alpha=1}^{N} q_\alpha(\omega)\,\mathbf{u}_\alpha(\mathbf{x}), \tag{16}$$

in which $\mathbf{q}(\omega) = (q_1(\omega),\ldots,q_N(\omega))$ is the complex-valued vector of the generalized coordinates which verifies the matrix equation on $\mathbb{C}^N$,

$$(-\omega^2\,[\mathcal{M}\,] + i\,\omega\,[\mathcal{D}] + [\,\mathcal{K}\,])\,\mathbf{q}(\omega) = \boldsymbol{\mathcal{F}}(\omega), \tag{17}$$

in which $[\mathcal{M}\,]$, $[\mathcal{D}]$, and $[\,\mathcal{K}\,]$ are $(N \times N)$ real symmetric positive definite matrices (for a free or a fixed structure). Matrices $[\mathcal{M}]$ and $[\,\mathcal{K}\,]$ are diagonal and such that

$$\begin{aligned}[\mathcal{M}\,]_{\alpha\beta} &= m(\mathbf{u}_\beta,\mathbf{u}_\alpha) = \mu_\alpha\,\delta_{\alpha\beta},\\ [\,\mathcal{K}\,]_{\alpha\beta} &= k(\mathbf{u}_\beta,\mathbf{u}_\alpha) = \mu_\alpha\,\omega_\alpha^2\delta_{\alpha\beta}.\end{aligned} \tag{18}$$

The damping matrix $[\mathcal{D}]$ is not sparse (fully populated) and the component $\mathcal{F}_\alpha$ of the complex-valued vector of the generalized forces $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1,\ldots,\mathcal{F}_N)$ are such that

$$[\mathcal{D}]_{\alpha\beta} = d(\mathbf{u}_\beta,\mathbf{u}_\alpha), \quad \mathcal{F}_\alpha(\omega) = f(\omega;\mathbf{u}_\alpha). \tag{19}$$

The reduced-order model is defined by (16)–(19).

Convergence of the solution constructed with the reduced-order model. For all real ω, (17) has a unique solution $\mathbf{u}^N(\omega)$ which is convergent in $\mathcal{C}_S$ when N goes to infinity. Quasi-static correction terms can be introduced to accelerate the convergence with respect to N.

Remarks concerning the diagonalization of the damping operator. When damping operator is diagonalized by the elastic structural modes, matrix $[\mathcal{D}]$ defined by (19), is an $(N \times N)$ diagonal matrix which can be written as $[\mathcal{D}]_{\alpha\beta} = d(\mathbf{u}_\beta,\mathbf{u}_\alpha) = 2\,\mu_\alpha\,\omega_\alpha\,\xi_\alpha\,\delta_{\alpha\beta}$ in which μ_α and ω_α are defined by (18). The critical damping rate ξ_α of elastic structural mode $\mathbf{u}_\alpha$ is a positive real number. A weakly damped structure is a structure such that $0 < \xi_\alpha \ll 1$ for all α in $\{1,\ldots,N\}$. Several algebraic expressions exist for diagonalizing the damping bilinear form with the elastic structural modes.

Bibliographical Comments

The mathematical aspects related to the variational formulation, existence and uniqueness, and finite element discretization of boundary value problems for elastodynamics can be found in Dautray and Lions [6], Oden and Reddy [11], and Hughes [9]. More details concerning the finite element method can be found in Zienkiewicz and Taylor [14]. Concerning the time integration algorithms in nonlinear computational dynamics, the readers are referred to Belytschko et al. [3], and Har and Tamma [8]. General mechanical formulations in computational structural dynamics, vibration, eigenvalue algorithms, and substructuring techniques can be found in Argyris and Mlejnek [1], Geradin and Rixen [7], Bathe and Wilson [2], and Craig and Bampton [5]. For computational structural dynamics in the low- and the medium-frequency ranges and extensions to structural acoustics, we refer the reader to Ohayon and Soize [12]. Various formulations for the high-frequency range can be found in Lyon and Dejong [10] for Statistical Energy Analysis, and in Chap. 4 of Bensoussan et al. [4] for diffusion of energy and transport equations. Concerning uncertainty quantification (UQ) in computational structural dynamics, we refer the reader to Soize [13].

References

1. Argyris, J., Mlejnek, H.P.: Dynamics of Structures. North-Holland, Amsterdam (1991)
2. Bathe, K.J., Wilson, E.L.: Numerical Methods in Finite Element Analysis. Prentice-Hall, New York (1976)
3. Belytschko, T., Liu, W.K., Moran, B.: Nonlinear Finite Elements for Continua and Structures. Wiley, Chichester (2000)
4. Bensoussan, A., Lions, J.L., Papanicolaou, G.: Asymptotic Analysis for Periodic Structures. AMS Chelsea Publishing, Providence (2010)
5. Craig, R.R., Bampton, M.C.C.: Coupling of substructures for dynamic analysis. AIAA J. **6**, 1313–1319 (1968)
6. Dautray, R., Lions, J.L.: Mathematical Analysis and Numerical Methods for Science and Technology. Springer, Berlin (2000)
7. Geradin, M., Rixen, D.: Mechanical Vibrations: Theory and Applications to Structural Dynamics, 2nd edn. Wiley, Chichester (1997)
8. Har, J., Tamma, K.: Advances in Computational Dynamics of Particles, Materials and Structures. Wiley, Chichester (2012)

9. Hughes, T.J.R.: The Finite Element Method: Linear Static and Dynamic Finite Element Analysis. Dover, New York (2000)
10. Lyon, R.H., Dejong, R.G.: Theory and Application of Statistical Energy Analysis, 2nd edn. Butterworth-Heinemann, Boston (1995)
11. Oden, J.T., Reddy, J.N.: An Introduction to the Mathematical Theory of Finite Elements. Dover, New York (2011)
12. Ohayon, R., Soize, C.: Structural Acoustics and Vibration. Academic, London (1998)
13. Soize, C.: Stochastic Models of Uncertainties in Computational Mechanics, vol. 2. Engineering Mechanics Institute (EMI) of the American Society of Civil Engineers (ASCE), Reston (2012)
14. Zienkiewicz, O.C., Taylor, R.L.: The Finite Element Method: The Basis, vol. 1, 5th edn. Butterworth- Heinemann, Oxford (2000)

Subdivision Schemes

Nira Dyn
School of Mathematical Sciences, Tel-Aviv
University, Tel-Aviv, Israel

Mathematics Subject Classification

Primary: 65D07; 65D10; 65D17. Secondary: 41A15; 41A25

Definition

A subdivision scheme is a method for generating a continuous function from discrete data, by repeated applications of refinement rules. A refinement rule operates on a set of data points and generates a denser set using local mappings. The function generated by a convergent subdivision scheme is the limit of the sequence of sets of points generated by the repeated refinements.

Description

Subdivision schemes are efficient computational methods for the design, representation, and approximation of curves and surfaces in 3D and for the generation of refinable functions, which are instrumental in the construction of wavelets.

The "classical" subdivision schemes are stationary and linear, applying the same linear refinement rule at each refinement level. The theory of these schemes is well developed and well understood; see [2, 7, 9, 15], and references therein.

Nonlinear schemes were designed for the approximation of piecewise smooth functions (see, e.g., [1,4]), for taking into account the geometry of the initial points in the design of curves/surfaces (see, e.g., [5, 8, 11]), and for manifold-valued data (see, e.g., [12, 14, 16]). These schemes were studied at a later stage, and in many cases their analysis is based on their *proximity* to linear schemes.

Linear Schemes

A linear refinement rule operates on a set of points in $\mathbb{R}^d$ with topological relations among them, expressed by relating the points to the vertices of a regular grid in $\mathbb{R}^s$. In the design of 3D surfaces, $d = 3$ and $s = 2$.

The refinement consists of a rule for refining the grid and a rule for defining the new points corresponding to the vertices of the refined grid. The most common refinement is binary, and the most common grid is $\mathbb{Z}^s$.

For a set of points $\mathcal{P} = \{P_i \in \mathbb{R}^d : i \in \mathbb{Z}^s\}$, related to $2^{-k}\mathbb{Z}^s$, the binary refinement rule $\mathcal{R}$ generates points related to $2^{-k-1}\mathbb{Z}^s$, of the form

$$(\mathcal{R}\mathcal{P})_i = \sum_{j \in \mathbb{Z}^s} a_{i-2j} P_j, \quad i \in \mathbb{Z}^s, \qquad (1)$$

with the point $(\mathcal{R}\mathcal{P})_i$ related to the vertex $i2^{-k-1}$. The set of coefficients $\{a_i \in \mathbb{R} : i \in \mathbb{Z}^s\}$ is called the mask of the refinement rule, and only a finite number of the coefficients are nonzero, reflecting the locality of the refinement.

When the same refinement rule is applied in all refinement levels, the scheme is called *stationary*, while if different linear refinement rules are applied in different refinement levels, the scheme is called *nonstationary* (see, e.g., [9]).

A stationary scheme is defined to be convergent (in the L_∞-norm) if for any initial set of points $\mathcal{P}$, there exists a continuous function F defined on $(\mathbb{R}^s)^d$ such that

$$\lim_{k\to\infty}\sup_{i\in\mathbb{Z}^s}|F(i2^{-k})-(\mathcal{R}^k\mathcal{P})_i|=0, \tag{2}$$

and if for at least one initial set of points, $F \not\equiv 0$. A similar definition of convergence is used for all other types of subdivision schemes, with $\mathcal{R}^k$ replaced by the appropriate product of refinement rules.

Although the refinement (1) is defined on all $\mathbb{Z}^s$, the finite support of the mask guarantees that the limit of the subdivision scheme at a point is affected only by a finite number of initial points.

When the initial points are samples of a smooth function, the limit function of a convergent linear subdivision scheme approximates the sampled function. Thus, a convergent linear subdivision scheme is a linear approximation operator.

As examples, we give two prototypes of stationary schemes for $s = 1$. Each is the simplest of its kind, converging to C^1 univariate functions. The first is the Chaikin scheme [3], called also "Corner cutting," with limit functions which "preserve the shape" of the initial sets of points. The refinement rule is

$$\begin{aligned}(\mathcal{RP})_{2i} &= \frac{3}{4}P_i+\frac{1}{4}P_{i+1},\\ (\mathcal{RP})_{2i+1} &= \frac{1}{4}P_i+\frac{3}{4}P_{i+1},\quad i\in\mathbb{Z}.\end{aligned} \tag{3}$$

The second is the 4-point scheme [6, 10], which interpolates the initial set of points (and all the sets of points generated by the scheme). It is defined by the refinement rule

$$\begin{aligned}(\mathcal{RP})_{2i}=P_i,\ \ (\mathcal{RP})_{2i+1}&=\frac{9}{16}(P_i+P_{i+1})\\ &-\frac{1}{16}(P_{i-1}+P_{i+2}),\ i\in\mathbb{Z}.\end{aligned} \tag{4}$$

While the limit functions of the Chaikin scheme can be written in terms of B-splines of degree 2, the limit functions of the 4-point scheme for general initial sets of points have a fractal nature and are defined only procedurally by (4).

For the design of surfaces in 3D, $s = 2$ and the common grids are $\mathbb{Z}^2$ and regular triangulations. The latter are refined by dividing each triangle into four equal ones. Yet regular grids (with each vertex belonging to four squares in the case of $\mathbb{Z}^2$ and to six triangles in the case of a regular triangulation) are not sufficient for representing surfaces of general topology, and a finite number of *extraordinary points* are required [13].

The analysis on regular grids of the convergence of a stationary linear scheme, and of the smoothness of the generated functions, is based on the coefficients of the mask. It requires the computation of the joint spectral radius of several finite dimensional matrices with mask coefficients as elements in specific positions (see, e.g., [2]) or an equivalent computation in terms of the Laurent polynomial $a(z) = \sum_{i\in\mathbb{Z}^s} a_i z^i$ (see, e.g., [7]).

When dealing with the design of surfaces, this analysis applies only in all parts of the grids away from the extraordinary points. The analysis at these points is local [13], but rather involved. It also dictates changes in the refinement rules that have to be made near extraordinary points [13, 17].

References

1. Amat, S., Dadourian, K., Liandart, J.: On a nonlinear subdivision scheme avoiding Gibbs oscillations and converging towards C. Math. Comput. **80**, 959–971 (2011)
2. Cavaretta, A.S., Dahmen, W., Micchelli, C.A.: Stationary Subdivision. Memoirs of MAS, vol. 93, p. 186. American Mathematical Society, Providence (1991)
3. Chaikin, G.M.: An algorithm for high speed curve generation. Comput. Graph. Image Process. **3**, 346–349 (1974)
4. Cohem, A., Dyn, N., Matei, B.: Quasilinear subdivision schemes with applications to ENO interpolation. Appl. Comput. Harmonic Anal. **15**, 89–116 (2003)
5. Deng, C., Wang, G.: Incenter subdivision scheme for curve interpolation. Comput. Aided Geom. Des. **27**, 48–59 (2010)
6. Dubuc, S.: Interpolation through an iterative scheme. J. Math. Anal. Appl. **114**, 185–204 (1986)
7. Dyn, N.: Subdivision schemes in computer-aided geometric design. In: Light, W. (ed.) Advances in Numerical Analysis – Volume II, Wavelets, Subdivision Algorithms and Radial Basis Functions, p. 36. Clarendon, Oxford (1992)
8. Dyn, N., Hormann, K.: Geometric conditions for tangent continuity of interpolatory planar subdivision curves. Comput. Aided Geom. Des. **29**, 332–347 (2012)
9. Dyn, N., Levin, D.: Subdivision schemes in geometric modelling. Acta-Numer. **11**, 73–144 (2002)
10. Dyn, N., Gregory, J., Levin, D.: A 4-point interpolatory subdivision scheme for curve design. Comput. Aided Geom. Des. **4**, 257–268 (1987)
11. Dyn, N., Floater, M.S., Hormann, K.: Four-point curve subdivision based on iterated chordal and centripetal parametrizations. Comput. Aided Geom. Des. **26**, 279–286 (2009)

12. Grohs, P.: A general proximity analysis of nonlinear subdivision schemes. SIAM J. Math. Anal. **42**, 729–750 (2010)
13. Peters, J., Reif, U.: Subdivision Surfaces, p. 204. Springer, Berlin/Heidelberg (2008)
14. Wallner, J., Navayazdani, E., Weinmann, A.: Convergence and smoothness analysis of subdivision rules in Riemannian and symmetric spaces. Adv. Comput. Math. **34**, 201–218 (2011)
15. Warren, J., Weimer, H.: Subdivision Methods for Geometric Design, p. 299. Morgan Kaufmann, San Francisco (2002)
16. Xie, G., Yu, T.: Smoothness equivalence properties of general manifold-valued data subdivision schemes. Multiscale Model. Simul. **7**, 1073–1100 (2010)
17. Zorin, D.: Smoothness of stationary subdivision on irregular meshes. Constructive Approximation. **16**, 359–397 (2000)

Symbolic Computing

Ondřej Čertík[1], Mateusz Paprocki[2], Aaron Meurer[3], Brian Granger[4], and Thilina Rathnayake[5]
[1]Los Alamos National Laboratory, Los Alamos, NM, USA
[2]refptr.pl, Wroclaw, Poland
[3]Department of Mathematics, New Mexico State University, Las Cruces, NM, USA
[4]Department of Physics, California Polytechnic State University, San Luis Obispo, CA, USA
[5]Department of Computer Science, University of Moratuwa, Moratuwa, Sri Lanka

Synonyms

Computer algebra system; Symbolic computing; Symbolic manipulation

Keywords

Symbolic computing; Computer algebra system

Glossary/Definition Terms

Numerical computing Computing that is based on finite precision arithmetic.

Symbolic computing Computing that uses symbols to manipulate and solve mathematical formulas and equations in order to obtain mathematically exact results.

Definition

Scientific computing can be generally divided into two subfields: *numerical computation*, which is based on finite precision arithmetic (usually single or double precision), and *symbolic computing* which uses symbols to manipulate and solve mathematical equations and formulas in order to obtain mathematically exact results.

Symbolic computing, also called *symbolic manipulation* or *computer algebra system* (CAS), typically includes systems that can focus on well-defined computing areas such as polynomials, matrices, abstract algebra, number theory, or statistics (symbolic), as well as on calculus-like manipulation such as limits, differential equations, or integrals. A full-featured CAS should have all or most of the listed features. There are systems that focus only on one specific area like polynomials – those are often called CAS too.

Some authors claim that symbolic computing and computer algebra are two views of computing with mathematical objects [1]. According to them, symbolic computation deals with expression trees and addresses problems of determination of expression equivalence, simplification, and computation of canonical forms, while computer algebra is more centered around the computation of mathematical quantities in well-defined algebraic domains. The distinction between *symbolic computing* and *computer algebra* is often not made; the terms are used interchangeably. We will do so in this entry as well.

History

Algorithms for Computer Algebra [2] provides a concise description about the history of symbolic computation. The invention of LISP in the early 1960s had a great impact on the development of symbolic computation. *FORTRAN* and *ALGOL* which existed at the time were primarily designed for numerical computation. In 1961, James Slagle at MIT (Massachusetts Institute of Technology) wrote a heuristics-based LISP program for Symbolic Automatic INTegration (*SAINT*) [3]. In 1963, Martinus J. G. Veltman developed *Schoonschip* [4, 5] for particle physics calculations. In 1966, Joel Moses (also from MIT) wrote a program called *SIN* [6] for the same purpose as *SAINT*, but he used a more efficient algorithmic approach. In 1968, *REDUCE* [7] was developed by Tony Hearn at Stanford University for physics calculations. Also in 1968, a specialized CAS called *CAMAL* [8] for handling Poisson series in celestial mechanics was developed by John Fitch and David Barton from the University of Cambridge. In 1970, a general purposed system called *REDUCE 2* was introduced.

In 1971 *Macsyma* [9] was developed with capabilities for algebraic manipulation, limit calculations, symbolic integration, and the solution of equations. In the late 1970s *muMATH* [10] was developed by the University of Hawaii and it came with its own programming language. It was the first CAS to run on widely available IBM PC computers. With the development of computing in 1980s, more modern CASes began to emerge. *Maple* [11] was introduced by the University of Waterloo with a small compiled kernel and a large mathematical library, thus allowing it to be used powerfully on smaller platforms. In 1988, *Mathematica* [12] was developed by Stephen Wolfram with better graphical capabilities and integration with graphical user interfaces. In the 1980s more and more CASes were developed like *Macaulay* [13], *PARI* [14], *GAP* [15], and *CAYLEY* [16] (which later became *Magma* [17]). With the popularization of open-source software in the past decade, many open-source CASes were developed like *Sage* [18], *SymPy* [19], etc. Also, many of the existing CASes were later open sourced; for example, *Macsyma* became *Maxima* [20]; *Scratchpad* [21] became *Axiom* [22].

Overview

A common functionality of all computer algebra systems typically includes at least the features mentioned in the following subsections. We use *SymPy* 0.7.5 and *Mathematica* 9 as examples of doing the same operation in two different systems, but any other full-featured CAS can be used as well (e.g., from the Table 1) and it should produce the same results functionally.

To run the *SymPy* examples in a *Python* session, execute the following first:

```
from sympy import *
x, y, z, n, m = symbols('x, y, z,
                         n, m')
f = Function('f')
```

To run the *Mathematica* examples, just execute them in a *Mathematica Notebook*.

Arbitrary Formula Representation

One can represent arbitrary expressions (not just polynomials). SymPy:

```
In [1]: (1+1/x)**x
Out[1]: (1 + 1/x)**x
In [2]: sqrt(sin(x))/z
Out[2]: sqrt(sin(x))/z
```

Mathematica:

```
In[1]:= (1+1/x)^x
Out[1]= (1+1/x)^x
```

Symbolic Computing, Table 1 Implementation details of various computer algebra systems

Program	License	Internal implementation language	CAS language
Mathematica [12]	Commercial	C/C++	Custom
Maple [11]	Commercial	C/C++	custom
Symbolic MATLAB toolbox [23]	Commercial	C/C++	Custom
Axiom[a] [22]	BSD	Lisp	Custom
SymPy [19]	BSD	Python	Python
Maxima [20]	GPL	Lisp	Custom
Sage [18]	GPL	C++/Cython/Lisp	Python[b]
Giac/Xcas [24]	GPL	C++	Custom

[a] The same applies to its two forks *FriCAS* [25] and *OpenAxiom* [26]

[b] The default environment in *Sage* actually extends the *Python* language using a preparser that converts things like `2^3` into `Integer(2)**Integer(3)`, but the preparser can be turned off and one can use *Sage* from a regular *Python* session as well

```
In[2]:= Sqrt[Sin[x]]/z
Out[2]= Sqrt[Sin[x]]/z
```

Limits

SymPy:

```
In [1]: limit(sin(x)/x, x, 0)
Out[1]: 1
In [2]: limit((2-sqrt(x))/(4-x),x,4)
Out[2]: 1/4
```

Mathematica:

```
In[1]:= Limit[Sin[x]/x,x->0]
Out[1]= 1
In[2]:= Limit[(2-Sqrt[x])/(4-x),x->4]
Out[2]= 1/4
```

Differentiation

SymPy:

```
In [1]: diff(sin(2*x), x)
Out[1]: 2*cos(2*x)
In [1]: diff(sin(2*x), x, 10)
Out[1]: -1024*sin(2*x)
```

Mathematica:

```
In[1]:= D[Sin[2 x],x]
Out[1]= 2 Cos[2 x]
In[2]:= D[Sin[2 x],{x,10}]
Out[2]= -1024 Sin[2 x]
```

Integration

SymPy:

```
In [1]: integrate(1/(x**2+1), x)
Out[1]: atan(x)
In [1]: integrate(1/(x**2+3), x)
Out[1]: sqrt(3)*atan(sqrt(3)*x/3)/3
```

Mathematica:

```
In[1]:= Integrate[1/(x^2+1),x]
Out[1]= ArcTan[x]
In[2]:= Integrate[1/(x^2+3),x]
Out[2]= ArcTan[x/Sqrt[3]]/Sqrt[3]
```

Polynomial Factorization

SymPy:

```
In [1]: factor(x**2*y + x**2*z
                + x*y**2 + 2*x*y*z
                + x*z**2 + y**2*z
                + y*z**2)
Out[1]: (x + y)*(x + z)*(y + z)
```

Mathematica:

```
In[1]:= Factor[x^2 y+x^2 z+x
            y^2+2 x y z+x
            z^2+y^2 z+y z^2]
Out[1]= (x+y) (x+z) (y+z)
```

Algebraic and Differential Equation Solvers

Algebraic equations, SymPy:

```
In [1]: solve(x**4+x**2+1, x)
Out[1]: [-1/2 - sqrt(3)*I/2, -1/2
            + sqrt(3)*I/2,
          1/2 - sqrt(3)*I/2,  1/2
            + sqrt(3)*I/2]
```

Mathematica:

```
In[1]:= Reduce[1+x^2+x^4==0,x]
Out[1]= x==-(-1)^(1/3)||x
         ==(-1)^(1/3)||
        x==-(-1)^(2/3)||x
         ==(-1)^(2/3)
```

and differential equations, SymPy:

```
In [1]: dsolve(f(x).diff(x, 2)
            +f(x), f(x))
Out[1]: f(x) == C1*sin(x)
            + C2*cos(x)
In [1]: dsolve(f(x).diff(x, 2)
            +9*f(x), f(x))
Out[1]: f(x) == C1*sin(3*x)
            + C2*cos(3*x)
```

Mathematica:

```
In[1]:= DSolve[f’’[x]+f[x]==0,f[x],x]
Out[1]= {{f[x]->C[1] Cos[x]
                 +C[2] Sin[x]}}
In[2]:= DSolve[f’’[x]+9 f[x]
                 ==0,f[x],x]
Out[2]= {{f[x]->C[1] Cos[3 x]
                 +C[2] Sin[3 x]}}
```

Formula Simplification

Simplification is not a well-defined operation (i.e., there are many ways how to define the complexity of an expression), but typically the CAS is able to simplify, for example, the following expressions in an expected way, SymPy:

```
In [1]: simplify(-1/(2*(x**2 + 1))
      - 1/(4*(x + 1))+1/(4*(x - 1))
      - 1/(x**4-1))
```

```
Out[1]: 0
```

Mathematica:

```
In[1]:= Simplify[-1/(2(x^2+1))
               -1/(4(x+1))+1/(4(x-1))
               -1/(x^4-1)]
Out[1]= 0
```

or, SymPy:

```
In [1]: simplify((x - 1)/(x**2 - 1))
Out[1]: 1/(x + 1)
```

Mathematica:

```
In[1]:= Simplify[(x-1)/(x^2-1)]
Out[1]= 1/(1+x)
```

Numerical Evaluation

Exact expressions like $\sqrt{2}$, constants, sums, integrals, and symbolic expressions can be evaluated to a desired accuracy using a CAS. For example, in SymPy:

```
In [1]: N(sqrt(2),30)
Out[1]: 1.41421356237309504880168872421
In [2]: N(Sum(1/n**n, (n,1,oo)),30)
Out[2]: 1.29128599706266354040728259060
```

Mathematica:

```
In[1]:= N[Sqrt[2],30]
Out[1]= 1.41421356237309504880168872421
In[2]:= N[Sum[1/n^n,{n,1,
        Infinity}],30]
Out[2]= 1.29128599706266354040728259060
```

Symbolic Summation

There are circumstances where it is mathematically impossible to get an explicit formula for a given sum. When an explicit formula exists, getting the exact result is usually desirable. SymPy:

```
In [1]: Sum(n, (n, 1, m)).doit()
Out[1]: m**2/2 + m/2
In [2]: Sum(1/n**6,(n,1,oo)).doit()
Out[2]: pi**6/945
In [3]: Sum(1/n**5,(n,1,oo)).doit()
Out[3]: zeta(5)
```

Mathematica:

```
In[1]:= Sum[n,{n,1,m}]
Out[1]= 1/2 m (1+m)
In[2]:= Sum[1/n^6,{n,1,Infinity}]
Out[2]= Pi^6/945
In[3]:= Sum[1/n^5,{n,1,Infinity}]
Out[3]= Zeta[5]
```

Software

A computer algebra system (CAS) is typically composed of a high-level (usually interpreted) language that the user interacts with in order to perform calculations. Many times the implementation of such a CAS is a mix of the high-level language together with some low-level language (like C or C++) for efficiency reasons. Some of them can easily be used as a library in user's programs; others can only be used from the custom CAS language.

A comprehensive list of computer algebra software is at [27]. Table 1 lists features of several established computer algebra systems. We have only included systems that can handle at least the problems mentioned in the Overview section.

Besides general full-featured CASes, there exist specialized packages, for Example, *Singular* [28] for very fast polynomial manipulation or *GiNaC* [29] that can handle basic symbolic manipulation but does not have integration, advanced polynomial algorithms, or limits. *Pari* [14] is designed for number theory computations and *Cadabra* [30] for field theory calculations with tensors. *Magma* [17] specializes in algebra, number theory, algebraic geometry, and algebraic combinatorics.

Finally, a CAS also usually contains a notebook like interface, which can be used to enter commands or programs, plot graphs, and show nicely formatted equations. For Python-based CASes, one can use *IPython Notebook* [31] or *Sage Notebook* [18], both of which are interactive web applications that can be used from a web browser. C++ CASes can be wrapped in Python, for example, *GiNaC* has several Python wrappers: *Swiginac* [32], *Pynac* [33], etc. These can then be used from Python-based notebooks. *Mathematica* and *Maple* also contain a notebook interface, which accepts the given CAS high-level language.

Applications of Symbolic Computing

Symbolic computing has traditionally had numerous applications. By 1970s, many CASes were used for celestial mechanics, general relativity, quantum electrodynamics, and other applications [34]. In this section we present a few such applications in more detail, but necessarily our list is incomplete and is only meant as a starting point for the reader.

Many of the following applications and scientific advances related to them would not be possible without symbolic computing.

Code Generation

One of the frequent use of a CAS is to derive some symbolic expression and then generate C or Fortran code that numerically evaluates it in a production high-performance code. For example, to obtain the best rational function approximation (of orders 8, 8) to a modified Bessel function of the first kind of half-integer argument $I_{9/2}(x)$ on an interval [4, 10], one can use (in *Mathematica*):

```
In[1]:= Needs["FunctionApproximations`"]
In[2]:= FortranForm[HornerForm[MiniMaxApproximation[
            BesselI[9/2, x]*Sqrt[Pi*x/2]/Exp[x],
            {x,{4,10},8,8},WorkingPrecision->30][[2,1]]]]
Out[2]//FortranForm=
      (0.000395502959013236968661582656143 +
    x*(-0.001434648369704841686633794071 +
    x*(0.002487834745835034731351436444434 +
    x*(-0.00274477921388295929464613063609 +
    x*(0.002162750181076572737255897404 99 +
    x*(-0.000236779926184242197820134964535 +
    x*(0.0000882030507076791807159699814428 +
      (-4.62078105288798755556136693122e-6 +
          8.23671374777791529292655504214e-7*x)*x)))))))
      /
      (1. + x*(0.504839286873735708062045336271 +
    x*(0.176683950009401712892997268723 +
    x*(0.0438594911840609324095487447279 +
    x*(0.00829753062428409331123592322788 +
    x*(0.00111693697900468156881720995034 +
    x*(0.000174719963536517752971223459247 +
      (7.22885338737473776714257581233e-6 +
          1.64737453771748367647332279826e-6*x)*x)))))))
```

The result can be readily used in a Fortran code (we reformatted the white space in the output Out[2] to better fit into the page).

Particle Physics

The application of symbolic computing in particle physics typically involves generation and then calculation of Feynman diagrams (among other things that involves doing fast traces of Dirac gamma matrices and other tensor operations). The first CAS that was designed for this task was *Schoonschip* [4, 5], and in 1984 *FORM* [35] was created as a successor. *FORM* has built-in features for manipulating formulas in particle physics, but it can also be used as a general purpose system (it keeps all expressions in expanded form, so it cannot do factorization; it also does not have more advanced features like series expansion, differential equations, or integration).

Many of the scientific results in particle physics would not be possible without a powerful CAS; *Schoonschip* was used for calculating properties of the

W boson in the 1960s and *FORM* is still maintained and used to this day.

Another project is *FeynCalc* [36], originally written for *Macsyma* (*Maxima*) and later *Mathematica*. It is a package for algebraic calculations in elementary particle physics, among other things; it can do tensor and Dirac algebra manipulation, Lorentz index contraction, and generation of Feynman rules from a Lagrangian, Fortran code generation. There are hundreds of publications that used FeynCalc to perform calculations.

Similar project is *FeynArts* [37], which is also a *Mathematica* package that can generate and visualize Feynman diagrams and amplitudes. Those can then be calculated with a related project *FormCalc* [38], built on top of *FORM*.

PyDy

PyDy, short for Python Dynamics, is a work flow that utilizes an array of scientific tools written in the Python programming language to study multi-body dynamics [39]. *SymPy* mechanics package is used to generate symbolic equations of motion in complex multi-body systems, and several other scientific Python packages are used for numerical evaluation (*NumPy* [40]), visualization (*Matplotlib* [41]), etc. First, an idealized version of the system is represented (geometry, configuration, constraints, external forces). Then the symbolic equations of motion (often very long) are generated using the mechanics package and solved (integrated after setting numerical values for the parameters) using differential equation solvers in *SciPy* [42]. These solutions can then be used for simulations and visualizations. Symbolic equation generation guarantees no mistakes in the calculations and makes it easy to deal with complex systems with a large number of components.

General Relativity

In general relativity the CASes have traditionally been used to symbolically represent the metric tensor $g^{\mu\nu}$ and then use symbolic derivation to derive various tensors (Riemann and Ricci tensor, curvature, ...) that are present in the Einstein's equations [34]:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + g_{\mu\nu} \Lambda = \frac{8\pi G}{c^4} T_{\mu\nu} \,. \tag{1}$$

Those can then be solved for simple systems.

Celestial Mechanics

The equations in celestial mechanics are solved using perturbation theory, which requires very efficient manipulation of a Poisson series [8, 34, 43–50]:

$$\sum P(a, b, c, \ldots, h) \, {\sin \atop \cos} \, (\lambda u + \mu v + \cdots + \gamma z) \tag{2}$$

where $P(a, b, c, \ldots, h)$ is a polynomial and each term contains either sin or cos. Using trigonometric relations, it can be shown that this form is closed to addition, subtraction, multiplication, differentiation, and restricted integration. One of the earliest specialized CASes for handling Poisson series is *CAMAL* [8]. Many others were since developed, for example, *TRIP* [43].

Quantum Mechanics

Quantum mechanics is well known for many tedious calculations, and the use of a CAS can aid in doing them. There have been several books published with many worked-out problems in quantum mechanics done using *Mathematica* and other CASes [51, 52].

There are specialized packages for doing computations in quantum mechanics, for example, *SymPy* has extensive capabilities for symbolic quantum mechanics in the `sympy.physics.quantum` subpackage. At the base level, this subpackage has *Python* objects to represent the different mathematical objects relevant in quantum theory [53]: states (bras and kets), operators (unitary, Hermitian, etc.), and basis sets as well as operations on these objects such as tensor products, inner products, outer products, commutators, anticommutators, etc. The base objects are designed in the most general way possible to enable any particular quantum system to be implemented by subclassing the base operators to provide system specific logic. There is a general purpose `qapply` function that is capable of applying operators to states symbolically as well as simplifying a wide range of symbolic expressions involving different types of products and commutator/anticommutators. The state and operator objects also have a rich API for declaring their representation in a particular basis. This includes the ability to specify a basis for a multidimensional system using a complete set of commuting Hermitian operators.

On top of this base set of objects, a number of specific quantum systems have been implemented. First, there is traditional algebra for quantum angular momentum [54]. This allows the different spin operators (S_x, S_y, S_z) and their eigenstates to be represented in any basis and for any spin quantum number. Facilities for Clebsch-Gordan coefficients, Wigner coefficients, rotations, and angular momentum coupling are also present in their symbolic and numerical forms. Other examples of particular quantum systems that are implemented include second quantization, the simple harmonic oscillator (position/momentum and raising/lowering forms), and continuous position/momentum-based systems.

Second there is a full set of states and operators for symbolic quantum computing [55]. Multidimensional qubit states can be represented symbolically and as vectors. A full set of one (X, Y, Z, H, etc.) and two qubit (*CNOT, SWAP, CPHASE,* etc.) gates (unitary operators) are provided. These can be represented as matrices (sparse or dense) or made to act on qubits symbolically without representation. With these gates, it is possible to implement a number of basic quantum circuits including the quantum Fourier transform, quantum error correction, quantum teleportation, Grover's algorithm, dense coding, etc.

There are other packages that specialize in quantum computing, for example, [56].

Number Theory

Number theory provides an important base for modern computing, especially in cryptography and coding theory [57]. For example, LLL [58] algorithm is used in integer programming; primality testing and factoring algorithms are used in cryptography [59]. CASes are heavily used in these calculations.

Riemann hypothesis [60, 61] which implies results about the distribution of prime numbers has important applications in computational mathematics since it can be used to estimate how long certain algorithms take to run [61]. Riemann hypothesis states that all nontrivial zeros of the Riemann zeta function, defined for complex variable s defined in the half-plane $\Re(s) > 1$ by the absolutely convergent series $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$, have real part equal to $\frac{1}{2}$. In 1986, this was proven for the first 1,500,000,001 nontrivial zeros using computational methods [62]. Sebastian Wedeniwski using ZettaGrid (a distributed computing project to find roots of the zeta function) verified the result for the first 400 billion zeros in 2005 [63].

Teaching Calculus and Other Classes

Computer algebra systems are extremely useful for teaching calculus [64] as well as other classes where tedious symbolic algebra is needed, such as many physics classes (general relativity, quantum mechanics and field theory, symbolic solutions to partial differential equations, e.g., in electromagnetism, fluids, plasmas, electrical circuits, etc.) [65].

Experimental Mathematics

One field that would not be possible at all without computer algebra systems is called "experimental mathematics" [66], where CASes and related tools are used to "experimentally" verify or suggest mathematical relations. For example, the famous Bailey–Borwein–Plouffe (BBP) formula

$$\pi = \sum_{k=0}^{\infty} \left[\frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right) \right] \tag{3}$$

was first discovered experimentally (using arbitrary-precision arithmetic and extensive searching using an integer relation algorithm), only then proved rigorously [67].

Another example is in [68] where the authors first numerically discovered and then proved that for rational x, y, the 2D Poisson potential function satisfies

$$\psi(x, y) = \frac{1}{\pi^2} \sum_{a,b \text{ odd}} \frac{\cos(a\pi x)\cos(b\pi y)}{a^2 + b^2} = \frac{1}{\pi} \log \alpha \tag{4}$$

where α is algebraic (a root of an integer polynomial).

Acknowledgements This work was performed under the auspices of the US Department of Energy by Los Alamos National Laboratory.

References

1. Watt, S.M.: Making computer algebra more symbolic. In: Dumas J.-G. (ed.) Proceedings of Transgressive Computing 2006: A Conference in Honor of Jean Della Dora, Facultad de Ciencias, Universidad de Granada, pp 43–49, April 2006
2. Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for computer algebra. In: Introduction to Computer Algebra. Kluwer Academic, Boston (1992)

3. Slagle, J.R.: A heuristic program that solves symbolic integration problems in freshman calculus. J. ACM **10**(4), 507–520 (1963)
4. Veltman, M.J.G.: From weak interactions to gravitation. Int. J. Mod. Phys. A **15**(29), 4557–4573 (2000)
5. Strubbe, H.: Manual for SCHOONSCHIP a CDC 6000/7000 program for symbolic evaluation of algebraic expressions. Comput. Phys. Commun. **8**(1), 1–30 (1974)
6. Moses, J.: Symbolic integration. PhD thesis, MIT (1967)
7. REDUCE: A portable general-purpose computer algebra system. http://reduce-algebra.sourceforge.net/ (2013)
8. Fitch, J.: CAMAL 40 years on – is small still beautiful? Intell. Comput. Math., Lect. Notes Comput. Sci. **5625**, 32–44 (2009)
9. Moses, J.: Macsyma: a personal history. J. Symb. Comput. **47**(2), 123–130 (2012)
10. Shochat, D.D.: Experience with the musimp/mumath-80 symbolic mathematics system. ACM SIGSAM Bull. **16**(3), 16–23 (1982)
11. Bernardin, L., Chin, P. DeMarco, P., Geddes, K.O., Hare, D.E.G., Heal, K.M., Labahn, G., May, J.P., McCarron, J., Monagan, M.B., Ohashi, D., Vorkoetter, S.M.: Maple Programming Guide. Maplesoft, Waterloo (2011)
12. Wolfram, S.: The Mathematica Book. Wolfram Research Inc., Champaign (2000)
13. Grayson, D.R., Stillman, M.E.: Macaulay 2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/
14. The PARI Group, Bordeaux: PARI/GP version `2.7.0`. Available from http://pari.math.u-bordeaux.fr/ (2014)
15. The GAP Group: GAP – groups, algorithms, and programming, version 4.4. http://www.gap-system.org (2003)
16. Cannon, J.J.: An introduction to the group theory language cayley. Comput. Group Theory **145**, 183 (1984)
17. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language. J. Symb. Comput. **24**(3–4), 235–265 (1997). Computational Algebra and Number Theory. London (1993)
18. Stein, W.A., et al.: Sage Mathematics Software (Version 6.4.1). The Sage Development Team (2014). http://www.sagemath.org
19. SymPy Development Team: SymPy: Python library for symbolic mathematics. http://www.sympy.org (2013)
20. Maxima: Maxima, a computer algebra system. Version 5.34.1. http://maxima.sourceforge.net/ (2014)
21. Jenks, R.D., Sutor, R.S., Watt, S.M.: Scratchpad ii: an abstract datatype system for mathematical computation. In: Janßen, R. (ed.) Trends in Computer Algebra. Volume 296 of Lecture Notes in Computer Science, pp. 12–37. Springer, Berlin/Heidelberg (1988)
22. Jenks, R.D., Sutor, R.S.: Axiom – The Scientific Computation System. Springer, New York/Berlin etc. (1992)
23. MATLAB: Symbolic math toolbox. http://www.mathworks.com/products/symbolic/ (2014)
24. Parisse, B.: Giac/xcas, a free computer algebra system. Technical report, Technical report, University of Grenoble (2008)
25. FriCAS, an advanced computer algebra system: http://fricas.sourceforge.net/ (2015)
26. OpenAxiom, The open scientific computation platform: http://www.open-axiom.org/ (2015)
27. Wikipedia: List of computer algebra systems. http://en.wikipedia.org/wiki/List_of_computer_algebra_systems (2013)
28. Decker, W., Greuel, G.-M., Pfister, G., Schönemann, H.: SINGULAR 3-1-6 – a computer algebra system for polynomial computations. http://www.singular.uni-kl.de (2012)
29. Bauer, C., Frink, A., Kreckel, R.: Introduction to the ginac framework for symbolic computation within the c++ programming language. J. Symb. Comput. **33**(1), 1–12 (2002)
30. Peeters, K.: Introducing cadabra: a symbolic computer algebra system for field theory problems. arxiv:hep-th/0701238; A field-theory motivated approach to symbolic computer algebra. Comput. Phys. Commun. **176**, 550 (2007). [arXiv:cs/0608005]. - 14
31. Pérez, F., Granger, B.E.: IPython: a system for interactive scientific computing. Comput. Sci. Eng. **9**(3), 21–29 (2007)
32. Swiginac, Python interface to GiNaC: http://sourceforge.net/projects/swiginac.berlios/ (2015)
33. Pynac, derivative of GiNaC with Python wrappers: http://pynac.org/ (2015)
34. Barton, D., Fitch, J.P.: Applications of algebraic manipulation programs in physics. Rep. Prog. Phys. **35**(1), 235–314 (1972)
35. Vermaseren, J.A.M.: New features of FORM. Math. Phys. e-prints (2000). ArXiv:ph/0010025
36. Mertig, R., Böhm, M., Denner, A.: Feyn calc – computer-algebraic calculation of feynman amplitudes. Comput. Phys. Commun. **64**(3), 345–359 (1991)
37. Hahn, T.: Generating feynman diagrams and amplitudes with feynarts 3. Comput. Phys. Commun. **140**(3), 418–431 (2001)
38. Hahn, T., Pérez-Victoria, M.: Automated one-loop calculations in four and d dimensions. Comput. Phys. Commun. **118**(2–3), 153–165 (1999)
39. Gede, G., Peterson, D.L., Nanjangud, A.S., Moore, J.K., Hubbard, M.: Constrained multibody dynamics with python: from symbolic equation generation to publication. In: ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, pp. V07BT10A051–V07BT10A051. American Society of Mechanical Engineers, San Diego (2013)
40. NumPy: The fundamental package for scientific computing in Python. http://www.numpy.org/ (2015)
41. Hunter, J.D.: Matplotlib: a 2d graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007)
42. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: open source scientific tools for Python (2001–). http://www.scipy.org/
43. Gastineau, M., Laskar, J.: Development of TRIP: fast sparse multivariate polynomial multiplication using burst tries. In: Computational Science – ICCS 2006, University of Reading, pp. 446–453 (2006)
44. Biscani, F.: Parallel sparse polynomial multiplication on modern hardware architectures. In: Proceedings of the 37th International Symposium on Symbolic and Algebraic Computation – ISSAC '12, (1), Grenoble, pp. 83–90 (2012)
45. Shelus, P.J., Jefferys III, W.H.: A note on an attempt at more efficient Poisson series evaluation. Celest. Mech. **11**(1), 75–78 (1975)
46. Jefferys, W.H.: A FORTRAN-based list processor for Poisson series. Celest. Mech. **2**(4), 474–480 (1970)

47. Broucke, R., Garthwaite, K.: A programming system for analytical series expansions on a computer. Celest. Mech. **1**(2), 271–284 (1969)
48. Fateman, R.J.: On the multiplication of poisson series. Celest. Mech. **10**(2), 243–247 (1974)
49. Danby, J.M.A., Deprit, A., Rom, A.R.M.: The symbolic manipulation of poisson series. In: SYMSAC '66 Proceedings of the First ACM Symposium on Symbolic and Algebraic Manipulation, New York, pp. 0901–0934 (1965)
50. Bourne, S.R.: Literal expressions for the co-ordinates of the moon. Celest. Mech. **6**(2), 167–186 (1972)
51. Feagin, J.M.: Quantum Methods with Mathematica. Springer, New York (2002)
52. Steeb, W.-H., Hardy, Y.: Quantum Mechanics Using Computer Algebra: Includes Sample Programs in C++, SymbolicC++, Maxima, Maple, and Mathematica. World Scientific, Singapore (2010)
53. Sakurai, J.J., Napolitano, J.J.: Modern Quantum Mechanics. Addison-Wesley, Boston (2010)
54. Zare, R.N.: Angular Momentum: Understanding Spatial Aspects in Chemistry and Physics. Wiley, New York (1991)
55. Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information. Cambridge University Press, Cambridge (2011)
56. Gerdt, V.P., Kragler, R., Prokopenya, A.N.: A Mathematica Package for Simulation of Quantum Computation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5743 LNCS, pp. 106–117. Springer, Berlin/Heidelberg (2009)
57. Shoup, V.: A Computational Introduction to Number Theory and Algebra. Cambridge University Press, Cambridge (2009)
58. Lenstra, A.K., Lenstra, H.W., Lovász, L.: Factoring polynomials with rational coefficients. Mathematische Annalen **261**(4), 515–534 (1982)
59. Cohen, H.: A Course in Computational Algebraic Number Theory, vol. 138. Springer, Berlin/New York (1993)
60. Titchmarsh, E.C.: The Theory of the Riemann Zeta-Function, vol. 196. Oxford University Press, Oxford (1951)
61. Mazur, B., Stein, W.: Prime Numbers and the Riemann Hypothesis. Cambridge University Press, Cambridge (2015)
62. van de Lune, J., Te Riele, H.J.J., Winter, D.T.: On the zeros of the riemann zeta function in the critical strip. iv. Math. Comput. **46**(174), 667–681 (1986)
63. Wells, D.: Prime Numbers: The Most Mysterious Figures in Math. Wiley, Hoboken (2011)
64. Palmiter, J.R.: Effects of computer algebra systems on concept and skill acquisition in calculus. J. Res. Math. Educ. **22**, 151–156 (1991)
65. Bing, T.J., Redish, E.F.: Symbolic manipulators affect mathematical mindsets. Am. J. Phys. **76**(4), 418–424 (2008)
66. Bailey, D.H., Borwein, J.M.: Experimental mathematics: examples, methods and implications. Not. AMS **52**, 502–514 (2005)
67. Bailey, D.H., Borwein, P., Plouffe, S.: On the rapid computation of various polylogarithmic constants. Math. Comput. **66**(218), 903–913 (1996)
68. Bailey, D.H., Borwein, J.M.: Lattice sums arising from the Poisson equation. J. Phys. A: Math. Theor. **3**, 1–18 (2013)

Symmetric Methods

Philippe Chartier
INRIA-ENS Cachan, Rennes, France

Synonyms

Time reversible

Definition

This entry is concerned with *symmetric methods* for solving ordinary differential equations (ODEs) of the form

$$\dot{y} = f(y) \in \mathbb{R}^n, \quad y(0) = y_0. \tag{1}$$

Throughout this article, we denote by $\varphi_{t,f}(y_0)$ the flow of equation (1) with vector field f, i.e., the exact solution at time t with initial condition $y(0) = y_0$, and we assume that the conditions for its well definiteness and smoothness for $(y_0, |t|)$ in an appropriate subset Ω of $\mathbb{R}^n \times \mathbb{R}_+$ are satisfied. Numerical methods for (1) implement numerical flows $\Phi_{h,f}$ which, for **small enough** *stepsizes* h, approximate $\varphi_{h,f}$. Of central importance in the context of symmetric methods is the concept of *adjoint method*.

Definition 1 The adjoint method $\Phi^*_{h,f}$ is the inverse of $\Phi_{t,f}$ with reversed time step $-h$:

$$\Phi^*_{h,f} := \Phi^{-1}_{-h,f} \tag{2}$$

A numerical method Φ_h is then said to be symmetric if $\Phi_{h,f} = \Phi^*_{h,f}$.

Overview

Symmetry is an essential property of numerical methods with regard to the *order* of accuracy and *geometric* properties of the solution. We briefly discuss the implications of these two aspects and refer to the corresponding sections for a more involved presentation:

- A method $\Phi_{h,f}$ is said to be of order p if

$$\Phi_{h,f}(y) = \varphi_{h,f}(y) + \mathcal{O}(h^{p+1}),$$

S

$$\begin{tikzcd}
\mathbb{R}^n \arrow[r, "f"] \arrow[d, "\rho"'] & \mathbb{R}^n \arrow[d, "\rho"] & \mathbb{R}^n \arrow[r, "\varphi_{t,f}"] \arrow[d, "\rho"'] & \mathbb{R}^n \arrow[d, "\rho"] & \mathbb{R}^n \arrow[r, "\Phi_{h,f}"] \arrow[d, "\rho"'] & \mathbb{R}^n \arrow[d, "\rho"] \\
\mathbb{R}^n & \mathbb{R}^n \arrow[l, "(-f)"] & \mathbb{R}^n & \mathbb{R}^n \arrow[l, "\varphi_{t,f}^{-1}"] & \mathbb{R}^n & \mathbb{R}^n \arrow[l, "\Phi_{h,f}^{-1}"]
\end{tikzcd}$$

Symmetric Methods, Fig. 1 ρ-reversibility of f, $\varphi_{t,f}$ and $\Phi_{h,f}$

and, if the *local error* has the following first-term expansion

$$\Phi_{h,f}(y) = \varphi_{h,f}(y) + h^{p+1}C(y) + \mathcal{O}(h^{p+2}),$$

then straightforward application of the implicit function theorem leads to

$$\Phi^*_{h,f}(y) = \varphi_{h,f}(y) - (-h)^{p+1}C(y) + \mathcal{O}(h^{p+2}).$$

This implies that **a symmetric method is necessarily of even order** $p = 2q$, since $\Phi_{h,f}(y) = \Phi^*_{h,f}(y)$ means that $(1 + (-1)^{p+1})C(y) = 0$. This property plays a key role in the construction of *composition* methods by *triple jump techniques* (see section on "Symmetric Methods Obtained by Composition"), and this is certainly no coincidence that Runge-Kutta methods of *optimal* order (Gauss methods) are symmetric (see section on "Symmetric Methods of Runge-Kutta Type"). It also explains why symmetric methods are used in conjunction with (Richardson) extrapolation techniques.

- The exact flow $\varphi_{t,f}$ is itself symmetric owing to the *group property* $\varphi_{s+t,f} = \varphi_{s,f} \circ \varphi_{t,f}$. Consider now an isomorphism ρ of the vector space $\mathbb{R}^n$ (the phase space of (1)) and assume that the vector field f satisfies the relation $\rho \circ f = -f \circ \rho$ (see Fig. 1). Then, $\varphi_{t,f}$ is said to be ρ-reversible, that is it to say the following equality holds:

$$\rho \circ \varphi_{t,f} = \varphi_{t,f}^{-1} \circ \rho \qquad (3)$$

Example 1 Hamiltonian systems

$$\dot{y} = \frac{\partial H}{\partial z}(y,z)$$
$$\dot{z} = -\frac{\partial H}{\partial y}(y,z)$$

with a Hamiltonian function $H(q,p)$ satisfying $H(y,-z) = H(y,z)$ are ρ-reversible for $\rho(y,z) = (y,-z)$.

Definition 2 A method Φ_h, applied to a ρ-reversible ordinary differential equation, is said to be ρ-reversible if

$$\rho \circ \Phi_{h,f} = \Phi_{h,f}^{-1} \circ \rho.$$

Note that if $\Phi_{h,f}$ is symmetric, it is ρ-reversible if and only if the following condition holds:

$$\rho \circ \Phi_{h,f} = \Phi_{-h,f} \circ \rho. \qquad (4)$$

Besides, if (4) holds for an invertible ρ, then $\Phi_{h,f}$ is ρ-reversible if and only if it is symmetric.

Example 2 The trapezoidal rule, whose flow is defined by the *implicit* equation

$$\Phi_{h,f}(y) = y + hf\left(\frac{1}{2}y + \frac{1}{2}\Phi_{h,f}(y)\right), \qquad (5)$$

is symmetric and is ρ-reversible when applied to ρ-reversible f.

Since most numerical methods satisfy relation (4), symmetry is the required property for numerical methods to share with the exact flow not only time reversibility but also ρ-reversibility. This illustrates that **a symmetric method mimics geometric properties of the exact flow**. *Modified differential equations* sustain further this assertion (see next section) and allow for the derivation of deeper results for *integrable reversible* systems such as the **preservation of invariants and the linear growth of errors** by symmetric methods (see section on "Reversible Kolmogorov-Arnold–Moser Theory").

Modified Equations for Symmetric Methods

Constant stepsize backward error analysis. Considering a numerical method Φ_h (not necessarily symmetric) and the sequence of approximations obtained by application of the formula $y_{n+1} = \Phi_{h,f}(y_n), n = 0, 1, 2, \ldots$, from the initial value y_0, the idea of *backward error analysis* consists in searching for a *modified vector field* f_h^N such that

$$\varphi_{h,f_h^N}(y_0) = \Phi_{h,f}(y_0) + \mathcal{O}(h^{N+2}), \qquad (6)$$

where the modified vector field, *uniquely* defined by a Taylor expansion of (6), is of the form

$$f_h^N(y) = f(y) + hf_1(y) + h^2 f_2(y) + \ldots + h^N f_N(y). \tag{7}$$

Theorem 1 *The modified vector field of a symmetric method $\Phi_{h,f}$ has an expansion in even powers of h, i.e., $f_{2j+1} \equiv 0$ for $j = 0, 1, \ldots$ Moreover, if f and $\Phi_{h,f}$ are ρ-reversible, then f_h^N is ρ-reversible as well for any $N \geq 0$.*

Proof. Reversing the time step h in (6) and taking the inverse of both sides, we obtain

$$(\varphi_{-h,f_{-h}^N})^{-1}(y_0) = (\Phi_{-h,f})^{-1}(y_0) + \mathcal{O}(h^{N+2}).$$

Now, the group property of exact flows implies that $(\varphi_{-h,f_{-h}^N})^{-1}(y_0) = \varphi_{h,f_{-h}^N}(y_0)$, so that

$$\varphi_{h,f_{-h}^N}(y_0) = \Phi_{h,f}^*(y_0) + \mathcal{O}(h^{N+2}),$$

and by uniqueness, $(f_h^N)^* = f_{-h}^N$. This proves the first statement. Assume now that f is ρ-reversible, so that (4) holds. It follows from $f_{-h}^N = f_h^N$ that

$$\begin{aligned} \rho \circ \varphi_{-h,f_h^N} &= \rho \circ \varphi_{-h,f_{-h}^N} \overset{\mathcal{O}(h^{N+2})}{=} \rho \circ \Phi_{-h,f} \\ &= \Phi_{h,f} \circ \rho \overset{\mathcal{O}(h^{N+2})}{=} \varphi_{h,f_h^N} \circ \rho, \end{aligned}$$

where the second and last equalities are valid up to $\mathcal{O}(h^{N+2})$-error terms. Yet the group property then implies that $\rho \circ \varphi_{-nh,f_h^N} = \varphi_{nh,f_h^N} \circ \rho + \mathcal{O}_n(h^{N+2})$ where the constant in the $\mathcal{O}_n$-term depends on n and an interpolation argument shows that for fixed N and small $|t|$

$$\rho \circ \varphi_{-t,f_h^N} = \varphi_{t,f_h^N} \circ \rho + \mathcal{O}(h^{N+1}),$$

where the $\mathcal{O}$-term depends smoothly on t and on N. Finally, differentiating with respect to t, we obtain

$$\begin{aligned} -\rho \circ f_h^N &= \frac{d}{dt}\rho \circ \varphi_{-t,f_h^N}\Big|_{t=0} = \frac{d}{dt}\varphi_{t,f_h^N} \circ \rho\Big|_{t=0} \\ &\quad + \mathcal{O}(h^{N+2}) = f_h^N \circ \rho + \mathcal{O}(h^{N+1}), \end{aligned}$$

and consequently $-\rho \circ f_h^N = f_h^N \circ \rho$. □

Remark 1 The expansion (7) of the modified vector field f_h^N can be computed explicitly at any order N with the *substitution product* of *B-series* [2].

Example 3 Consider the Lotka-Volterra equations in Poisson form

$$\begin{pmatrix} \dot u \\ \dot v \end{pmatrix} = \begin{pmatrix} 0 & uv \\ -uv & 0 \end{pmatrix} \begin{pmatrix} \nabla_u H(u,v) \\ \nabla_v H(u,v) \end{pmatrix},$$

$$H(u,v) = \log(u) + \log(v) - u - v,$$

i.e., $y' = f(y)$ with $f(y) = (u(1-v), v(u-1))^T$. Note that $\rho \circ f = -f \circ \rho$ with $\rho(u,v) = (v,u)$. The modified vector fields $f_{h,\mathrm{iE}}^2$ for the *implicit Euler* method and $f_{h,\mathrm{mr}}^2$ for the implicit midpoint rule read (with $N = 2$)

$$f_{h,\mathrm{iE}}^2 = f + \frac{1}{2}hf'f + \frac{h^2}{12}f''(f,f) + \frac{h^2}{3}f'f'f$$

$$\text{and } f_{h,\mathrm{mr}}^2 = f - \frac{h^2}{24}f''(f,f) + \frac{h^2}{12}f'f'f.$$

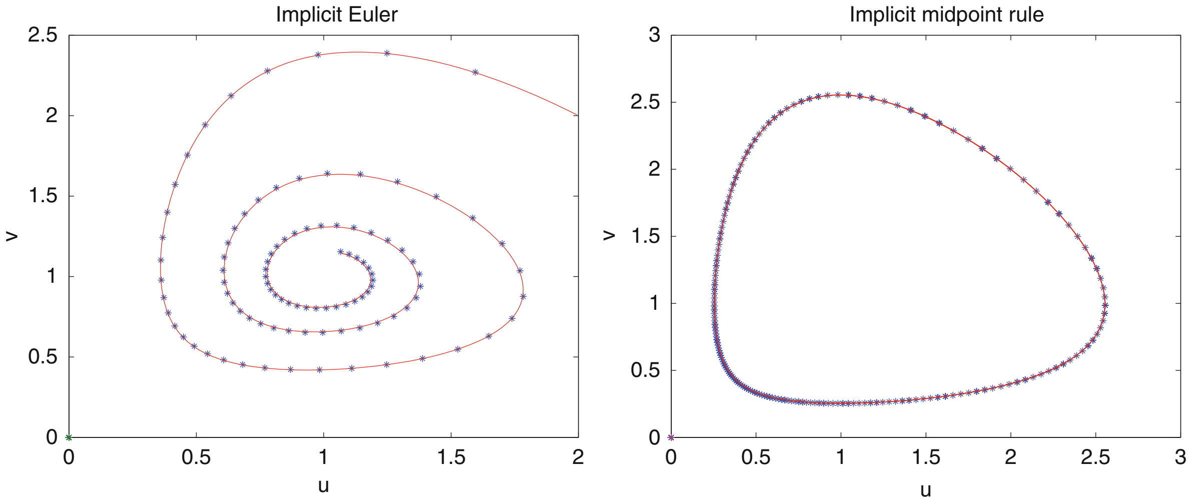
The exact solutions of the modified ODEs are plotted on Fig. 2 together with the corresponding numerical solution. Though the modified vector fields are truncated only at second order, the agreement is excellent. The difference of the behavior of the two solutions is also striking: only the symmetric method captures the periodic nature of the solution. (The good behavior of the midpoint rule cannot be attributed to its *symplecticity* since the system is a noncanonical Poisson system.) This will be further explored in the next section.

Variable stepsize backward error analysis. In practice, it is often fruitful to resort to variable stepsize implementations of the numerical flow $\Phi_{h,f}$. In accordance with [17], we consider stepsizes that are proportional to a function $\epsilon s(y,\epsilon)$ depending only on the current state y and of a parameter ϵ prescribed by the user and aimed at controlling the error. The approximate solution is then given by

$$y_{n+1} = \Phi_{\epsilon s(y_n,\epsilon),f}(y_n), \quad n = 0, \ldots.$$

A remarkable feature of this algorithm is that it preserves the symmetry of the exact solution as soon as $\Phi_{h,f}$ is symmetric and s satisfies the relation

$$s(\Phi_{\epsilon s(y,\epsilon),f}(y), -\epsilon) = s(y,\epsilon)$$



Symmetric Methods, Fig. 2 Exact solutions of modified equations (*red lines*) versus numerical solutions by implicit Euler and midpoint rule (*blue points*)

and preserves the ρ-reversibility as soon as $\Phi_{h,f}$ is ρ-reversible and satisfies the relation

$$s(\rho^{-1} \circ \Phi_{\epsilon s(y,\epsilon),f}(y), -\epsilon) = s(y, \epsilon).$$

A result similar to Theorem 1 then holds with h replaced by ϵ.

Remark 2 A recipe to construct such a function s, suggested by Stoffer in [17], consists in requiring that the local error estimate is kept constantly equal to a tolerance parameter. For the details of the implementation, we refer to the original paper or to Chap. VIII.3 of [10].

Reversible Kolmogorov-Arnold-Moser Theory

The theory of *integrable Hamiltonian* systems has its counterpart for *reversible integrable* ones. A reversible system

$$\dot{y} = f(y,z),\ \dot{z} = g(y,z) \quad \text{where} \quad \rho \circ (f,g)$$
$$= -(f,g) \circ \rho \quad \text{with} \quad \rho(y,z) = (y,-z), \qquad (8)$$

is reversible integrable if it can be brought, through a reversible transformation $(a,\theta) = (I(y,z), \Theta(y,z))$, to the *canonical* equations

$$\dot{a} = 0, \quad \dot{\theta} = \omega(a).$$

An interesting instance is the case of *completely integrable Hamiltonian* systems:

$$\dot{y} = \frac{\partial H}{\partial z}(y,z), \quad \dot{z} = -\frac{\partial H}{\partial y}(y,z),$$

with first integrals I_j's in involution (That is to say such that $(\nabla_y I_i)\cdot(\nabla_z I_j) = (\nabla_z I_i)\cdot(\nabla_y I_j)$ for all i, j.) such that $I_j \circ \rho = I_j$. In the conditions where Arnold-Liouville theorem (see Chap. X.1.3. of [10]) can be applied, then, under the additional assumption that

$$\exists (y^*, 0) \in \{(y,z), \forall j, I_j(y,z) = I_j(y_0,z_0)\}, \qquad (9)$$

such a system is reversible integrable. In this situation, ρ-reversible methods constitute a very interesting way around symplectic method, as the following result shows:

Theorem 2 *Let $\Phi_{h,(f,g)}$ be a reversible numerical method of order p applied to an integrable reversible system (8) with real-analytic f and g. Consider $a^\bullet = (I_1(y^\bullet,z^\bullet),\dots,I_d(y^\bullet,z^\bullet))$: If the condition*

$$\forall k \in \mathbb{Z}^d/\{0\},\ |k \cdot \omega(a^\bullet)| \geq \gamma \left(\sum_{i=1}^{d} |k_i|\right)^{-\nu}$$

is satisfied for some positive constants γ and ν, then there exist positive C, c, and h_0 such that the following assertion holds:

$$\forall h \leq h_0, \forall (x_0, y_0) \textit{ such that } \max_{j=1,\ldots,d} |I_j(y_0, z_0) - a^\bullet| \leq c|\log h|^{-\nu-1}, \tag{10}$$

$$\forall t = nh \leq h^{-p}, \begin{cases} \|\Phi^n_{h,(f,g)}(x_0, y_0) - (y(t), z(t))\| \leq Cth^p \\ |I_j(\Phi^n_{h,(f,g)}(y_0, z_0)) - I_j(y_0, z_0)| \leq Ch^p \textit{ for all } j. \end{cases}$$

Analogously to symplectic methods, ρ-reversible methods thus preserve invariant tori $I_j = cst$ over long intervals of times, and the error growth is linear in t. Remarkably and in contrast with symplectic methods, this result remains valid for reversible variable stepsize implementations (see Chap. X.I.3 of [10]). However, it is important to note that for a Hamiltonian reversible system, the Hamiltonian ceases to be preserved when condition (9) is not fulfilled. This situation is illustrated on Fig. 3 for the Hamiltonian system with $H(q, p) = \frac{1}{2}p^2 + \cos(q) + \frac{1}{5}\sin(2q)$, an example borrowed from [4].

Symmetric Methods of Runge-Kutta Type

Runge-Kutta methods form a popular class of numerical integrators for (1). Owing to their importance in applications, we consider general systems (1) and subsequently partitioned systems.

Methods for general systems. We start with the following:

Definition 3 Consider a matrix $A = (a_{i,j}) \in \mathbb{R}^s \times \mathbb{R}^s$ and a vector $b = (b_j) \in \mathbb{R}^s$. The Runge-Kutta method denoted (A, b) is defined by

$$Y_i = y + h \sum_{j=1}^{s} a_{i,j} f(Y_j), \quad i = 1, \ldots, s \tag{11}$$

$$\tilde{y} = y + h \sum_{j=1}^{s} b_j f(Y_j). \tag{12}$$

Note that strictly speaking, the method is properly defined only for small $|h|$. In this case, the corresponding numerical flow $\Phi_{h,f}$ maps y to $\tilde{y}$. Vector Y_i approximates the solution at intermediate point $t_0 + c_i h$, where $c_i = \sum_j a_{i,j}$, and it is customary since [1] to represent a method by its *tableau*:

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \ldots & a_{1,s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & \ldots & a_{s,s} \\ \hline & b_1 & \ldots & b_s \end{array} \tag{13}$$

Runge-Kutta methods automatically satisfy the ρ-compatibility condition (4): changing h into $-h$ in (11) and (12), we have indeed by linearity of ρ and by using $\rho \circ f = -f \circ \rho$

$$\rho(Y_i) = \rho(y) - h \sum_{j=1}^{s} a_{i,j} f\Big(\rho(Y_j)\Big), \quad i = 1, \ldots, s$$

$$\rho(\tilde{y}) = \rho(y) - h \sum_{j=1}^{s} b_j f\Big(\rho(Y_j)\Big).$$

By construction, this is $\rho(\Phi_{-h,f}(y))$ and by previous definition $\Phi_{h,f}(\rho(y))$. As a consequence, ρ-reversible Runge-Kutta methods coincide with symmetric methods. Nevertheless, symmetry requires an additional algebraic condition stated in the next theorem:

Theorem 3 *A Runge-Kutta method (A, b) is symmetric if*

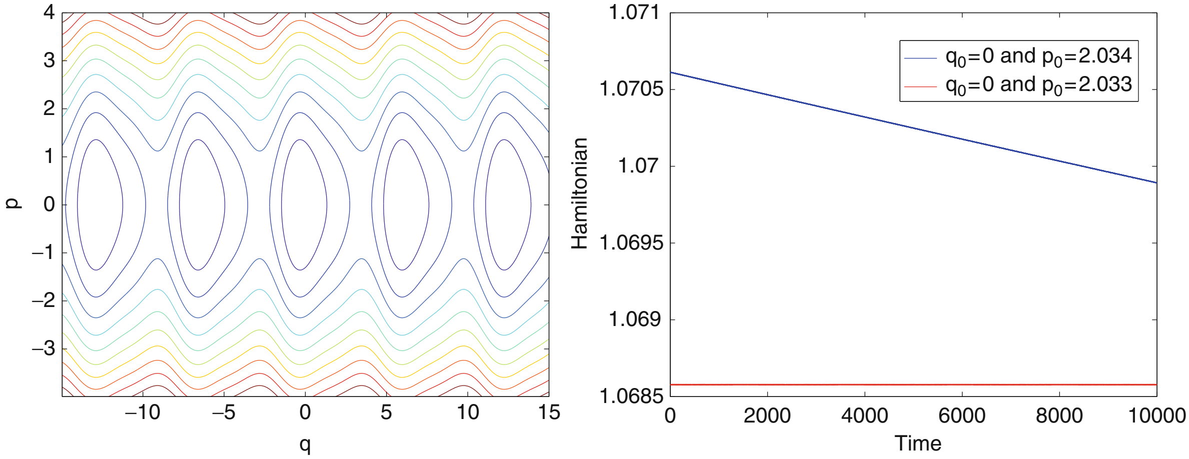
$$PA + AP = eb^T \textit{ and } b = Pb, \tag{14}$$

where $e = (1, \ldots, 1)^T \in \mathbb{R}^s$ and P is the permutation matrix defined by $p_{i,j} = \delta_{i,s+1-j}$.

Proof. Denoting $Y = \Big(Y_1^T, \ldots, Y_s^T\Big)^T$ and $F(Y) = \Big(f(Y_1)^T, \ldots, f(Y_s)^T\Big)^T$, a more compact form for (11) and (12) is

$$Y = e \otimes y + h(A \otimes I)F(Y), \tag{15}$$

$$\tilde{y} = y + h(b^T \otimes I)F(Y). \tag{16}$$



Symmetric Methods, Fig. 3 Level sets of H (*left*) and evolution of H w.r.t. time for two different initial values

On the one hand, premultiplying (15) by $P \otimes I$ and noticing that

$$(P \otimes I)F(Y) = F\Big((P \otimes I)Y\Big),$$

it is straightforward to see that $\Phi_{h,f}$ can also be defined by coefficients PAP^T and Pb. On the other hand, exchanging h and $-h$, y, and $\tilde{y}$, it appears that $\Phi^*_{h,f}$ is defined by coefficients $A^* = eb^T - A$ and $b^* = b$. The flow $\Phi_{h,f}$ is thus symmetric as soon as $eb^T - A = PAP$ and $b = Pb$, which is nothing but condition (14). □

Remark 3 For methods without redundant stages, condition (14) is also necessary.

Example 4 The *implicit midpoint rule*, defined by $A = \frac{1}{2}$ and $b = 1$, is a symmetric method of order 2. More generally, the s-stage Gauss collocation method based on the roots of the sth shifted Legendre polynomial is a symmetric method of order $2s$. For instance, the 2-stage and 3-stage Gauss methods of orders 4 and 6 have the following coefficients:

$$\begin{array}{c|cc}
\frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\
\frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}$$

$$\begin{array}{c|ccc}
\frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\
\frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array} \tag{17}$$

Methods for partitioned systems. For systems of the form

$$\dot{y} = f(z), \quad \dot{z} = g(y), \tag{18}$$

it is natural to apply two different Runge-Kutta methods to variables y and z: Written in compact form, a partitioned Runge-Kutta method reads:

$$\begin{aligned}
Y &= e \otimes y + h(A \otimes I)F(Z), \\
Z &= e \otimes y + h(\hat{A} \otimes I)G(Y), \\
\tilde{y} &= y + h(b^T \otimes I)F(Z), \\
\tilde{z} &= y + h(\hat{b}^T \otimes I)G(Y),
\end{aligned}$$

and the method is symmetric if both (A, b) and $(\hat{A}, \hat{b})$ are. An important feature of partitioned Runge-Kutta method is that they can be symmetric and *explicit* for systems of the form (18).

Example 5 The Verlet method is defined by the following two Runge-Kutta tableaux:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \text{and} \quad \begin{array}{c|cc} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad (19)$$

The method becomes explicit owing to the special structure of the partitioned system:

$$\begin{array}{llll} Y_1 = y_0, & & Z_1 = z_0 + \frac{h}{2} f(Y_1), \\ Y_2 = y_0 + hg(Z_1), & & Z_2 = Z_1, \\ y_1 = Y_2, & & z_1 = z_0 + \frac{h}{2}\Big(f(Y_1) + f(Y_2)\Big) \end{array}$$

The Verlet method is the most elementary method of the class of partitioned Runge-Kutta methods known as Lobatto IIIA-IIIB. Unfortunately, methods of higher orders within this class are no longer explicit in general, even for the equations of the form (18). It is nevertheless possible to construct symmetric explicit Runge-Kutta methods, which turn out to be equivalent to compositions of Verlet's method and whose introduction is for this reason postponed to the next section.

Note that a particular instance of partitioned systems are second-order differential equations of the form

$$\dot{y} = z, \quad \dot{z} = g(y), \qquad (20)$$

which covers many situations of practical interest (for instance, mechanical systems governed by Newton's law in absence of friction).

Symmetric Methods Obtained by Composition

Another class of symmetric methods is constituted of symmetric *compositions* of low-order methods. The idea consists in applying a basic method $\Phi_{h,f}$ with a sequence of prescribed stepsizes: Given s real numbers $\gamma_1, \ldots, \gamma_s$, its composition with stepsizes $\gamma_1 h, \ldots, \gamma_s h$ gives rise to a new method:

$$\Psi_{h,f} = \Phi_{\gamma_s h,f} \circ \ldots \circ \Phi_{\gamma_1 h,f}. \qquad (21)$$

Noticing that the local error of $\Psi_{h,f}$, defined by $\Psi_{h,f}(y) - \varphi_{h,f}(y)$, is of the form

$$(\gamma_1^{p+1} + \ldots + \gamma_s^{p+1}) h^{p+1} C(y) + \mathcal{O}(h^{p+2}),$$

as soon as $\gamma_1 + \ldots + \gamma_s = 1$, $\Psi_{h,f}$ is of order at least $p+1$ if

$$\gamma_1^{p+1} + \ldots + \gamma_s^{p+1} = 0.$$

This observation is the key to *triple jump* compositions, as proposed by a series of authors [3,5,18,21]: Starting from a symmetric method $\Phi_{h,f}$ of (even) order $2q$, the new method obtained for

$$\gamma_1 = \gamma_3 = \frac{1}{2 - 2^{1/(2q+1)}} \quad \text{and}$$
$$\gamma_2 = \frac{2^{1/(2q+1)}}{2 - 2^{1/(2q+1)}}$$

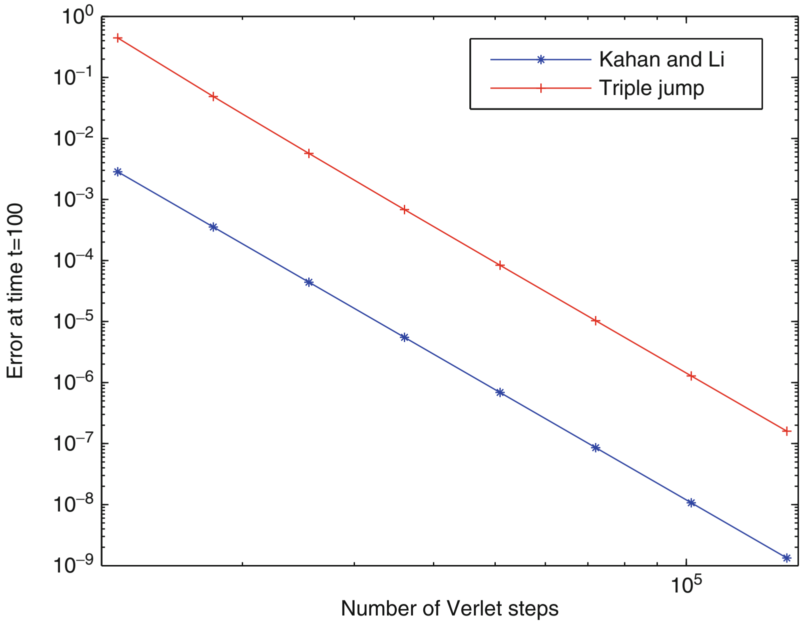
is symmetric

$$\begin{aligned} \Psi^*_{h,f} &= \Phi^*_{\gamma_1 h,f} \circ \Phi^*_{\gamma_2 h,f} \circ \Phi^*_{\gamma_3 h,f} \\ &= \Phi_{\gamma_3 h,f} \circ \Phi_{\gamma_2 h,f} \circ \Phi_{\gamma_1 h,f} = \Psi_{h,f} \end{aligned}$$

and of order at least $2q+1$. Since the order of a symmetric method is even, $\Psi_{h,f}$ is in fact of order $2q+2$. The procedure can then be repeated recursively to construct arbitrarily high-order symmetric methods of orders $2q+2, 2q+4, 2q+6, \ldots$., with respectively $3, 9, 27, \ldots$, compositions of the original basic method $\Phi_{h,f}$. However, the construction is far from being the most efficient, for the combined coefficients become large, some of which being negatives. A partial remedy is to envisage compositions with $s = 5$. We hereby give the coefficients obtained by Suzuki [18]:

$$\gamma_1 = \gamma_2 = \gamma_4 = \gamma_5 = \frac{1}{4 - 4^{1/(2q+1)}} \quad \text{and}$$
$$\gamma_3 = -\frac{4^{1/(2q+1)}}{4 - 4^{1/(2q+1)}}$$

which give rise to very efficient methods for $q = 1$ and $q = 2$. The most efficient high-order composition methods are nevertheless obtained by solving the full system of order conditions, i.e., by raising the order directly from 2 to 8, for instance, without going through the intermediate steps described above. This requires much more effort though, first to derive the order conditions and then to solve the resulting polynomial system. It is out of the scope of this article to describe the two steps involved, and we rather refer to the paper [15] on the use of ∞B-series for order conditions and to Chap. V.3.2. of [10] for various examples and numerical comparisons. An excellent method of order 6 with 9 stages has been obtained by Kahan and Li [12] and we reproduce here its coefficients:



$$\gamma_1 = \gamma_9 = 0.3921614440073141,$$
$$\gamma_2 = \gamma_8 = 0.3325991367893594,$$
$$\gamma_3 = \gamma_7 = -0.7062461725576393,$$
$$\gamma_4 = \gamma_6 = 0.0822135962935508,$$
$$\gamma_5 = 0.7985439909348299.$$

For the sake of illustration, we have computed the solution of Kepler's equations with this method and the method of order 6 obtained by the triple jump technique. In both cases, the basic method is Verlet's scheme. The gain offered by the method of Kahan and Li is impressive (it amounts to two digits of accuracy on this example). Other methods can be found, for instance, in [10, 14].

Remark 4 It is also possible to consider symmetric compositions of nonsymmetric methods. In this situation, raising the order necessitates to compose the basic method and its adjoint.

Symmetric Methods for Highly Oscillatory Problems

In this section, we present methods aimed at solving problems of the form

$$\ddot{q} = -\nabla V_{fast}(q) - \nabla V_{slow}(q) \tag{22}$$

where V_{fast} and V_{slow} are two potentials acting on different time scales, typically such that $\nabla^2 V_{fast}$ is positive semi-definite and $\|\nabla^2 V_{fast}\| >> \|\nabla^2 V_{slow}\|$. Explicit standard methods suffer from severe stability restrictions due to the presence of high oscillations at the slow time scale and necessitate small steps and many evaluations of the forces. Since slow forces $-\nabla V_{slow}$ are in many applications much more expensive to evaluate than fast ones, efficient methods in this context are thus devised to require significantly fewer evaluations per step of the slow force.

Example 6 In applications to molecular dynamics, for instance, fast forces deriving from V_{fast} (short-range interactions) are much cheaper to evaluate than slow forces deriving from V_{slow} (long-range interactions). Other examples of applications are presented in [11].

Methods for general problems with nonlinear fast potentials. Introducing the variable $p = \dot{q}$ in (22), the equation reads

$$\underbrace{\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix}}_{\dot{y}} = \underbrace{\begin{pmatrix} p \\ 0 \end{pmatrix}}_{f_K(y)} + \underbrace{\begin{pmatrix} 0 \\ -\nabla_q V_{fast}(q) \end{pmatrix}}_{f_F(y)}$$

$$+\underbrace{\begin{pmatrix} 0 \\ -\nabla_q V_{slow}(q) \end{pmatrix}}_{f_S(y)}.$$

The usual Verlet method [20] would consist in composing the flows $\varphi_{h,(f_F+f_S)}$ and φ_{h,f_K} as follows:

$$\varphi_{\frac{h}{2},(f_F+f_S)} \circ \varphi_{h,f_K} \circ \varphi_{\frac{h}{2},(f_F+f_S)}$$

or, if necessary, numerical approximations thereof and would typically be restricted to very small stepsizes. The impulse method [6,8,19] combines the three pieces of the vector field differently:

$$\varphi_{\frac{h}{2},f_S} \circ \varphi_{h,(f_K+f_F)} \circ \varphi_{\frac{h}{2},f_S}.$$

Note that φ_{h,f_S} is explicit

$$\varphi_{h,f_S}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q \\ p - h\nabla_q V_{slow}(q) \end{pmatrix}$$

while $\varphi_{h,(f_K+f_F)}$ may require to be approximated by a numerical method $\Phi_{h,(f_K+f_F)}$ which uses stepsizes that are fractions of h. If $\Phi_{h,(f_K+f_F)}$ is symmetric (and/or symplectic), the overall method is symmetric as well (and/or symplectic) and allows for larger stepsizes. However, it still suffers from resonances and a better option is given by the mollified impulse methods, which considers the mollified potential $\bar{V}_{slow}(q) = V_{slow}(a(q))$ in loco of $V_{slow}(q)$, where $a(q)$ and $a'(q)$ are *averaged* values given by

$$a(q) = \frac{1}{h}\int_0^h x(s)ds, \quad a'(q) = \frac{1}{h}\int_0^h X(s)ds$$

where

$$\begin{aligned} \ddot{x} &= -\nabla V_{fast}(x), x(0) = q, \dot{x}(0) = p, \\ \ddot{X} &= -\nabla^2 V_{fast}(x)X, X(0) = I, \ \dot{X}(0) = 0. \end{aligned} \qquad (23)$$

The resulting method uses the mollified force $-a'(q)^T(\nabla_q V_{slow})(a(q))$ and is still symmetric (and/or symplectic) provided (23) is solved with a symmetric (and/or symplectic) method.

Methods for problems with quadratic fast potentials. In many applications of practical importance, the potential V_{fast} is quadratic of the form $V_{fast}(q) = \frac{1}{2}q^T\Omega^2 q$. In this case, the mollified impulse method falls into the class of trigonometric symmetric methods of the form

$$\Phi_h\begin{pmatrix} p \\ q \end{pmatrix} = R(h\Omega)\begin{pmatrix} p \\ q \end{pmatrix} - \frac{1}{2}h\begin{pmatrix} \psi_0(h\Omega)\nabla V_{slow}\Big(\phi(h\Omega)q_0\Big) + \psi_1(h\Omega)\nabla V_{slow}\Big(\phi(h\Omega)q_1\Big) \\ h\psi(h\Omega)\nabla V_{slow}\Big(\phi(h\Omega)q_0\Big) \end{pmatrix}$$

where $R(h\Omega)$ is the block matrix given by

$$R(h\Omega) = \begin{pmatrix} \cos(h\Omega) & -\Omega\sin(h\Omega) \\ \Omega^{-1}\sin(h\Omega) & \cos(h\Omega) \end{pmatrix}$$

and the functions ϕ, ψ, ψ_0 and ψ_1 are even functions such that

$$\begin{aligned} &\psi(z) = \frac{\sin(z)}{z}\psi_1(z), \ \psi_0(z) = \cos(z)\psi_1(z), \ \text{and} \\ &\psi(0) = \phi(0) = 1. \end{aligned}$$

Various choices of functions ψ and ϕ are possible and documented in the literature. Two particularly interesting ones are $\psi(z) = \frac{\sin^2(z)}{z^2}$, $\phi(z) = 1$ (see [9]) or $\psi(z) = \frac{\sin^3(z)}{z^3}$, $\phi(z) = \frac{\sin(z)}{z}$ (see [7]).

Conclusion

This entry should be regarded as an introduction to the subject of symmetric methods. Several topics have not been exposed here, such as symmetric projection for ODEs on manifolds, DAEs of index 1 or 2, symmetric multistep methods, symmetric splitting methods, and symmetric Lie-group methods, and we refer the interested reader to [10, 13, 16] for a comprehensive presentation of these topics.

References

1. Butcher, J.C.: Implicit Runge-Kutta processes. Math. Comput. **18**, 50–64 (1964)
2. Chartier, P., Hairer, E., Vilmart, G.: Algebraic structures of B-series. Found. Comput. Math. **10**(4), 407–427 (2010)
3. Creutz, M., Gocksch, A.: Higher-order hybrid Monte Carlo algorithms. Phys. Rev. Lett. **63**, 9–12 (1989)
4. Faou, E., Hairer, E., Pham, T.L.: Energy conservation with non-symplectic methods: examples and counter-examples. BIT **44**(4), 699–709 (2004)
5. Forest, E.: Canonical integrators as tracking codes. AIP Conf. Proc. **184**, 1106–1136 (1989)
6. García-Archilla, B., Sanz-Serna, J.M., Skeel, R.D.: Long-time-step methods for oscillatory differential equations. SIAM J. Sci. Comput. **20**, 930–963 (1999)
7. Grimm, V., Hochbruck, M.: Error analysis of exponential integrators for oscillatory second-order differential equations. J. Phys. A **39**, 5495–5507 (2006)
8. Grubmüller, H., Heller, H., Windemuth, A., Tavan, P.: Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. Mol. Sim. **6**, 121–142 (1991)
9. Hairer, E., Lubich, C.: Long-time energy conservation of numerical methods for oscillatory differential equations. SIAM J. Numer. Anal. **38**, 414–441 (2000)
10. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer Series in Computational Mathematics, vol. 31. Springer, Berlin (2006)
11. Jia, Z., Leimkuhler, B.: Geometric integrators for multiple time-scale simulation. J. Phys. A **39**, 5379–5403 (2006)
12. Kahan, W., Li, R.C.: Composition constants for raising the orders of unconventional schemes for ordinary differential equations. Math. Comput. **66**, 1089–1099 (1997)
13. Leimkuhler, B., Reich, S.: Simulating Hamiltonian Dynamics. Cambridge Monographs on Applied and Computational Mathematics, vol. 14. Cambridge University Press, Cambridge (2004)
14. McLachlan, R.I., Quispel, G.R.W.: Splitting methods. Acta Numer. **11**, 341–434 (2002)
15. Murua, A., Sanz-Serna, J.M.: Order conditions for numerical integrators obtained by composing simpler integrators. Philos. Trans. R. Soc. Lond. A **357**, 1079–1100 (1999)
16. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems. Chapman & Hall, London (1994)
17. Stoffer, D.: On reversible and canonical integration methods. Technical Report SAM-Report No. 88-05, ETH-Zürich (1988)
18. Suzuki, M.: Fractal decomposition of exponential operators with applications to many-body theories and monte carlo simulations. Phys. Lett. A **146**, 319–323 (1990)
19. Tuckerman, M., Berne, B.J., Martyna, G.J.: Reversible multiple time scale molecular dynamics. J. Chem. Phys. **97**, 1990–2001 (1992)
20. Verlet, L.: Computer "experiments" on classical fluids. i. Thermodynamical properties of Lennard-Jones molecules. Phys. Rev. **159**, 98–103 (1967)
21. Yoshida, H.: Construction of higher order symplectic integrators. Phys. Lett. A **150**, 262–268 (1990)

Symmetries and FFT

Hans Z. Munthe-Kaas
Department of Mathematics, University of Bergen,
Bergen, Norway

Synonyms

Fourier transform; Group theory; Representation theory; Symmetries

Synopsis

The fast Fourier transform (FFT), group theory, and symmetry of linear operators are mathematical topics which all connect through group representation theory. This entry provides a brief introduction, with emphasis on computational applications.

Symmetric FFTs

The finite Fourier transform maps functions on a periodic lattice to a dual Fourier domain. Formally, let $\mathbb{Z}_{\mathbf{n}} = \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \cdots \times \mathbb{Z}_{n_\ell}$ be a finite abelian (commutative) group, where $\mathbb{Z}_{n_j}$ is the cyclic group $\mathbb{Z}_{n_j} = \{0, 1, \ldots, n_j - 1\}$ with group operation $+(\text{mod } n_j)$ and $\mathbf{n} = (n_1, \ldots, n_\ell)$ is a multi-index with $|\mathbf{n}| = n_1 n_2 \cdots n_\ell$. Let $\mathbb{C}\mathbb{Z}_{\mathbf{n}}$ denote the linear space of complex-valued functions on $\mathbb{Z}_{\mathbf{n}}$. The *primal domain* $\mathbb{Z}_{\mathbf{n}}$ is an ℓ-dimensional lattice periodic in all directions, and the *Fourier domain* is $\widehat{\mathbb{Z}}_{\mathbf{n}} = \mathbb{Z}_{\mathbf{n}}$ (this is the Pontryagin dual group [12]). For infinite abelian groups, the primal and Fourier domains in general differ, such as Fourier series on the circle where $\widehat{\mathbb{R}/\mathbb{Z}} = \mathbb{Z}$. The discrete Fourier transform (DFT) is an (up to scaling) unitary map $\mathcal{F}\colon \mathbb{C}\mathbb{Z}_{\mathbf{n}} \to \mathbb{C}\widehat{\mathbb{Z}}_{\mathbf{n}}$. Letting $F(f) \equiv \widehat{f}$, we have

$$\widehat{f}(\mathbf{k}) = \sum_{\mathbf{j}\in\mathbb{Z}_\mathbf{n}} f(\mathbf{j})e^{2\pi i\left(\frac{k_1 j_1}{n_1}+\cdots+\frac{k_\ell j_\ell}{n_\ell}\right)},$$
$$\text{for } \mathbf{k} = (k_1,\ldots,k_\ell) \in \widehat{Z}_\mathbf{n}, \tag{1}$$

$$f(\mathbf{j}) = \frac{1}{|\mathbf{n}|}\sum_{\mathbf{k}\in\widehat{\mathbb{Z}}_\mathbf{n}} \widehat{f}(\mathbf{k})e^{-2\pi i\left(\frac{k_1 j_1}{n_1}+\cdots+\frac{k_\ell j_\ell}{n_\ell}\right)},$$
$$\text{for } \mathbf{j} = (j_1,\ldots,j_\ell) \in Z_\mathbf{n}. \tag{2}$$

A *symmetry* for a function $f \in \mathbb{C}\mathbb{Z}_\mathbf{n}$ is an $\mathbb{R}$-linear map $S\colon \mathbb{C}\mathbb{Z}_\mathbf{n} \to \mathbb{C}\mathbb{Z}_\mathbf{n}$ such that $Sf = f$. As examples, consider real symmetry $S_R f(\mathbf{j}) = \overline{f(\mathbf{j})}$, even symmetry $S_e f(\mathbf{j}) = f(-\mathbf{j})$ and odd symmetry $S_o f(\mathbf{j}) = -f(-\mathbf{j})$. If f has a symmetry S, then $\widehat{f}$ has an adjoint symmetry $\widehat{S} = \mathcal{F}S\mathcal{F}^{-1}$, example $\widehat{S_e} = S_e$, $\widehat{S_o} = S_o$ and $\widehat{S_R}\widehat{f}(\mathbf{k}) = \overline{\widehat{f}(-\mathbf{k})}$. The set of all symmetries of f forms a *group*, i.e., the set of symmetries is closed under composition and inversion. Equivalently, the symmetries can be specified by defining an abstract group G and a map $R\colon G \to \mathrm{Lin}_\mathbb{R}(\mathbb{C}\mathbb{Z}_\mathbf{n})$, which for each $g \in G$ defines an $\mathbb{R}$-linear map $R(g)$ on $\mathbb{C}\mathbb{Z}_\mathbf{n}$, such that $R(gg') = R(g)R(g')$ for all $g, g' \in G$. R is an example of a real *representation* of G.

The DFT on $\mathbb{Z}_\mathbf{n}$ is computed by the fast Fourier transform (FFT), costing $\mathcal{O}(|\mathbf{n}|\log(|\mathbf{n}|))$ floating point operations. It is possible to exploit may symmetries in the computation of the FFT, and for large classes of symmetry groups, savings a factor $|G|$ can be obtained compared to the nonsymmetric FFT.

Equivariant Linear Operators and the GFT

Representations

Let G be a finite group with $|G|$ elements. A d_R-dimensional unitary *representation* of G is a map $R\colon G \to U(d_R)$ such that $R(gh) = R(g)R(h)$ for all $g, h \in G$, where $U(d_R)$ is the set of $d_R \times d_R$ unitary matrices. More generally, a representation is a linear action of a group on a vector space. Two representations R and $\widetilde{R}$ are *equivalent* if there exists a matrix X such that $\widetilde{R}(g) = XR(g)X^{-1}$ for all $g \in G$. A representation R is *reducible* if it is equivalent to a block diagonal representation; otherwise it is *irreducible*. For any finite group G, there exists a complete list of nonequivalent irreducible representations $\mathcal{R} = \{\rho_1, \rho_2, \ldots, \rho_n\}$, henceforth called *irreps*, such that $\sum_{\rho\in\mathcal{R}} d_\rho^2 = |G|$. For example, the cyclic group $Z_n = \{0, 1, \ldots, r-1\}$ with group operation $+(\mathrm{mod}\, n)$ has exactly n 1-dimensional irreps given as $\rho_k(j) = \exp(2\pi i kj/n)$. A matrix A is *equivariant* with respect to a representation R of a group G if $AR(g) = R(g)A$ for all $g \in G$. Any representation R can be block diagonalized, with irreducible representations on the diagonal. This provides a change of basis matrix F such that FAF^{-1} is block diagonal for any R-equivariant A. This result underlies most of computational Fourier analysis and will be exemplified by convolutional operators.

Convolutions in the Group Algebra

The group algebra $\mathbb{C}G$ is the complex $|G|$-dimensional vector space where the elements of G are the basis vectors; equivalently $\mathbb{C}G$ consists of all complex-valued functions on G. The product in G extends linearly to the convolution product $*\colon \mathbb{C}G \times \mathbb{C}G \to \mathbb{C}G$, given as $(a * b)(g) = \sum_{h\in G} a(h)b(h^{-1}g)$ for all $g \in G$. The *right regular representation* of G on $\mathbb{C}G$ is, for every $h \in G$, a linear map $R(h)\colon \mathbb{C}G \to \mathbb{C}G$ given as right translation $R(h)a(g) = a(gh)$. A linear map $A\colon \mathbb{C}G \to \mathbb{C}G$ is *convolutional* (i.e., there exists an $a \in \mathbb{C}G$ such that $Ab = a * b$ for all $b \in \mathbb{C}G$) if and only if A is equivariant with respect to the right regular representation.

The Generalized Fourier Transform (GFT)

The *generalized Fourier transform* [6, 10] and the inverse are given as

$$\widehat{a}(\rho) = \sum_{g\in G} a(g)\rho(g) \in \mathbb{C}^{d_\rho\times d_\rho}, \text{ for all } \rho \in \mathcal{R} \tag{3}$$

$$a(g) = \frac{1}{|G|}\sum_{\rho\in\mathcal{R}} d_\rho \mathrm{trace}\left(\rho(g^{-1})\widehat{a}(\rho)\right), \text{ for all } g \in G. \tag{4}$$

From the convolution formula $\widehat{a * b}(\rho) = \widehat{a}(\rho)\widehat{b}(\rho)$, we conclude: *The GFT block-diagonalizes convolutional operators on* $\mathbb{C}G$. The blocks are of size d_ρ, the dimensions of the irreps.

Equivariant Linear Operators

More generally, consider a linear operator $A\colon \mathcal{V} \to \mathcal{V}$ where $\mathcal{V}$ is a finite-dimensional vector space and A is equivariant with respect to a linear right action of G

on $\mathcal{V}$. If the action is free and transitive, then A is convolutional on $\mathbb{C}G$. If the action is not transitive, then $\mathcal{V}$ splits in s separate orbits under the action of G, and A is a block-convolutional operator. In this case, the GFT block diagonalizes A with blocks of size approximately sd_ρ. The theory generalizes to infinite compact Lie groups via the Peter–Weyl theorem and to certain important non-compact groups (unimodular groups), such as the group of Euclidean transformations acting on $\mathbb{R}^n$; see [13].

Applications

Symmetric FFTs appear in many situations, such as real sine and cosine transforms. The 1-dim real cosine transform has four symmetries generated by S_R and S_e, and it can be computed four times faster than the full complex FFT. This transform is central in Chebyshev approximations. More generally, multivariate Chebyshev polynomials possess symmetries of kaleidoscopic reflection groups acting upon $\mathbb{Z}_\mathbf{n}$ [9, 11, 14].

Diagonalization of equivariant linear operators is essential in signal and image processing, statistics, and differential equations. For a cyclic group $\mathbb{Z}_n$, an equivariant A is a Toeplitz circulant, and the GFT is given by the discrete Fourier transform, which can be computed fast by the FFT algorithm. More generally, a finite abelian group $\mathbb{Z}_\mathbf{n}$ has $|\mathbf{n}|$ one-dimensional irreps, given by the exponential functions. $\mathbb{Z}_\mathbf{n}$-equivariant matrices are block Toeplitz circulant matrices. These are diagonalized by multidimensional FFTs. Linear differential operators with constant coefficients typically lead to discrete linear operators which commute with translations acting on the domain. In the case of periodic boundary conditions, this yields block circulant matrices. For more complicated boundaries, block circulants may provide useful approximations to the differential operators.

More generally, many computational problems possess symmetries given by a (discrete or continuous) group acting on the domain. For example, the Laplacian operator commutes with any isometry of the domain. This can be discretized as an equivariant discrete operator. If the group action on the discretized domain is free and transitive, the discrete operator is a convolution in the group algebra. More generally, it is a block-convolutional operator. For computational efficiency, it is important to identify the symmetries (or approximate symmetries) of the problem and employ the irreducible characters of the symmetry group and the GFT to (approximately) block diagonalize the operators. Such techniques are called *domain reduction* techniques [7]. Block diagonalization of linear operators has applications in solving linear systems, eigenvalue problems, and computation of matrix exponentials. The GFT has also applications in image processing, image registration, and computational statistics [1–5, 8].

References

1. Åhlander, K., Munthe-Kaas, H.: Applications of the generalized Fourier transform in numerical linear algebra. BIT Numer. Math. **45**(4), 819–850 (2005)
2. Allgower, E., Böhmer, K., Georg, K., Miranda, R.: Exploiting symmetry in boundary element methods. SIAM J. Numer. Anal. **29**, 534–552 (1992)
3. Bossavit, A.: Symmetry, groups, and boundary value problems. A progressive introduction to noncommutative harmonic analysis of partial differential equations in domains with geometrical symmetry. Comput. Methods Appl. Mech. Eng. **56**(2), 167–215 (1986)
4. Chirikjian, G., Kyatkin, A.: Engineering Applications of Noncommutative Harmonic Analysis. CRC, Boca Raton (2000)
5. Diaconis, P.: Group Representations in Probability and Statistics. Institute of Mathematical Statistics, Hayward (1988)
6. Diaconis, P., Rockmore, D.: Efficient computation of the Fourier transform on finite groups. J. Am. Math. Soc. **3**(2), 297–332 (1990)
7. Douglas, C., Mandel, J.: An abstract theory for the domain reduction method. Computing **48**(1), 73–96 (1992)
8. Fässler, A., Stiefel, E.: Group Theoretical Methods and Their Applications. Birkhauser, Boston (1992)
9. Hoffman, M., Withers, W.: Generalized Chebyshev polynomials associated with affine Weyl groups. Am. Math. Soc. **308**(1), 91–104 (1988)
10. Maslen, D., Rockmore, D.: Generalized FFTs—a survey of some recent results. Publisher: In: Groups and Computation II, Am. Math. Soc. vol. 28, pp. 183–287 (1997)
11. Munthe-Kaas, H.Z., Nome, M., Ryland, B.N.: Through the Kaleidoscope; symmetries, groups and Chebyshev approximations from a computational point of view. In: Foundations of Computational Mathematics. Cambridge University Press, Cambridge (2012)
12. Rudin, W.: Fourier Analysis on Groups. Wiley, New York (1962)
13. Serre, J.: Linear Representations of Finite Groups, vol 42. Springer, New York (1977)
14. Ten Eyck, L.: Crystallographic fast Fourier transforms. Acta Crystallogr Sect. A Crystal Phys. Diffr. Theor. General Crystallogr. **29**(2), 183–191 (1973)

Symplectic Methods

J.M. Sanz-Serna
Departamento de Matemática Aplicada, Universidad de Valladolid, Valladolid, Spain

Definition

This entry, concerned with the practical task of integrating numerically Hamiltonian systems, follows up the entry ▶ Hamiltonian Systems and keeps the notation and terminology used there.

Each one-step numerical integrator is specified by a smooth map $\Psi^H_{t_{n+1},t_n}$ that advances the numerical solution from a time level t_n to the next t_{n+1}

$$(p^{n+1}, q^{n+1}) = \Psi^H_{t_{n+1},t_n}(p^n, q^n); \qquad (1)$$

the superscript H refers to the Hamiltonian function $H(p, q; t)$ of the system being integrated. For instance for the explicit Euler rule

$$(p^{n+1}, q^{n+1}) = (p^n, q^n) + (t_{n+1} - t_n)\big(f(p^n, q^n; t_n),\; g(p^n, q^n; t_n)\big);$$

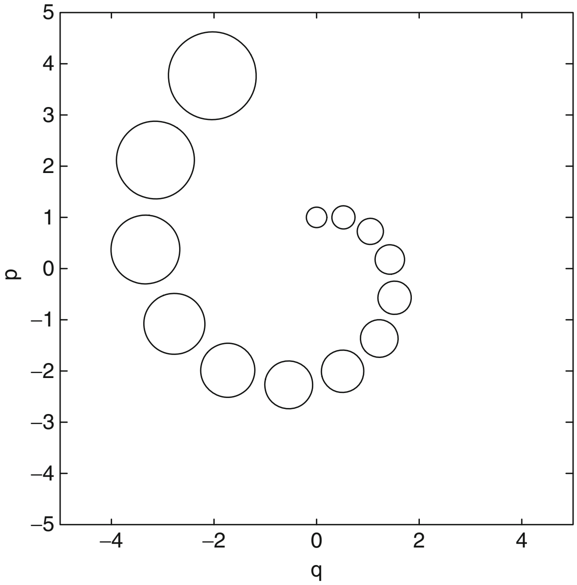
here and later f and g denote the d-dimensional real vectors with entries $-\partial H/\partial q_i$, $\partial H/\partial p_i$ (d is the number of degrees of freedom) so that (f, g) is the canonical vector field associated with H (in simpler words: the right-hand side of Hamilton's equations). For the integrator to make sense, $\Psi^H_{t_{n+1},t_n}$ has to approximate the solution operator $\Phi^H_{t_{n+1},t_n}$ that advances the true solution from its value at t_n to its value at t_{n+1}:

$$\big(p(t_{n+1}), q(t_{n+1})\big) = \Phi^H_{t_{n+1},t_n}\big(p(t_n), q(t_n)\big).$$

For a method of (consistency) order ν, $\Psi^H_{t_{n+1},t_n}$ differs from $\Phi^H_{t_{n+1},t_n}$ in terms of magnitude $\mathcal{O}\big((t_{n+1} - t_n)^{\nu+1}\big)$.

The solution map $\Phi^H_{t_{n+1},t_n}$ is a canonical (symplectic) transformation in phase space, an important fact that substantially constrains the dynamics of the true solution $\big(p(t), q(t)\big)$. If we wish the approximation Ψ^H to retain the "Hamiltonian" features of Φ^H, we should insist on Ψ^H also being a symplectic transformation. However, most standard numerical integrators – including explicit Runge–Kutta methods, regardless of their order ν – replace Φ^H by a nonsymplectic mapping Ψ^H. This is illustrated in Fig. 1 that corresponds to the Euler rule as applied to the harmonic oscillator $\dot{p} = -q$, $\dot{q} = p$. The (constant) step size is $t_{n+1} - t_n = 2\pi/12$. We have taken as a family of initial conditions the points of a circle centered at $p = 1, q = 0$ and seen the evolution after $1, 2, \ldots, 12$ steps. Clearly the circle, which should move clockwise without changing area, gains area as the integration proceeds: The numerical Ψ^H is not symplectic. As a result, the origin, a center in the true dynamics, is turned by the discretization procedure into an unstable spiral point, i.e., into something that cannot arise in Hamiltonian dynamics. For the implicit Euler rule, the corresponding integration loses area and gives rise to a family of smaller and smaller circles that spiral toward the origin. Again, such a stable focus is incompatible with Hamiltonian dynamics.

This failure of well-known methods in mimicking Hamiltonian dynamics motivated the consideration of integrators that generate a symplectic mapping Ψ^H when applied to a Hamiltonian problem. Such methods are called *symplectic* or *canonical.* Since symplectic transformations also preserve volume, symplectic integrators applied to Hamiltonian problems are automatically *volume preserving*. On the other hand, while many important symplectic integrators are time-



Symplectic Methods, Fig. 1 The harmonic oscillator integrated by the explicit Euler method

S

reversible (symmetric), reversibility is neither sufficient nor necessary for a method to be symplectic ([8], Remark 6.5).

Even though early examples of symplectic integration may be traced back to the 1950s, the systematic exploration of the subject started with the work of Feng Kang (1920–1993) in the 1980s. An early short monograph is [8] and later books are the comprehensive [5] and the more applied [6]. Symplectic integration was the first step in the larger endeavor of developing structure-preserving integrators, i.e., of what is now often called, following [7], *geometric integration*.

Limitations of space restrict this entry to one-step methods and *canonical* Hamiltonian problems. For noncanonical Hamiltonian systems and multistep integrators the reader is referred to [5], Chaps. VII and XV.

Integrators Based on Generating Functions

The earliest systematic approaches by Feng Kang and others to the construction of symplectic integrators (see [5], Sect. VI.5.4 and [8], Sect. 11.2) exploited the following well-known result of the canonical formalism: The canonical transformation $\Phi^H_{t_{n+1},t_n}$ possesses a generating function S_2 that solves an initial value problem for the associated Hamilton–Jacobi equation. It is then possible, by Taylor expanding that equation, to obtain an approximation $\widetilde{S}_2$ to S_2. The transformation $\Psi^H_{t_{n+1},t_n}$ generated by $\widetilde{S}_2$ will automatically be canonical and therefore will define a symplectic integrator. If $\widetilde{S}_2$ differs from S_2 by terms $\mathcal{O}\big((t_{n+1}-t_n)^{\nu+1}\big)$, the integrator will be of order ν. Generally speaking, the high-order methods obtained by following this procedure are more difficult to implement than those derived by the techniques discussed in the next two sections.

Runge–Kutta and Related Integrators

In 1988, Lasagni, Sanz-Serna, and Suris (see [8], Chap. 6) discovered independently that some well-known families of numerical methods contain symplectic integrators.

Runge–Kutta Methods

Symplecticness Conditions

When the Runge–Kutta (RK) method with s stages specified by the tableau

$$\begin{array}{|ccc} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{ss} \\ \hline b_1 & \cdots & b_s \end{array} \tag{2}$$

is applied to the integration of the Hamiltonian system with Hamiltonian function H, the relation (1) takes the form

$$p^{n+1} = p^n + h_{n+1}\sum_{i=1}^{s} b_i\, f(P_i, Q_i; t_n + c_i h_{n+1}),$$

$$q^{n+1} = q^n + h_{n+1}\sum_{i=1}^{s} b_i\, g(P_i, Q_i; t_n + c_i h_{n+1}),$$

where $c_i = \sum_j a_{ij}$ are the abscissae, $h_{n+1} = t_{n+1} - t_n$ is the step size and $P_i, Q_i, i = 1, \ldots, s$ are the internal stage vectors defined through the system

$$P_i = p^n + h_{n+1}\sum_{j=1}^{s} a_{ij}\, f(P_j, Q_j; t_n + c_j h_{n+1}), \tag{3}$$

$$Q_i = q^n + h_{n+1}\sum_{j=1}^{s} a_{ij}\, g(P_j, Q_j; t_n + c_j h_{n+1}). \tag{4}$$

Lasagni, Sanz-Serna, and Suris proved that if the coefficients of the method in (2) satisfy

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \quad i, j = 1, \ldots, s, \tag{5}$$

then the method is symplectic. Conversely ([8], Sect. 6.5), the relations (5) are essentially necessary for the method to be symplectic. Furthermore for symplectic RK methods the transformation (1) is in fact *exact symplectic* ([8], Remark 11.1).

Order Conditions

Due to symmetry considerations, the relations (5) impose $s(s+1)/2$ independent equations on the $s^2 + s$

elements of the RK tableau (2), so that there is no shortage of symplectic RK methods. The available free parameters may be used to increase the accuracy of the method. It is well known that the requirement that an RK formula has a target order leads to a set of nonlinear relations (order conditions) between the elements of the corresponding tableau (2). For order $\geq \nu$ there is an order condition associated with each rooted tree with $\leq \nu$ vertices and, if the a_{ij} and b_i are free parameters, the order conditions are mutually independent. For symplectic methods however the tableau coefficients are constrained by (5), and Sanz-Serna and Abia proved in 1991 that then there are redundancies between the order conditions ([8], Sect. 7.2). In fact to ensure order $\geq \nu$ when (5) holds it is necessary and sufficient to impose an order condition for each so-called nonsuperfluous (nonrooted) tree with $\leq \nu$ vertices.

Examples of Symplectic Runge–Kutta Methods

Setting $j = i$ in (5) shows that explicit RK methods (with $a_{ij} = 0$ for $i \leq j$) cannot be symplectic.

Sanz-Serna noted in 1988 ([8], Sect. 8.1) that the *Gauss method* with s stages, $s = 1, 2, \dots,$ (i.e., the unique method with s stages that attains the maximal order $2s$) *is* symplectic. When $s = 1$ the method is the familiar *implicit midpoint rule*. Since for all Gauss methods the matrix (a_{ij}) is full, the computation of the stage vectors P_i and Q_i require, at each step, the solution of the system (3) and (4) that comprises $s \times 2d$ scalar equations. In non-stiff situations this system is readily solved by functional iteration, see [8] Sects. 5.4 and 5.5 and [5] Sect. VIII.6, and then the Gauss methods combine the advantages of symplecticness, easy implementation, and high order with that of being applicable to all canonical Hamiltonian systems.

If the system being solved is stiff (e.g., it arises through discretization of the spatial variables of a Hamiltonian partial differential equation), Newton iteration has to be used to solve the stage equations (3) and (4), and for high-order Gauss methods the cost of the linear algebra may be prohibitive. It is then of interest to consider the possibility of *diagonally implicit* symplectic RK methods, i.e., methods where $a_{ij} = 0$ for $i < j$ and therefore (3) and (4) demand the successive solution of s systems of dimension $2d$, rather than that of a single $(s \times 2d)$–dimensional system. It turns out ([8], Sect. 8.2) that such methods are necessarily composition methods (see below) obtained by concatenating implicit midpoint sub-steps of lengths $b_1 h_{n+1}, \dots, b_s h_{n+1}$. The determination of the free parameters b_i is a task best accomplished by means of the techniques used to analyze composition methods.

The B-series Approach

In 1994, Calvo and Sanz-Serna ([5], Sect. VI.7.2) provided an indirect technique for the derivation of the symplecticness conditions (5). The first step is to identify conditions for the symplecticness of the associated B-series (i.e., the series that expands the transformation (1)) in powers of the step size. Then the conditions (on the B-series) obtained in this way are shown to be equivalent to (5). This kind of approach has proved to be very powerful in the theory of geometric integration, where extensive use is made of formal power series.

Partitioned Runge–Kutta Methods

Partitioned Runge–Kutta (PRK) methods differ from standard RK integrators in that they use *two* tableaux of coefficients of the form (2): one to advance p and the other to advance q. Most developments of the theory of symplectic RK methods are easily adapted to cover the partitioned situation, see e.g., [8], Sects. 6.3, 7.3, and 8.4.

The main reason ([8], Sect. 8.4) to consider the class of PRK methods is that it contains integrators that are both *explicit* and symplectic when applied to *separable* Hamiltonian systems with $H(p, q; t) = T(p) + V(q; t)$, a format that often appears in the applications. It turns out ([8], Remark 8.1, [5], Sect. VI.4.1, Theorem 4.7) that such explicit, symplectic PRK methods may always be viewed as splitting methods (see below). Moreover it is advantageous to perform their analysis by interpreting them as splitting algorithms.

Runge–Kutta–Nyström Methods

In the special but important case where the (separable) Hamiltonian is of the form $H = (1/2)p^T M^{-1} p + V(q; t)$ (M a positive-definite symmetric matrix) the canonical equations

$$\frac{d}{dt} p = -\nabla V(q; t), \qquad \frac{d}{dt} q = M^{-1} p \tag{6}$$

lead to

$$\frac{d^2}{dt^2} q = -M^{-1} \nabla V(q; t),$$

a second-order system whose right-hand side is independent of $(d/dt)q$. Runge–Kutta–Nyström (RKN) methods may then be applied to the second-order form and are likely to improve on RK integrations of the original first-order system (6).

There are *explicit, symplectic* RKN integrators ([8], Sect. 8.5). However their application (see [8], Remark 8.5) is always equivalent to the application of an explicit, symplectic PRK method to the first-order equations (6) and therefore – in view of a consideration made above – to the application of a splitting algorithm.

Integrators Based on Splitting and Composition

The related ideas of splitting and composition are extremely fruitful in deriving practical symplectic integrators in many fields of application. The corresponding methods are typically *ad hoc* for the problem at hand and do not enjoy the universal off-the-shelf applicability of, say, Gaussian RK methods; however, when applicable, they may be highly efficient. In order to simplify the exposition, we assume hereafter that the Hamiltonian H is *time-independent* $H = H(p,q)$; we write $\phi^H_{h_{n+1}}$ and $\psi^H_{h_{n+1}}$ rather than $\Phi^H_{t_{n+1},t_n}$ and $\Psi^H_{t_{n+1},t_n}$. Furthermore, we shall denote the time step by h omitting the possible dependence on the step number n.

Splitting

Simplest Splitting

The easiest possibility of splitting occurs when the Hamiltonian H may be written as $H_1 + H_2$ and the Hamiltonian systems associated with H_1 and H_2 may be explicitly integrated. If the corresponding flows are denoted by $\phi^{H_1}_t$ and $\phi^{H_2}_t$, the recipe (Lie–Trotter splitting, [8], Sect. 12.4.2, [5], Sect. II.5)

$$\psi^H_h = \phi^{H_2}_h \circ \phi^{H_1}_h \tag{7}$$

defines the map (1) of a first-order integrator that is symplectic (the mappings being composed in the right-hand side are Hamiltonian flows and therefore symplectic). Splittings of H in more than two pieces are feasible but will not be examined here.

A particular case of (7) of great practical significance is provided by the *separable* Hamiltonian $H(p,q) = T(p) + V(q)$ with $H_1 = T$, $H_2 = V$; the flows associated with H_1 and H_2 are respectively given by

$$(p,q) \mapsto (p, q + t\nabla T(p)), \ (p,q) \mapsto (p - t\nabla V(q), q).$$

Thus, in this particular case the scheme (7) reads

$$p^{n+1} = p^n - h\nabla V(q^{n+1}), \quad q^{n+1} = q^n + h\nabla T(p^n), \tag{8}$$

and it is sometimes called the *symplectic Euler* rule (it is obviously possible to interchange the roles of p and q). Alternatively, (8) may be considered as a one-stage, explicit, symplectic PRK integrator as in [8], Sect. 8.4.3.

As a second example of splitting, one may consider (nonseparable) formats $H = H_1(p,q) + V^*(q)$, where the Hamiltonian system associated with H_1 can be integrated in closed form. For instance, H_1 may correspond to a set of uncoupled harmonic oscillators and $V^*(q)$ represent the potential energy of the interactions between oscillators. Or H_1 may correspond to the Keplerian motion of a point mass attracted to a fixed gravitational center and V^* be a potential describing some sort of perturbation.

Strang Splitting

With the notation in (7), the symmetric Strang formula ([8], Sect. 12.4.3, [5], Sect. II.5)

$$\bar{\psi}^H_h = \phi^{H_2}_{h/2} \circ \phi^{H_1}_h \circ \phi^{H_2}_{h/2} \tag{9}$$

defines a time-reversible, *second-order* symplectic integrator $\bar{\psi}^H_h$ that improves on the first order (7).

In the separable Hamiltonian case $H = T(p) + V(q)$, (9) leads to

$$\begin{aligned} p^{n+1/2} &= p^n - \frac{h}{2}\nabla V(q^n), \\ q^{n+1} &= q^n + h\nabla T(p^{n+1/2}), \\ p^{n+1} &= p^{n+1/2} - \frac{h}{2}\nabla V(q^{n+1}). \end{aligned}$$

This is the Störmer–Leapfrog–Verlet method that plays a key role in molecular dynamics [6]. It is also possible to regard this integrator as an explicit, symplectic PRK with two stages ([8], Sect. 8.4.3).

More Sophisticated Formulae

A further generalization of (7) is

$$\phi^{H_2}_{\beta_s h} \circ \phi^{H_1}_{\alpha_s h} \circ \phi^{H_2}_{\beta_{s-1} h} \circ \cdots \circ \phi^{H_2}_{\beta_1 h} \circ \phi^{H_1}_{\alpha_1 h} \tag{10}$$

where the coefficients α_i and β_i, $\sum_i \alpha_i = 1$, $\sum_i \beta_i = 1$, are chosen so as to boost the order ν of the method. A systematic treatment based on trees of the required order conditions was given by Murua and Sanz-Serna in 1999 ([5], Sect. III.3). There has been much recent activity in the development of accurate splitting coefficients α_i, β_i and the reader is referred to the entry ▶ Splitting Methods in this encyclopedia.

In the particular case where the splitting is given by $H = T(p) + V(q)$, the family (10) provides the most general explicit, symplectic PRK integrator.

Splitting Combined with Approximations

In (7), (9), or (10) use is made of the exact solution flows $\phi^{H_1}_t$ and $\phi^{H_2}_t$. Even if one or both of these flows are not available, it is still possible to employ the idea of splitting to construct symplectic integrators. A simple example will be presented next, but many others will come easily to mind.

Assume that we wish to use a Strang-like method but $\phi^{H_1}_t$ is not available. We may then advance the numerical solution via

$$\phi^{H_2}_{h/2} \circ \widehat{\psi}^{H_1}_h \circ \phi^{H_2}_{h/2}, \tag{11}$$

where $\widehat{\psi}^{H_1}_h$ denotes a consistent method for the integration of the Hamiltonian problem associated with H_1. If $\widehat{\psi}^{H_1}_h$ is time-reversible, the composition (11) is also time-reversible and hence of order $\nu = 2$ (at least). And if $\widehat{\psi}^{H_1}_h$ is symplectic, (11) will define a symplectic method.

Composition

A step of a composition method ([5], Sect. II.4) consists of a concatenation of a number of sub-steps performed with one or several simpler methods. Often the aim is to create a high-order method out of low-order integrators; the composite method automatically inherits the conservation properties shared by the methods being composed. The idea is of particular appeal within the field of geometric integration, where it is frequently not difficult to write down first- or second-order integrators with good conservation properties.

A useful example, due to Suzuki, Yoshida, and others (see [8], Sect. 13.1), is as follows. Let ψ^H_h be a time-reversible integrator that we shall call the *basic* method and define the composition method $\widehat{\psi}^H_h$ by

$$\widehat{\psi}^H_h = \psi^H_{\alpha h} \circ \psi^H_{(1-2\alpha)h} \circ \psi^H_{\alpha h};$$

if the basic method is symplectic, then $\widehat{\psi}^H_h$ will obviously be a symplectic method. It may be proved that, if $\alpha = (1/3)(2 + 2^{1/3} + 2^{-1/3})$, then $\widehat{\psi}^H_h$ will have order $\nu = 4$. By using this idea one may perform symplectic, fourth-order accurate integrations while really implementing a simpler second-order integrator. The approach is particularly attractive when the direct application of a fourth-order method (such as the two-stage Gauss method) has been ruled out on implementation grounds, but a suitable basic method (for instance the implicit midpoint rule or a scheme derived by using Strang splitting) is available.

If the (time-reversible) basic method is of order 2μ and $\alpha = \left(2-2^{1/(2\mu+1)}\right)^{-1}$ then $\widehat{\psi}^H_h$ will have order $\nu = 2\mu+2$; the recursive application of this idea shows that it is possible to reach arbitrarily high orders starting from a method of order 2.

For further possibilities, see the entry ▶ Composition Methods and [8], Sect. 13.1, [5], Sects. II.4 and III.3.

The Modified Hamiltonian

The properties of symplectic integrators outlined in the next section depend on the crucial fact that, when a symplectic integrator is used, a numerical solution of the Hamiltonian system with Hamiltonian H may be viewed as an (almost) exact solution of a Hamiltonian system whose Hamiltonian function $\widetilde{H}$ (the so-called modified Hamiltonian) is a perturbation of H.

An example. Consider the application of the symplectic Euler rule (8) to a one-degree-of-freedom system with separable Hamiltonian $H = T(p) + V(q)$. In order to describe the behavior of the points (p^n, q^n) computed by the algorithm, we could just say that they approximately behave like the solutions $(p(t_n), q(t_n))$ of the Hamiltonian system $\mathcal{S}$ being integrated. This would not be a very precise description because the true flow ϕ_h^H and its numerical approximation ψ_h^H differ in $\mathcal{O}(h^2)$ terms. Can we find *another* differential system $\mathcal{S}_2$ (called a modified system) so that (8) is consistent of the *second* order with $\mathcal{S}_2$? The points (p^n, q^n) would then be closer to the solutions of $\mathcal{S}_2$ than to the solutions of the system $\mathcal{S}$ we want to integrate. Straightforward Taylor expansions ([8], Sect. 10.1) lead to the following expression for $\mathcal{S}_2$ (recall that $f = -\partial H/\partial q$, $g = \partial H/\partial p$)

$$\frac{d}{dt}p = f(q) + \frac{h}{2}g(p)f'(q), \qquad \frac{d}{dt}q = g(p) - \frac{h}{2}g'(p)f(q), \tag{12}$$

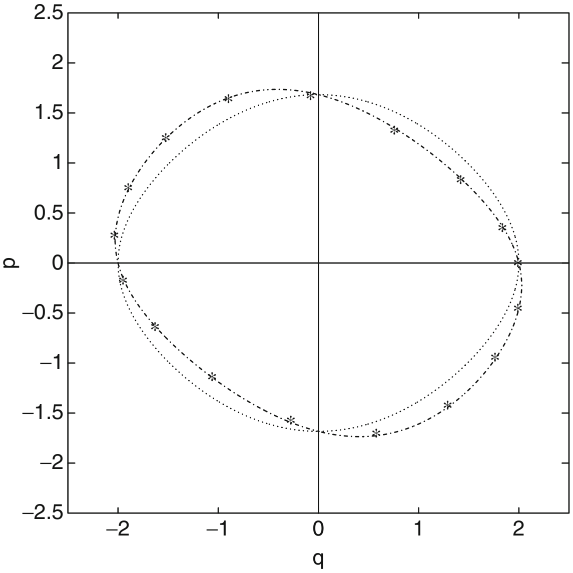
where we recognize the Hamiltonian system with (h-dependent!) Hamiltonian

$$\widetilde{H}_2^h = T(p) + V(q) + \frac{h}{2}T'(p)V'(q) = H + \mathcal{O}(h). \tag{13}$$

Figure 2 corresponds to the pendulum equations $g(p) = p$, $f(q) = -\sin q$ with initial condition $p(0) = 0$, $q(0) = 2$. The stars plot the numerical solution with $h = 0.5$. The dotted line H = constant provides the true pendulum solution. The dash–dot line $\widetilde{H}_2^h$ = constant gives the solution of the modified system (12). The agreement of the computed points with the modified trajectory is very good.

The origin is a center of the modified system (recall that a small Hamiltonian perturbation of a Hamiltonian center is still a center); this matches the fact that, in the plot, the computed solution does not spiral in or out. On the other hand, the analogous modified system for the (nonsymplectic) integration in 1 is found not be a Hamiltonian system, but rather a system with negative dissipation: This agrees with the spiral behavior observed there.

By adding extra $\mathcal{O}(h^2)$ terms to the right-hand sides of (12), it is possible to construct a (more accurate) modified system $\mathcal{S}_3$ so that (8) is consistent of the *third* order with $\mathcal{S}_3$; thus, $\mathcal{S}_3$ would provide an even better description of the numerical solution. The procedure may be iterated to get modified systems $\mathcal{S}_4, \mathcal{S}_5, \ldots$ and all of them turn out to be Hamiltonian.



Symplectic Methods, Fig. 2 Computed points, true trajectory (*dotted line*) and modified trajectory (*dash–dot line*)

General case. Given an arbitrary Hamiltonian system with a smooth Hamiltonian H, a consistent symplectic integrator ψ_h^H and an arbitrary integer $\rho > 0$, it is possible ([8], Sect. 10.1) to construct a modified Hamiltonian system $\mathcal{S}_\rho$ with Hamiltonian function $\widetilde{H}_\rho^h$, such that ψ_h^H differs from the flow $\phi_h^{\widetilde{H}_\rho^h}$ in $\mathcal{O}(h^{\rho+1})$ terms. In fact, $\widetilde{H}_\rho^h$ may be chosen as a polynomial of degree $< \rho$ in h; the term independent of h coincides with H (cf. (13)) and for a method of order ν the terms in $h, \ldots, h^{\nu-1}$ vanish.

The polynomials in h $\widetilde{H}_\rho^h$, $\rho = 2, 3, \ldots$ are the partial sums of a series in powers of h. Unfortunately this series does not in general converge for fixed h, so that, in particular, the modified flows $\phi_h^{\widetilde{H}_\rho^h}$ cannot converge to ψ_h^H as $\rho \uparrow \infty$. Therefore, in general, it is impossible to find a Hamiltonian $\widetilde{H}^h$ such that $\phi_h^{\widetilde{H}^h}$ coincides *exactly* with the integrator ψ_h^H. Neishtadt ([8], Sect. 10.1) proved that by retaining for each $h > 0$ a suitable number $N = N(h)$ of terms of the series it is possible to obtain a Hamiltonian $\widetilde{H}^h$ such that $\phi_h^{\widetilde{H}^h}$ differs from ψ_h^H in an exponentially small quantity.

Here is the conclusion for the practitioner: For a symplectic integrator applied to an autonomous

Hamiltonian system, modified autonomous Hamiltonian problems exist so that the computed points lie "very approximately" on the exact trajectories of the modified problems. This makes possible a backward error interpretation of the numerical results: The computed solutions are solving "very approximately" a nearby Hamiltonian problem. In a modeling situation where the exact form of the Hamiltonian H may be in doubt, or some coefficients in H may be the result of experimental measurements, the fact that integrating the model numerically introduces perturbations to H comparable to the uncertainty in H inherent in the model is the most one can hope for.

On the other hand, when a nonsymplectic formula is used the modified systems are not Hamiltonian: The process of numerical integration perturbs the model in such a way as to take it out of the Hamiltonian class.

Variable steps. An important point to be noted is as follows: *The backward error interpretation only holds if the numerical solution after n steps is computed by iterating n times one and the same symplectic map.* If, alternatively, one composes n symplectic maps (one from t_0 to t_1, a different one from t_1 to t_2, etc.) the backward error interpretation is lost, because the modified system changes at each step ([8], Sect. 10.1.3).

As a consequence, *most favorable properties of symplectic integrators (and of other geometric integrators) are lost when they are naively implemented with variable step sizes.* For a complete discussion of this difficulty and of ways to circumvent it, see [5], Sects. VIII 1–4.

Finding explicitly the modified Hamiltonians. The *existence* of a modified Hamiltonian system is a general result that derives directly from the symplecticness of the transformation ψ_h^H ([8], Sect. 10.1) and does not require any hypothesis on the particular nature of such a transformation. However, much valuable information may be derived from the *explicit construction* of the modified Hamiltonians. For RK and related methods, a way to compute systematically the $\widetilde{H}_\rho^h$'s was first described by Hairer in 1994 and then by Calvo, Murua, and Sanz-Serna ([5], Sect. IX.9). For splitting and composition integrators, the $\widetilde{H}_\rho^h$'s may be obtained by use of the Baker–Campbell–Hausdorff formula ([8], Sect. 12.3, [5], Sect. III.4) that provides a means to express as a single flow the composition of two flows. This kind of research relies very much on concepts and techniques from the theory of Lie algebras.

Properties of Symplectic Integrators

We conclude by presenting an incomplete list of favorable properties of symplectic integrators. Note that the advantage of symplecticness become more prominent as the integration interval becomes longer.

Conservation of energy. For autonomous Hamiltonians, the value of H is of course a conserved quantity and the invariance of H usually expresses conservation of physical energy. Ge and Marsden proved in 1988 ([8], Sect. 10.3.2) that the requirements of symplecticness and *exact* conservation of H cannot be met simultaneously by a *bona fide* numerical integrator. Nevertheless, symplectic integrators have very good energy behavior ([5], Sect. IX.8): Under very general hypotheses, for a symplectic integrator of order ν: $H(p^n, q^n) = H(p^0, q^0) + \mathcal{O}(h^\nu)$, where the constant implied in the $\mathcal{O}$ notation is independent of n over exponentially long time intervals $nh \leq \exp\big(h_0/(2h)\big)$.

Linear error growth in integrable systems. For a Hamiltonian problem that is integrable in the sense of the Liouville–Arnold theorem, it may be proved ([5], Sect. X.3) that, in (long) time intervals of length proportional to $h^{-\nu}$, the errors in the action variables are of magnitude $\mathcal{O}(h^\nu)$ and remain bounded, while the errors in angle variables are $\mathcal{O}(h^\nu)$ and exhibit a growth that is only linear in t. By implication the error growth in the components of p and q will be $\mathcal{O}(h^\nu)$ and grow, at most, linearly. Conventional integrators, including explicit Runge–Kutta methods, typically show *quadratic* error growth in this kind of situation and therefore cannot be competitive in a sufficiently long integration.

KAM theory. When the system is closed to integrable, the KAM theory ([5], Chap. X) ensures, among other things, the existence of a number of invariant tori that contribute to the stability of the dynamics (see [8], Sect. 10.4 for an example). On each invariant torus the motion is quasiperiodic. Symplectic integrators ([5], Chap. X, Theorem 6.2) possess invariant tori $\mathcal{O}(h^\nu)$ close to those of the system being integrated and furthermore the dynamics on each invariant torus is conjugate to its exact counterpart.

Linear error growth in other settings. Integrable systems are not the only instance where symplectic integrators lead to linear error growth. Other cases include, under suitable hypotheses, periodic orbits, solitons, relative equilibria, etc., see, among others, [1–4].

Cross-References

- ▶ B-Series
- ▶ Composition Methods
- ▶ Euler Methods, Explicit, Implicit, Symplectic
- ▶ Gauss Methods
- ▶ Hamiltonian Systems
- ▶ Molecular Dynamics
- ▶ Nyström Methods
- ▶ One-Step Methods, Order, Convergence
- ▶ Runge–Kutta Methods, Explicit, Implicit
- ▶ Symmetric Methods

References

1. Cano, B., Sanz-Serna, J.M.: Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems. SIAM J. Numer. Anal. **34**, 1391–1417 (1997)
2. de Frutos, J., Sanz-Serna, J.M.: Accuracy and conservation properties in numerical integration: the case of the Korteweg-deVries equation. Numer. Math. **75**, 421–445 (1997)
3. Duran, A., Sanz-Serna, J.M.: The numerical integration of relative equilibrium solution. Geometric theory. Nonlinearity **11**, 1547–1567 (1998)
4. Duran, A., Sanz-Serna, J.M.: The numerical integration of relative equilibrium solutions. The nonlinear Schroedinger equation. IMA. J. Numer. Anal. **20**, 235–261 (1998)
5. Hairer, E., Lubich, Ch., Wanner, G.: Geometric Numerical Integration, 2nd edn. Springer, Berlin (2006)
6. Leimkuhler, B., Reich, S.: Simulating Hamiltonian Dynamics. Cambridge University Press, Cambridge (2004)
7. Sanz-Serna, J.M.: Geometric integration. In: Duff, I.S., Watson, G.A. (eds.) The State of the Art in Numerical Analysis. Clarendon Press, Oxford (1997)
8. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems. Chapman & Hall, London (1994)

Systems Biology, Minimalist vs Exhaustive Strategies

Jeremy Gunawardena
Department of Systems Biology, Harvard Medical School, Boston, MA, USA

Introduction

Systems biology may be defined as the study of how physiology emerges from molecular interactions [11]. Physiology tells us about function, whether at the organismal, tissue, organ or cellular level; molecular interactions tell us about mechanism. How do we relate mechanism to function? This has always been one of the central problems of biology and medicine but it attains a particular significance in systems biology because the molecular realm is the base of the biological hierarchy. Once the molecules have been identified, there is nowhere left to go but up.

This is an enormous undertaking, encompassing, among other things, the development of multicellular organisms from their unicellular precursors, the hierarchical scales from molecules to cells, tissues, and organs, and the nature of malfunction, disease, and repair. Underlying all of this is evolution, without which biology can hardly be interpreted. Organisms are not designed to perform their functions, they have evolved to do so—variation, transfer, drift, and selection have tinkered with them over 3.5×10^9 years—and this has had profound implications for how their functions have been implemented at the molecular level [12].

The mechanistic viewpoint in biology has nearly always required a strongly quantitative perspective and therefore also a reliance on quantitative models. If this trend seems unfamiliar to those who have been reared on molecular biology, it is only because our historical horizons have shrunk. The quantitative approach would have seemed obvious to physiologists, geneticists, and biochemists of an earlier generation. Moreover, quantitative methods wax and wane within an individual discipline as new experimental techniques emerge and the focus shifts between the descriptive and the functional. The great Santiago Ramón y Cajal, to whom we owe the conception of the central nervous system as a network of neurons, classified “theorists” with “contemplatives, bibliophiles and polyglots, megalomaniacs, instrument addicts, misfits" [3]. Yet, when Cajal died in 1934, Alan Hodgkin was already starting down the road that would lead to the Hodgkin-Huxley equations.

In a similar way, the qualitative molecular biology of the previous era is shifting, painfully and with much grinding of gears, to a quantitative systems biology. What kind of mathematics will be needed to support this? Here, we focus on the level of abstraction for modelling cellular physiology at the molecular level. This glosses over many other relevant and hard problems but allows us to bring out some of the distinctive challenges of molecularity. We may caricature the current situation in two extreme views. One approach, mindful of the enormous complexity at the molecular level, strives to encompass that complexity, to dive into it, and to be exhaustive; the other, equally mindful of the complexity but with a different psychology, strives to abstract from it, to rise above it, and to be minimalist. Here, we examine the requirements and the implications of both strategies.

Models as Dynamical Systems

Many different kinds of models are available for describing molecular systems. It is convenient to think of each as a dynamical system, consisting of a description of the state of the system along with a description of how that state changes in time. The system state typically amalgamates the states of various molecular components, which may be described at various levels of abstraction. For instance, Boolean descriptions are often used by experimentalists when discussing gene expression: this gene is ON, while that other is OFF. Discrete dynamic models can represent time evolution as updates determined by Boolean functions. At the other end of the abstraction scale, gene expression may be seen as a complex stochastic process that takes place at an individual promoter site on DNA: states may be described by the numbers of mRNA molecules and the time evolution may be described by a stochastic master equation. In a different physiological context, Ca^{2+} ions are a "second messenger" in many key signalling networks and show complex spatial and temporal behaviour within an individual cell. The state may need to be described as a concentration that varies in space and time and the time evolution by a partial differential equation. In the most widely-used form of dynamical system, the state is represented by the (scalar) concentrations of the various molecular components in specific cellular compartments and the time evolution by a system of coupled ordinary differential equations (ODEs).

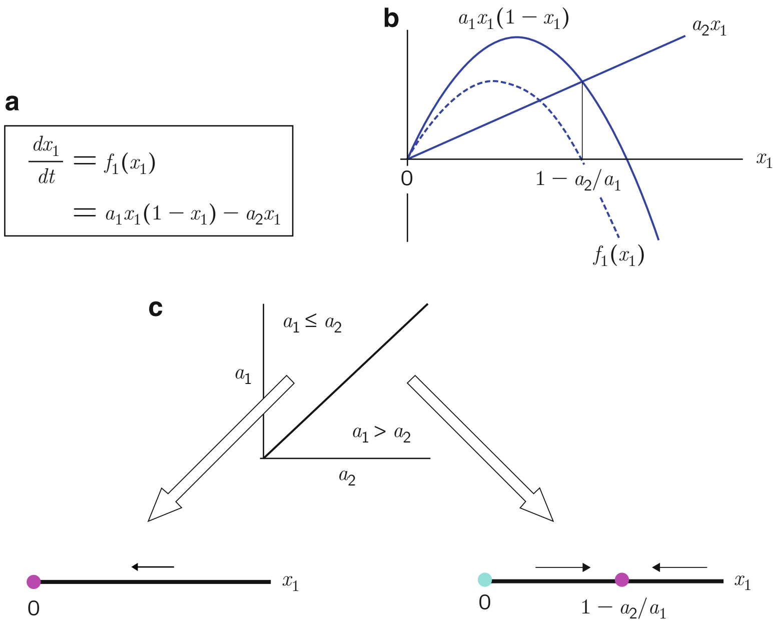
What is the right kind of model to use? That depends entirely on the biological context, the kind of biological question that is being asked and on the experimental capabilities that can be brought to bear on the problem. Models in biology are not objective descriptions of reality; they are descriptions of our assumptions about reality [2].

For our purposes, it will be simplest to discuss ODE models. Much of what we say applies to other classes of models. Assume, therefore, that the system is described by the concentrations within specific cellular compartments of n components, $x_1, \cdots, x_n$, and the time evolution is given, in vector form, by $dx/dt = f(x; a)$. Here, $a \in \mathbb{R}^m$ is a vector of parameters. These may be quantities like proportionality constants in rate laws. They have to take numerical values before the dynamics on the state space can be fully defined and thereby arises the "parameter problem" [6]. In any serious model, most of the parameter values are not known, nor can they be readily determined experimentally. (Even if some of them can, there is always a question of whether an in-vitro measurement reflects the in-vivo context.)

The dynamical behaviour of a system may depend crucially on the specific parameter values. As these values change through a *bifurcation*, the qualitative "shape" of the dynamics may alter drastically; for instance, steady states may alter their stability or appear or disappear [22]. In between bifurcations, the shape of the dynamics only alters in a quantitative way while the qualitative portrait remains the same. The geography of parameter space therefore breaks up into regions; within each region the qualitative portrait of the dynamics remains unaltered, although its quantitative details may change, while bifurcations take place between regions resulting in qualitative changes in the dynamics (Fig. 1).

Parameterology

We see from this that parameter values matter. They are typically determined by fitting the model to experimental data, such as time series for the concentrations of



Systems Biology, Minimalist vs Exhaustive Strategies, Fig. 1 The geography of parameter space. (**a**) A dynamical system with one state variable, x_1 and two parameters, a_1, a_2. (**b**) Graphs of $f_1(x_1)$ (*dashed curve*) and of the terms in it (*solid curves*), showing the steady states, where $dx_1/dt = f_1(x_1) = 0$. (**c**) Parameter space breaks up into two regions: (*1*) $a_1 \leq a_2$, in which the state space has a single stable steady state at $x_1 = 0$ to which any positive initial condition tends (*arrow*); and (*2*) $a_1 > a_2$, in which there are two steady states, a positive stable state at $x_1 = 1 - a_2/a_1$, to which any positive initial conditions tends (*arrows*), and an unstable state at $x_1 = 0$. Here, the state space is taken to be the nonnegative real line. A magenta dot indicates a stable state and a cyan dot indicates an unstable state. The dynamics in the state space undergoes a *transcritical bifurcation* at $a_1 = a_2$ [22]

some of the components. The fitting may be undertaken by minimizing a suitable measure of discrepancy between the calculated and the observed data, for which several nonlinear optimization algorithms are available [10]. Empirical studies on models with many (>10) parameters have revealed what could be described as a "80/20" rule [9, 19]. Roughly speaking, 20 % of the parameters are well constrained by the data or "stiff": They cannot be individually altered by much without significant discrepancy between calculated and observed data. On the other hand, 80 % of the parameters are poorly constrained or "sloppy," they can be individually altered by an order of magnitude or more, without a major impact on the discrepancy. The minimization landscape, therefore, does not have a single deep hole but a flat valley with rather few dimensions orthogonal to the valley. The fitting should have localized the valley within one of the parameter regions. At present, no theory accounts for the emergence of these valleys.

Two approaches can be taken to this finding. On the one hand, one might seek to constrain the parameters further by acquiring more data. This raises an interesting problem of how best to design experiments to efficiently constrain the data. Is it better to get more of the same data or to get different kinds of data? On the other hand, one might seek to live with the sloppiness, to acknowledge that the fitted parameter values may not reflect the actual ones but nevertheless seek to draw testable conclusions from them. For instance, the stiff parameters may suggest experimental interventions whose effects are easily observable. (Whether they are also biological interesting is another matter.) There may also be properties of the system that are themselves insensitive, or "robust," to the parameter differences. One can, in any case, simply draw conclusions based on the fits and seek to test these experimentally.

A successful test may provide some encouragement that the model has captured aspects of the mechanism that are relevant to the question under study. However, models are working hypotheses, not explanations. The conclusion that is drawn may be correct but that may be an accident of the sloppy parameter values or the particular assumptions made. It may be correct for the wrong reasons. Molecular complexity is such that there may well be other models, based on different assumptions, that lead to the same conclusions. Modellers often think of models as finished entities. Experimentalists know better. A model is useful only as a basis for making a better model. It is through repeated tests and revised assumptions that a firmly grounded, mechanistic understanding of molecular behaviour slowly crystallizes.

Sometimes, one learns more when a test is not successful because that immediately reveals a problem with the assumptions and stimulates a search to correct them. However, it may be all too easy, because of the sloppiness in the fitting, to refit the model using the data that invalidated the conclusions and then claim that the newly fitted model "accounts for the new data." From this, one learns nothing. It is better to follow Popperian principles and to specify in advance how a model is to be rejected. If a model cannot be rejected, it cannot tell you anything. In other areas of experimental science, it is customary to set aside some of the data to fit the model and to use another part of the data, or newly acquired data, to assess the quality of the model. In this way, a rejection criterion can be quantified and one can make objective comparisons between different models.

The kind of approaches sketched above only work well when modelling and experiment are intimately integrated [18,21]. As yet, few research groups are able to accomplish this, as both aspects require substantial, but orthogonal, expertise as well as appropriate integrated infrastructure for manipulating and connecting data and models.

Model Simplification

As pointed out above, complexity is a relative matter. Even the most complex model has simplifying assumptions: components have been left out; posttranslational modification states collapsed; complex interactions aggregated; spatial dimensions ignored; physiological context not made explicit. And these are just some of the things we know about, the "known unknowns." There are also the "unknown unknowns." We hope that what has been left out is not relevant to the question being asked. As always, that is a hypothesis, which may or may not turn out to be correct.

The distinction between exhaustive and minimal models is therefore more a matter of scale than of substance. However, having decided upon a point on the scale, and created a model of some complexity, there are some systematic approaches to simplifying it. Here, we discuss just two.

One of the most widely used methods is separation of time scales. A part of the system is assumed to be working significantly faster than the rest. If the faster part is capable of reaching a (quasi) steady state, then the slower part is assumed to see only that steady state and not the transient states that led to it. In some cases, this allows variables within the faster part to be eliminated from the dynamics.

Separation of time scales appears in Michaelis and Menten's pioneering study of enzyme-catalysed reactions. Their famous formula for the rate of an enzyme arises by assuming that the intermediate enzyme-substrate complex is in quasi-steady state [8]. Although the enzyme-substrate complex plays an essential role, it has been eliminated from the formula. The King-Altman procedure formalizes this process of elimination for complex enzyme mechanisms with multiple intermediates. This is an instance of a general method of linear elimination underlying several well-known formulae in enzyme kinetics, in protein allostery and in gene transcription, as well as more modern simplifications arising in chemical reaction networks and in multienzyme posttranslational modification networks [7].

An implicit assumption is often made that, after elimination, the behaviour of the simplified dynamical system approximates that of the original system. The mathematical basis for confirming this is through a singular perturbation argument and Tikhonov's Theorem [8], which can reveal the conditions on parameter values and initial conditions under which the approximation is valid. It must be said that, aside from the classical Michaelis–Menten example, few singular perturbation analyses have been undertaken. Biological intuition can be a poor guide to the right conditions. In the Michaelis–Menten case, for example, the intuitive basis for the quasi-steady state assumption is that under in vitro conditions, substrate, S, is in excess over enzyme, E: $S_{tot} \gg E_{tot}$. However, singular perturbation reveals a broader region, $S_{tot} + K_M \gg E_{tot}$, where K_M is the Michaelis–Menten constant, in which the quasi-steady state approximation remains valid [20]. Time-scale separation has been widely used but cannot be expected to provide dramatic reductions in complexity; typically, the number of components are reduced twofold, not tenfold.

The other method of simplification is also based on an old trick: linearization in the neighbourhood of a steady state. The Hartman–Grobman Theorem for a dynamical system states that, in the local vicinity of a hyperbolic steady state—that is, one in which none of the eigenvalues of the Jacobian have zero real part—the nonlinear dynamics is qualitatively similar to the dynamics of the linearized system, $dy/dt = (Jf)_{ss}y$, where $y = x - x_{ss}$ is the offset relative to the steady state, x_{ss}, and $(Jf)_{ss}$ is the Jacobian matrix for the nonlinear system, $dx/dt = f(x)$, evaluated at the steady state. Linearization simplifies the dynamics but does not reduce the number of components.

Straightforward linearization has not been particularly useful for analysing molecular networks, because it loses touch with the underlying network structure. However, control engineering provides a systematic way to interrogate the linearized system and, potentially, to infer a simplified network. Such methods were widely used in physiology in the cybernetic era [5], and are being slowly rediscovered by molecular systems biologists. They are likely to be most useful when the steady state is homeostatically maintained. That is, when the underlying molecular network acts like a thermostat to maintain some internal variable within a narrow range, despite external fluctuations. Cells try to maintain nutrient levels, energy levels, pH, ionic balances, etc., fairly constant, as do organisms in respect of Claude Bernard's *"milieu intérieure"*; chemotaxing *E. coli* return to a constant tumbling rate after perturbation by attractants or repellents [23]; *S. cerevisiae* cells maintain a constant osmotic pressure in response to external osmotic shocks [16].

The internal structure of a linear control system can be inferred from its frequency response. If a stable linear system is subjected to a sinusoidal input, its steady-state output is a sinuosoid of the same frequency but possibly with a different amplitude and phase. The amplitude gain and the phase shifts, plotted as functions of frequency—the so-called Bode plots, after Hendrik Bode, who developed frequency analysis at Bell Labs—reveal a great deal about the structure of the system [1]. More generally, the art of systems engineering lies in designing a linear system whose frequency response matches specified Bode plots.

The technology is now available to experimentally measure approximate cellular frequency responses in the vicinity of a steady state, at least under simple conditions. Provided the amplitude of the forcing is not too high, so that a linear approximation is reasonable, and the steady state is homeostatically maintained, reverse engineering of the Bode plots can yield a simplified linear control system that may be an useful abstraction of the complex nonlinear molecular network responsible for the homeostatic regulation [15]. Unlike time-scale separation, the reduction in complexity can be dramatic. As always, this comes at the price of a more abstract representation of the underlying biology but, crucially, one in which some of the control structure is retained. However, at present, we have little idea how to extend such frequency analysis to large perturbations, where the nonlinearities become significant, or to systems that are not homeostatic.

Frequency analysis, unlike separation of time scales, relies on data, reinforcing the point made previously that integrating modelling with experiments and data can lead to powerful synergies.

Looking Behind the Data

Experimentalists have learned the hard way to develop their conceptual understanding from experimental data. As the great Otto Warburg advised, *"Solutions usually have to be found by carrying out innumerable experiments without much critical hesitation"* [13]. However, sometimes the data you need is not the data you get, in which case conceptual interpretation can become risky. For instance, signalling in mammalian cells has traditionally relied on grinding up 10^6 cells and running Western blots with antibodies against specific molecular states. Such data has told us a great deal, qualitatively. However, a molecular network operates in a single cell. Quantitative data aggregated over a cell population is only meaningful if the distribution of responses in the population is well represented by its average. Unfortunately, that is not always the case, most notoriously when responses are oscillatory. The averaged response may look like a damped oscillation, while individual cells actually have regular oscillations but at different frequencies and phases [14, 17]. Even when the response is not oscillatory single-cell analysis may reveal a bimodal response, with two apparently distinct sub-populations of cells [4]. In both cases, the very concept of an "average" response is a statistical fiction that may be unrelated to the behaviour of any cell in the population.

One moral of this story is that one should always check whether averaged data is representative of the individual, whether individual molecules, cells, or organisms. It is surprising how rarely this is done. The other moral is that data interpretation should always be mechanistically grounded. No matter how intricate the process through which it is acquired, the data always arises from molecular interactions taking place in individual cells. Understanding the molecular mechanism helps us to reason correctly, to know what data are needed and how to interpret the data we get. Mathematics is an essential tool in this, just as it was for Michaelis and Menten. Perhaps one of the reasons that biochemists of the Warburg generation were so successful, without "critical hesitation," was because Michaelis and others had already provided a sound mechanistic understanding of how individual enzymes worked. These days, systems biologists confront extraordinarily complex multienzyme networks and want to know how they give rise to cellular physiology. We need all the mathematical help we can get.

References

1. Åström, K.J., Murray, R.M.: Feedback systems. An Introduction for Scientists and Engineers. Princeton University Press, Princeton (2008)
2. Black, J.: Drugs from emasculated hormones: the principles of syntopic antagonism. In: Frängsmyr, T. (ed.) Nobel Lectures, Physiology or Medicine 1981–1990. World Scientific, Singapore (1993)
3. Cajal, S.R.: Advice for a Young Investigator. MIT Press, Cambridge (2004)
4. Ferrell, J.E., Machleder, E.M.: The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. Science **280**, 895–898 (1998)
5. Grodins, F.S.: Control Theory and Biological Systems. Columbia University Press, New York (1963)
6. Gunawardena, J.: Models in systems biology: the parameter problem and the meanings of robustness. In: Lodhi, H., Muggleton, S. (eds.) Elements of Computational Systems Biology. Wiley Book Series on Bioinformatics. Wiley, Hoboken (2010)
7. Gunawardena, J.: A linear elimination framework. http://arxiv.org/abs/1109.6231 (2011)
8. Gunawardena, J.: Modelling of interaction networks in the cell: theory and mathematical methods. In: Egelmann, E. (ed.) Comprehensive Biophysics, vol. 9. Elsevier, Amsterdam (2012)
9. Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P.: Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. **3**, 1871–1878 (2007)
10. Kim, K.A., Spencer, S.L., Ailbeck, J.G., Burke, J.M., Sorger, P.K., Gaudet, S., Kim, D.H.: Systematic calibration of a cell signaling network model. BMC Bioinformatics. **11**, 202 (2010)
11. Kirschner, M.: The meaning of systems biology. Cell **121**, 503–504 (2005)
12. Kirschner, M.W., Gerhart, J.C.: The Plausibility of Life. Yale University Press, New Haven (2005)
13. Krebs, H.: Otto Warburg: Cell Physiologist, Biochemist and Eccentric. Clarendon, Oxford (1981)
14. Lahav, G., Rosenfeld, N., Sigal, A., Geva-Zatorsky, N., Levine, A.J., Elowitz, M.B., Alon, U.: Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat. Genet. **36**, 147–150 (2004)
15. Mettetal, J.T., Muzzey, D., Gómez-Uribe, C., van Oudenaarden, A.: The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. Science **319**, 482–484 (2008)
16. Muzzey, D., Gómez-Uribe, C.A., Mettetal, J.T., van Oudenaarden, A.: A systems-level analysis of perfect adaptation in yeast osmoregulation. Cell **138**, 160–171 (2009)
17. Nelson, D.E., Ihekwaba, A.E.C., Elliott, M., Johnson, J.R., Gibney, C.A., Foreman, B.E., Nelson, G., See, V., Horton, C.A., Spiller, D.G., Edwards, S.W., McDowell, H.P., Unitt, J.F., Sullivan, E., Grimley, R., Benson, N., Broomhead, D., Kell, D.B., White, M.R.H.: Oscillations in NF-κB control the dynamics of gene expression. Science **306**, 704–708 (2004)
18. Neumann, L., Pforr, C., Beaudoin, J., Pappa, A., Fricker, N., Krammer, P.H., Lavrik, I.N., Eils, R.: Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. Mol. Syst. Biol. **6**, 352 (2010)
19. Rand, D.A.: Mapping the global sensitivity of cellular network dynamics: sensitivity heat maps and a global summation law. J. R. Soc Interface. 5(Suppl 1), S59–S69 (2008)
20. Segel, L.: On the validity of the steady-state assumption of enzyme kinetics. Bull. Math. Biol. **50**, 579–593 (1988)
21. Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M., Sorger, P.K.: Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature **459**, 428–433 (2009)
22. Strogatz, S.H.: Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry and Engineering. Perseus Books, Cambridge (2001)
23. Yi, T.M., Huang, Y., Simon, M.I., Doyle, J.: Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc. Natl. Acad. Sci. U.S.A. **97**, 4649–4653 (2000)