IMA Journal of Numerical Analysis (2015) **35**, 989–1010
doi:10.1093/imanum/dru025
Advance Access publication on June 17, 2014

Numerical analysis of very weakly well-posed hyperbolic Cauchy problems

F. COLOMBINI*
Dipartimento di Matematica, Università di Pisa, 56127 Pisa, Italy
*Corresponding author: colombini@dm.unipi.it

AND

J. RAUCH
Department of Mathematics, University of Michigan, Ann Arbor 48109, MI, USA
rauch@umich.edu

[Received on 3 May 2013; revised on 7 April 2014]

This paper analyses the approximate solution of very weakly well-posed hyperbolic Cauchy problems. These problems have very sensitive dependence on initial data. We treat a single family of such problems showing that, in spite of the sensitive dependence, approximate solutions with desired precision ε can be computed in finite-precision arithmetic with cost growing polynomially in $1/\varepsilon$. The sensitive dependence requires high finite precision. The analysis required a new Gevrey stability estimate for the leapfrog scheme. The latter depends on a new discrete Glaeser inequality. The cost of calculating solutions with features on a scale $\ell \ll 1$ grows as $e^{C\ell^{-1/2}}$.

Keywords: weakly hyperbolic, sensitive dependence, Glaeser inequality, Gevery classes, spectral leap frog.

1. Introduction

For the Prandtl equations describing the profiles of incompressible boundary layers, the underlying dynamic equations are very weakly well posed. Gerard-Varet's examples show that there cannot be continuous dependence for data that are only infinitely differentiable (Gérard-Varet & Dormy, 2010; Guo & Nguyen, 2011; Gérard-Varet & Nguyen, 2012). In spite of this, the Cauchy problem has recently been proved well set in the framework of Gevrey spaces (Gérard-Varet & Masmoudi, 2013). In an analogous vein, the analysis of Landau damping by Mouhot & Villani (2011) applies to initial data that are merely of Gevrey class. It is suspected that there is nonexistence for data that is merely C_0^∞. Both of these important examples are strongly nonlinear.

These problems, though well posed for Gevrey-class data, are extremely sensitive to initial data. A typical example of a mapping that is not continuous from $\mathcal{S}(\mathbb{R}^d)$ to $\mathcal{S}'(\mathbb{R}^d)$ but is continuous on Gevrey spaces is the Fourier multiplier with symbol $e^{\langle\xi\rangle^\alpha}$ with $0 < \alpha < 1$.

From the point of view of numerical computation such an operator poses a daunting challenge. At the very least, to achieve an approximation with k significant digits requires data with many more significant digits. One might think that the problem is intractable in the sense that computational costs would grow unreasonably fast with desired accuracy.

This paper shows that this pessimistic outlook is not correct. For a nontrivial linear hyperbolic Cauchy problem with variable coefficients, we prove that numerical approximation is feasible with polynomially growing costs.

Consider space-time with variables $(t,x)\in\mathbb{R}^{1+d}$, dual variables (τ,ξ) and standard differential operators

$$D_t := \frac{1}{\mathrm{i}}\frac{\partial}{\partial t}, \quad D_j := D_{x_j} := \frac{1}{\mathrm{i}}\frac{\partial}{\partial x_j}, \quad 1\leqslant j\leqslant d.$$

Suppose that $P(\tau,\xi)$ is a polynomial of degree $\mu\geqslant 1$. Denote by P_μ the principal part homogeneous of degree μ. Suppose that $\{t=0\}$ is noncharacteristic, that is, $P_\mu(1,0)\neq 0$. Consider first the constant coefficient initial value problem

$$P(D)u=0, \quad \partial_t^j u(0,x) = \begin{cases} g(x) & \text{for } j=0,\\ 0 & \text{for } 1\leqslant j\leqslant \mu-1.\end{cases}$$

Solving by Fourier transformation in x yields an explicit formula for the partial Fourier transform with respect to x:

$$\hat{u}(t,\xi) = M(t,\xi)\hat{g}(\xi). \tag{1.1}$$

The Fourier multiplier $M(t,\xi)$ is for each ξ the unique solution of the ordinary differential initial value problem in t,

$$P(D_t,\xi)M=0, \quad \frac{\partial^j M(0,\xi)}{\partial t^j} = \begin{cases} 1 & \text{for } j=0,\\ 0 & \text{for } 1\leqslant j\leqslant \mu-1.\end{cases}$$

In order for (1.1) to make sense, the multiplier M must not grow too rapidly as $\xi\to\infty$. Without any hypothesis on P,

$$|M|\lesssim e^{c\langle\xi\rangle}, \quad \langle\xi\rangle := (1+|\xi|^2)^{1/2}.$$

Then (1.1) makes sense only for $\hat{g}$ decaying exponentially. These g are analytic. The values of g are uniquely determined by their restriction to any open set in $\mathbb{R}^d$. In this sense the Cauchy problem is not solvable for a sufficiently large set of g to be useful for applications since initial values on disjoint open sets should be essentially independent in order to consider the Cauchy problem well posed.

A case that can be analysed completely is that of homogeneous constant coefficient operators P.

Definition 1.1 A homogeneous polynomial $P(\tau,\xi)$ is *hyperbolic* with timelike variable t if and only if $P(1,0)\neq 0$ and for each $\xi\in\mathbb{R}^d$ the roots τ of $P(\tau,\xi)=0$ are all real.

If a homogeneous P is not hyperbolic, then M will grow exponentially in some directions ξ and admissible initial data will have undesirable analyticity properties. In the case of P, homogeneous or not, that is hyperbolic in the sense of Gårding (1951), one has

$$\sup_{[0,T]\times\mathbb{R}^d} |M(t,\xi)| \leqslant C\langle\xi\rangle^{\mu-1}.$$

For $g\in H^s(\mathbb{R}^d)$ there is a unique solution with $u\in C(\mathbb{R}:H^{s-(\mu-1)}(\mathbb{R}^d))$. There is a loss of no more than $\mu-1$ derivatives.

Example 1.2 Consider $P:=\partial_t^2$ for which $u(t)=tu_t(0)+u(0)$ is no smoother than $u_t(0)$ showing a loss of one derivative. Next consider the lower-order perturbation $\partial_t^2+a\partial_x$. In that case the solution is

given by

$$\hat{u}(t,\xi) = M(t,\xi)\hat{u}(0,\xi), \quad \text{with } M(t,\xi) = \frac{\sin(t(ia\xi)^{1/2})}{(ia\xi)^{1/2}}.$$

Since sin is an odd function, the expression is independent of the square root taken. In this case,

$$|M(t,\xi)| \lesssim e^{c|t|\sqrt{|\xi|}}.$$

The solution makes sense only for Cauchy data whose transform decays faster than $e^{-c|t|\sqrt{|\xi|}}$. These are initial data in Gevrey classes.

More generally, if the principal part of a constant coefficient operator is hyperbolic, Rouché's theorem implies that the roots τ of $P(\tau,\xi) = 0$ with ξ real satisfy $|\mathrm{Im}\tau| \lesssim \langle\xi\rangle^{(\mu-1)/\mu}$; so

$$|M(t,\xi)| \lesssim e^{c|t||\xi|^{(\mu-1)/\mu}}, \quad \text{for } 0 \leqslant t \leqslant T.$$

In one sense this estimate is very weak. It does not justify the solution of the Cauchy problem when $g \in C_0^\infty(\mathbb{R}^d)$. For that one would need the Gårding–Petrowski condition $M \lesssim \langle\xi\rangle^N$ for some N. That stronger condition is the standard definition of hyperbolicity for nonhomogeneous P. As the example shows, it is not invariant under lower-order perturbations. Remarkably the weaker condition, requiring only hyperbolicity of the principal symbol, is not only invariant under lower-order perturbations but also is sufficient for variable coefficient problems. Standard hyperbolic problems lose only a finite number of derivatives. However, the C^∞ existence theorem does not pass to the case of variable coefficients. The problems whose principal symbol is required to be hyperbolic lose an infinite number of derivatives. Equivalently M grows faster than a polynomial and the fundamental solution is not a distribution in the sense of Laurent Schwartz.

For a precise result, consider the case of equations and solutions defined for $(t,x) \in \mathbb{R} \times \mathbb{T}^d$ that are 2π-periodic in x. This avoids uniformity hypotheses for $|x|$ large.

Definition 1.3 For $1 < s < \infty$, the space $\mathcal{G}^s$ of Gevrey s periodic functions consists of $u(\theta) = \sum \hat{u}(n)\, e^{in\theta}$ that satisfy

$$\exists c, C, \quad \forall n \in \mathbb{Z}, \quad |\hat{u}(n)| \leqslant C\, e^{-c|n|^{1/s}}.$$

The family $\mathcal{G}^s(c)$ for $c > 0$ is the subset of functions such that $\hat{u}(n)\exp(c|n|^{1/s}) \in \ell^\infty(\mathbb{Z})$. It is a Banach space with the natural norm. Then $\mathcal{G}^s := \cup_{c=1}^\infty \mathcal{G}^s(c)$.

The next result in the case of coefficients that are Gevrey in t,x was proved in a fundamental paper of Bronshtein (1980). The generalizations to coefficients that are finitely smooth in time can be found in Nishitani (1983a,b) and Ohya & Tarama (1986, 2006).

Theorem 1.4 Suppose that $P(t,x,\partial_t,\partial_x)$ is an μth-order linear scalar partial differential operator whose coefficients are smooth functions valued in Gevrey class $\mathcal{G}^s$ with $1 < s < \mu/(\mu-1)$, $P_\mu(t,x,1,0,\dots,0) = 1$ with principal symbol $P_\mu(t,x,\tau,\xi)$ that is hyperbolic for each t,x. Then, for $s < \underline{s} < \mu/(\mu-1)$ and Cauchy data in $\cup_c \mathcal{G}^{\underline{s}}(c)$, the Cauchy problem has a global solution smooth in time with values in $\mathcal{G}^{\underline{s}}$.

These Cauchy problems have a finite speed of propagation computed in the usual way from the local propagation cones (see Joly *et al.*, 2005 for coefficients that are Gevrey in t,x, and Colombini & Rauch, 2011 when the coefficients are merely smooth in time).

As candidate models in science, it is desirable that approximate solutions for the Cauchy problems of Bronshtein be computable without excessive computational cost.

The numerical solution of weakly well-posed hyperbolic partial differential equations has until very recently received little attention. The introduction of Bérenger's perfectly matched layers for the Maxwell equations changed that situation (Bérenger, 1994, 1996). Though the Maxwell equations are strongly well posed with no loss of derivatives, the split equations of Bérenger in the absorbing layers define a hyperbolic system whose solutions are one derivative less regular than the initial data when measured in the standard scale of Sobolev spaces (Abarbanel & Gottlieb, 1997). That standard methods, notably the Yee scheme, converge even though the underlying problem loses a derivative is proved in the thesis Petit-Bergez (2006). In addition the thesis introduces the notion of *gap*, the difference between the number of derivatives lost by a numerical method and the possibly smaller number lost by the Cauchy problem itself.

This paper continues the theme of looking at problems that are much more sensitive to initial data.

As a test case for the question of computability, consider the variable coefficient equations studied in Colombini *et al.* (1983):

$$u_{tt} - a(t)u_{\theta\theta} = 0, \quad 0 \leqslant a \in C^2(\mathbb{R}). \tag{1.2}$$

The analysis of these authors has two great advantages over the general result of Bronshtein. First, the analysis is relatively elementary. Second, it gives an accurate assessment of the regularity required of a for various levels of continuity and discontinuity of the evolution operator. The positive results and counterexamples are sharply paired. The analysis though elementary is subtle. From the numerical perspective it is not at all obvious that numerical methods would behave in similar ways.

Subtracting a constant from u, we may restrict attention to solutions with vanishing mean:

$$\int_0^{2\pi} u(t,\theta)\,\mathrm{d}\theta = 0. \tag{1.3}$$

We prove that *for these very weakly well-posed initial value problems it is feasible to compute solutions achieving accuracy with a given tolerance ε with a computational cost that grows at most polynomially in ε^{-1}.*

This problem is intermediate between the paradigm of the standard numerical analysis of problems that are at worst well posed with the loss of a finite number of derivatives and problems that are ill posed. For the latter, the fundamental analysis of John (1960) showed that in spite of ill-posedness, if data are given for which a solution is known to exist with a given *a priori* bound, then often solutions can be computed, with polynomially growing cost, even for strongly ill-posed problems. The existence of a solution with a bound is such a strong restriction on the data that a stable computation is possible. Our problems also benefit from this effect. John did not discuss discretization. We have the additional difficulty of proving stability bounds for a time discretization that are comparable to the very weak stability of the continuous-in-time problem.

The analysis falls into three parts. The hardest part is to show that a spectral leapfrog method has the same qualitative behaviour as the exact solution. That is, for Gevrey data the approximations are bounded in Gevrey norms. The Gevrey well-posedness of (1.2) is a consequence of the following highly nontrivial ordinary differential equation estimate of Colombini *et al.* (1983) and Jannelli (1984). In addition they show that for $a \in C^k$ with $k > 2$, weaker Gevrey regularity suffices.

THEOREM 1.5 Suppose that $0 \leqslant a \in C^2([0,1])$ and $m \geqslant 1$. Then there is a constant C depending only on $\|a\|_{C^2([0,1])}$ such that solutions of $x'' + a(t)m^2x = 0$ satisfy

$$\forall 0 \leqslant t \leqslant 1, \quad |x(t)| + |x'(t)| \leqslant C\,e^{Ctm^{1/2}}(|x(0)| + |x'(0)|). \tag{1.4}$$

If the derivative $a'(t)$ changes sign at most a finite number of times, then there is polynomial growth in m.

Estimate (1.4) implies existence and continuous dependence of solutions for $0 \leqslant t \leqslant 1$, provided that initial data belong to $\mathcal{G}^2(c)$ with $c > C$. Since the coefficient of $m^{1/2}$ in $e^{Ctm^{1/2}}$ grows with time, one only gets finite time existence for data in $\mathcal{G}^2(c)$ with c fixed. For data belonging to $\cap_{c>0}\mathcal{G}^2(c)$ there is global existence with data remaining in $\cap_{c>0}\mathcal{G}^2(c)$.

We analyse a mixed spectral leapfrog method. Expand mean-zero solutions u of (1.2) in a Fourier series:

$$u(t,\theta) = \sum_{m \neq 0} \hat{u}(t,m)\, e^{\mathrm{i}m\theta}.$$

Equation (1.2) is equivalent to the family of ordinary differential equations

$$\frac{\mathrm{d}^2\hat{u}(t,m)}{\mathrm{d}t^2} + a(t)m^2\hat{u}(t,m) = 0. \tag{1.5}$$

Our method truncates the Fourier expansion at $|m| \leqslant M$.

The difficult part of the method is the time discretization of the variable coefficient ordinary differential equations (1.5). The second time derivative is discretized by a centred second difference, yielding the leapfrog scheme. The centred difference is a natural and simple discretization of $\mathrm{d}^2/\mathrm{d}t^2$ and most importantly the leapfrog scheme has a discrete energy identity for constant a that is an analogue of the energy identity for the continuous-in-time problem. We expect that the phenomenon exposed in this paper is of much wider validity, but choose the simplest nontrivial test case.

Using a new discrete Glaeser inequality from Section 2.2, we prove a stability theorem for the leapfrog scheme analogous to Theorem 1.5. It asserts that the solutions of the leapfrog scheme grow at most as $e^{(Ct|m|^{1/2})}$. A crude estimate would yield $\exp(Ct|m|^{1/2})$. The crude estimate is sufficient for analysing only analytic initial data.

The third difficulty is that the large amplification factors oblige one to approximate initial data very accurately so that the solutions after a finite time maintain sufficient precision. This means that one is obliged to work in very high-precision arithmetic. This is a potentially costly constraint. Fortunately there is a compensating saving in computational effort that comes from the fact that the data are Gevrey class. The Fourier coefficients decay very rapidly, so that one attains accuracy using very few coefficients. The main theorem is a precise version of the following theorem. The precise statement in Section 4.2 includes a computability hypothesis for $a(t)$.

THEOREM 1.6 For the model problem (1.2) with computable $a(t)$, the spectral leapfrog method has computational cost that grows polynomially in the desired precision when implemented in carefully chosen multiprecision floating-point arithmetic.

The corresponding analysis in exact real arithmetic is easier and is presented in Section 4.1. A key first step, in Section 3.3, is to estimate how large M needs to be.

Open problems:

1. Prove analogous results for general operators with Gevrey coefficients and hyperbolic principal part. Any such result requires a stability result for the numerical method that in turn implies well-posedness of the Cauchy problem in Gevrey classes. A proof cannot be easier than a proof of well-posedness.

2. A second open problem is to find a proof of Bronshtein's theorem that is easier than the current ones. Such a proof would be likely to simplify the first problem.

2. Analysis of the leapfrog scheme for $x'' + m^2 a(t)\, x = 0$, $a \geqslant 0$

When $a(t) \geqslant \delta > 0$ the solutions of the leapfrog scheme are bounded on $0 \leqslant t \leqslant T$ uniformly in m and $k := \Delta t$. For $a \geqslant 0$ crude estimates yield growth $\lesssim \exp(ct|m|)$. This section shows that the growth is no worse than $\lesssim \exp(ct\sqrt{|m|})$, the same qualitative behaviour as the continuous-in-time problem.

2.1 *Energy laws for the leapfrog scheme*

The ordinary differential equations (1.5) are replaced by difference equations. The bound on the amplification is proved by an energy argument. We use a leapfrog scheme. For a time step $\Delta t = k$ define $a^n := a(nk)$. The approximate values $x^n \approx x(nk)$ are constructed as solutions of

$$\frac{x^{n+1} - 2x^n + x^{n-1}}{k^2} + m^2 \frac{a^{n+1}x^{n+1} + a^{n-1}x^{n-1}}{2} = 0. \tag{2.1}$$

The coefficient of x^{n+1} is

$$\frac{1}{k^2} + \frac{m^2 a^{n+1}}{2} > 0,$$

so the difference equation determines x^{n+1}.

Complex solutions are estimated by estimating their real and imaginary parts separately. It suffices to consider real solutions for which the energy identities are a bit simpler. Define the discrete ε-dependent energy of a real sequence x^n by

$$E^n_\varepsilon := \frac{1}{2}\left(\frac{x^n - x^{n-1}}{k}\right)^2 + \frac{m^2 a^n_\varepsilon (x^n)^2}{2}, \quad a^n_\varepsilon := a(nk) + \varepsilon. \tag{2.2}$$

The main result of this paper is the following stability estimate for the leapfrog scheme. The important fact is that the exponent is $\lesssim |m|^{1/2}t$. The difference scheme has the same Gevrey stability possessed by the continuous problem.

In the estimate below the $g(n)$ will be nonzero because the exact solution is an approximate solution of the leapfrog scheme with small residual. There are also contributions from round-off in the case of finite-precision arithmetic.

REMARK 2.1 The solutions of our ordinary differential equation vary on the length scale $\sim |m|^{-1}$ so that k small compared with $1/|m|$ is a reasonable constraint on the time step. For $|m| \gg 1$ it is much stronger than (2.24) imposed in the next theorem.

THEOREM 2.2 There is a constant $C_1 > 0$ such that, with $\varepsilon := |m|^{-1/2}$ and $k \leqslant 1/|m|$ satisfying (2.24), the real solutions of the inhomogeneous leapfrog scheme

$$\frac{x^{n+1} - 2x^n + x^{n-1}}{k^2} + m^2 \frac{a^{n+1}x^{n+1} + a^{n-1}x^{n-1}}{2} = g(n) \tag{2.3}$$

satisfy, for all $|m| > 0$ and $n \geqslant 0$,

$$E^n_\varepsilon \leqslant C_1\, e^{C_1|m|^{1/2}t_n}(E^2_\varepsilon + E^1_\varepsilon + g^2), \quad t_n := nk, \quad g := \max_n |g(n)|, \tag{2.4}$$

where the quantities E^n_ε are defined in (2.2).

A novelty is the use of a new discrete Glaeser inequality from Proposition 2.10.

REMARK 2.3 With $\varepsilon := |m|^{-1/2}$, one has $m^2 a_\varepsilon^n \geqslant m^2\varepsilon = |m|^{3/2}$, so

$$\frac{1}{2}\left(\frac{x^n - x^{n-1}}{k}\right)^2 + \frac{|m|^{3/2}(x^n)^2}{2} \leqslant E_\varepsilon^n \leqslant \frac{1}{2}\left(\frac{x^n - x^{n-1}}{k}\right)^2 + \frac{(1 + \max_{[0,1]}|a|)m^2(x^n)^2}{2}. \tag{2.5}$$

The proof occupies the rest of Section 2. It begins with the local energy law for an appropriately modified spring energy. We then derive some standard comparison estimates and the new Glaeser inequality. Theorem 2.2 is proved by combining these elements.

The ordinary differential equation *with constant a* has the energy conservation identity

$$x'[x'' + am^2 x] = \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{(x')^2}{2} + am^2\frac{x^2}{2}\right].$$

Part (i) of the next lemma gives a discrete analogue where the time derivative on the right is replaced by a difference. Adding yields a telescoping sum.

LEMMA 2.4

(i) For constant a one has

$$\begin{aligned}&\frac{x^{n+1} - x^{n-1}}{2k}\left(\frac{x^{n+1} - 2x^n + x^{n-1}}{k^2} + am^2\frac{x^{n+1} + x^{n-1}}{2}\right)\\ &\quad = \frac{1}{2k}\left[\left(\frac{x^{n+1} - x^n}{k}\right)^2 - \left(\frac{x^n - x^{n-1}}{k}\right)^2\right] + \frac{am^2}{2}\frac{(x^{n+1})^2 - (x^{n-1})^2}{2k}.\end{aligned} \tag{2.6}$$

(ii) For $\underline{a}(t)$ not necessarily constant,

$$\begin{aligned}&\frac{x^{n+1} - x^{n-1}}{2k}\left(\frac{x^{n+1} - 2x^n + x^{n-1}}{k^2} + m^2\frac{\underline{a}^{n+1}x^{n+1} + \underline{a}^{n-1}x^{n-1}}{2}\right)\\ &\quad = \frac{1}{2k}\left[\left(\frac{x^{n+1} - x^n}{k}\right)^2 - \left(\frac{x^n - x^{n-1}}{k}\right)^2\right] + \frac{m^2}{2}\frac{\underline{a}^{n+1}(x^{n+1})^2 - \underline{a}^{n-1}(x^{n-1})^2}{2k}\\ &\qquad - \frac{m^2}{2}\frac{(\underline{a}^{n+1} - \underline{a}^{n-1})}{2k}x^{n-1}x^{n+1}.\end{aligned} \tag{2.7}$$

Proof.

(i) For the first term we write

$$x^{n+1} - 2x^n + x^{n-1} = (x^{n+1} - x^n) - (x^n - x^{n-1}).$$

The second is easy.

(ii) Compute

$$(x^{n+1}-x^{n-1})(\underline{a}^{n+1}x^{n+1}+\underline{a}^{n-1}x^{n-1})$$
$$=(\underline{a}^{n+1}(x^{n+1})^2-\underline{a}^{n-1}(x^{n-1})^2)-(\underline{a}^{n+1}-\underline{a}^{n-1})x^{n-1}x^{n+1}.$$

□

LEMMA 2.5 If the sequence x^n satisfies the leapfrog equation (2.3), and $a^n_\varepsilon := a^n + \epsilon$, then one has

$$\left[\left(\frac{x^{n+1}-x^n}{k}\right)^2-\left(\frac{x^n-x^{n-1}}{k}\right)^2\right]+\frac{m^2}{2}(a^{n+1}_\varepsilon(x^{n+1})^2-a^{n-1}_\varepsilon(x^{n-1})^2)=2kF^n+\frac{x^{n+1}-x^{n-1}}{2k}g(n), \tag{2.8}$$

$$F^n := \frac{x^{n+1}-x^{n-1}}{2k}\varepsilon m^2\frac{x^{n+1}+x^{n-1}}{2}+\frac{m^2}{2}\frac{(a^{n+1}-a^{n-1})}{2k}x^{n-1}x^{n+1}. \tag{2.9}$$

Proof. Start with (2.7) and $\underline{a}(t) := a(t) + \varepsilon$. Equation (2.3) implies that the contribution of the $a(t)$ terms on the left-hand side of (2.7) is equal to $g(n)(x^{n+1} - x^{n-1})/2k$ yielding the second term on the right of (2.8). The ε terms contribute the first term on the right of (2.9). □

2.2 *Calculus of finite differences*

Consider real-valued functions

$$k\mathbb{Z} \ni nk \mapsto u^n \in \mathbb{R}$$

defined on the discrete real line $k\mathbb{Z}$.

DEFINITION 2.6 For the map $nk \mapsto u^n$ the first difference quotient is defined as

$$(\delta u)^n := \frac{u^n - u^{n-1}}{k}.$$

The second difference δ^2 is the difference of the first difference. And so on.

The second difference at n is given by

$$(\delta^2 u)^n := \frac{1}{k}\left(\frac{u^n-u^{n-1}}{k}-\frac{u^{n-1}-u^{n-2}}{k}\right)=\frac{u^n-2u^{n-1}+u^{n-2}}{k^2}.$$

EXAMPLE 2.7

(i) The difference of a constant function is equal to 0.

(ii) The difference of the function $u^n := nk$ is equal to 1. The second difference is equal to 0.

(iii) The difference of the function $(nk)^2$ is

$$\delta(nk)^2=\frac{(nk)^2-((n-1)k)^2}{k}=\frac{(nk)^2-(n^2-2n+1)k^2}{k}=2nk-k. \tag{2.10}$$

(iv) Since $\delta(nk^2) = k$, the map

$$nk \mapsto \frac{(nk)^2 + nk^2}{2} = \frac{nk(nk+k)}{2}$$

vanishes with its first difference at $n = 0$ and has second difference identically equal to 1.

(v) Given a map $nk \mapsto u^n$, the map

$$nk \mapsto u^0 + (\delta u)^0 nk$$

is linear in n and has at the origin the same value and same first difference as u^n.

2.3 *Comparison principles*

PROPOSITION 2.8 Suppose that $nk \mapsto u^n$ is a map from $k\mathbb{Z} \to \mathbb{R}$.

(i) If $u^0 \geqslant 0$ and $(\delta u)^n \geqslant 0$ for $n \geqslant 1$, then $u^n \geqslant 0$ for all $0 \leqslant n \in \mathbb{Z}$.

(ii) If $u^0 \geqslant 0$, $(\delta u)^1 \geqslant 0$ and $(\delta^2 u)^n \geqslant 0$ for $n \geqslant 2$, then $u^n \geqslant 0$ for all $0 \leqslant n \in \mathbb{Z}$.

(iii) If $(\delta^2 u)^n \geqslant 0$ for all n, $u^0 \geqslant 0$ and $(\delta u)^0 = 0$, then $u^n \geqslant 0$ for all $n \in \mathbb{Z}$.

Proof.

(i) The hypothesis $(\delta u)^n \geqslant 0$ is equivalent to the fact that u^n is nondecreasing.

(ii) Apply (i) to $w^n := (\delta u)^n$ to find that $w^n \geqslant 0$ for all $n \geqslant 1$. Then apply (i) to u^n.

(iii) Apply (ii) to conclude that $u^n \geqslant 0$ for all $n \geqslant 0$.
Define $w^n := u^{-n}$. Compute, for all n,

$$k^2(\delta^2)w^{n+1} = w^{n+1} - 2w^n + w^{n-1} = u^{-n-1} - 2u^{-n} + u^{-n+1} = k^2(\delta^2 u)^{-n+1} \geqslant 0.$$

Since $(\delta w)^1 = (\delta u)^0 = 0$ and $w^0 = u^0 \geqslant 0$, (ii) implies that $w^n \geqslant 0$ for all $n \geqslant 0$. □

2.4 *Discrete Glaeser inequality*

The statement and proof of the discrete Glaeser inequality are modelled on the well-known continuous case, presented for convenience.

PROPOSITION 2.9 (Glaeser inequality) If $0 \leqslant a \in C^2(\mathbb{R})$ and $\|a''(t)\|_{L^\infty(\mathbb{R})} \leqslant K$, then for all t,

$$|a'(t)|^2 \leqslant 2Ka(t).$$

Proof. It suffices to treat $t = 0$. Taylor's theorem with remainder implies

$$0 \leqslant a(t+s) \leqslant a(0) + a'(0)s + \frac{Ks^2}{2}.$$

The polynomial

$$q(s) := a(0) + a'(0)s + \frac{Ks^2}{2} = \frac{K}{2}\left[\left(s + \frac{a'(0)}{K}\right)^2 - \left(\frac{a'(0)^2 - 2a(0)K}{K^2}\right)\right]$$

attains its minimum value at $s^* = -a'(0)/K$. At the minimum,

$$0 \leqslant q(s^*) = a(0) - \frac{1}{2K}(a'(0))^2,$$

proving the proposition. □

The discrete analogue that follows recovers the continuous case in the limit $k \to 0$.

PROPOSITION 2.10 (Discrete Glaeser inequality) Suppose $0 < k$ and that $nk \mapsto a^n \geqslant 0$ is a non-negative sequence defined on $k\mathbb{Z}$ with $\|\delta^2 a^n\|_{\ell^\infty} \leqslant K$. Then for all n,

$$(\delta a^n + Kk/2)^2 \leqslant 2Ka^n + K^2k^2/4. \tag{2.11}$$

Proof. The function

$$nk \mapsto a^n - \left(\frac{K}{2}nk(nk+k) + \delta a^0 nk + a^0\right) := w^n$$

satisfies

$$w^0 = \delta w^0 = 0, \quad \delta^2 w^n \leqslant 0.$$

The comparison principle yields $w^n \leqslant 0$ for all n. Equivalently,

$$a^n \leqslant \frac{K}{2}\left(nk(nk+k) + \frac{2\delta a^0}{K}nk + \frac{2a^0}{K}\right) = \frac{K}{2}\left((nk)^2 + \frac{2\delta a^0 + Kk}{K}nk + \frac{2a^0}{K}\right).$$

Completing the square inside the parentheses and using $0 \leqslant a^n$ yields

$$0 \leqslant \left(nk + \frac{2\delta a^0 + Kk}{2K}\right)^2 - \left(\frac{(2\delta a^0 + Kk)^2}{4K^2} - \frac{8Ka^0}{4K^2}\right).$$

The minimum over n of the right-hand side is achieved at the n that minimizes the first term. At that n the first term is no larger than $(k/2)^2$, so

$$\left(\frac{(2\delta a^0 + Kk)^2}{4K^2} - \frac{8Ka^0}{4K^2}\right) \leqslant \frac{k^2}{4}, \quad \text{so } (2\delta a^0 + Kk)^2 - 8Ka^0 \leqslant K^2k^2.$$

Thus, $p(\delta a^0) \leqslant 0$, where $p(s) := (2s + Kk)^2 - 8Ka^0 - K^2k^2$. The polynomial p is $\leqslant 0$ only between its two real roots

$$s_\pm := \frac{-Kk \pm \sqrt{8Ka^0 + K^2k^2}}{2} = -\frac{Kk}{2} \pm \sqrt{2Ka^0 + K^2k^2/4}.$$

Therefore $s_- \leqslant \delta a^0 \leqslant s_+$. Adding $Kk/2$ yields

$$-\sqrt{2Ka^0 + K^2k^2/4} \leqslant \delta a^0 + Kk/2 \leqslant \sqrt{2Ka^0 + K^2k^2/4},$$

proving the proposition. □

2.5 *Gevrey stability of the leapfrog scheme for* $x'' + m^2 a(t)x = 0$

The stability is proved by summing the energy identity (2.8). The left-hand sum is telescopic. We need to estimate the sum of the source terms on the right. These estimates are discrete analogues of the proof in Colombini *et al.* (1983). The discrete Glaeser inequality is needed in (iii).

LEMMA 2.11

(i)
$$\left|\frac{x^{n+1}-x^{n-1}}{2k}\right| \leqslant \frac{1}{\sqrt{2}}\left(\sqrt{E_\varepsilon^{n+1}}+\sqrt{E_\varepsilon^{n}}\right). \tag{2.12}$$

(ii)
$$\left|\varepsilon m^2\,\frac{x^{n+1}+x^{n-1}}{2}\right| \leqslant \frac{|m|\sqrt{\varepsilon}}{\sqrt{2}}\left(\sqrt{E_\varepsilon^{n+1}}+\sqrt{E_\varepsilon^{n-1}}\right). \tag{2.13}$$

(iii) For all $k \leqslant \sqrt{\varepsilon}$ one has, with K as in Proposition 2.10,
$$\left|m^2\frac{a_\varepsilon^{n+1}-a_\varepsilon^{n-1}}{2k}x^{n+1}x^{n-1}\right| \leqslant \frac{4\sqrt{2K+K^2}}{\sqrt{\varepsilon}}\sqrt{E_\varepsilon^{n+1}}\sqrt{E_\varepsilon^{n-1}}. \tag{2.14}$$

Proof.

(i) Use the triangle inequality:
$$\begin{aligned}\left|\frac{x^{n+1}-x^{n-1}}{2k}\right| &= \frac{1}{2}\left|\frac{x^{n+1}-x^{n}}{k}+\frac{x^{n}-x^{n-1}}{k}\right| \\ &\leqslant \frac{1}{2}\left|\frac{x^{n+1}-x^{n}}{k}\right|+\frac{1}{2}\left|\frac{x^{n}-x^{n-1}}{k}\right| \\ &\leqslant \frac{1}{\sqrt{2}}\left(\sqrt{E_\varepsilon^{n+1}}+\sqrt{E_\varepsilon^{n}}\right).\end{aligned} \tag{2.15}$$

(ii) Estimate using the energy:
$$m^2(x^{n+1})^2 = \frac{m^2 a_\varepsilon^{n+1}(x^{n+1})^2}{a_\varepsilon^{n+1}} \leqslant \frac{2E_\varepsilon^{n+1}}{a_\varepsilon^{n+1}} \leqslant \frac{2E_\varepsilon^{n+1}}{\varepsilon}, \quad |m|\,|x^{n+1}| \leqslant \frac{\sqrt{2E_\varepsilon^{n+1}}}{\sqrt{\varepsilon}}.$$

Therefore
$$m^2\frac{x^{n+1}+x^{n-1}}{2} \leqslant \frac{|m|}{2}\,\frac{\sqrt{2E_\varepsilon^{n+1}}+\sqrt{2E_\varepsilon^{n-1}}}{\sqrt{\varepsilon}}, \tag{2.16}$$

proving (ii).

(iii) Use $m^2 a_\varepsilon^n (x^n)^2 \leqslant 2E_\varepsilon^n$ to find

$$|m^2 x^{n+1} x^{n-1}| = \frac{|m|\sqrt{a_\varepsilon^{n+1}}|x^{n+1}|\,|m|\sqrt{a_\varepsilon^{n-1}}|x^{n-1}|}{\sqrt{a_\varepsilon^{n+1}}\sqrt{a_\varepsilon^{n-1}}} \leqslant \frac{\sqrt{2E_\varepsilon^{n+1}}\sqrt{2E_\varepsilon^{n-1}}}{\sqrt{a_\varepsilon^{n+1}}\sqrt{a_\varepsilon^{n-1}}}. \tag{2.17}$$

Next show that for $k^2 \leqslant \varepsilon$,

$$\left|\frac{a_\varepsilon^{n+1} - a_\varepsilon^{n-1}}{2k}\right| \leqslant 2\sqrt{(2K + K^2)a_\varepsilon^{n+1}}. \tag{2.18}$$

The discrete Glaeser inequality implies that

$$\left|\frac{a_\varepsilon^{n+1} - a_\varepsilon^{n-1}}{2k} + \frac{K(2k)}{2}\right| \leqslant \left(2Ka_\varepsilon^{n+1} + \frac{K^2(2k)^2}{4}\right)^{1/2},$$

so

$$\left|\frac{a_\varepsilon^{n+1} - a_\varepsilon^{n-1}}{2k}\right| \leqslant (2Ka_\varepsilon^{n+1} + K^2k^2)^{1/2} + Kk \leqslant 2(2Ka_\varepsilon^{n+1} + K^2k^2)^{1/2}.$$

Using $k^2 \leqslant \varepsilon \leqslant a_\varepsilon^{n+1}$ yields (2.18).

The last and easiest ingredient is

$$\frac{1}{\sqrt{a_\varepsilon^{n-1}}} \leqslant \frac{1}{\sqrt{\varepsilon}}. \tag{2.19}$$

Combining (2.17–2.19) proves (iii). □

Proof of Theorem 2.2. Sum equation (2.8) from $n = 2$ to n. The left-hand side is telescopic. For $n > 2$, one finds

$$\begin{aligned} E_\varepsilon^{n+1} &= \left(\frac{x^2 - x^1}{k}\right)^2 + \frac{m^2 a_\varepsilon^1 (x^1)^2}{2} + \sum_{\ell=2}^{n}\left(2kF^\ell + \frac{x^{\ell+1} - x^{\ell-1}}{2k}\,g(\ell)\right) \\ &\leqslant E_\varepsilon^2 + E_\varepsilon^1 + \sum_{\ell=2}^{n}\left(2k|F^\ell| + \left|\frac{x^{\ell+1} - x^{\ell-1}}{2k}g(\ell)\right|\right). \end{aligned} \tag{2.20}$$

Use Lemma 2.11(i) to estimate the last term on the right of (2.8):

$$\left|\frac{x^{\ell+1} - x^{\ell-1}}{2k}g(\ell)\right| \leqslant \left(\sqrt{E_\varepsilon^{\ell+1}} + \sqrt{E_\varepsilon^{\ell}}\right)|g(\ell)| \leqslant \frac{1}{10}(E^{\ell+1} + E^\ell) + C|g(\ell)|^2. \tag{2.21}$$

The terms in (2.9) are bounded directly using the estimates of Lemma 2.11. With constants C independent of ε, k, m, ℓ one has

$$|F^\ell| \lesssim \left(|m|\sqrt{\varepsilon} + \frac{1}{\sqrt{\varepsilon}}\right)(E_\varepsilon^{\ell+1} + E_\varepsilon^\ell + E_\varepsilon^{\ell-1}).$$

Use the definition of g in (2.4) and optimize by taking $\varepsilon := |m|^{-1/2}$ to find, for $n > 2$,

$$E_\varepsilon^{n+1} \leqslant E_\varepsilon^2 + E_\varepsilon^1 + Cg^2 + \sum_{\ell=2}^{n} \left(\frac{1}{10} + 2Ck|m|^{1/2} \right) (E_\varepsilon^{\ell+1} + E_\varepsilon^\ell + E_\varepsilon^{\ell-1}). \tag{2.22}$$

The term E_ε^{n+1} appears on the right of (2.22) with coefficient $2Ck|m|^{1/2} + \frac{1}{10}$. Combine that term with the one on the left. The terms E_ε^ℓ with $2 \leqslant \ell \leqslant n$ appear on the right either two or three times. Taking all of them three times yields the upper bound for $\ell > 2$:

$$\left(\frac{9}{10} - 2Ck|m|^{1/2} \right) E_\varepsilon^{n+1} \leqslant C \left(E_\varepsilon^2 + E_\varepsilon^1 + g^2 + \sum_{\ell=2}^{n} k|m|^{1/2} E_\varepsilon^\ell \right). \tag{2.23}$$

Constrain k to satisfy

$$2Ck|m|^{1/2} < \frac{9}{20}. \tag{2.24}$$

With k constrained as in the theorem statement, the coefficient of E_ε^{n+1} is strictly positive. With

$$A := \frac{20C}{9}(E_\varepsilon^2 + E_\varepsilon^1 + g^2), \quad B := \frac{20C|m|^{1/2}}{9},$$

for $|m| \geqslant 2$, one finds

$$E_\varepsilon^{n+1} \leqslant A + Bk \sum_{\ell=2}^{n} E_\varepsilon^\ell.$$

By induction on n it follows that for all $n \geqslant 2$ one has $E_\varepsilon^n \leqslant Y_\varepsilon^n$, where Y_ε^n is the solution of the equation

$$Y_\varepsilon^{n+1} = A + Bk \sum_{l=2}^{n} Y_\varepsilon^l, \quad Y_\varepsilon^2 = E_\varepsilon^2.$$

The solution Y_ε^n is computed by taking the difference quotient to find

$$\frac{Y_\varepsilon^{n+1} - Y_\varepsilon^n}{k} = BY_\varepsilon^n.$$

Simplifying and an induction yields

$$Y_\varepsilon^{n+1} = (1 + kB)Y_\varepsilon^n, \quad Y^n = (1 + kB)^{n-2} Y_\varepsilon^2 = (1 + kB)^{n-2} E_\varepsilon^2.$$

Therefore, for $n \geqslant 2$,

$$E_\varepsilon^n \leqslant Y_\varepsilon^n = (1 + kB)^{n-2} E_\varepsilon^2 \leqslant e^{k(n-2)B} E_\varepsilon^2.$$

This implies (2.4), hence completing the proof. □

3. Analysis of the spectral leapfrog method

3.1 *Definition of the method*

Approximate solutions at the times t_n are defined by

$$u_{\rm app}(t_n,\theta) := \sum_{|m|\leqslant M} \hat{u}_{\rm app}(t_n,m)\, e^{{\rm i}m\theta}.$$

The coefficient $\hat{u}_{\rm app}(t_n,m)$ is an approximate solution of the ordinary differential equation

$$\left(\frac{{\rm d}^2}{{\rm d}t^2} + a(t)m^2\right)\hat{u}(t,m) = 0 \tag{3.1}$$

computed by the leapfrog scheme with time step k. This is a mixed spectral finite difference scheme.

The values $u(0,\theta)$ and $u_t(0,\theta)$ are given. Define $\hat{u}_{\rm app}(0,m)$ for $|m|\leqslant M$ to be the exact initial Fourier coefficients.

To start the leapfrog scheme one needs values for $u_{\rm app}(t_1,m)$. These are obtained by approximating $u(t_1,\theta)$ by a Taylor expansion at $t=0$ and then taking Fourier coefficients. This yields

$$\hat{u}_{\rm app}(t_1,m) := \hat{u}(0,m) + k\hat{u}_t(0,m). \tag{3.2}$$

3.2 *Residual error for the spectral leapfrog*

As is usual in the analysis of difference schemes, the solution of the differential equation is viewed as an approximate solution of the difference scheme and the residual is estimated.

The mth Fourier component of the exact solution at time step n is $\hat{u}(t_n,m)$. Next estimate the extent to which $\hat{u}(t_n,m)$ satisfies the leapfrog difference scheme. The error, denoted by $g(n,m)$, is given as

$$\frac{\hat{u}(t_{n+1},m) - 2\hat{u}(t_n,m) + \hat{u}(t_{n-1},m)}{k^2} + m^2\frac{a^{n+1}\hat{u}(t_{n+1},m) + a^{n-1}\hat{u}(t_{n-1},m)}{2} := g(n,m). \tag{3.3}$$

LEMMA 3.1 There is a constant C such that for all k,m,n satisfying $n\geqslant 1$ and $t_n = kn \leqslant 1$, the residual error (3.3) satisfies

$$|g(n,m)| \leqslant Ck^2m^4 \max_{0\leqslant t\leqslant 1} |\hat{u}(t,m)|. \tag{3.4}$$

Proof. Use (1.5) to write

$$g(n,m) = g(n,m) - (\hat{u}_{tt}(t_n,m) + a(t)m^2\hat{u}(t_n,m)).$$

The triangle inequality yields

$$\begin{aligned} |g(n,m)| \leqslant & \left|\frac{\hat{u}(t_{n+1},m) - 2\hat{u}(t_n,m) + \hat{u}(t_{n-1},m)}{k^2} - \hat{u}_{tt}(t_n,m)\right| \\ & + m^2\left|\frac{a^{n+1}\hat{u}(t_{n+1},m) + a^{n-1}\hat{u}(t_{n-1},m)}{2} - a(t)\hat{u}(t_n,m)\right|. \end{aligned} \tag{3.5}$$

The standard estimate for the second difference quotient is

$$\left|\frac{\hat{u}(t_{n+1},m)-2\hat{u}(t_n,m)+\hat{u}(t_{n-1},m)}{k^2}-\hat{u}_{tt}(t_n,m)\right|\leqslant Ck^2\max|\partial_t^4\hat{u}(t,m)|. \tag{3.6}$$

Estimate the second term in (3.5) by the standard estimate for the centred average:

$$\left|\frac{a^{n+1}\hat{u}(t_{n+1},m)+a^{n-1}\hat{u}(t_{n-1},m)}{2}-a(t_n)\hat{u}(t_n,m)\right|\leqslant Ck^2\max|\partial_t^2(a(t)\hat{u}(t,m))|. \tag{3.7}$$

Combining yields

$$|g(n,m)|\leqslant C(k^2\max|\partial_t^4\hat{u}(t,m)|+m^2k^2\max|\partial_t^2(a(t)\hat{u}(t,m))|).$$

Use the differential equation to write $\partial_t^4\hat{u}(t,m)=\partial_t^2(m^2a(t)\hat{u})$. Use this to find

$$|g(n,m)|\leqslant Cm^2k^2\max|\partial_t^2(a(t)\hat{u}(t,m))|.$$

Use the fact that a and its derivatives up to order 2 are bounded in order to estimate

$$\max(|\partial_t^2(a(t)\hat{u}(t,m))|+|a(t)\hat{u}(t,m)|)\leqslant C\max(|\hat{u}(t,m)|+|\hat{u}_t(t,m)|+|\hat{u}_{tt}(t,m)|).$$

By interpolation the middle term on the right can be omitted. Then the u_{tt} term is replaced using the differential equation to yield

$$\max(|a(t)\hat{u}_{\theta\theta}(t,m)|+|\hat{u}(t,m)|)\leqslant Cm^2\max|\hat{u}(t,m)|.$$

Therefore

$$|g(n,m)|\leqslant Ck^2m^4\max|\hat{u}(t,m)|,$$

completing the proof. □

3.3 *Approximation error in Fourier truncation*

One computes approximate values for a finite number of Fourier coefficients. A first step in studying these methods is to estimate the approximation error by Fourier truncation.

Suppose that $u\in\mathcal{G}^2(c)$ so that $|\hat{u}(m)|\leqslant A\,e^{-c\sqrt{|m|}}$, the best constant A given by the norm of u in $\mathcal{G}^2(c)$.

An approximation is given by truncation of the Fourier expansion:

$$u^M(x):=\sum_{0<|m|\leqslant M}\hat{u}(m)\,e^{imx}.$$

The error in this approximation is equal to

$$\varepsilon^M(x):=u-u^M=\sum_{|m|>M}\hat{u}(m)\,e^{imx}.$$

The values $\|\varepsilon^M\|_{\mathcal{G}^2(c)}$ need not converge to zero as $M\to\infty$. The error does converge to zero in all $\mathcal{G}^2(\underline{c})$ with $\underline{c}<c$.

LEMMA 3.2 Given $0 < \underline{c} < c$, the norm of the map $u \mapsto u^M$ from $\mathcal{G}^2(\underline{c})$ to $\mathcal{G}^2(c)$ satisfies

$$\|u - u^M\|_{\mathcal{G}^2(\underline{c})} \leqslant e^{(\underline{c}-c)M^{1/2}} \|u\|_{\mathcal{G}^2(c)}. \tag{3.8}$$

Proof. The definition of the norm in $\mathcal{G}^2(\underline{c})$ yields

$$\|u - u^M\|_{\mathcal{G}^2(\underline{c})} = \sup_{|m|>M} e^{\underline{c}|m|^{1/2}} |\hat{u}(m)|.$$

Write

$$e^{\underline{c}|m|^{1/2}} \hat{u}(m) = \frac{e^{\underline{c}|m|^{1/2}}}{e^{c|m|^{1/2}}} e^{c|m|^{1/2}} \hat{u}(m) = e^{(\underline{c}-c)|m|^{1/2}} e^{c|m|^{1/2}} \hat{u}(m).$$

For $|m| > M$, estimate $e^{(\underline{c}-c)|m|^{1/2}} \leqslant e^{(\underline{c}-c)M^{1/2}}$. Taking the supremum over $|m| > M$ yields (3.8). □

EXAMPLE 3.3 If u is a solution of (1.2), then

$$\sup_{0\leqslant t\leqslant 1} \left\| u(t) - \sum_{0<|m|\leqslant M} \hat{u}(t,m)\, e^{\mathrm{i}m\theta} \right\|_{\mathcal{G}^2(\underline{c})} \leqslant e^{(\underline{c}-c)\sqrt{M}} \sup_{0\leqslant t\leqslant 1} \|u(t)\|_{\mathcal{G}^2(c)}. \tag{3.9}$$

3.4 *Error estimate for the spectral leapfrog method*

LEMMA 3.4 Denote by C_1 the constant in Theorem 2.2 and suppose that k satisfies the constraints of that theorem. Then there is a constant C_2 independent of m, n, k such that, so long as $n \geqslant 1$ satisfies $t_n = kn \leqslant 1$, the error in the approximate solution satisfies, for all M and all $|m| \leqslant M$,

$$\begin{aligned} &\frac{1}{2}\left|\frac{(\hat{u}-\hat{u}_{\mathrm{app}})(t_n,m) - (\hat{u}-\hat{u}_{\mathrm{app}})(t_{n-1},m)}{k}\right|^2 + \frac{|m|^{3/2}|(\hat{u}-\hat{u}_{\mathrm{app}})(t_n,m)|^2}{2} \\ &\quad\leqslant C_2\, e^{C_1|m|^{1/2}} k^4 m^8 \max_{[0,1]} |\hat{u}(t,m)|^2. \end{aligned} \tag{3.10}$$

Proof. Considering the real and imaginary parts separately, it is sufficient to consider the case of real-valued $\hat{u}$. Using Theorem 2.2 with $\varepsilon = |m|^{-1/2}$ together with the residual estimate in Lemma 3.1 yields

$$\begin{aligned} E^n_\varepsilon(\hat{u}_{\mathrm{app}}(t_n) - \hat{u}(t_n)) \leqslant C_1\, e^{C_1|m|^{1/2}} &(E^2_\varepsilon(\hat{u}_{\mathrm{app}} - \hat{u}) \\ &+ E^1_\varepsilon(\hat{u}_{\mathrm{app}} - \hat{u}) + k^4 m^8 \max_{[0,1]} |\hat{u}(t,m)|^2). \end{aligned} \tag{3.11}$$

Next estimate the initial errors $E^2_\varepsilon(\hat{u}_{\mathrm{app}} - \hat{u})$ and $E^1_\varepsilon(\hat{u}_{\mathrm{app}} - \hat{u})$.

Formula (3.2) together with Taylor's theorem and $t_1 = k$ yields

$$u(t_1,\theta) - u_{\mathrm{app}}(t_1,\theta) = \int_0^k \frac{(t_1-s)^2}{2} u_{tt}(s,\theta)\, \mathrm{d}s = \int_0^k \frac{(t_1-s)^2}{2} a(s) u_{\theta\theta}(s,\theta)\, \mathrm{d}s.$$

Therefore, with C_2 denoting a constant independent of M, n, m, that can change from line to line,

$$|\hat{u}(t_1,m) - \hat{u}_{\mathrm{app}}(t_1,m)| \leqslant C_2 m^2 k^2 \max_{[0,k]} |\hat{u}(t,m)|. \tag{3.12}$$

At $t = 0$ the exact and approximate solutions have equal Fourier coefficients for $|m| \leqslant M$. Subtract (3.3) with $n = 1$ from the difference equation defining $\hat{u}_{\text{app}}(t_2, m)$ to find

$$\hat{u}(t_2, m) - \hat{u}_{\text{app}}(t_2, m) = 2(\hat{u}(t_1, m) - \hat{u}_{\text{app}}(t_1, m)) + k^2 g(1, m). \tag{3.13}$$

Estimate $g(1, m)$ as in Lemma 3.1. Instead of (3.6) use

$$\left| \frac{\hat{u}(t_2, m) - 2\hat{u}(t_1, m) + \hat{u}(t_0, m)}{k^2} \right| \leqslant C_2 \max_{[0,2k]} |\partial_t^2 \hat{u}(t, m)|$$
$$= C_2 \max_{[0,2k]} |a(t)\hat{u}_{\theta\theta}(t, m)| \leqslant C_2 m^2 \max_{[0,2k]} |\hat{u}(t, m)|.$$

This yields

$$|g(1, m)| \leqslant C_2 m^2 \max_{[0,2k]} |\hat{u}(t, m)|. \tag{3.14}$$

Combining (3.12–3.14), and using the upper bound on the right of (2.5) shows that

$$E_\varepsilon^2(\hat{u}_{\text{app}} - \hat{u}) + E_\varepsilon^1(\hat{u}_{\text{app}} - \hat{u}) \leqslant C_2 k^4 m^8 \max_{[0,1]} |\hat{u}(t, m)|^2.$$

This yields the inequality

$$E_\varepsilon^n(\hat{u}_{\text{app}}(t_n, m) - \hat{u}(t_n, m)) \leqslant C_2\, e^{C_1 |m|^{1/2}} k^4 m^8 \max_{[0,1]} |\hat{u}(t, m)|^2. \tag{3.15}$$

The left-hand inequality of (2.5) completes the proof with the final step increasing the value of C_2 by 1. □

4. Computational cost of the spectral leapfrog method

4.1 *Approximate solution with prescribed error tolerance*

PROPOSITION 4.1 Suppose that C_1, C_2 are as in Lemma 3.4 and c, c' satisfy $c' > c + C_1/2$. Suppose in addition that mean-zero Cauchy data in $\mathcal{G}^s(c')$ and an error tolerance $0 < \eta$ are given. The approximate solution $v(t_n, x) := \sum_{|m| \leqslant M} \hat{u}_{\text{app}}(t_n, m)\, e^{\mathrm{i}m\theta}$ is constructed with $0 < M$ satisfying

$$e^{(c-c')M^{1/2}} = \frac{\eta}{2}, \quad \text{equivalently, } M^{1/2} = \frac{|\ln \eta/2|}{c' - c}, \tag{4.1}$$

and $\hat{u}_{\text{app}}(t_n, m)$ equal to the solutions of the leapfrog scheme with k chosen such that

$$k^2 \max_{1 \leqslant |m| < M} (2C_2)^{1/2}\, e^{(C_1/2 + c - c')|m|^{1/2}} |m|^{13/4} = \frac{\eta}{2}. \tag{4.2}$$

The value $\hat{u}_{\text{app}}(0, m)$ is exact and $\hat{u}_{\text{app}}(t_1, m)$ is given by (3.2). Then the error satisfies

$$\sup_{0 \leqslant t_n := kn \leqslant 1} \|u(t_n) - v(t_n)\|_{\mathcal{G}^2(c)} \leqslant \eta \sup_{[0,1]} \|u(t)\|_{\mathcal{G}^2(c')}. \tag{4.3}$$

Proof. **Step (i)**. Equation (4.1) together with estimate (3.9) implies that the error committed in Fourier truncation satisfies

$$\sup_{[0,1]} \left\| u(t) - \sum_{0<|m|\leqslant M} \hat{u}(t,m)\, e^{im\theta} \right\|_{\mathcal{G}^2(c)} \leqslant \frac{\eta}{2} \sup_{[0,1]} \|u(t)\|_{\mathcal{G}^2(c')}. \tag{4.4}$$

Step (ii). The computed values are complex solutions of the leapfrog approximation to solutions of the spring equation. Their real and imaginary parts are real solutions, and so satisfy (3.15).

Using only the potential energy term and adding the estimates for real and imaginary parts yields, so long as $t_n \leqslant 1$,

$$|m|^{3/2} \frac{|\hat{u}_{\text{app}}(t_n,m) - \hat{u}(t_n,m)|^2}{2} \leqslant C_2\, e^{C_1|m|^{1/2}} k^4 m^8 \max_{[0,1]} |\hat{u}(t,m)|^2.$$

Taking the square root yields

$$|\hat{u}_{\text{app}}(t_n,m) - \hat{u}(t_n,m)| \leqslant (2C_2)^{1/2}\, e^{(C_1/2)|m|^{1/2}} k^2 |m|^{13/4} \max_{[0,1]} |\hat{u}(t,m)|.$$

Thus,

$$\begin{aligned} e^{c|m|^{1/2}} |\hat{u}_{\text{app}}(t_n,m) - \hat{u}(t_n,m)| &\leqslant k^2 (2C_2)^{1/2}\, e^{(C_1/2+c)|m|^{1/2}} |m|^{13/4} \max_{[0,1]} |\hat{u}(t,m)| \\ &= k^2 (2C_2)^{1/2}\, e^{(C_1/2+c-c')|m|^{1/2}} |m|^{13/4} \max_{[0,1]} |e^{c'|m|^{1/2}} \hat{u}(t,m)|. \end{aligned}$$

The choice (4.2) guarantees that for $|m| \leqslant M$ one has

$$k^2 (2C_2)^{1/2}\, e^{(C_1/2+c-c')|m|^{1/2}} |m|^{13/4} \leqslant \frac{\eta}{2}.$$

Therefore

$$\left\| \sum_{0<|m|\leqslant M} \hat{u}(t_n,m)\, e^{im\theta} - \hat{u}_{\text{app}}(t_n,m)\, e^{im\theta} \right\|_{\mathcal{G}^2(c)} \leqslant \frac{\eta}{2} \max_{[0,1]} \|u(t)\|_{\mathcal{G}^2(c')}. \tag{4.5}$$

Combining (4.4) and (4.5) completes the proof. □

4.2 *Computational cost with high-precision floating-point arithmetic*

One needs a hypothesis asserting that the computation of $a(t)$ in floating point is not too costly.

HYPOTHESIS 4.2 For all c_1, c_2, there exist c_3, q, so that for all N, if

$$\{t_n := nk \,:\, k := c_1(2^{-N}) \text{ and } n \leqslant 1/k\}$$

is the discrete interval $[0,1]$ with spacing $\Delta t = k$, then the values $a(t_n)$ can be calculated with error at most $2^{-N(1+c_2)}$ in a number of arithmetic operations no larger than $c_3 |\ln k|^q / k$.

REMARK 4.3 (i) The quantity $k = \Delta t$ is on the one hand the time increment and on the other hand proportional to the number of significant digits. (ii) There are $1/k$ values computed, each one with a cost bounded polynomially in the number of significant digits required.

EXAMPLE 4.4 For $0 < \underline{t} < 1$ and $2 < \gamma \in \mathbb{R}$ define

$$a(t) := \begin{cases} 0 & \text{for } t \leqslant \underline{t}, \\ (t - \underline{t})^{\gamma} & \text{for } t \geqslant \underline{t}. \end{cases}$$

If $\underline{t}$ is rational, then this function satisfies the hypothesis. However, if $\underline{t}$ is an irrational number whose binary expansion is difficult to compute, then $a(t)$ does not satisfy the hypothesis. For example, suppose that $\underline{t}$ has binary expansion $0.\alpha_1\alpha_2\ldots$ and the digit $\alpha_j \in \{0, 1\}$ is computed by solving a problem that requires $j!$ arithmetic operations. If $k = 2^{-L}$, then to compute a at the points t_n requires a determination of the first L binary digits and therefore at least $L!$ operations. Hypothesis 4.2 authorizes at most $CL^q 2^L$ operations.

EXAMPLE 4.5 The following classes of functions satisfy the hypothesis: (1) polynomials with rational coefficients; (2) polynomials with easily computed coefficients; (3) quotients of the above with poles off the real axis; (4) $a(t) = e^{p(t)}$ with p one of the above; (5) functions defined by piecing together a finite number of good functions with rational transition points or transition points with easily computed binary expansion.

The next example shows that the standard examples of functions $a(t)$ with Cauchy problems that lose an infinite number of derivatives satisfy the hypothesis.

EXAMPLE 4.6 Suppose that $\underline{t} \in]0, 1[$ has easily computed binary expansion and $4 < \gamma \in \mathbb{Q}$. Then

$$a(t) := (t - \underline{t})^{\gamma} [\sin(2\pi/(t - \underline{t}))]^2$$

satisfies the hypothesis. To see this, first compute the decimal expansion $0.\alpha_1\alpha_2 \cdots \alpha_j$ up to $j = L$ with $2^{-L} < k$. For each n one deduces the $N(n)$ such that $N \leqslant 1/(t_n - \underline{t}) < N + 1$. Then

$$\sin\left(\frac{1}{t_n - \underline{t}}\right) = \sin\left(\frac{1}{t_n - \underline{t}} - 2\pi N\right), \quad 0 < \frac{1}{t_n - \underline{t}} - 2\pi N < 1.$$

The sin on the right can then be computed to the desired precision from a truncated Taylor series. Similarly, one treats the smooth example with $(t - \underline{t})^{\gamma}$ replaced by $\exp[-1/(t - \underline{t})]$.

THEOREM 4.7 If $a(t)$ satisfies Hypothesis 4.2 and C_1 is the constant from Theorem 2.2, then there exist constants q, C_3 and $C_4 > C_1$ so that to obtain an approximation with error satisfying (4.4) with finite-precision arithmetic it is sufficient to use multiprecision arithmetic with the number of binary digits $N \geqslant C_3 |\ln \eta|$ and to choose k and M as in Proposition 4.1 with C_1 replaced by C_4. Then (4.4) holds and the number of floating-point operations is bounded by

$$C_3 |\ln \eta|^q \eta^{-1/2}. \tag{4.6}$$

Proof. One effect of finite-precision arithmetic is to add an error due to round-off to each of the residuals $g(n, m)$. A second is to add error to the initial values of E^2_ε and E^1_ε.

First estimate the round-off error in computing one step of the iteration

$$x^{n+1} = \left[1 + \frac{k^2 m^2 a(t_{n+1})}{2}\right]^{-1} \left[2x^n - \left(1 + \frac{k^2 m^2 a(t_{n-1})}{2}\right) x^{n-1}\right]. \tag{4.7}$$

Since $k|m| \leqslant 1$, formula (4.7) involves the inverse of a number of size ~ 1, and the subtraction of a number close to x^{n-1} from $2x^n$. As $x^n \approx x^{n-1}$, this formula is not subject to unusual round-off errors.

Consider the factors $k^2 m^2 a$. Since the time step k and m are related by $k^2 m^2 \leqslant 1$ and the function a is assumed bounded, the round-off error committed in computing these quantities in finite precision is $\leqslant C\rho$, where $\rho \lesssim 2^{-N(1+c_1)}$, if one computes in binary arithmetic with $N(1 + c_1)$ digits. Then the finite-precision representation of a real number r has round-off error $\sim |r|\rho$.

The two terms in the second pair of square brackets of (4.7) have size $\lesssim \max(|x^n|, |x^{n-1}|)$ and their finite-precision computations have round-off $\lesssim \rho \max(|x^n|, |x^{n-1}|)$. Combining with the analysis of the first term shows that the residual for the computation of $\hat{u}(t_{n+1}, m)$ with round-off is bounded by

$$|g(n, m)| + C\rho \max(|\hat{u}(t_n, m)|,\ |\hat{u}(t_{n-1}, m)|). \tag{4.8}$$

The residual estimate (3.4) shows that without round-off, $|g| \leqslant Ck^2 m^4 \max |\hat{u}(t, m)|$. If the second term in (4.8), coming from round-off, is no larger than a constant times the first, then the qualitative behaviour will not be affected. This is guaranteed by choosing the number of digits N such that ρ satisfies

$$\rho \leqslant k^2. \tag{4.9}$$

Then the residual is no larger than a constant times the estimate for g without round-off.

The additional round-off effect to estimate is from the steps leading to $\hat{u}(t_1, 0)$ and $\hat{u}(t_2, 0)$. In exact arithmetic the error at these times is controlled by (3.12) and the estimate that follows it by $\lesssim m^2 k^2 \max_{[0,2k]} |\hat{u}(t, m)|$. The round-off in making this computation modifies the answer by a quantity $\lesssim \rho \max_{[0,2k]} |\hat{u}(t, m)|$. This is no larger than the estimate for the first, provided ρ is chosen satisfying (4.9).

With these choices the proof of accuracy in Proposition 4.1 shows that error with round-off is bounded by

$$C\rho + \eta \sup_{[0,1]} \|u(t)\|_{\mathcal{G}^2(c')}.$$

Accuracy is guaranteed by choosing the number of significant digits N to satisfy

$$C2^{-N} \leqslant \eta \sup_{[0,1]} \|u(t)\|_{\mathcal{G}^2(c')}, \quad N \lesssim \ln(1/\eta). \tag{4.10}$$

Next estimate the computational cost. The cost is the sum on m of the cost of approximately solving the ordinary differential equation for the mth Fourier coefficient. For $0 < |m| \leqslant M$ and $0 \leqslant t_n \leqslant 1$, we must compute $x^n = \hat{u}_{\text{app}}(t_n, m)$ using (4.7). The formula for x^{n+1} requires an evaluation of $a(t_{n+1})$ with precision $\lesssim 2^{-N(1+c_1)}$ and a small number of arithmetic operations. The cost grows no faster than polynomially in N. To compute x^n at the $1/k$ discrete times $t_n \leqslant 1$ has cost $\lesssim N^p/k$.

This price must be paid $2M$ times, once for each Fourier coefficient, for a total of CMN^p/k. The defining formulas yield

$$M = C|\ln \eta|, \quad k = C\eta^{1/2}.$$

The number of arithmetic operations is bounded by $C\,|\ln \eta\,|^{p+1}\,\eta^{-1/2}$. □

REMARK 4.8

(i) Since $k \leqslant C\sqrt{\eta}$, the constraint (4.9) requires $\rho \leqslant C\eta$. The conclusion is that *one must employ a number of digits N that grows as* $C|\ln \eta|$. This is comparable to the $\ln \eta$ digits required to achieve an accurate finite-arithmetic representation of the answer.

(ii) With $k = \Delta T \sim \sqrt{\eta}$ there are $1/\sqrt{\eta}$ time steps to reach time 1. The cost of the computation is comparable to the cost of accurately solving one ordinary differential equation. This is consistent with the fact that very few Fourier coefficients are needed. Multiprecision arithmetic is required but the number of digits grows at most logarithmically with the precision.

EXAMPLE 4.9 We estimate the number of computations necessary to compute the solutions with data of the form $e^{i\nu\theta}$, $\nu \in \mathbb{N}$. The solution is concentrated on a single Fourier mode and only one ordinary differential equation needs to be solved. The cost to achieve an approximate solution with two or three significant decimal digits grows very rapidly with ν. The subtlety is that the norms on the two sides of (4.3) are very different.

To compute to N decimal places of accuracy means

$$\frac{\|u - u_{\text{app}}\|_{L^\infty}}{\|u\|_{L^\infty}} \approx 10^{-N}.$$

One expects

$$\|u - u_{\text{app}}\|_{\mathcal{G}^2(c)} \approx C\, e^{c\sqrt{\nu}} \|u - u_{\text{app}}\|_{L^\infty}, \quad \|u\|_{\mathcal{G}^2(c')} \approx C\, e^{c'\sqrt{\nu}} \|u\|_{L^\infty}.$$

Therefore estimate (4.3) yields

$$\frac{\|u - u_{\text{app}}\|_{L^\infty}}{\|u\|_{L^\infty}} \approx C \frac{\|u - u_{\text{app}}\|_{\mathcal{G}^2(c)}}{\|u\|_{\mathcal{G}^2(c')}}\, e^{(c'-c)\sqrt{\nu}} \lesssim \eta\, e^{(c'-c)\sqrt{\nu}}.$$

To achieve N digits requires $\eta \lesssim 10^{-N}\, e^{(c-c')\sqrt{\nu}}$. The number of operations from (4.6) is then $\lesssim 10^{Nq}\, e^{C\sqrt{\nu}}$. This increases rapidly but subexponentially with ν. The growth reflects the very weak well-posedness of the problem. It is only upon considering data with structures on small spatial scales that the costs become apparent.

Acknowledgement

The authors thank two referees for many wise suggestions.

Funding

The first author is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The second author thanks the GNAMPA, the Centro De Giorgi, the Laboratorio Fibonacci of the CNRS and the National Science Foundation grant NSF DMS 0807600 for their support.

REFERENCES

ABARBANEL, S. & GOTTLIEB, D. (1997) A mathematical analysis of the PML method. *J. Comput. Phys.*, **134**, 357–363.

Bérenger, J.-P. (1994) A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, **114**, 185–200.

Bérenger, J.-P. (1996) Three-dimensional perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, **127**, 363–379.

Bronshtein, M. D. (1980) The Cauchy problem for hyperbolic operators with characteristics of variable multiplicity. *Tr. Mosk. Mat. Obs.*, **41**, 83–99. (English transl. (1982) *Trans. Moscow. Math. Soc.*, **1**, 87–103.)

Colombini, F., Jannelli, E. & Spagnolo, S. (1983) Well-posedness in the Gevrey classes of the Cauchy problem for non-strictly hyperbolic equations with coefficients depending on time. *Ann. Sc. Norm. Super. Pisa*, **10**, 291–312.

Colombini, F. & Rauch, J. (2011) Sharp finite speed for hyperbolic problems well posed in Gevrey classes. *Comm. Partial Differential Equations*, **36**, 1–9.

Gårding, L. (1951) Linear hyperbolic partial differential equations with constant coefficients. *Acta Math.*, **85**, 1–62.

Gérard-Varet, D. & Dormy, E. (2010) On the ill-posedness of the Prandtl equation. *J. Amer. Math. Soc.*, **23**, 591–609.

Gérard-Varet, D. & Masmoudi, N. (2013) Well-posedness for the Prandtl system without analyticity or monotonicity. *Preprint*, arXiv:1305.0221.

Gérard-Varet, D. & Nguyen, T. (2012) Remarks on the ill-posedness of the Prandtl equation. *Asymptot. Anal.*, **77**, 71–88.

Guo, Y. & Nguyen, T. (2011) A note on Prandtl boundary layers. *Comm. Pure Appl. Math.*, **64**, 1416–1438.

Jannelli, E. (1984) Gevrey well-posedness for a class of weakly hyperbolic equations. *J. Math. Kyoto Univ.*, **24**, 763–778.

John, F. (1960) Continuous dependence on data for solutions of partial differential equations with a prescribed bound. *Comm. Pure Appl. Math.*, **13**, 551–585.

Joly, J. L., Métivier, G. & Rauch, J. (2005) Hyperbolic domains of determinacy and Hamilton–Jacobi equations. *J. Hyperbolic Differ. Equ.*, **2**, 713–744.

Mouhot, C. & Villani, C. (2011) On Landau damping. *Acta Math.*, **207**, 29–201.

Nishitani, T. (1983a) Energy inequality for non-strictly hyperbolic operators in Gevrey class. *J. Math. Kyoto Univ.*, **23**, 739–773.

Nishitani, T. (1983b) Sur les équations hyperboliques à coefficients hölderiens en t et de classe Gevrey en x. *Bull. Sci. Math.*, **107**, 113–138.

Ohya, Y. & Tarama, S. (1986) Le problème de Cauchy à caractéristiques multiples dans la classe de Gevrey. I. Coefficients hölderiens en t. *Hyperbolic Equations and Related Topics* (Katata/Kyoto, 1984). Boston, MA: Academic Press, pp. 273–306.

Ohya, Y. & Tarama, S. (2006) A note on: the Cauchy problem with multiple characteristics in the Gevrey class. I. Hölder coefficients in t. *Sci. Math. Jpn*, **63**, 287–304.

Petit-Bergez, S. (2006) Problèmes faiblement bien posés: discrétisation et applications. France: Thèse de l'Université Paris 13.