Highlights

High-Order Schemes of Exponential Time Differencing for Stiff Systems with Nondiagonal Linear Part

Evelina V. Permyakova, Denis S. Goldobin

- An approach to implementation of exponential time differencing methods is presented
- The approach allows for simulation of equation systems with nondiagonal linear part
- Performance gain for high-precision calculations is up to several orders of magnitude
- Efficient for PDEs, Fokker-Planck equations, and large ensembles of oscillators
High-Order Schemes of Exponential Time Differencing for Stiff Systems with Nondiagonal Linear Part

Evelina V. Permyakova\textsuperscript{a}, Denis S. Goldobin\textsuperscript{a, b, ∗}

\textsuperscript{a}Institute of Continuous Media Mechanics UB RAS, 1 Akad. Koroleva street, Perm 614013, Russia
\textsuperscript{b}Institute of Physics and Mathematics, Perm State University, 15 Bukireva street, Perm 614990, Russia

Abstract

Exponential time differencing methods is a power tool for high-performance numerical simulation of computationally challenging problems in condensed matter physics, fluid dynamics, chemical and biological physics, where mathematical models often possess fast oscillating or decaying modes—in other words, are stiff systems. Practical implementation of these methods for the systems with nondiagonal linear part of equations is exacerbated by infeasibility of an analytical calculation of the exponential of a nondiagonal linear operator; in this case, the coefficients of the exponential time differencing scheme cannot be calculated analytically. We suggest an approach, where these coefficients are numerically calculated with auxiliary problems. We rewrite the high-order Runge–Kutta type schemes in terms of the solutions to these auxiliary problems and practically examine the accuracy and computational performance of these methods for a heterogeneous Cahn–Hilliard equation, a sixth-order spatial derivative equation governing pattern formation in the presence of an additional conservation law, and a Fokker–Planck equation governing macroscopic dynamics of a network of neurons.

Keywords:
exponential time differencing, direct numerical simulation, nonlinear partial differential equations, stiff systems

1. Introduction

In condensed matter, chemical and biological physics, many systems are governed by the mathematical models which can be called ‘stiff’ systems—the systems where some modes are fast oscillating or decaying. The dynamics of wave function in atom lattices with frozen parametric disorder in potential is fundamentally contributed by the fast oscillating modes and this dynamics gives rise to the phenomenon of Anderson localization \cite{1,2,3,4,5,6}. The spinodal decomposition \cite{7,8,9,10,11,12} of two-component mixtures and heat-mass-transfer in active media \cite{13,14,15,16,17,18,19,20,21,22} is governed by partial differential equations with high-order spatial derivatives; these derivatives drive a fast decay of short wavelength perturbations. The systems near a SNIPER (saddle-node infinite period) bifurcation of chemical oscillations \cite{23} are subject to thermal noise and can be subject to other noise sources; at the bifurcation point, these systems can be described by the same mathematical models as the quadratic integrate-and-fire neurons with noise and synaptic coupling \cite{24,25,26,27,28,29,30}. The numerical simulation of the Fokker–Planck equation of these chemical oscillation and neuronal systems is again the case of a stiff system.

Generally, fast oscillating or decaying modes impose severe limitation on the time stepsize, which has to be very small. For oscillating modes it is needed to maintain a demanded accuracy of simulation \cite{6}. Fast decaying modes are dying-out and do not influence the dynamics of the chemical/physical system; therefore, there is no need for a high precision of their simulation. However, for a finite stepsize, fast decay can produce numerical instability if the stepsize in not small enough. As a result, the numerical simulation of stiff system requires the employment of \textit{ad hoc} approaches and implicit schemes, which require a sophisticated individual mathematical preliminary work, is costly.
or infeasible in higher dimensions, and are helpless for fast oscillatory modes in conservative systems. Otherwise, the simulation of these problems should be performed with a tiny time stepsize and becomes extremely CPU (central processing unit) time consuming. An alternative approach is a relatively new class of methods of exponential time differencing (ETD) \cite{31,32,33,34,35,36,37,38,39,40}.

In ETD methods \cite{31,32,33} the linear part of equations, which gives the origin to the fast modes, is solved exactly. For the nonlinear part, one constructs an exact solution for a quasistatic approximation of these terms during one time step in the case of a first-order scheme. For an \( n \)-th-order scheme, one effectively constructs the exact solution for the case where the nonlinear part in one time step is an \((n - 1)\)-order polynomial of time \( \tau \). This approach provides high resolution for temporal dynamics of fast modes and prevents numerical instabilities even for a quite large values of time stepsize.

The application of ETD methods is straightforward when the linear part is diagonal, which is abundant for spectral methods for the problems with periodic boundary conditions and Fourier basis. However, for spectral methods with other basis functions (e.g., Chebyshev polynomials) the linear part can become nondiagonal. Moreover, for problems with frozen heterogeneity of parameters \cite{26,41,42,43,44,45}, the linear part becomes essentially nondiagonal. In this case the analytical calculation of the coefficients of ETD schemes (considered in Sec. 2) can become not just difficult but completely impossible.

In this paper we develop an approach to calculation of the coefficients of high-order schemes of exponential time differencing for systems with a nondiagonal linear part. In this approach the coefficients are calculated by means of numerical integration of auxiliary problems. This integration is carried out with a plain standard method, like the predictor–corrector one, and a very small time stepsize \( \tau \), ensuring numerical stability, but over a short time interval—the time step \( \tau \) of the ETD scheme. The basic concept was suggested in \cite{46} and here we develop and apply the approach to the Runge–Kutta type schemes \cite{31} of 3rd and 4th order. For the sake of illustration, we consider the numerical simulation of the discrete-space versions of partial differential equations: Cahn–Hilliard equation \cite{7,8} and Matthews–Cox equation \cite{13,14}, and the Fourier projection of a Fokker–Planck equation \cite{24,27,47}. The performance gain varies from two to many orders of magnitude. The technical implementation of the suggested approach and the bulk of the text below do not employ advanced topics from Linear algebra and Tensor calculus.

The paper is organized as follows. In Sec. 2 we provide three exponential differencing schemes of the Runge–Kutta type and rewrite them in terms of the solutions of the auxiliary problems. The procedure for constructing the solutions of the auxiliary problems is provided. In Sec. 3 we considers examples of the application of the ETD schemes to three generic mathematical models: Cahn–Hilliard equation (Sec. 3.1), sixth-order spatial derivative Matthews–Cox equation (Sec. 3.2), and Fokker–Planck equation for a network of theta-neurons (Sec. 3.3). The accuracy and CPU time performance of the ETD schemes is analysed. In Sec. 4 we derive conclusions. At Appendix A we examine the “strong” accuracy of the suggested approach to numerical solution of the auxiliary problems.

2. Exponential time differencing

We deal with the problem of numerical integration of the equations system of the following form:

\[
\dot{u} = L \cdot u + f(u, t),
\]

where \( u(t) \) is an \( N \)-component vector, \( L \) is an \([N \times N]\)-matrix with time-independent elements, \( f(u, t) \) is the nonlinear part of equations. The decomposition of the equations into the linear part with time-independent coefficients, \( L \cdot u \), and the nonlinear part is not unique and guided by the merit of convenience: \( f(u, t) \) can even contain a time-independent linear part. However, it is desirable that all the dominating sources of numerical instability should be collected into \( L \cdot u \) as fully as possible.

In this paper we consider implementation of three exponential differencing schemes of the Runge–Kutta type \cite{31}; for equation system \( \dot{u} \) these schemes take the following form:

(ETD2RK) two-step scheme:

\[
\begin{align*}
a &= e^{L \tau} \cdot u(t) + L^{-1} \cdot (e^{L \tau} - I) \cdot f(u(t), t), \\
u(t + \tau) &= a + L^{-2} \cdot (e^{L \tau} - I - L \tau) \cdot \frac{f(a, t + \tau) - f(u(t), t)}{\tau},
\end{align*}
\]
where $\tau$ is the time stepsize of the numerical scheme, $a$ is a preliminary approximation of $u$ for the next time step $t+\tau$, $I$ is the unitary matrix, the exponential of a matrix is defined by the series $\exp A = I + A + A^2/2! + A^3/3! + \ldots$; (ETD3RK) three-step scheme:

$$a = e^{L\tau/2} \cdot u(t) + L^{-1} \cdot (e^{L\tau/2} - I) \cdot f(u(t), t),$$

$$b = e^{L\tau} \cdot u(t) + L^{-1} \cdot (e^{L\tau} - I) \cdot [2f(a, t + \tau/2) - f(u(t), t)],$$

$$u(t + \tau) = e^{L\tau} \cdot u(t) + \frac{L^{-3}}{\tau^3} \cdot \left( \left[ -4 - L\tau + e^{L\tau}(4 - 3L\tau + L^2\tau^2) \right] \cdot f(u(t), t)
+ 4 \left[ 2 + L\tau + e^{L\tau}(-2 + L\tau) \right] \cdot f(a, t + \tau/2) + \left[ -4 - 3L\tau - L^2\tau^2 + e^{L\tau}(4 - L\tau) \right] \cdot f(b, t + \tau) \right).$$

where $a$ and $b$ are the preliminary approximations of $u$ for $t + \tau/2$ and $t + \tau$, respectively; (ETD4RK) four-step scheme:

$$a = e^{L\tau/2} \cdot u(t) + L^{-1} \cdot (e^{L\tau/2} - I) \cdot f(u(t), t),$$

$$b = e^{L\tau/2} \cdot u(t) + L^{-1} \cdot (e^{L\tau/2} - I) \cdot f(a, t + \tau/2),$$

$$c = e^{L\tau/2} \cdot a + L^{-1} \cdot (e^{L\tau/2} - I) \cdot [2f(b, t + \tau/2) - f(u(t), t)],$$

$$u(t + \tau) = e^{L\tau} \cdot u(t) + \frac{L^{-3}}{\tau^3} \cdot \left( \left[ -4 - L\tau + e^{L\tau}(4 - 3L\tau + L^2\tau^2) \right] \cdot f(u(t), t)
+ 2 \left[ 2 + L\tau + e^{L\tau}(-2 + L\tau) \right] \cdot (f(a, t + \tau/2) + f(b, t + \tau/2))
+ \left[ -4 - 3L\tau - L^2\tau^2 + e^{L\tau}(4 - L\tau) \right] \cdot f(c, t + \tau) \right).$$

The error of the ETD2RK scheme on one time step is $-\tau^3 f/12$ [3], the ETD3RK scheme error is $\propto \tau^4 d^4 f/d^4 t$, the ETD4RK scheme error is $\propto \tau^4 d^4 f/d^4 t$.

In the numerical schemes above, the reciprocal matrix $L^{-1}$ is introduced for the brevity of expressions using the exponential; wherever the matrix $L^{-1}$ appears in equations, it is multiplied by expressions the se ries representation of which starts from the matrix $L^n$, $m \geq n$. For instance, $L^{-1} \cdot (e^{L\tau} - I)$ is actually a formal expression for the series $\pi L/2 + \pi^2 L^2/2! + \pi^3 L^3/3! + \ldots$. Therefore, the zero eigenvalues of the matrix $L$, which would make $L^{-1} \rightarrow \infty$, are not an issue; the divergence of $L^{-1}$ only prevents the usage of a shorter form of equations. Albeit, in this paper, we are handling the case of a nondiagonal shape of matrix $L$, for which analytical calculation of the matrix $e^{L\tau}$ can be impossible or problematic.

More convenient for our approach is to recast the numerical schemes in terms of the solutions of the following auxiliary problems [4]. The general solution to the Cauchy problem

$$\dot{u} = L \cdot u$$

is given by the matrix $Q(\tau) \equiv e^{L\tau}$ such that

$$u(t = \tau u(0), f = 0) = Q(\tau) \cdot u(0).$$

Hence, one obtains the definition of the matrix $Q(\tau)$, which is convenient for numerical simulations:

$$Q_{\beta}(\tau) = u_j(t = \tau u(0) = \delta_{\beta}, f = 0).$$

The solutions to the problems

$$\dot{u} = L \cdot u + g t^\alpha, \quad u(0) = 0, \quad g = \text{const}$$

are given by the matrices $M_\alpha(\tau) \equiv \int_0^\tau e^{L(t - \tau)} t^\alpha - 1 dt$:

$$u(t = \tau u(0) = 0, f(t) = g t^{\alpha - 1}) = M_\alpha(\tau) \cdot g.$$

Hence, one obtains the definition of the matrix $M_\alpha(\tau)$, which is convenient for numerical simulations:

$$\left(M_\alpha(\tau)\right)_\beta = u_j(t = \tau u(0) = 0, f(t) = \delta_{\beta} t^{\alpha - 1}).$$
Using the recurrent relationship $M_{n+1}(\tau) = L^{-1} \cdot (\tau^n I + nM_n(\tau))$, one can obtain

$$
M_1(\tau) = L^{-1} \cdot (\tau^1 I - I), \quad (9a)
$$

$$
M_2(\tau) = L^{-2} \cdot (\tau^2 I - L\tau), \quad (9b)
$$

$$
M_3(\tau) = L^{-3} \cdot (2\tau^3 I - 2L\tau - L^2\tau^2), \quad (9c)
$$

As proposed in [46], for an essentially nondiagonal shape of matrix $L$, the evaluation of matrices $Q$ and $M_n$ defined by equations (6) and (8) can be conducted via the direct numerical integration of problems (5) and (7) with a very small time stepsize but on a short time interval — one step of the ETD scheme $\tau$.

In terms of matrices $M_n$ the schemes (2), (3), and (4) acquire the following form:

**(ETD2RK):**

$$
a = Q \cdot u(t) + M_1 \cdot f(u(t), t),
$$

$$
u(t + \tau) = a + \tau^{-1}M_2 \cdot [f(a, t + \tau) - f(u(t), t)];
$$

**(ETD3RK):**

$$
a = Q_2 \cdot u(t) + M_{1, z} \cdot f(u(t), t),
$$

$$
b = Q \cdot u(t) + M_1 \cdot [2f(a, t + \tau/2) - f(u(t), t)],
$$

$$
u(t + \tau) = Q \cdot u(t) + \left[\frac{2M_3}{\tau} - \frac{3M_2}{\tau} + M_1\right] \cdot f(u(t), t)
$$

$$
- \left[\frac{4M_3}{\tau^2} - \frac{4M_2}{\tau}\right] \cdot f(a, t + \tau/2) + \left[\frac{2M_3}{\tau^2} - \frac{M_2}{\tau}\right] \cdot f(b, t + \tau),
$$

where $Q_2 = Q(\tau/2)$, $M_{1, z} = M_1(\tau/2)$, and for the full stepsize $\tau$ the corresponding subscript is omitted;

**(ETD4RK):**

$$
a = Q_2 \cdot u(t) + M_{1, z} \cdot f(u(t), t),
$$

$$
b = Q_2 \cdot u(t) + M_{1, z} \cdot f(a, t + \tau/2),
$$

$$
c = Q_2 \cdot a + M_{1, z} \cdot [2f(b, t + \tau/2) - f(u(t), t)],
$$

$$
u(t + \tau) = Q \cdot u(t) + \left[\frac{2M_3}{\tau^2} - \frac{3M_2}{\tau} + M_1\right] \cdot f(u(t), t)
$$

$$
- \left[\frac{2M_3}{\tau^3} - \frac{2M_2}{\tau}\right] \cdot f(a, t + \tau/2) + \left[\frac{2M_3}{\tau^3} - \frac{M_2}{\tau}\right] \cdot f(c, t + \tau).
$$

In the literature, for the difficult case of non-diagonalizable $L$, the matrices $Q$ and $M_n$ are calculated approximately from formulas (9) either by means of the Taylor expansions and projection into Krylov subspaces [38, 39] or with discretized contour integrals on the complex plane [40]. Additionally, for small eigenvalues of matrix $L$ (the presence of which is abundant), the computations with formula $(e^L - I)/(\tau L)$ [41] suffer from the cancellation errors which rapidly grow for $M_n$ as $n$ increases further above 1. Tackling the latter issue is involved but doable [31, 40]; this issue does not appear with the algorithm we consider in this paper. Moreover, the technical implementation of this algorithm and the text below do not employ any advanced topics from Linear algebra and Tensor calculus.

3. Examples

3.1. 1D Cahn–Hilliard equation

For a broad class of active media systems the pattern formation in thin layers is governed by the Cahn–Hilliard equation [13, 14, 18] (CHE), which as well describes the spinodal decomposition of two-component mixtures [6, 8].
For the generality of consideration, we will also allow for an advective transfer along the layer [8, 18]. In the one-dimensional case (the patterns are homogeneous along the second direction in the layer plane), CHE for the field $u(x, t)$ with advection velocity $v$ reads

$$\frac{\partial u}{\partial t} = -v \frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial x^2} \left[ q(x) + \frac{\partial^2 u}{\partial x^2} - u^3 \right],$$  \hspace{1cm} (13)

where $q(x)$ is the local deviation of the bifurcation parameter from the instability threshold of the homogeneous infinite layer (the patterns are excited for $q > 0$).

For an example, we consider CHE in a domain $0 < x < L$ with trivial boundary conditions:

$$u(0) = \frac{\partial u}{\partial x} \bigg|_{x=0} = u(L) = \frac{\partial u}{\partial x} \bigg|_{x=L} = 0.$$

The local excitability parameter $q(x)$ is assumed to be positive within certain excitation zone and negative beyond it:

$$q(x) = \begin{cases} 
2.5, & \text{for } \frac{3L}{10} < x < \frac{7L}{10}; \\
-3, & \text{otherwise}.
\end{cases}$$ \hspace{1cm} (14)

For a localized excitation $q(x)$, strong enough advection $v$ results in an oscillatory behavior (e.g., see [45]). For $v = 1$ and $q(x)$ given by (14), the oscillatory solution of Eq. (13) is presented in Fig. 1.

3.1.1. Numerical scheme

The spatially-discrete version of CHE (13) for vector $\mathbf{u}(t) \equiv [u_j(t)]_{j = 1, 2, ..., N}$, where $u_j(t) = u(jh_x, t)$, $h_x = L/N$, can be written as Eq. (15) with

$$(\mathbf{L} \cdot \mathbf{u})_j = v \frac{u_{j-1} - u_{j+1}}{2h_x} - \frac{q_j u_{j+1} - 2q_j u_j + q_{j-1} u_{j-1}}{h_x^2} - \frac{u_{j+2} - 4u_{j+1} + 6u_j - 4u_{j-1} + u_{j-2}}{h_x^2},$$ \hspace{1cm} (15a)

$$f_j(\mathbf{u}, t) = \frac{u^3_{j+1} - 2u^3_j + u^3_{j-1}}{h_x^2},$$ \hspace{1cm} (15b)

where $q_j = q(jh_x)$ \hspace{1cm} (14). The trivial boundary conditions we consider imply that one can formally set $u_{-1} = u_0 = u_{N+1} = u_{N+2} = 0$ in the right-hand parts of (15).

The problem (15) can be numerically simulated with the basic Euler scheme or the predictor–corrector (PC) one. For both of these schemes, the forth-order $x$-derivative term imposes the limitation on the time stepsize, $\tau_1 < h_x^4/8$; for a bigger time stepsize the direct numerical simulation becomes unstable. More sophisticated Runge–Kutta type
Figure 2: For Cahn–Hilliard equation (13), matrices $\log_{10}|Q_{jk}|$ (panel a), $\log_{10}|M_{1,jk}|$ (panel b), $\log_{10}|(\tau^{-1}M_{2,jk})|$ (panel c), $\log_{10}|(\tau^{-2}M_{3,jk})|$ (panel d) are plotted versus indices $k$ and $j$ for $\tau = 2.5 \times 10^{-4}$, $L = 10$, $N = 200$; $\tau_1 = 0.1h_4^2 = 6.25 \times 10^{-7}$.

schemes or any other high-order methods are not needed here as the PC scheme with such a small time stepsize already warrants an excessive accuracy of the numerical integration in time. Henceforth, we will employ the PC scheme:

$$F(t) = L \cdot u(t) + f(u(t), t),$$

$$a = u(t) + F(t)\tau_1,$$

$$F(t + \tau_1) = L \cdot a + f(a, t),$$

$$u(t + \tau_1) = u(t) + \frac{F(t) + F(t + \tau_1)}{2}\tau_1.$$  \hspace{1cm} (16a)  \hspace{1cm} (16b)  \hspace{1cm} (16c)  \hspace{1cm} (16d)

For problem (13), the PC scheme (16) was employed to integrate the auxiliary problems (5) and (7) for $t \in [0, \tau]$ to evaluate matrices $Q_{zk}$, $Q$, $M_{1,z}$, $M_n$, $n = 1, 2, 3$ (8), respectively. The time stepsize for the direct numerical integration of the auxiliary problems was $\tau_1 = 0.1h_4^2$. In Fig. 2, a sample structure of these matrices can be seen.

3.1.2. Accuracy and performance of numerical simulation with exponential time differencing schemes

The ETD2RK, ETD3RK, and ETD4RK schemes with matrices $Q_{zk}$, $Q$, $M_{1,z}$, $M_n$, $n = 1, 2, 3$ calculated as described in Sec. 3.1.1 were employed for the numerical simulation of CHE (13) with $\nu = 1$ and $q(x)$ given by (14) in the domain of length $L = 10$ (see Fig. 1). In Figs. 3 and 4, the dependencies of the accuracy and the simulation performance versus the stepsize $\tau$ of the ETD schemes are presented.

In Fig. 3, one can compare the performance of ETD methods to that of the PC method (blue solid circles). The numerical error of a specific scheme is evaluated as the deviation of the result from the result calculated with the PC
Figure 3: The error rate and the CPU time for the numerical simulation of Cahn–Hilliard equation (13) with several ETD schemes over the time interval $0 < t < 50$ are plotted versus the ETD stepsize $\tau$ for $N = 200$ (i.e. $h_x = 0.05$). Solid squares: ETD2RK, solid up-triangles: ETD3RK, solid down-triangles: ETD4RK, solid circles: predictor–corrector scheme with fixed time stepsize $\tau_1 = 0.1h_x^2 = 6.25 \times 10^{-7}$. CPU times for the preliminary (preparatory) calculation of the matrices $Q$ and $M_n$ required for the respective ETD scheme are plotted in the right panel with open symbols. CPU times are provided for the processor Intel(R) Core(TM) i7-4790K CPU 4.00 GHz, disabled hypertrading; RAM: DDR3 16 GB; program in FORTRAN.

Figure 4: The error rate and the CPU time for the numerical simulation of CHE (13) with several ETD schemes are plotted versus the ETD stepsize $\tau$ for $N = 400$ (i.e. $h_x = 0.025$); $\tau_1 = 0.1h_x^2 = 3.90625 \times 10^{-8}$. See Caption to Fig. 3 for notations and values of other parameters.

The accuracy of the PC method is excessive, but it is dictated by the maximal admissible value of the time stepsize $h_x^2/8$. In turn, the ETD methods can provide decent accuracy even for large values of $\tau$, where they perform much faster (see the right panel) than the PC method. For example, with a required accuracy of $10^{-6}$ (see crests in Fig. 3), one can pick-up as large time stepsize as $\tau \approx 0.005$ for the ETD2RK scheme, 0.02 for ETD3RK, and 0.04 for ETD4RK. The biggest performance gain is achieved with ETD4RK scheme and exceeds a factor of 100 for all three schemes. For an increased accuracy of the spatial discretization (see Fig. 4 where $N$ is doubled compared to the case of Fig. 3), the performance gain provided by ETD schemes significantly increases.

Note, the CPU time for the preliminary calculations of $Q$ and $M_n$ increases linearly with $\tau$ and can become nonsmall (see open symbols in the right panels of Figs. 3 and 4). This CPU time can be significant for the problems where simulation over short or moderate intervals of time is sufficient. For these problems, an optimal performance is achieved for the values of $\tau$, which provide approximately equal CPU times for the preliminary calculation of $Q$ and $M_n$ and the run of the ETD scheme (see the crossings of the lines marked by solid and open symbols in Figs. 3 and 4). The overall performance gain is decreased, but still large. The ETD2RK scheme becomes the most efficient one in this case. However, the most computationally demanding problems are those of complex dynamics [8, 16, 19, 20, 21, 48, 49, 51, 52, 54, 55] (including spatiotemporal chaos) and dynamics in the presence of frozen parametric disorder [41, 43, 44, 45], where the CPU time for the preliminary calculations can be neglected.

In Appendix A our analysis is complemented with the examination of the accuracy of approximate calculation of
Figure 5: For Matthews–Cox equation (17), matrices \(\log_{10}|Q|\) (panel a), \(\log_{10}|M_1|\) (panel b), \(\log_{10}|(\tau^{-1}M_2)|\) (panel c), \(\log_{10}|(\tau^{-2}M_3)|\) (panel d) are plotted versus indices \(k\) and \(j\) for \(\tau = 1.6 \times 10^{-4}, L = 10, N = 100, r_1 = 0.02L^3 = 2 \times 10^{-3}\).

\[Q\] and \(M_n\). The inaccuracies in \(Q\) and \(M_n\) are primarily contributed by fast decaying modes, for which the contribution into the inaccuracy of a numerically simulated solution \(u(t)\) is exponentially suppressed. The impact of inaccuracies of \(Q\) and \(M_n\) related to different decaying perturbation modes onto the inaccuracy of \(u(t)\) is practically impossible to unpack. The net error of the numerical method is contributed by two sources: the error of the ETD scheme and the error of approximate calculation of matrices \(Q\) and \(M_n\). With all the importance of the information on the inaccuracy of \(Q\) and \(M_n\), the net error decides the practical applicability of the approach; this error is reported in the left panels of Figs. 3 and 4.

3.2. 1D pattern formation with a conservation law

The conservation laws (of the chemical specie mass, etc.) are structurally stable (persistent) properties of systems and influence the general form of equations governing the pattern formation. Matthews and Cox [15] studied the model equation governing pattern formation with an additional conservation law:

\[
\frac{\partial u}{\partial t} = -v \frac{\partial u}{\partial x} - \frac{\partial^3 u}{\partial x^3} \left[ q(x)u - 2 \frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial x^4} - u^3 \right],
\]

(17)

where the terms inside the brackets are just those of the well-studied Swift–Hohenberg equation [56] and an additional advective \(v\)-term is introduced [16]. Let us consider the Matthews–Cox equation (MCE) as an example of the pattern formation equation with the sixth-order spatial derivative.
The local excitability parameter $q$ where $v$ enough advection $u$.

Figure 7: The error rate and the CPU time for the numerical simulation of MCE (17) with several ETD schemes are plotted versus the ETD stepsize $\tau$ for $N = 50$ (i.e. $h_x = 0.2$); $\tau_i = 0.02h_x^2 = 1.6 \times 10^{-1}$. See Caption to Fig 3 for notations and values of other parameters.

Figure 6: The error rate and the CPU time for the numerical simulation of MCE (17) with several ETD schemes are plotted versus the ETD stepsize $\tau$.

3.2.1. Numerical scheme

Eq. (17) is presented in Fig. 1.

We consider MCE in a domain $0 < x < L$ with trivial boundary conditions:

$$u(0) = \frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial x^2} = u(L) = \frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial x^2} = 0.$$  

The local excitability parameter $q(x)$ is assumed to be the same as for CHE (14). For a localized excitation $q(x)$, strong enough advection $v$ results in an oscillatory behavior. For $v = 1$ and $q(x)$ given by (14), the oscillatory solution of Eq. (17) is presented in Fig. 1.

3.2.1. Numerical scheme

The spatially-discrete version of MCE (17) for vector $u(t) \equiv \{u_j(t)\}_{j = 1, 2, \ldots, N}$, where $u_j(t) = u(jh_x, t)$, $h_x = L/N$, can be written as Eq. (1) with

$$f_j(u, t) = \frac{u^{j+1}_t - u^{j+2}}{2h_x^2} - \frac{q_j u^{j+1}_u + q_{j-1} u^{j-1}_u}{2h_x} + \frac{u^{j+2}_u - 4u^{j+1}_u + 6u^{j+1}_u - 4u^{j-1}_u + u^{j-2}_u}{h_x^2} - \frac{u^{j+1}_u - 6u^{j+2}_u + 15u^{j+1}_u - 20u^{j}_u + 15u^{j+1}_u - 6u^{j+2}_u + u^{j+3}_u}{h_x^2}, \quad (18a)$$

where $q_j = q(jh_x)$ (14). The trivial boundary conditions we consider imply that one can formally set $u_{-2} = u_{-1} = u_0 = u_{N+1} = u_{N+2} = u_{N+3} = 0$ in the right-hand parts of (15).
The problem (11,18) can be numerically simulated with the basic Euler scheme or the predictor–corrector one. For both of these schemes, the sixth-order x-derivative term imposes the limitation on the time stepsize, \( \tau_1 < h_{\text{opt}}^6 / 32 \); for a bigger time stepsize the direct numerical simulation becomes unstable. Henceforth, we will employ the PC scheme (16).

For problem (11,18), the PC scheme (16) was employed to integrate the auxiliary problems (5) and (7) for \( t \in [0, \tau] \) to evaluate matrices \( Q_2, Q, M_1, \ldots, M_n, n = 1, 2, 3 \). The time stepsize for the direct numerical integration of the auxiliary problems was \( \tau_1 = 0.02 h_{\text{opt}}^6 \). In Fig. 5 a sample structure of these matrices can be seen.

3.2.2. Accuracy and performance of numerical simulation with exponential time differencing schemes

The ETD2RK, ETD3RK, and ETD4RK schemes with matrices \( Q_2, Q, M_1, \ldots, M_n, n = 1, 2, 3 \) calculated as described in Sec. (12,13) were employed for the numerical simulation of MCE (17) with \( v = 1 \) and \( q(x) \) given by (13) in the domain of length \( L = 10 \) (see Fig. 1). In Figs. 6 and 7 the dependencies of the accuracy and the simulation performance versus the stepsize \( \tau \) of the ETD schemes are presented. One can see even more dramatic performance gain than for the numerical simulation of CHE (14).

Generally, for the partial differential equations with the highest-order spatial derivative \( \partial^m / \partial x^m \), the performance gain for the ETD methods is \( \propto h_{\text{opt}}^{m-2} \), as anticipated in (46). However, in Figs. 2 and 5 one can see that matrix elements rapidly decay away from the diagonal. For nonlarge \( \tau \), the absolute values of a significant fraction of elements are below \( \sim 10^{-16} \) which is the level of the double precision computer accuracy. The elements smaller than the required error level can be set to zero without any damage to the scheme accuracy. The optimization of a program code for the sparseness of matrices \( Q \) and \( M_n \) allows for a noticeable performance acceleration. For a given stepsize \( \tau \), the highest derivative \( \partial^m / \partial x^m \) creates a bigger spread of nonsmall values of matrix elements (compare Fig. 2 to Fig. 5 where the stepsize \( \tau \) is even somewhat smaller). Therefore, this additional optimization becomes less beneficial for high-order derivatives, where the gain is actually large without any optimization.

3.3. Macroscopic dynamics of population of quadratic integrate-and-fire neurons with noise

The next example we consider is the macroscopic dynamics of a large population of coupled oscillators. The quadratic integrate-and-fire neuron (QIF) (57) is not only a useful toy model for the excitable behavior of neurons, but also a normal for the Class I excitability neurons near the bifurcation point (58). Networks of QIFs can exhibit sophisticated macroscopic dynamics; these dynamics and the effect of endogenic noise on them can be often studied and well understood within the framework of the Fokker–Planck equation (24, 25, 26, 27, 28, 29). However, the problem of reasonably accurate direct numerical simulation of it for some important collective regimes becomes a challenging task. Below we provide the equations for a large recurrent synaptic network of QIFs with global coupling and endogenic Gaussian noise, explain typical accuracy breakdown of its direct numerical simulation, and test three ETD schemes for it.

3.3.1. Governing equations

In this section we provide a concise derivation and interpretation for equation system (25) and (27), the direct numerical simulation of which can be a challenging task tackled by means of the ETD-methods.

We consider the population of QIFs with endogenic noise:

\[ V_n = V_n^2 + I_n, \]
\[ I_n = \eta_n + \sigma \zeta_n(t) + J s(t) + I(t), \]

where \( V_n \) is the membrane voltage of the \( n \)th neuron, \( \eta_n \) is the excitability parameter of individual neuron (an isolated neuron is excitable for \( \eta_n < 0 \) and periodically spiking otherwise), \( I(t) \) is the external input current, \( \sigma \zeta_n(t) \) are independent Gaussian endogenic noises: \( \langle \zeta_n(t) \zeta_m(t + \tau) \rangle = 2 \delta_{nm} \delta(\tau) \), \( \delta_{nm} \) is the Kronecker delta which is 1 for \( n = m \) and 0 otherwise. When \( V_n \) reaches \(+\infty\) it is reset to \(-\infty\) and a synaptic spike is generated (58). The input synaptic current from other neurons \( J s(t) \) is characterised by the coupling coefficient \( J \), which is negative for an inhibitory coupling, and a common field \( s(t) \) equals to the neuron firing rate \( r(t) \) for a thermodynamically large population with instantaneous synaptic spikes.
and parameters \( \alpha^2 = 10 \), and periodically modulated external input current \( I(t) = 0.3 \sin 10t \). The numerical errors \( \|z_j\| \) are plotted for the simulations with ETD schemes with \( r = 0.025 \) and \( r_t = 1.25 \times 10^{-6} \).

One can introduce a phase variable \( \phi \),

\[ V_n = \tan \frac{\phi_n}{2}, \]

and rewrite Eq. (19) as

\[ \phi_n = (1 - \cos \phi_n) + (1 + \cos \phi_n)[\eta_n + a_\eta(t) + J_s(t) + I(t)]. \]

Let us index the QIFs with the value of their parameter \( \eta_n \). The Fokker–Planck equation for the probability density \( w_\eta(\phi, t) \) of stochastic system (21) reads

\[ \frac{\partial w_\eta}{\partial t} + \frac{\partial}{\partial \phi} \left[ (1 - \cos \phi + (1 + \cos \phi)(\eta + J_s + I(t))w_\eta \right] = \sigma^2 \left( \frac{\partial}{\partial \phi} (1 + \cos \phi) \frac{\partial}{\partial \phi} (1 + \cos \phi) w_\eta \right). \]

Now we calculate the firing rate \( r \) in terms of \( \phi \). Given the distribution of \( \eta \) as \( g(\eta) \), in the thermodynamic limit of large population, the firing rate equals \( r(t) = \int q_\eta(\phi = \pi) g(\eta) d\eta \), where \( q_\eta \) is the probability density flux,

\[ q_\eta = [1 - \cos \phi + (1 + \cos \phi)(\eta + J_s + I(t))]w_\eta - \sigma^2(1 + \cos \phi) \frac{\partial}{\partial \phi} (1 + \cos \phi) w_\eta. \]

For \( \phi = \pi \), the flux \( q_\eta(\phi = \pi) = 2w_\eta(\pi) \) and one finds

\[ r(t) = \int q_\eta(\pi) g(\eta) d\eta = 2 \int w_\eta(\pi) g(\eta) d\eta. \]

In the Fourier space, \( w_\eta(\phi, t) = \frac{1}{2\pi} \{ 1 + \sum_{j=1}^\infty (a_j(t) e^{-i j \phi} + c.c.) \} \), where “c.c.” stands for complex conjugate, and Eq. (22) takes the following form:

\[ \dot{a}_j(\eta) = \frac{i}{2} a_j + 2 A_\eta (a_j - 1) + j^2 A_\eta (a_j - j^2 A_\eta (a_j) \right) \]

\[ - \sigma^2 \left[ \frac{3}{2} j^2 a_j + \left( j^2 - \frac{j}{2} \right) j^2 a_j - 1 + \left( j^2 + \frac{j}{2} \right) a_j + \frac{j(j-1)}{4} a_{j-2} + \frac{j(j+1)}{4} a_{j+2} \right], \]

where \( A_\eta = \eta + J_s + I(t) - 1 \); \( a_0 = 1 \) and \( a_{-j} = a_j^* \), by definition. For a heterogeneous population with Lorentzian distribution \( g(\eta) = \gamma / [\pi(\gamma^2 + (\eta - \eta_0)^2)] \) with median \( \eta_0 \) and half-width at half maximum \( \gamma \), one can employ the Residue theorem [59,60,61] to derive from Eq. (24) the infinite chain of dynamics equations for the Kuramoto–Daido order parameters [62].

\[ z_j = \int d\eta g(\eta) a_j(\eta) = \int d\eta \int d\phi v_\phi(\phi) e^{ij \phi}. \]
Figure 9: For equation system (26) and (27), matrices $\log_{10}|Q_{jk}|$ (panel a), $\log_{10}|(M_1)_{jk}|$ (panel b), $\log_{10}|(r^{-1}M_2)_{jk}|$ (panel c), $\log_{10}|(r^{-2}M_3)_{jk}|$ (panel d) are plotted versus indices $k$ and $j$ for $r = 10^{-4}$, $q_0 = -2$, $\gamma = 1$, $J = 10$, $\sigma^2 = 10$, $\tau_1 = 1.25 \times 10^{-6}$.

One finds

$$\dot{z}_j = \frac{j}{2} \left[ 2iz_j + \frac{iA_0 - \gamma}{2} (z_{j-1} + 2z_j + z_{j+1}) \right] - \sigma^2 \left[ \frac{3}{2} j^2 z_j + \left( j^2 - \frac{j}{2} \right) z_{j-1} + \left( j^2 + \frac{j}{2} \right) z_{j+1} + \frac{j(j - 1)}{4} z_{j-2} + \frac{j(j + 1)}{4} z_{j+2} \right]$$

$$\equiv \left( L \cdot z \right)_j + f_j(z, t),$$

(25)

where $z = \{z_0, z_1, z_2, \ldots \}$ (which does not contradict $z_0 = 1$),

$$f_j = ij \frac{Jr + I(t)}{2} (z_{j-1} + 2z_j + z_{j+1}),$$

and everything else is collected in the linear part with constant coefficients. Notice, the $Jr$-term of $f$ is quadratic with respect to $z$, as the firing rate $r(t)$ is a function of the system state. Namely, substituting the Fourier series of $w_0(\phi)$ into Eq. (23), one finds the neuron firing rate [63]

$$r = \frac{1}{\pi} \left( 1 - z_1 - z_1^* + z_2 + z_2^* - z_3 - z_3^* + z_4 + z_4^* + \ldots \right)$$

$$= \mathrm{Re} W,$$

$$W = 1 - 2z_1 + 2z_2 - 2z_3 + 2z_4 + \ldots .$$

(27)

To summarize, the macroscopic dynamics of a recurrent synaptic network of QIFs with all-to-all coupling is governed by the system of equations (25) and (27); its Eq. (1) form is given by (26) and (27).
3.3.2. Direct numerical simulation challenges and ETD methods

For perfect synchrony regimes, the distribution \( w(\phi, t) \) is a travelling delta-function \( \delta(\phi - \varphi(t)) \), the Fourier spectrum of which is \( z_j = e^{jk\varphi(t)} \) and does not decay with \( j \). In physically meaningful set-ups the synchrony will be imperfect, but still high-synchrony regimes are of interest. The series \( z_j \) for these regimes decay slowly and may require as many as 1000–2000 elements for a reasonably accurate simulation of the dynamics of observables \( z_j(t), n(t), \) etc. \( 47 \) In what follows, we estimate how much demanding for the numerical time stepsize are Eqs. \( 25 \) for large \( j \).

Typical explicit schemes which are stable for infinitesimally small time stepsize can become unstable for a larger stepsize \( \tau_* \). For a perturbation \( \delta z_j \sim C_\ell e^{ik\ell} \), where only \( -\pi < k < \pi \) make sense as \( j \) is integer, the reference value of \( \tau_* \) can be estimated from the condition \( |\delta z_j\tau_*|/|\delta z_j| \sim 1 \) (details vary from scheme to scheme, but the order of magnitude is 1). For Eq. \( 25 \) and large \( j \), this condition reads:

\[
\left| j \left( 2i + (iA_\gamma - \gamma)(1 + \cos k) \right) - j\sigma^2 \left( \frac{3}{2} + 2 \cos k + \frac{1}{2} \cos 2k \right) + i \sin k + \frac{i}{2} \sin 2k \right| \tau_* \sim 1.
\]

Keeping only principal contributions and maximizing over \( k \), one finds approximately

\[
2j \sqrt{(1 + A_\gamma)^2 + \gamma^2 + 4j^2\sigma^2} \tau_* \sim 1.
\]

The Runge–Kutta–Merson and simplest predictor-corrector schemes yield stability threshold values of the same order of magnitude. For computationally challenging cases of \( j \sim 1000 \), the maximal admissible time stepsize for both conventional schemes becomes as small as \( \tau_* \sim 10^{-7}/(\sigma^2 + 0.0005(1 + A_\gamma)^2 + \gamma^2) \); however, this and even bigger number of modes is required for some high-synchrony regimes.

For the application of an ETD method for this problem, one practically cannot make the matrix \( L \) diagonal or calculate matrix \( \exp(L\tau) \) analytically in some other way. On the other hand, one can calculate matrices \( Q, M_\alpha, \) etc. numerically with the Runge–Kutta–Merson or simplest predictor-corrector schemes. In Fig. 8 one can see the results of the direct numerical simulation of system \( 25 \) and \( 27 \) with ETD methods for \( N = 200 \) modes \( z_j \). The reference "accurate" solution is calculated with the predictor-corrector scheme \( 16 \) and small time step \( \tau = 1.25 \times 10^{-7} \). The ETD scheme stepsize is set to a large value \( \tau = 0.025 \) in order to illustrate the accuracy order of three employed ETD schemes. In Fig. 9 the structure of numerically calculated matrices \( Q, M_\alpha \) is presented for \( \tau = 10^{-4} \).

In Table 1 the performance of three ETD schemes is compared against the background of the conventional Runge–Kutta–Merson and predictor-corrector \( 15 \) schemes. For typical studies, as many modes as \( N = 200–400 \) is sufficient for accurate resolution of fine elements of phase diagrams. For short-term simulations, optimal stepsize was found to be around \( \tau = 0.0005 \) and ETD schemes provide the performance gain in range 3–10; for long-term simulations, where one can neglect the time for calculation of the scheme coefficients, a bigger stepsize \( \tau = 0.005 \) allows one to have decent accuracy alongside with the performance gain up to two orders of magnitude.

4. Conclusion

We have suggested a practical approach for the implementation of the exponential time differencing schemes of the Runge–Kutta type \( 31 \) for numerical simulation of stiff systems with nondiagonal linear part. In this approach, the numerical schemes are written in terms of the solutions of auxiliary problems which are preliminary integrated numerically by means of an explicit basic method with a very small time step but over a short interval of time—one step of the ETD scheme. As a result, one can not only use the ETD methods for the systems where an analytical calculation of the exponential of a nondiagonal linear operator is infeasible, but also employ the same program code for different equation systems, the only part of code which is subject to change is the subroutine with the equations to be simulated.

\[ ^1 \text{The stepsize around } \tau \sim 10^{-7} \text{ was found to provide the best accuracy with double precision calculations for the given set of parameter values. Here, the accumulation of the relative machine error } \sim 10^{-15} \text{ at each time step sums up over the unit time to a value much smaller than } \sim 10^{-15}/10^{-7} = 10^{-8} \text{ because of the fast decay of the leading perturbation modes; as a result, the net numerical error rate per the unit time is well below } 10^{-12}. \]
Table 1: CPU time for the direct numerical simulation of equation system \( \text{(25)} \) and \( \text{(27)} \) is averaged over 100 runs on the dimensionless time interval \( t \in [0; 5] \) for \( q_0 = -2, \gamma = 1, J = 10, \sigma^2 = 10, H(t) = 0.3 \sin 10t \); “preparation”: time for calculation of \( Q \) and \( M_0 \) [sec]; “run”: time of simulation with a numerical scheme with pre-calculated matrices \( Q \) and \( M_0 \) [sec]; stepsize \( \tau \) was varied for the ETD schemes, and for the RKM- and PC-ones, the maximal values allowing for a numerically stable simulations were used; relative error \( \varepsilon = \frac{\sum |z_j|}{\sum |z_j|} \). For the CPU specifications see Caption to Fig. 3.

| \( N \) | \( \tau \) | preparation | run | \( \varepsilon \) |
|-------|-------|-----------|-----|-------|
| Runge–Kutta–Merson method | 200 | 2.3 \times 10^{-6} | 16.1 | 6.1 \times 10^{-11} |
|       | 300 | 10^{-6} | 53.0 | 1.4 \times 10^{-11} |
|       | 400 | 0.5 \times 10^{-6} | 139 | 4.9 \times 10^{-10} |
| predictor-corrector [16] | 200 | 1.2 \times 10^{-6} | 13.8 | 5.1 \times 10^{-8} |
|       | 300 | 5.5 \times 10^{-7} | 49.9 | 5.0 \times 10^{-6} |
|       | 400 | 3.1 \times 10^{-7} | 106 | 5.1 \times 10^{-8} |
| ETD2RK | 200 | 0.0005 | 0.752 | 1.1 \times 10^{-10} |
|       | 300 | 0.005 | 7.51 | 1.0 \times 10^{-7} |
|       | 400 | 0.0005 | 4.23 | 1.9 \times 10^{-10} |
| ETD3RK | 200 | 0.0005 | 1.25 | 17.6 | 4.3 \times 10^{-10} |
|       | 300 | 0.0005 | 12.5 | 3.23 | 1.0 \times 10^{-7} |
|       | 400 | 0.0005 | 159 | 3.23 | 1.0 \times 10^{-7} |
| ETD4RK | 200 | 0.0005 | 1.25 | 2.59 | 5.6 \times 10^{-11} |
|       | 300 | 0.005 | 12.5 | 3.97 | 4.2 \times 10^{-10} |
|       | 400 | 0.0005 | 7.05 | 6.02 | 1.5 \times 10^{-10} |

The employment of this approach for Cahn–Hilliard equation \([13]\) and sixth-order spatial derivative Matthews–Cox equation \([17]\) demonstrated that one can have a performance gain by two orders of magnitude for the former and by three orders for the latter. Moreover, for partial differential equations, the performance gain grows as one increases the required accuracy of spatial resolution; the basic gain for an equation with the highest-order spatial derivative \( \partial^m \partial x^m \) is \( h \) \( h \). For nonlarge \( m \), one can introduce code optimization accounting for the sparseness of the matrices \( Q \) and \( M_0 \), which gives some further performance acceleration on top of the basic one (this acceleration is by far not that large as the basic gain). For the simulation of a long-term evolution, as required for complex spatiotemporal dynamics or Anderson localization phenomena, the maximal performance gain is achieved with the 4th-order Runge–Kutta type ETD scheme; for a short-time simulation, the usage of the 2nd-order ETD scheme is somewhat more efficient.

The employment of the approach for spectral methods is illustrated with 1D Fokker–Planck equation governing the macroscopic dynamics of populations of quadratic integrate-and-fire neurons with endogenic noise. Here, one can have a controllable accuracy and achieve the performance gain up to one order of magnitude for time-independent regimes, where short-term simulations are sufficient, and several orders of magnitude for time-dependent regimes where long-term simulations are demanded. Our approach is also efficient for numerical simulation of other classical systems such as populations of coupled active rotators \([53, 54]\).

The suggested approach can be extended into multiple dimensions in a straightforward way \([14]\). However, the basic version of the code for a partial differential equation, say, in 2d requires matrices \( Q \) and \( M_0 \) of size \([N \times N] \) with
In this section we evaluate the “strong” accuracy of the approximate calculation of matrices $Q$ and $M_n$ with the suggested algorithm. Below we will also discuss why one distinguishes the “strong” accuracy of calculation of these coefficients and the accuracy of the numerical solution $u(t)$, which is the “soft” accuracy of the numerical simulation of dynamical system (1) and often higher; significantly higher in the examples we considered in this paper.
Matrices $Q = e^{\tau L}$ and $M_n$ given by Eqs. (2) are determined by their series:

$$Q = e^{\tau L} = I + \tau L + \frac{\tau^2}{2!} L^2 + \frac{\tau^3}{3!} L^3 + \ldots,$$

(A.1)

$$M_n = \frac{\tau^n}{n!} I + \frac{\tau^{n+1}}{m(n+1)} L^2 + \frac{\tau^{n+2}}{m(n+1)(n+2)} L^3 + \cdots + \frac{\tau^{n+m}}{m(n+1)\ldots(n+m)} L^m + \ldots.$$

(A.2)

Notice, the set of formulas (2) is just a short way to write series (A.1) and (A.2) and they are well defined for the case where matrix $L$ possesses zero eigenvalues (which are inherent to conservation laws in physical systems and some bifurcation points) and $L^{-1}$ diverges.

A mathematically idealistic way to compute these matrices is related to their diagonalization or decomposition into the basis of eigenvectors $y_j (j = 1, 2, \ldots, N)$ of $L$, defined by $L \cdot y_j = \lambda_j y_j$, with the corresponding Hermitian adjoint problems $v_j^* \cdot L = \lambda_j y_j$, where the superscript “$*$” indicates a transposed vector/matrix and the orthonormalization condition is fulfilled: $v_j^* \cdot y_m = \delta_{jm}$. With a solved eigenvalue problem, one can substitute $L = \sum_{j=1}^N \lambda_j y_j v_j^T$ into series (A.1) and (A.2) and calculate

$$Q = \sum_{j=1}^N e^{\lambda_j \tau} y_j v_j^T,$$

(A.3)

$$M_n = \sum_{j=1}^N \frac{(n-1)!}{(\lambda_j)^n} \left( e^{\lambda_j \tau} - \sum_{m=0}^{n-1} \frac{(\lambda_j \tau)^m}{m!} \right) y_j v_j^T.$$

(A.4)

The $j$-th summand in $M_n$ is again correctly defined also for $\lambda_j \rightarrow 0$: one takes the limit $\lambda_j \rightarrow 0$ and finds $(\tau^n/n) y_j v_j^T$.

In practice, the exact solution of the eigenvalue problem is not always feasible and approaches for high-precision approximate calculations are available in the literature (e.g., employing the projection into Krylov subspaces [38, 39]).

Our task in this section is to compute the “exact” matrices $Q$ and $M_n$ with a controlled accuracy (arbitrary small error) and examine the results of their calculations with the algorithm of this paper. Therefore, we perform a calculation with Taylor series (A.1) and (A.2) with as many elements and decimal digits as needed to have the results with the standard double precision accuracy—not less than 16 decimal digits. The Taylor series of the exponentials of $(\lambda_j \tau)$ in Eqs. (A.3) and (A.4) with respect to $\tau$ are equivalent to series (A.1) and (A.2). Since the exponential does converge for all argument values, the series we have to compute always converges even though a gigantic number of elements can be required.

The most operation-number efficient way of computing $Q$ employs the formula (A.1) in a form $Q = I + \tau L \cdot (I + \frac{\tau}{2} L \cdot (I + \frac{\tau}{3} L \cdot (I + \ldots)))$. Accordingly to this form, for a Taylor series truncated after the $M$-th element, one can write $Q_M = I$ and iteratively compute

$$Q_{m-1} = I + \frac{\tau}{m} L \cdot Q_m,$$

(A.5)

descending from $m = M$ to $m = 1$. This iterative procedure yields all required matrices,

$$Q = Q_0, \quad M_n = \frac{\tau^n}{n} Q_n,$$

(A.6)

and provides the slowest accumulation rate for the machine cancellation error.

The core purpose of the ETD methods is to iterate a numerical scheme with time step size $\tau$ for which $|\lambda_{\text{fast}}| \gg 1$, where $|\lambda_{\text{fast}}|$ is the eigenvalue of the fastest decaying or oscillating mode. Hence, the practically relevant cases require the calculation of Taylor series with as many elements as needed to resolve $e^{\lambda_{\text{fast}} \tau}$ [see Eqs. (A.3) and (A.4), which are equivalent to (A.1) and (A.2) for a large $|\lambda_{\text{fast}}|$ or at least avoid its numerical explosion. To avoid the numerical instability of the fastest mode caused by the Taylor series truncation, one needs the order of magnitude of the truncated terms for this mode $|\lambda_{\text{fast}}|^{-M}/M! < 1$. With the Stirling’s approximation $M! \approx \sqrt{2\pi M}(M/e)^M$, one can estimate the minimal required truncation length $M_e$:

$$\frac{e^{|\lambda_{\text{fast}}| \tau}}{M_e^{(2\pi M_e)^{\frac{1}{M_e}}} \sim 1}.$$
For CHE (13), the short wavelength modes \( \delta u(x, t) \propto c_s(t)e^{i\kappa x} \) with the wavenumber \( \kappa \gg 1 \) and amplitude \( c_s(t) \) are the fastest decaying ones. Their decay rate is dominated by the last term in (15a) and, substituting the perturbation \( c_s(t)e^{i\kappa x} \) into (15a), one finds: \( \lambda_c c_s(t)e^{i\kappa x} = \left[ \cdots - (e^{i\kappa h/2} - e^{-i\kappa h/2})^4/\hbar^4 \right] c_s(t)e^{i\kappa x} = \left[ \cdots - (16/\hbar^2) \sin^2(kh/2) \right] c_s(t)e^{i\kappa x} \).

Hence, \( \lambda_{\text{fast}} \approx -16/\hbar^4 \) and Eq. (A.7) yields

\[
M_s = 16\tau/\hbar^4.
\]

For the example in Fig. 2 with \( \tau = 2.5 \times 10^{-4} \) and the same values of other parameters, we find \( M_s = 1740 \) and take truncation length \( M = 1.2M_s = 2088 \) to be on a safe side with the series convergence. We use the Maple analytical calculation package for computations with a variable number of decimal digits to handle the problem of cancellation error accumulation. For 256 decimal digits the first 20 digits of \( Q_{jk} \) are still affected by the variation of the mantissa length (cancellation error). We conduct the calculations with 384 digits and find the first 20 digits of the results to be the same as with 512 for \( M = 1.2M_s \) and elongated series \( M = 1.4M_s = 2436 \). Therefore, we trust the first 20 digits of the “exact” calculated matrices \( Q \) and \( M \).

In Fig. 10, one can see that the relative error of the significant parts of \( Q_{jk} \) and \( (M_{jk}) \) is somewhat smaller than 10^{-6}.

With the approach of this paper the fine dynamics of the fastest decaying modes is resolved accurately enough to avoid the numerical instability but not with a high precision. The corresponding contributions into \( Q \) and \( M \) are not very accurate. These contributions result in the “strong” accuracy of calculation of the matrices \( \lesssim 10^{-6} \) for the example in Figs. 10 and 2. However, the inaccuracy of the fastest decaying perturbations makes an exponentially small contribution into the numerically simulated system state; the accuracy of numerical solutions is practically important and is much higher as one can see in Figs. 4 and 6. The latter can be referred to as the “soft” accuracy of the numerical scheme.

Similarly, for MCE (17), the short wavelength modes \( \delta u(x, t) \propto c_s(t)e^{i\kappa x} \) are the fastest decaying ones. Their decay rate is dominated by the last term in (18a) and one finds: \( \lambda_c c_s(t)e^{i\kappa x} = \left[ \cdots + (e^{i\kappa h/2} - e^{-i\kappa h/2})^4/\hbar^4 \right] c_s(t)e^{i\kappa x} = \left[ \cdots - (64/\hbar^2) \sin^2(kh/2) \right] c_s(t)e^{i\kappa x} \).

Hence, \( \lambda_{\text{fast}} \approx -64/\hbar^4 \) and Eq. (A.7) yields

\[
M_s = 64\tau/\hbar^4.
\]

For the example in Fig. 5 with \( \tau = 3.2 \times 10^{-4} \) and the same values of other parameters, we find \( M_s = 870 \) and take truncation length \( M = 1.2M_s = 1044 \). For calculations with 128 decimal digits the first 20 digits of \( Q_{jk} \) are still affected by the variation of the mantissa length. We conduct the calculations with 192 digits and find the first 20 digits of the results to be the same as with 256 for \( M = 1.2M_s \) and \( M = 1.4M_s = 1218 \). Therefore, we trust the first 20 digits of the “exact” calculated matrices \( Q \) and \( M \). In Fig. 10, one can see that the relative error of the significant parts of \( Q_{jk} \) and \( (M_{jk}) \) is still \( 10^{-6} \).

The “soft” accuracy of the numerical scheme is higher, as one can see in Figs. 6 and 7.

For Fokker–Planck equation (25), the fastest decaying modes \( \delta z_j(t) \propto c_s(t)e^{i\kappa j} \) are localized at largest \( j \approx N \). Their decay rate is dominated by the \( \sigma^2 \)-term in (25) and one finds: \( \lambda_c c_s(t)e^{i\kappa j} = \left[ \cdots - \sigma^2 \hat{f}(\hat{f}/2 + e^{-i\kappa j/2})^4/4 \right] c_s(t)e^{i\kappa j} = \left[ \cdots - 4\hat{f}^2\sigma^2 \cos^4(\kappa/2) \right] c_s(t)e^{i\kappa j} \).

Hence, \( \lambda_{\text{fast}} \approx -4N^2\sigma^2 \) and Eq. (A.7) yields

\[
M_s = 4N^2\sigma^2\tau.
\]

For the example in Fig. 9 with \( \tau = 10^{-4} \) and the same values of other parameters, we find \( M_s = 435 \) and take truncation length \( M = 1.2M_s = 522 \). For calculations with 64 decimal digits the first 20 digits of \( Q_{jk} \) are still affected by the cancellation error. We conduct the calculations with 96 digits and find the first 20 digits of the results to be the same as with 128 for \( M = 1.2M_s \) and \( M = 1.4M_s = 609 \). Therefore, we trust the first 20 digits of the “exact” calculated matrices \( Q \) and \( M \). In Fig. 11, one can see that the relative error of the significant parts of \( Q_{jk} \) and \( (M_{jk}) \) is somewhat smaller than 10^{-5}.

The “soft” accuracy of the numerical scheme is still better, as one can see in Fig. 8 and Table 1.
editing (supporting). Denis S. Goldobin: Conceptualization (lead), Methodology (equal), Software (equal), Validation (equal), Formal analysis (equal), Visualization (lead), Writing – original draft (lead), Writing – review & editing (lead).

Declaration of competing interest

The authors declare no competing interests to disclose.

Data availability statement

The data that supports the findings of this study are available within the article in the graphic form. The data sheets for the graphs and the program codes in FORTRAN are available on request from the authors.

Acknowledgements

The work was carried out as part of a major scientific project (Agreement no. 075-15-2024-535 by April 23, 2024).

References

[1] P. W. Anderson, Absence of diffusion in certain random lattices, Phys. Rev. 109 (5) (1958) 1492–1505. doi:{10.1103/PhysRev.109.1492}
[2] J. Fröhlich, T. Spencer, A rigorous approach to Anderson localization, Physics Reports 103 (1) (1984) 9–25. doi:{10.1016/0370-1573(84)90061-9}
[3] I. M. Lifshitz, S. A. Gredeskul, L. A. Pastur, Introduction to the Theory of Disordered Systems, Wiley, New York, 1988.
[4] R. Blümel, S. Fishman, U. Smilansky, Excitation of molecular rotation by periodic microwave pulses. a testing ground for anderson localization, J. Chem. Phys. 84 (5) (1985) 2604–2614. doi:{10.1063/1.450330}
[5] T. Schwartz, G. Bartal, S. Fishman, M. Segev, Transport and Anderson localization in disordered two-dimensional photonic lattices, Nature 446 (7131) (2007) 52–55. doi:{10.1038/nature05623}
[6] A. S. Pikovsky, D. L. Shepelyansky, Destruction of Anderson localization by a weak nonlinearity, Phys. Rev. Lett. 100 (9) (2008) 094101. doi:{10.1103/PhysRevLett.100.094101}
[7] J. W. Cahn, J. E. Hilliard, Free energy of a nonuniform system. I. Interfacial free energy, J. Chem. Phys. 28 (2) (1958) 258–267. doi:{10.1063/1.1744102}
[8] A. A. Golovin, A. A. Nepomnyashchy, S. H. Davis, M. A. Zaks, Convective Cahn-Hilliard Models: From Coarsening to Roughening, Phys. Rev. Lett. 86 (2001) 1550–1553. doi:{10.1103/PhysRevLett.86.1550}
[9] A. Podolny, M. A. Zaks, B. Y. Rubinstein, A. A. Golovin, A. A. Nepomnyashchy, Dynamics of domain walls governed by the convective Cahn–Hilliard equation, Phys. D 201 (3) (2005) 291–305. doi:{10.1016/j.physd.2005.01.003}
[10] T. Speck, A. M. Menzel, J. Bialké, H. Löwen, Dynamical mean-field theory and weakly non-linear analysis for the phase separation of active Brownian particles, J. Chem. Phys. 142 (22) (2015) 224109. doi:{10.1063/1.4922324}
[11] Y. Kuramoto, T. Tsuzuki, Persistent propagation of concentration waves in dissipative media far from thermal equilibrium, Prog. Theor. Phys. 55 (2) (1976) 356–369. doi:{10.1143/PTP.55.356}
[12] E. Knobloch, Pattern selection in long-wavelength convection, Phys. D 41 (3) (1990) 450–479. doi:{10.1016/0167-2789(90)90008-D}
[13] L. Shilman, G. Sivashinsky, Hexagonal structure of large-scale Marangoni convection, Phys. Nonlin. Phenom. 52 (2–3) (1958) 477–488. doi:{10.1016/0167-2789(51)90130-8}
(1969) 160–164.
URL "http://jetp.ras.ru/cgi-bin/dn/e_028_01_0160.pdf"

[60] M. I. Rabinovich, D. I. Trubetskov, Oscillations and Waves: In Linear and Nonlinear Systems, Springer Netherlands, 1989.

[61] J. D. Crawford, Amplitude expansions for instabilities in populations of globally-coupled oscillators, J. Stat. Phys. 74 (1994) 1047–1084. doi:{10.1007/BF02188217} URL https://doi.org/10.1007/BF02188217

[62] H. Daido, Onset of cooperative entrainment in limit-cycle oscillators with uniform all-to-all interactions: bifurcation of the order function, Phys. D 91 (1) (1996) 24–66. doi:{10.1016/0167-2789(95)00260-X} URL https://doi.org/10.1016/0167-2789(95)00260-X

[63] E. Montbrió, D. Pazo, A. Roxin, Macroscopic Description for Networks of Spiking Neurons, Phys. Rev. X 5 (2015) 021028. doi:{10.1103/PhysRevX.5.021028} URL https://doi.org/10.1103/PhysRevX.5.021028