Asymptotic analysis of finite difference methods

Michael Junk * Zhaoxia Yang

FB Mathematik, Universität Kaiserslautern, Erwin-Schrödingerstraße, 67663 Kaiserslautern, Germany

Abstract

With this article, we want to advocate the use of asymptotic methods for the analysis of finite difference schemes. We present several examples to demonstrate the applicability of the approach. Advantages over the modified equation and truncation error analysis are pointed out.

Key words: finite difference methods, initial value problems, boundary value problems, modified equations, asymptotic error analysis, multiscale expansions, matched asymptotic expansions
MSC2000: 65M06, 39A11

1 Introduction

If we consider a finite difference method simply as a set of equations containing a small parameter h (the grid spacing), it is evident that the tools of asymptotic analysis can give us useful information about the method. This is the essence of the approach presented in this article.

Of course, the idea to use asymptotic methods for the analysis of finite difference schemes is not new (we give some review below) but it certainly has not gained the same status as the truncation error analysis (for consistency), the von Neumann analysis (for stability), or the modified equation approach (for consistency and, in some cases, stability). With this article, we try to

* corresponding author
Email addresses: junk@mathematik.uni-kl.de (Michael Junk), yang@mathematik.uni-kl.de (Zhaoxia Yang).
URL: http://www.mathematik.uni-kl.de/~junk (Michael Junk).

raise the status of the asymptotic analysis approach by showing its usefulness in predicting the behavior of finite difference schemes. To this end, we consider several well known schemes for simple ordinary and partial differential equations whose behavior should be known to the reader so that the relevance of the obtained results can easily be assessed. However, this choice of examples should not indicate that the method is only applicable in simple situations. In fact, our original motivation is the analysis of the lattice Boltzmann method LBM which is a class of finite difference schemes for incompressible Navier-Stokes and related equations (see [1] for a review). With the proposed method, we are now able to analyze LBM in the *same* way as any other finite difference scheme, for example the standard five point discretization for 2D Dirichlet-Poisson problems or the upwind discretization of the advection equation. Results of the analysis of boundary conditions for LBM will be published in a subsequent paper.

Following the tradition of classical asymptotic analysis, we will stay on a formal level in the sense that we do not prove the existence of asymptotic expansions. Such proofs require detailed stability information about the schemes and about the considered equations (see, for example, [2] for such an approach). If stability estimates are available, the analysis can be made rigorous, if they are not available, the analysis still gives valuable information (see, for example, section 2.4).

Let us now briefly review the application of asymptotic methods in the context of finite difference schemes (see also the review [3]). Already in 1910, Richardson had the idea to exploit an asymptotic expansion of finite difference solutions for the acceleration of convergence [4,5]. By evaluating the same method on two different grids, the leading error term could be removed by a suitable combination of the results. Similarly, the idea of deferred or iterated deferred correction [6,7] requires the existence of an asymptotic expansion of the error. These ideas have been taken up especially in numerical methods for ordinary differential equations (the existence of asymptotic expansions is discussed in every textbook on this subject – see, for example, [8,9]) but they are also applied to partial differential equations where the incompatibility between grid and general domain boundaries complicates the issue [2,10]. Apart from applications which aim at the construction of new methods, asymptotic expansions can also be useful to show convergence of finite difference schemes [11, 12]. However, all the examples mentioned above involve *regular* asymptotic expansions. In contrast to this, we propose the use of additional asymptotic methods to understand possible non-uniform or long time behavior of a scheme. This fruitful idea has been adopted from the series of papers [13–18], where multiscale expansions are used to correctly represent the long time behavior of the modified equation of difference schemes for ordinary differential equations. Another reference in this context is [19] where the authors investigate a general class of difference equations with a multiscale expansion and apply their

result to a multistep discretization of a stiff ordinary differential equation.

While [13–18] gives a complete picture for ordinary differential equations, a straight forward application to the case of boundary value problems for partial differential equations is not possible because the concept of modified equations leads to contradictions. These complications are avoided by direct expansion of the scheme without prior construction of a modified equation which is the approach presented in this article. We remark that the difficulties with the modified equation approach are related to the derivation of the modified equation which requires the existence of a function u which smoothly interpolates the discrete finite difference result. This function is then inserted in the difference equation, a Taylor expansion is performed and, possibly after some back substitutions, a differential equation for u is obtained (see [13, 20]). In many cases (typically one step schemes for pure initial value problems) this approach works quite well and the modified equation describes the scheme very accurately (see [21] for a list of references). However, when the finite difference scheme has an oscillatory error on the grid level, the prediction of the modified equation is poor (see [21,22]). This is not surprising because the basic assumption of a smooth interpolating function is violated. Unfortunately, such oscillations on the grid level inevitably occur in finite difference schemes for general multidimensional boundary value problems – a case which is of primary interest to us. With higher order interpolation at boundary nodes, the oscillatory behavior can be moved to higher orders in the error but it does not disappear. This situation has been carefully analyzed in connection with deferred correction methods (see, for example, [10]). While our method is essentially equivalent to the modified equation approach in smooth situations (see section 2.1), oscillatory behavior of the error does not create a problem. In fact, the method even predicts oscillations as demonstrated in sections 2.3 and 2.6.

Compared to the truncation error analysis, the asymptotic analysis is not disturbed by a lower order consistency at boundary nodes: it still predicts the overall consistency order correctly (see section 2.5 and 2.6).

To illustrate the asymptotic analysis of finite difference schemes, we consider several examples which have been selected according to the following criteria: 1) Both the differential equation and the difference scheme should be text book examples because our aim is not to present new models or schemes. By considering well known examples, it is easy to assess the predictions of the asymptotic analysis. 2) Each example should present a separate effect that can appear in finite difference discretizations. Specifically, we cover the following topics:

2.1) The basic approach
2.2) Investigation of long time behavior

2.3) Incompatible initial conditions
2.4) Formal stability analysis
2.5) Boundary value problems with grid-boundary incompatibility
2.6) Jumps and inner layers
2.7) Oscillations arising from non-aligned boundaries

2 Examples

The asymptotic analysis will be explained for finite difference methods on regular quadratic grids. In this case, the grid points are of the form

$$\boldsymbol{x}_{\boldsymbol{i}}(G) = h\boldsymbol{i} + \boldsymbol{\alpha}, \qquad \boldsymbol{i} \in \mathbb{Z}^d,\, G = (h, \boldsymbol{\alpha}) \in \mathbb{R}^+ \times \mathbb{R}^d,$$

where $d \in \mathbb{N}$ is the space dimension and $\boldsymbol{\alpha}$ is an offset vector. Since the regular grid is completely determined by h and $\boldsymbol{\alpha}$, we will frequently refer to $G = (h, \boldsymbol{\alpha})$ simply as *the grid*. The indices of the points which are contained in some open set $\Omega \subset \mathbb{R}^d$ are collected in the index set

$$I(G, \Omega) = \{\boldsymbol{i} \in \mathbb{Z}^d : \boldsymbol{x}_{\boldsymbol{i}}(G) \in \Omega\}.$$

Finally, a grid function v is a mapping from $I(G, \Omega)$ to $\mathbb{R}$ which assigns to each index $\boldsymbol{i}$ a value $v_{\boldsymbol{i}}(G, \Omega)$. In the following, we will drop the arguments $G = (h, \boldsymbol{\alpha})$, or (G, Ω) from all quantities, unless there is danger of misunderstanding or the arguments are changing, for example, when we consider grid sequences $G_n = (h_n, \boldsymbol{\alpha}_n)$ with $h_n \to 0$, $\boldsymbol{\alpha}_n \in \mathbb{R}^d$.

The basic idea of the analysis is now the following: if a grid function v is defined by finite difference relations, the corresponding method can be analyzed by carrying out an asymptotic analysis of v for $h \to 0$. In the simplest case of a regular expansion, we assume

$$v_{\boldsymbol{i}}(G) = u_0(\boldsymbol{x}_{\boldsymbol{i}}(G)) + hu_1(\boldsymbol{x}_{\boldsymbol{i}}(G)) + h^2 u_2(\boldsymbol{x}_{\boldsymbol{i}}(G)) + \cdots .$$

Here, the functions u_k are assumed to be sufficiently smooth with continuous derivatives up to the boundary. Other types of expansion will also be presented.

2.1 The Euler method for an autonomous equation

To demonstrate how the approach works, let us start with a simple example. We consider the explicit Euler discretization of the initial value problem

$$u'(x) = f(u(x)), \quad x \in (0, \infty), \qquad u(0) = \eta.$$

On the grid $G = (h, 0)$ with grid points $x_i = ih$, the scheme reads

$$\frac{v_{i+1} - v_i}{h} = f(v_i), \quad i \in I \cup \{0\}, \qquad v_0 = \eta. \tag{1}$$

We now insert $v_i = u_0(x_i) + hu_1(x_i) + \dots$ into (1). The initial condition immediately yields

$$u_0(0) = \eta, \qquad u_1(0) = 0,$$

and a Taylor expansion of the terms in the difference relation gives rise to

$$u_0'(x_i) + \frac{h}{2}u_0''(x_i) + hu_1'(x_i) = f(u_0(x_i)) + hf'(u_0(x_i))u_1(x_i) + \mathcal{O}(h^2). \tag{2}$$

Choosing a grid sequence $G_n = (h_n, 0)$ with $h_n \to 0$, we can find for any $\bar{x} > 0$ a sequence of indices $i_n \in I(G_n)$ such that $x_{i_n}(G_n) \to \bar{x}$ for $n \to \infty$. Taking the limit of (2) evaluated at x_{i_n}, we thus find

$$u_0'(\bar{x}) = f(u_0(\bar{x})), \quad \bar{x} \in (0, \infty).$$

Using this relation with $\bar{x} = x_i$ in (2) and dividing by h, we arrive at

$$u_1'(x_i) + \frac{1}{2}u_0''(x_i) = f'(u_0(x_i))u_1(x_i) + \mathcal{O}(h)$$

from which we deduce in the same way

$$u_1'(\bar{x}) = f'(u_0(\bar{x}))u_1(\bar{x}) - \frac{1}{2}u_0''(\bar{x}), \quad \bar{x} \in (0, \infty). \tag{3}$$

Similarly, equations for higher order expansion coefficients can be derived (see [8] for the resulting expansion in the general case of first order systems of ordinary differential equations). Despite the zero initial value for u_1, the coefficient will in general not vanish because of the source term $-u_0''/2$. Altogether we have $v_i = u_0(x_i) + hu_1(x_i) + \mathcal{O}(h^2) = u_0(x_i) + \mathcal{O}(h)$, from which we see that the Euler method is first order accurate because u_0 is the exact solution of the problem.

To illustrate the accuracy of the predicted coefficient u_1, let us consider the case $f(u) = \cos^2 u$

$$u'(x) = \cos^2 u(x), \qquad u(0) = \frac{\pi}{4}.$$

In this case, the expansion coefficients are

$$u_0(x) = \arctan(x + 1), \qquad u_1(x) = -\frac{1}{2}f(u_0(x)) \ln 2f(u_0(x)).$$

The left plot in figure 1 shows the norms $\sup_i |v_i - u_0(x_i)|$ and $\sup_i |v_i - u_0(x_i) - hu_1(x_i)|$ versus $1/h$ in double logarithmic scale. The right plot shows u_0, v_i

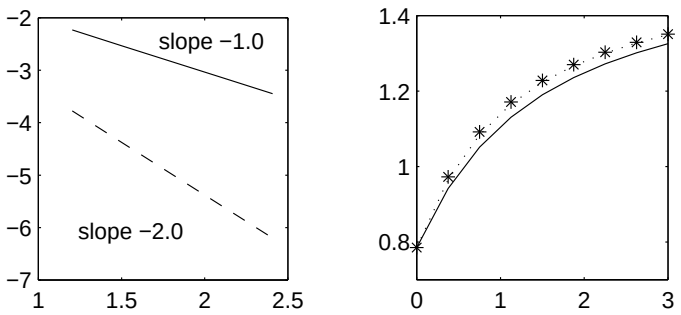

Fig. 1. Left: double logarithmic plot of maximal difference between v_i and u_0 (solid), respectively $u_0 + hu_1$ (dashed) versus $1/h$. The slopes confirm the predicted $\mathcal{O}(h)$ respectively $\mathcal{O}(h^2)$ behavior. Right: the exact solution u_0 (solid), the Euler solution v_i (stars), and the first order expansion $u_0 + hu_1$ (dotted) on the interval $[0, 3]$ with $h = 3/8$.

and $u_0 + hu_1$. Similar to the approach in [15], the knowledge of equation (3) for the leading error term allows us to construct a higher order method. The basic idea of this *direct correction* approach is to modify the original scheme in such a way that the leading order error term satisfies a homogeneous equation which only has the zero solution. In our example, we have to remove the source term $-u_0''/2 = -f'(u_0)f(u_0)/2$ in (3) which is easily accomplished by applying the Euler discretization to the problem

$$u' = \tilde{f}(u) = f(u) + \frac{1}{2}hf'(u)f(u), \qquad u(0) = \eta \tag{4}$$

with a grid $G = (h, 0)$ adapted to the parameter h in the equation. This leads to the scheme

$$\frac{v_{i+1} - v_i}{h} = f(v_i) + \frac{1}{2}hf'(v_i)f(v_i), \quad i \in I \cup \{0\}, \qquad v_0 = \eta$$

which is first order accurate to (4) but second order accurate with respect to our original problem. In fact, the asymptotic analysis yields

$$\begin{aligned} u_0'(x) &= f(u_0(x)), & u_0(0) &= \eta \\ u_1'(x) &= f'(u_0(x))u_1(x), & u_1(0) &= 0 \\ u_2'(x) &= f'(u_0(x))u_2(x) - u_0'''(x)/6, & u_2(0) &= 0 \end{aligned}$$

so that $v_i = u_0(x_i) + \mathcal{O}(h^2)$ because u_1 vanishes. Note that a repetition of the direct correction method for already corrected schemes leads to successively higher order methods (these schemes are also obtained with the Taylor expansion method [23]).

At this point, we can briefly comment on the relation between the modified equation approach and the direct asymptotic analysis presented here. The construction of modified equations is based on the assumption that a smooth

function $U(x,h)$ exists with the property that $U(x_i,h)$ satisfies the finite difference relation, i.e. (1) in the present case. Performing a Taylor expansion, we find that, at least at the grid points,

$$U' + \frac{h}{2}U' + \frac{h^2}{6}U''' + \cdots = f(U) \tag{5}$$

which is called (truncated) *first modified equation* in [13]. By taking suitable linear combinations of derivatives of (5), the higher order derivatives of U can be removed on the left hand side which gives rise to the (truncated) *second modified equation*

$$U' = f(U) - \frac{h}{2}\frac{d}{dx}f(U) + \frac{h^2}{12}\frac{d^2}{dx^2}f(U) + \ldots. \tag{6}$$

Finally, by using chain rule to work out the $f(U)$ derivatives and by successive removal of U derivatives with the help of (6), the truncated *third modified* or simply *modified equation* is obtained [13]

$$U' = f(U) - \frac{h}{2}f'(U)f(U) + \frac{h^2}{12}\left(f''(U)f(U)^2 + 4f'(U)^2 f(U)\right) + \ldots. \tag{7}$$

This is the form which is usually employed. Compared to (5) and (6) it has the advantage that no higher U derivatives appear which removes the problem of finding additional initial conditions.

To see how equations (5), (6), and (7) relate to the direct asymptotic analysis, we first note that the expansion coefficients for the Euler method (34) satisfy

$$\begin{aligned} u_0' &= f(u_0), \\ u_1' &= f'(u_0)u_1 - \frac{1}{2}u_0'', \\ u_2' &= f'(u_0)u_2 + \frac{1}{2}f''(u_0)u_1^2 - \frac{1}{6}u_0''' - \frac{1}{2}u_1'' \end{aligned} \tag{8}$$

Defining a function $u(x,h) = u_0(x) + hu_1(x) + h^2u_2(x)$, it is a straight forward calculation to see that u satisfies (5) up to terms of order h^3. In view of the u_0 equation in (8), we can also rewrite the u_1 equation as $u_1' = f'(u_0)u_1 - (f(u_0))'/2$ and similarly,

$$u_2' = f'(u_0)u_2 + \frac{1}{2}f''(u_0)u_1^2 - \frac{1}{2}\frac{d}{dx}\left(f'(u_0)u_1\right) + \frac{1}{12}\frac{d^2}{dx^2}f(u_0).$$

From this rewritten form, we easily find that u also satisfies (6) up to third order. Finally, by rewriting (8) in such a way that the derivatives of u_0 and u_1 on the right hand sides are removed, we see that u also satisfies (7) up to $\mathcal{O}(h^3)$ terms. More generally, we can say that $u = u_0 + hu_1 + \cdots + h^m u_m$ satisfies the truncated first, second, and third modified equations up to a remainder of

order h^{m+1}. Conversely, a regular asymptotic expansion of any of the modified equations necessarily leads to the same coefficients as in our approach. As already mentioned in [13], a detailed analysis of the modified equations anyhow requires methods of asymptotic analysis because of the appearance of the small parameter h in the equations. Hence, one can as well drop the unnecessary and strong assumption of a smooth function U with the required properties which is generally not justified (see also [22] and section 2.3, 2.6, and 2.7).

2.2 The symplectic Euler method

The purpose of this is example is to demonstrate the usefulness of multiscale expansions for understanding finite difference schemes (further examples can be found in [13]). We consider the harmonic oscillator problem on $\Omega = (0, \infty)$

$$U'(t) = u(t), \quad u'(t) = -U(t), \qquad U(0) = 1, u(0) = 0$$

which is discretized on a grid $G = (h, 0)$ by the symplectic Euler method (explicit Euler for the first and implicit Euler for the second equation – see [9])

$$\frac{V_{i+1} - V_i}{h} = v_i, \quad \frac{v_{i+1} - v_i}{h} = -V_{i+1}, \qquad V_0 = 1, v_0 = 0. \tag{9}$$

Performing a regular expansion as in the previous example, we find

$$\begin{aligned} U_0'(t) &= u_0(t), & u_0'(t) &= -U_0(t), \\ U_1'(t) &= u_1(t) - U_0''/2, & u_1'(t) &= -U_1(t) - u_0''/2 - u_0, \\ U_2'(t) &= u_2(t) - U_1''/2 - U_0''/6, & u_2'(t) &= -U_2(t) - u_1''/2 - u_1 - u_0'''/6, \end{aligned}$$

with $U_0(0) = 1$, $u_0(0) = 0$ and zero conditions for the higher order terms. The solutions are

$$\begin{pmatrix} U_0(t) \\ u_0(t) \end{pmatrix} = \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix}, \qquad \begin{pmatrix} U_1(t) \\ u_1(t) \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sin t \\ 0 \end{pmatrix},$$

and

$$\begin{pmatrix} U_2(t) \\ u_2(t) \end{pmatrix} = -\frac{1}{24} \begin{pmatrix} t \sin t \\ t \cos t + 3 \sin t \end{pmatrix}.$$

We conclude that the symplectic Euler method is first order accurate (the component v_i is actually second order accurate in our example) but we also see that this behavior can only be expected on bounded time intervals $[0, T]$ where $T \ll 1/h$: if t is of the order of $1/h$, the second order error contribution which is proportional to t (a so called *secular* term) influences the first order. This can be seen in the left plot of figure 2 where the asymptotic behavior of

the difference between (V_i, v_i) and the first order expansion is investigated on time intervals of different length. We conclude that the long time behavior of

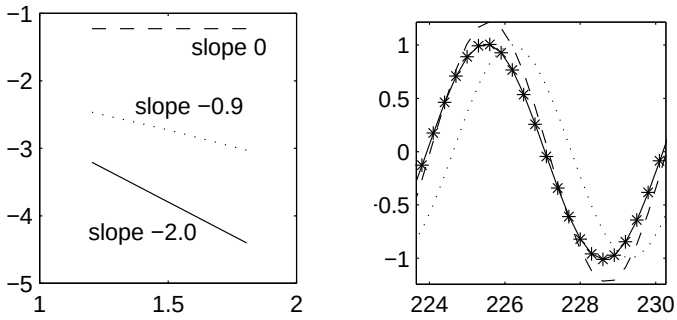

Fig. 2. Left: double logarithmic plot of maximal difference between numerical solution and first order expansion versus $1/h$ over the interval $[0, 1]$ (solid), on intervals $[0, 1/h]$ (dotted), and intervals $[0, 1/h^2]$ (dashed). Right: the exact solution U_0 (dotted), the Euler solution V_i (stars), the second order regular expansion $U_0+hU_1+h^2U_2$ (dashed) and the first order multiscale expansion (solid) around $t = 6/h^3$ with $h = 0.3$.

the scheme for fixed h is not properly represented by the regular expansion. The standard approach to deal with secular terms which are responsible for the non-uniform behavior of the regular expansion is the so called multiscale expansion (see, for example, [24])

$$V_i = U_0(t_i, h^2 t_i) + hU_1(t_i, h^2 t_i) + \mathcal{O}(h^2). \tag{10}$$

Here, $U_k(t, \tau)$ are smooth functions of two variables. We have chosen $\tau = h^2 t$ as slowly varying scale because of its appearance in the regular expansion. Inserting a two scale expansion like (10) for V_i and v_i into (9), expanding in h and collecting terms of same order, we find the expression

$$\begin{aligned} &\partial_t U_0 - u_0 + h(\partial_t U_1 + \partial_t^2 U_0/2 - u_1) \\ &\qquad + h^2(\partial_t U_2 + \partial_t^2 U_1/2 + \partial_\tau U_0 + \partial_t^3 U_0/6 - u_2) = \mathcal{O}(h^3) \end{aligned} \tag{11}$$

and a similar relation from the second part in (9). Note that all functions in (11) are evaluated at $(t_i, h^2 t_i)$ so that we cannot proceed as in the regular expansion: in the limit $h \to 0$, the leading order only yields $\partial_t U_0(t, 0) = u_0(t, 0)$ but no information about $\partial_t U_0(t, \tau) - u_0(t, \tau)$ for $\tau \neq 0$. Hence, we cannot remove the lowest order term in (11) to extract the first order condition. This problem is related to the non-uniqueness of the expansion coefficients in a multiscale expansion (see [24] for a general discussion). Usually, it is circumvented by *defining* the coefficients in such a way that the orders in (11) vanish separately. This choice meets the ultimate goal of the asymptotic analysis to specify coefficients $U_0, U_1, \ldots$ such that a certain expansion (like (10)) holds. It is convenient if the coefficients are determined uniquely but certainly not mandatory to achieve the goal.

Following this approach, we obtain for the lowest order

$$\partial_t U_0(t,\tau) = u_0(t,\tau), \quad \partial_t u_0(t,\tau) = -U_0(t,\tau), \qquad U_0(0,0) = 1, u_0(0,0) = 0$$

which yields

$$U_0(t,\tau) = A(\tau)\cos(t-\theta(\tau)), \qquad u_0(t,\tau) = -A(\tau)\sin(t-\theta(\tau))$$

with $A(0) = 1$, $\theta(0) = 0$. The undetermined coefficients will be fixed with the h^2-equations which contain τ-derivatives of U_0, u_0. In first order we have

$$\partial_t U_1 = u_1 + U_0/2, \quad \partial_t u_1 = -U_1 - u_0/2, \qquad U_1(0,0) = u_0(0,0) = 0$$

which leads to

$$\begin{aligned} U_1(t,\tau) &= B(\tau)\cos(t-\phi(\tau)) + A(\tau)\cos(t-\theta(\tau))/4, \\ u_1(t,\tau) &= -B(\tau)\sin(t-\phi(\tau)) - A(\tau)\sin(t-\theta(\tau))/4 \end{aligned}$$

with $B(0) = 1/4$, $\phi(0) = \pi/2$. In second order, the τ-derivative of the zero order terms appears

$$\begin{aligned} \partial_t U_2 &= u_2 + B\cos(t-\phi)/2 + (\theta' + 1/8)A\sin(t-\theta) - A'\cos(t-\theta), \\ \partial_t u_2 &= -U_2 + B\sin(t-\phi)/2 + (\theta' - 1/24)A\cos(t-\theta) + A'\sin(t-\theta). \end{aligned}$$

The additional degree of freedom in the parameters A, θ can now be used to suppress secular terms in the coefficients U_2, u_2. For example, the general solution for U_2 is

$$\begin{aligned} U_2(t,\tau) = t\Big(&-A'(\tau)\cos(t-\theta(\tau)) + (2\theta'(\tau) + 1/12)A(\tau)\sin(t-\theta(\tau))\Big) \\ &+ C(\tau)\cos(t-\lambda(\tau)) + B(\tau)\sin(t-\phi(\tau))/4 - A(\tau)\cos(t-\theta(\tau))/24 \end{aligned}$$

and the secular terms are removed with the conditions $A'(\tau) = 0$, $\theta'(\tau) = -1/24$. Taking the initial values for A and θ into account, we are led to $A(\tau) = 1$ and $\theta(\tau) = -\tau/24$. This choice also removes the secular terms in u_2. Similarly, $B(\tau) = 1/4$ and $\phi(\tau) = \pi/2 - \tau/24$ are fixed by the h^3 equation where one secular term remains because we have not included the scale $h^3 t$ into the lowest order of our multiscale expansion. Summarizing the result, we have

$$\begin{pmatrix} V_i \\ v_i \end{pmatrix} = \begin{pmatrix} \cos((1+h^2/24)t) \\ -\sin((1+h^2/24)t) \end{pmatrix} + \frac{h}{2}\begin{pmatrix} \sin((1+h^2/24)t) \\ 0 \end{pmatrix} + \mathcal{O}(h^2) \qquad (12)$$

which is valid up to $t = \mathcal{O}(1/h)$. There are two conclusions we can draw from the multiscale expansion: first, the amplitude of the symplectic Euler solution shows no divergent or convergent long-time behavior as, for example,

the explicit or implicit Euler solution. Second, the solution obtained with the symplectic Euler method has a slightly higher frequency $1 + h^2/24$ than the exact solution. This frequency leads to a varying phase difference. In the right plot of figure 2, the numerical solution and the twoscale expansion (12) are shown to be in very good coincidence even for large t. Moreover, the phase shift compared to the exact solution and the failure of the regular expansion can be observed.

2.3 A two step method

This example is taken from [21] where it was used as typical case in which the modified equation approach leads to wrong predictions. We consider again the initial value problem

$$u'(x) = f(u(x)), \quad x \in (0, \infty), \qquad u(0) = \eta$$

which is now discretized with the two step method on $G = (h, 0)$

$$\frac{v_{i+2} - v_i}{2h} = f(v_{i+1}), \qquad i \in I \cup \{0\}. \tag{13}$$

Apart form the initial value for v_0, this method needs also a starting step to determine v_1 which we take as explicit Euler step

$$v_0 = \eta, \qquad \frac{v_1 - v_0}{h} = f(v_0). \tag{14}$$

The regular expansion

$$v_i = u_0(x_i) + hu_1(x_i) + h^2 u_2(x_i) + \dots \tag{15}$$

inserted into relation (13) implies in the usual way

$$u_0'(x) = f(u_0(x)), \tag{16}$$
$$u_1'(x) = f'(u_0(x))u_1(x), \tag{17}$$
$$u_2'(x) = f'(u_0(x))u_2(x) + f''(u_0(x))u_1^2(x)/2 - u_0'''(x)/6, \tag{18}$$

and the initial condition yields

$$u_0(0) = \eta, \qquad u_1(0) = 0, \qquad u_2(0) = 0. \tag{19}$$

The new aspect in this example is the starting step which gives rise to conditions on u_k at a single point $x = h$. Since this node runs into the boundary point $x = 0$ for $h \to 0$, we Taylor expand the relation around the limit point

to make sure that the resulting equation is fully expanded in h. We obtain

$$u_0'(0) - f(u_0(0)) + h(u_1'(0) - f'(u_0(0))u_1(0) + u_0''(0)/2)$$
$$+h^2\bigg(u_2'(0) - f'(u_0(0))u_2(0) - f''(u_0(0))u_1'(0)^2/2 + u_0'''(0)/6 + u_1''(0)/2\bigg) = \mathcal{O}(h^3).$$

In view of (16) and the assumed smoothness up to the boundary, we can simplify this relation to

$$hu_0''(0)/2 + h^2 u_1''(0)/2 = \mathcal{O}(h^3). \tag{20}$$

which leads to additional boundary conditions

$$u_0''(0) = 0, \qquad u_1''(0) = 0.$$

While $u_1' = f'(u_0)u_1$, $u_1(0) = u_1''(0) = 0$ has $u_1(x) = 0$ as solution, the problem

$$u_0' = f(u_0), \qquad u_0(0) = \eta, \qquad u_0''(0) = 0,$$

is, in general, not solvable so that we arrive at a contradiction which tells us that the original expansion (15) with smooth coefficients is incompatible with the scheme (13), (14). In order to resolve this contradiction, we have to allow for non-smooth behavior. In fact, our expansion (15) is a particular case of a multiscale expansion of the grid function v_i where we assume a dependence only on the slow variable $hi = x_i$. More generally, we could also assume a dependence on the fast variable i which describes the behavior at the grid level (see also [19]). In our case, it suffices to introduce an additional expansion coefficient at second order to balance the term $hu_0''(0)/2$ in (20). Inserting

$$v_i = u_0(x_i) + h^2(u_2(x_i) + \tilde{u}_2(i)) + \mathcal{O}(h^3) \tag{21}$$

into (13), where u_0, u_2 satisfy (16) with initial conditions (19) (the term u_1 is equal to zero under these conditions and thus does not appear in the expansion), we find the following condition on $\tilde{u}_2$

$$\frac{\tilde{u}_2(i+2) - \tilde{u}_2(i)}{2h} = f'(u_0(x_{i+1}))\tilde{u}_2(i+1).$$

The Euler starting step implies

$$hu_0''(0)/2 + h^2\left(\frac{\tilde{u}_2(1) - \tilde{u}_2(0)}{h} - f'(u_0(0))\tilde{u}_2(0)\right) = \mathcal{O}(h^3) \tag{22}$$

and since the boundary condition $v_0 = \eta$ is already satisfied at lowest order, we have $\tilde{u}_2(0) = 0$ so that (22) is satisfied if

$$\tilde{u}_2(1) + u_0''(0)/2 = 0.$$

Altogether, we find a difference scheme for the grid function $\tilde{u}_2$

$$\frac{\tilde{u}_2(i+2)-\tilde{u}_2(i)}{2h} = f'(u_0(x_{i+1}))\tilde{u}_2(i+1), \quad \tilde{u}_2(0)=0,\ \frac{1}{2}\tilde{u}_2(1) = -u_0''(0). \tag{23}$$

We want to stress that this derivation, as many derivations in asymptotic analysis, should actually be read backwards: if u_0, u_1, u_2 satisfy (16) and (19) and if $\tilde{u}_2$ satisfies (23) and if all these coefficients are bounded independent of h (here, stability arguments for (23) and boundedness results for solutions of (16) are required which we implicitly assume in our formal analysis), then v_i satisfies the expansion (21).

From (23) we clearly see why the regular expansion leads to a contradiction. It simply neglects the function $\tilde{u}_2$ which vanishes at $x=0$ but has the value $-u_0''(0)/2$ at $x=h$; a behavior which is incompatible with a smooth, h-independent expansion coefficient.

The structure of $\tilde{u}_2$ can be investigated by noting that, due to linearity, we can write $\tilde{u}_2(i) = w_i + z_i$ where

$$\frac{w_{i+2}-w_i}{2h} = f'(u_0(x_{i+1}))w_{i+1}, \quad w_0 = -\frac{1}{4}u_0''(0),\ w_1 = w_0 \tag{24}$$

and

$$\frac{z_{i+2}-z_i}{2h} = f'(u_0(x_{i+1}))z_{i+1}, \quad z_0 = \frac{1}{4}u_0''(0),\ z_1 = -z_0. \tag{25}$$

Since the values w_0, w_1 do not differ, we can assume that, in leading order, w can be described by a smooth function. Performing an expansion, we find $w_i = \bar{w}(x_i) + \mathcal{O}(h)$ where

$$\bar{w}'(x) = f'(u_0(x))\bar{w}(x), \qquad \bar{w}(0) = -\frac{1}{4}u_0''(0).$$

In contrast to this, the grid function z is expected to be very oscillatory because z_0 and z_1 differ exactly by their sign. Guided by this observation, we assume an expansion of the form

$$z_i = (-1)^i\psi(x_i) + \mathcal{O}(h) \tag{26}$$

with a smooth function ψ. Inserting (26) into (25), we obtain in leading order after a Taylor expansion

$$\psi'(x) = f'(u_0(x))\psi(x), \qquad \psi(0) = \frac{1}{4}u_0''(0)$$

so that

$$\tilde{u}_2(i) = \bar{w}(x_i) + (-1)^i\psi(x_i) + \mathcal{O}(h).$$

This oscillatory behavior of the error of multistep schemes is well known [8,21]. To verify our predictions, we consider the case $f(u) = \cos^2 u$ where all relevant

functions can be expressed in terms of

$$u_0(x) = \arctan(x + \tan\eta) \qquad \text{and} \qquad \phi(x) = \frac{\tan^2\eta + 1}{(x + \tan\eta)^2 + 1}.$$

Specifically, we have

$$u_2(x) = \frac{1}{6} f(\eta)(2\eta + 2f'(\eta))\phi(x) - \frac{1}{6} f(u_0(x))\Big(2u_0(x) + 2f'(u_0(x))\Big),$$

and

$$\bar{w}(x) = -\frac{1}{4} u_0''(0)\phi(x), \qquad \psi(x) = \frac{1}{4} u_0''(0)\frac{1}{\phi(x)}.$$

The left plot in figure 3 shows the difference between v_i and $u_0 + h^2 u_2$, respectively $u_0 + h^2(u_2 + \tilde{u}_2)$. On the right, the error $(v_i - u_0(x_i))/h^2$ is shown which coincides with the predicted second order contribution

$$u_2(x_i) + \bar{w}(x_i) + (-1)^i \psi(x_i)$$

up to plotting accuracy.

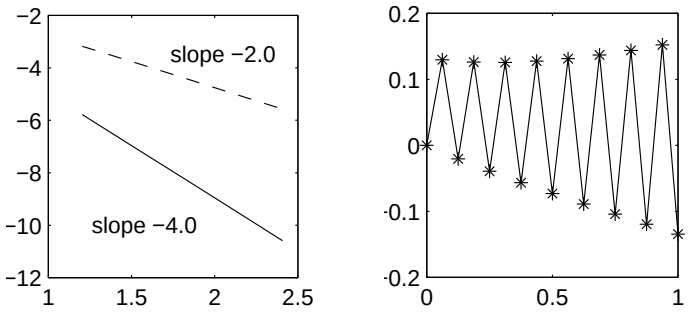

Fig. 3. Left: double logarithmic plot of maximal difference between v_i and $u_0 + h^2 u_2$ (dashed), respectively $u_0 + h^2(u_2 + \tilde{u}_2)$ (solid) versus $1/h$. Right: the oscillating error $(v_i - u_0(x_i))/h^2$ on the interval $[0, 1]$ with $h = 1/16$.

2.4 Upwind and downwind scheme

With this example, we want to show how a multiscale analysis of a finite difference scheme can give a first idea about its stability. We discretize the advection equation

$$\partial_t u(t,x) + \partial_x u(t,x) = 0, \qquad u(0,x) = \phi(x)$$

on $\Omega = (0,\infty) \times \mathbb{R}$ with both the upwind and the downwind method. Note that the advection velocity is positive so that the downwind method will be unstable – this has to be predicted by our analysis. Since our domain is now

two dimensional, the index $\boldsymbol{i} = i_1\boldsymbol{e}_1 + i_2\boldsymbol{e}_2$ has two components where the first corresponds to time and the second to space ($\boldsymbol{e}_1 = (1,0)$ and $\boldsymbol{e}_2 = (0,1)$ are the standard unit vectors). The point $\boldsymbol{x_i}$ corresponding to the index $\boldsymbol{i}$ is chosen as

$$\boldsymbol{x_i} = (t_{\boldsymbol{i}}, x_{\boldsymbol{i}}) = \lambda i_1 h\boldsymbol{e}_1 + i_2 h\boldsymbol{e}_2, \qquad \lambda > 0$$

which corresponds to a rectangular space-time grid if $\lambda \neq 1$. The upwind discretization is given by

$$v_{\boldsymbol{i}+\boldsymbol{e}_1} = v_{\boldsymbol{i}} - \lambda(v_{\boldsymbol{i}} - v_{\boldsymbol{i}-\boldsymbol{e}_2}), \qquad v_{k\boldsymbol{e}_2} = \phi(x_k) \tag{27}$$

and the downwind discretization differs only in the spatial difference

$$v_{\boldsymbol{i}+\boldsymbol{e}_1} = v_{\boldsymbol{i}} - \lambda(v_{\boldsymbol{i}+\boldsymbol{e}_2} - v_{\boldsymbol{i}}), \qquad v_{k\boldsymbol{e}_2} = \phi(x_k). \tag{28}$$

Starting with a regular expansion up to first order, we obtain in case of the upwind scheme

$$\begin{aligned} &\partial_t u_0 + \partial_x u_0 = 0, && u_0(0,x) = \phi(x), \\ &\partial_t u_1 + \partial_x u_1 = \frac{1-\lambda}{2}\partial_x^2 u_0, && u_1(0,x) = 0. \end{aligned} \tag{29}$$

To obtain the u_1 equation, we have used that $\partial_t^2 u_0 = \partial_x^2 u_0$ according to the equation satisfied by u_0. The solution to (29) is given by

$$u_0(t,x) = \phi(x-t), \qquad u_1(t,x) = t\frac{1-\lambda}{2}\phi''(x-t).$$

Similarly, for the downwind scheme (28), equations like (29) are obtained with the parameter $-(\lambda+1)/2$ instead of $(1-\lambda)/2$ in the equation for u_1. The corresponding solution is therefore

$$u_0(t,x) = \phi(x-t), \qquad u_1(t,x) = -t\frac{1+\lambda}{2}\phi''(x-t).$$

At this stage, both methods seem to be rather similar but we remark that the first order coefficient is a secular term because of the factor t. This indicates that the expansion is only valid for a short time. In figure 4, the numerical error of the schemes is compared with the predicted error for the case $\phi(x) = \exp(-x^2)$. One can check that, until the instability starts, the first order expansion of the downwind method really matches the numerical solution up to order h^2. To investigate the long time behavior of the scheme, we now use a two scale expansion. Inserting

$$v_{\boldsymbol{i}} = u_0(t_{\boldsymbol{i}}, ht_{\boldsymbol{i}}, x_{\boldsymbol{i}}) + hu_1(t_{\boldsymbol{i}}, ht_{\boldsymbol{i}}, x_{\boldsymbol{i}}) + \mathcal{O}(h^2)$$

into (27), we find conditions on the coefficients $u_k(t,\tau,x)$ by equating the expressions in different orders to zero. Specifically, we have

$$\partial_t u_0 + h\partial_\tau u_0 + \frac{\lambda}{2}h\partial_t^2 u_0 + h\partial_t u_1 = -\partial_x u_0 - h\partial_x u_1 + \frac{h}{2}\partial_x^2 u_0 + \mathcal{O}(h^2)$$

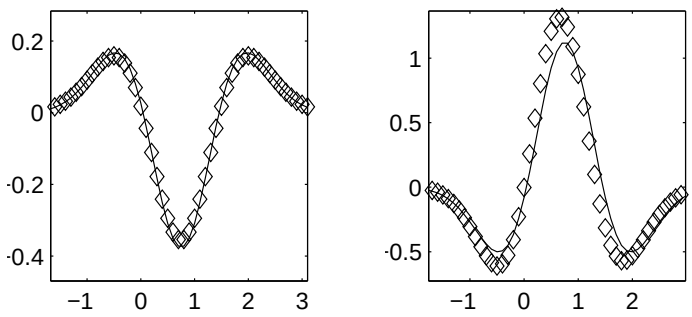

Fig. 4. Numerical error $(v_i - u_0(x_i))/h$ (diamonds) and predicted first order error u_1 (solid) for the upwind (left) and downwind discretization (right). The initial value is $\phi(x) = \exp(-x^2)$, $\lambda = 0.5$, $h = 0.1$ and the result corresponds to $t = 0.75$.

and we obtain in lowest order

$$\partial_t u_0 + \partial_x u_0 = 0, \qquad u_0(0, 0, x) = \phi(x).$$

with solution $u_0(t, \tau, x) = A(\tau, x-t)$ where $A(0, x) = \phi(x)$. Using $\partial_t^2 u_0 = \partial_x^2 u_0$, we find in first order

$$\partial_t u_1 + \partial_x u_1 = \frac{1-\lambda}{2}\partial_x^2 u_0 - \partial_\tau u_0, \qquad u_1(0, 0, x) = 0.$$

To avoid a secular term in the expression for u_1 we have to assume that

$$\partial_\tau u_0 = \frac{1-\lambda}{2}\partial_x^2 u_0$$

which is a diffusion problem for the undetermined coefficient A. For $\lambda < 1$, the problem is well posed and its solution is given by a convolution with the fundamental solution

$$A(\tau, x-t) = (\phi * G_\tau)(x-t), \qquad G_\tau(y) = \frac{1}{\sqrt{2\pi(1-\lambda)\tau}} \exp\left(-\frac{y^2}{2(1-\lambda)\tau}\right).$$

In the case $\lambda > 1$, the problem is ill posed which means that the secular term cannot be suppressed with a two scale expansion – an indication for problems in the long time behavior. The same problem arises if we consider the two scale expansion of the downwind scheme. All the calculations are exactly as in the upwind case, only the coefficient $(1-\lambda)/2$ is replaced by $-(1+\lambda)/2$ resulting again in a backward diffusion equation for the determination of the function A. We conclude that by considering the long time behavior of a finite difference scheme with a two scale expansion, we can obtain information about its stability. For the upwind scheme with $\lambda < 1$ we can illustrate the precision of the two scale expansion for the initial value $\phi(x) = \exp(-x^2)$. In this case,

the convolution integral can easily be evaluated

$$u_0(t,\tau,x) = \frac{1}{\sqrt{2(1-\lambda)\tau+1}} \exp\left(-\frac{(x-t)^2}{2(1-\lambda)\tau+1}\right).$$

In figure 5, the behavior of the difference $|v_{\boldsymbol{i}} - u_0(t_{\boldsymbol{i}}, ht_{\boldsymbol{i}}, x_{\boldsymbol{i}})|$ for $h \to 0$ is shown.

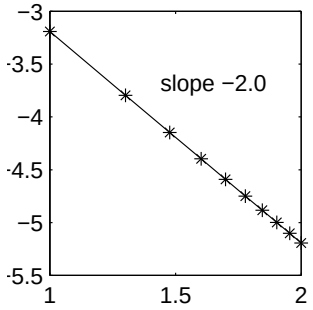

Fig. 5. Double logarithmic plot of maximal difference between v_i and two scale expansion u_0 versus $1/h$ on the time interval $[0, 1/h]$.

2.5 A two point boundary value problem

In the initial value problems considered up to now, we always chose the grid in such a way that the first node in time direction had exactly the grid distance h from the boundary $t = 0$. In fact, whenever Ω is an open interval, we can choose the grid $G = (h, \alpha)$ in such a way that the terminal nodes (those nodes in $I(G, \Omega)$ whose neighbors are outside Ω) have distance h to the boundary. However, this nice coincidence of grids and open sets is only possible in 1D situations (or for very special sets Ω in higher dimensions like half spaces or cubes). Already in two dimensions, we face the problem that a regular grid is generally incompatible with a curved boundary: the distance to the boundary changes from terminal node to terminal node and, eventually, the fluctuating distance information leads to fluctuations in the solution.

Before we consider this effect, let us first study a 1D situation which resembles the multidimensional case in the sense that the grid is not compatible with the domain, i.e. the terminal points do not have grid distance to the boundary. We discretize

$$u'' = f \quad \text{in } \Omega = (a,b), \qquad u = g \quad \text{on } \partial\Omega = \{a,b\} \tag{30}$$

on a general grid $G = (h, \alpha)$ Defining the indices of the left and right terminal point

$$l(G) = \min I(G,\Omega), \qquad r(G) = \max I(G,\Omega),$$

and the indices $I_{int}(G,\Omega) = \{l+1,\ldots,r-1\}$ of the interior nodes, we can formulate a finite difference method as

$$\frac{1}{h^2}\left(v_{i+1} - 2v_i + v_{i-1}\right) = f(x_i), \qquad i \in I_{int} \tag{31}$$

together with

$$\frac{1}{h^2}\left(v_{l+1} - 2v_l + g(a)\right) = f(x_l) \tag{32}$$

and

$$\frac{1}{h^2}\left(g(b) - 2v_r + v_{r-1}\right) = f(x_r). \tag{33}$$

The standard method to assess consistency of the scheme is the truncation error analysis: assuming that u is the solution of (30), the local truncation error at $x_i \in I_{int}$ is obtained by substituting the grid function $v_i = u(x_i)$ into the difference of left and right hand side of (31) and performing a Taylor expansion. We find

$$\frac{1}{h^2}\left(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\right) - f(x_i) = u''(x_i) - f(x_i) + \mathcal{O}(h^2) = \mathcal{O}(h^2)$$

which implies second order accuracy. If we assume that the distance between x_l and a is a fraction ph of the grid spacing, the same procedure applied to (32) yields at x_l

$$\begin{aligned}\frac{1}{h^2}&\left(u(x_{l+1}) - 2u(x_l) + u(x_l - ph)\right) - f(x_l)\\ &= \frac{1-p}{h}u'(x_l) + \frac{p^2-1}{2}f''(x_l) + \frac{1-p^3}{6}hu'''(x_l) + \mathcal{O}(h^2).\end{aligned}$$

In the favorable case $p = 1$, the consistency is again of second order. However, if $p < 1$, the discretization is actually inconsistent due to the $\mathcal{O}(h^{-1})$ behavior of the truncation error. Hence, the standard stability argument would not imply convergence and a more detailed analysis is required to show that the two exceptional points are not so important in the sense that convergence is actually of first order in h (see, for example, [25]). In any case, the simple rule: consistency order = convergence order is no longer applicable. As we will see below, this problem is circumvented with the asymptotic analysis – we obtain first order consistency.

Our analysis starts by substituting a regular expansion

$$v_i = u_0(x_i) + hu_1(x_i) + h^2u_2(x_i) + \mathcal{O}(h^3) \tag{34}$$

into the interior algorithm (31). Noting that for smooth functions u

$$\frac{1}{h^2}\left(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\right) = u''(x_i) + \frac{h^2}{12}u^{(4)}(x_i) + \mathcal{O}(h^4),$$

we obtain in the usual way

$$u_0'' = f, \quad u_1'' = 0, \quad u_2'' = -\frac{1}{12}u_0^{(4)}, \quad u_3'' = 0, \quad \text{in } \Omega. \tag{35}$$

In order to fully determine the coefficients u_k, we now derive boundary conditions for the equations (35). Details are given for the left boundary. The idea is to substitute the expansion (34) into the boundary relation (32) and to expand all functions around the boundary point $x = a$. Using

$$u(x_l) = u(a + ph) = u(a) + phu'(a) + \frac{1}{2}p^2h^2u''(a) + \dots,$$
$$u(x_{l+1}) = u(a + (1+p)h) = u(a) + (1+p)hu'(a) + \frac{1}{2}(1+p)^2h^2u''(a) + \dots,$$

we find

$$\frac{1}{h^2}(-u_0(a) + g(a)) + \frac{1}{h}(-u_1(a) + (1-p)u_0'(a)) = \mathcal{O}(1). \tag{36}$$

From the leading order, we conclude that $u_0(a) = g(a)$. A similar investigation of the right boundary yields $u_0(b) = g(b)$ so that u_0 is the solution of the original problem (30). Since $u_0(a) = g(a)$ removes the leading term in (36), we obtain after multiplication with h

$$-u_1(a) + (1-p)u_0'(a) = \mathcal{O}(h), \qquad -u_1(b) + (q-1)u_0'(b) = \mathcal{O}(h) \tag{37}$$

where the coefficients

$$p(G) = \frac{x_l(G) - a}{h}, \qquad q(G) = \frac{b - x_r(G)}{h}$$

incorporate details of the grid. It is clear that (37) can only lead to reasonable boundary conditions for u_1 if $p(G)$ and $q(G)$ are convergent along a grid sequence. In general, however, this cannot be expected. As example, we consider the case $\Omega = (0, 2\pi)$ and the grid sequence $G_n = (h_n, 0)$ with $h_n = 2^{-n}$. In this case, $q(G_n)$ does not converge

$$p(G_n) = 1, \qquad q(G_n) = 2^{n-1}\pi - \lfloor 2^{n-1}\pi \rfloor$$

so that $(1 - q(G_n))u_0'(b)$ in (37) cannot be balanced by the grid independent term $u_1(b)$. This conflict clearly shows that a grid independent function u_1 is, in general, incompatible and we have to modify our expansion. One possibility is to assume that u_1 depends on the grid via some function $m_G : \partial\Omega \to \mathbb{R}$ given by $m_G(a) = 1 - p(G)$ and $m_G(b) = q(G) - 1$. In this case, conditions (37) are satisfied by assuming

$$u_1(x, m) = m(x)u_0'(x), \qquad x \in \{a, b\}.$$

In connection with $u_1'' = 0$, we conclude that u_1 is, for every m, a linear function on Ω. Since u_0 is the exact solution of (30) and u_1 is generally different from

zero, we conclude with $v_i = u_0(x_i) + hu_1(x_i, m_G) + \cdots = u_0(x_i) + \mathcal{O}(h)$ that the method is first order consistent.

To underline the relevance of the modified expansion

$$v_i(G) = u_0(x_i(G)) + hu_1(x_i(G), m_G) + \cdots , \tag{38}$$

let us consider an example on the interval $\Omega = (0, 2\pi)$ with $g(0) = g(2\pi) = 0$ and source $\hat{f}(x) = \sin x$. The corresponding exact solution is $\hat{u}(x) = -\sin x$ and we expect the grid function to satisfy

$$\hat{v}_i = \hat{u}(x_i) + h\left(m_G(a) + \frac{m_G(b) - m_G(a)}{b-a}(x_i - a)\right) + \mathcal{O}(h^2).$$

Along a grid sequence with $h_n = 2^{-n}$, $\alpha_n = 0$, the behavior of $\sup_i |v_i - u_0(x_i)|$ and $\sup_i |v_i - u_0(x_i) - hu_1(x_i, m_G)|$ versus $1/h$ is shown in figure 6. Note that the error curves are oscillating because of the oscillating value $q(G_n)$. Since

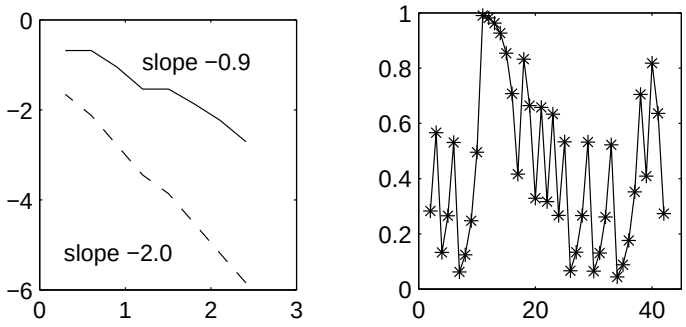

Fig. 6. Left: double logarithmic plot of maximal difference between v_i and u_0 (solid) and $u_0 + hu_1$ (dashed) versus $1/h$. The oscillations are due to the irregular behavior of $q(G_n)$ shown on the right for $n = 2 \ldots 42$.

$u_1'' = 0$, the scheme would be second order accurate if u_1 satisfied homogeneous Dirichlet conditions. This can be achieved by direct correction: we replace the boundary value g in (32), (33) by $g - hm_G u_0'$ where u_0' is discretized as

$$u_0'(a) \approx \frac{v_l - g(a)}{ph}, \qquad u_0'(b) \approx \frac{g(b) - v_r}{qh}$$

and obtain the modified boundary conditions

$$\frac{1}{h}\left(\frac{v_{l+1} - v_l}{h} - \frac{v_l - g(a)}{ph}\right) = f(x_l), \qquad \frac{1}{h}\left(\frac{g(b) - v_r}{qh} - \frac{v_r - v_{r-1}}{h}\right) = f(x_r). \tag{39}$$

The asymptotic analysis of (30) with (39) yields (35) with

$$u_0 = g, \qquad u_1 = 0, \qquad u_2 = m_G^{(2)} u_0'', \qquad \text{on } \partial\Omega$$

where

$$m_G^{(2)}(a) = \frac{p(p-1)}{2}, \qquad m_G^{(2)}(b) = \frac{q(1-q)}{2}$$

so that we expect second order accuracy with a grid dependent leading error term u_2. With an additional direct correction, also the boundary condition for u_2 can be made homogeneous which does not increase the order of the scheme (unless f is linear so that $u_0^{(4)} = 0$) but it removes the oscillation of the leading error term. Replacing g in (39) by $g - h^2 m_G^{(2)} u_0''$ or, equivalently in that order, by

$$g(a) - h^2 \frac{p(p-1)}{2} f(x_l), \qquad g(b) - h^2 \frac{q(1-q)}{2} f(x_r)$$

we obtain the boundary condition

$$\begin{aligned} \frac{2}{h(1+p)} \left(\frac{v_{l+1} - v_l}{h} - \frac{v_l - g(a)}{ph} \right) &= f(x_l), \\ \frac{2}{h(1+q)} \left(\frac{g(b) - v_r}{qh} - \frac{v_r - v_{r-1}}{h} \right) &= f(x_r) \end{aligned} \tag{40}$$

which is known as Shortley-Weller approximation [26]. The asymptotic analysis yields (35) with

$$u_0 = g, \qquad u_1 = 0, \qquad u_2 = 0, \qquad u_3 = m_G^{(3)} u_0'', \qquad \text{on } \partial\Omega$$

where

$$m_G^{(3)}(a) = \frac{(1-p^2)p}{6}, \qquad m_G^{(3)}(b) = \frac{(q^2-1)q}{6}. \tag{41}$$

In figure 7, the error of the scheme based on (39), respectively (40) is shown for the test case $f(x) = \cos(x/2)$ on $(0, 2\pi)$ with $g = 0$. For a linear source $\check{f}(x) = 1 + \frac{x}{2\pi}$, the Shortley-Weller approximation yields, as predicted, a third order accurate approximation of the exact solution

$$\check{u}(x) = \frac{x^3}{12\pi} + \frac{x^2}{2} - \frac{4\pi x}{3}.$$

One can actually check that the expansion $\check{v}_i = \check{u}(x_i) + h^3 \check{u}_3(x_i, m_G^{(3)})$ is exact in this case (see also figure 7).

2.6 A 2D hyperbolic problem

In the previous examples, we have seen that smoothness of the coefficients u_k is required to carry out the expansion. While this is a reasonable assumption in the case of elliptic equations, problems may arise in connection with hyperbolic equations where smoothness is a more delicate issue. We will show that, as far as the asymptotic analysis is concerned, a lack of smoothness

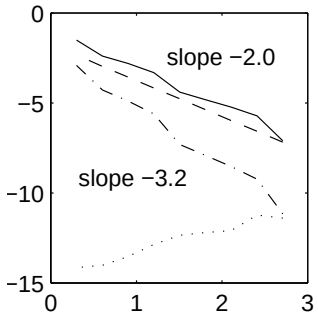

Fig. 7. Numerical error versus $1/h$ in double logarithmic scale: the Shortley-Weller approximation gives a straight line (dashed), the scheme (31) with (39) has an oscillating error curve. The Shortley-Weller approximation with a linear source is third order accurate (dash-dot) and the third order expansion is exact up to round off errors (dotted).

in the coefficients necessitates a more careful asymptotic analysis. For example, jump discontinuities give rise to interior layers which require a matched asymptotic expansion (see [24]). To illustrate this situation, let us consider the linear hyperbolic equation

$$\partial_x u + \partial_y u = 0, \qquad \text{in } \Omega = (0,1)^2. \tag{42}$$

On the west boundary $\Gamma_w = \{0\} \times (0,1)$ and the south boundary $\Gamma_s = (0,1) \times \{0\}$ of the unit square Ω, values for u are prescribed

$$u = g \qquad \text{on } \Gamma = \Gamma_w \cup \Gamma_s \cup \{0\}. \tag{43}$$

The function g is assumed to be smooth on $\mathbb{R}^2$. The solution of (42), (43) is easily determined with the method of characteristics

$$u(x,y) = g((x-y)_+, (y-x)_+), \qquad z_+ = \max(z,0).$$

In the following, we will investigate three situations

$$\hat{g}(x,y) = x - y, \qquad \tilde{g}(x,y) = \sin(2\pi(x-y)), \qquad \check{g}(x,y) = \sin(x)$$

with corresponding solutions

$$\hat{u}(x,y) = x - y, \qquad \tilde{u}(x,y) = \sin(2\pi(x-y)), \qquad \check{u}(x,y) = \sin((x-y)_+).$$

Note that $\hat{u}$ and $\tilde{u}$ are smooth while $\check{u}$ has a corner (jump in the derivative) along the diagonal $x = y$. As finite difference discretization for (42), we use the upwind method

$$\frac{v_{\boldsymbol{i}} - v_{\boldsymbol{i}-\boldsymbol{e}_1}}{h} + \frac{v_{\boldsymbol{i}} - v_{\boldsymbol{i}-\boldsymbol{e}_2}}{h} = 0, \qquad \boldsymbol{i} \in I. \tag{44}$$

For the nodes $\boldsymbol{x}_{\boldsymbol{i}}$ where the neighbors are missing, i.e. $\boldsymbol{x}_{\boldsymbol{i}-\boldsymbol{e}_k} \notin \Omega$, we define $v_{\boldsymbol{i}-\boldsymbol{e}_k}$ as suitable boundary value. Specifically, we take the value of g at the

point on the boundary where the half line $\boldsymbol{x_i} - t\boldsymbol{e}_k$, $t > 0$ intersects Γ. Introducing $d_{\Gamma,\boldsymbol{e}}$ as distance to Γ in direction $\boldsymbol{e}$ and $\Pi_{\Gamma,\boldsymbol{e}}$ as corresponding boundary projection

$$\Pi_{\Gamma,\boldsymbol{e}}(\boldsymbol{x}) = \boldsymbol{x} + d_{\Gamma,\boldsymbol{e}}(\boldsymbol{x})\boldsymbol{e}, \qquad \text{if } d_{\Gamma,\boldsymbol{e}} = \inf\{t > 0 : \boldsymbol{x} + t\boldsymbol{e} \in \Gamma\} < \infty$$

we can summarize the definition as

$$v_{\boldsymbol{i}-\boldsymbol{e}_k} := g\left(\Pi_{\Gamma,-\boldsymbol{e}_k}(\boldsymbol{x_i})\right), \qquad \boldsymbol{x}_{\boldsymbol{i}-\boldsymbol{e}_k} \notin \Omega. \tag{45}$$

Starting with the expansion

$$v_{\boldsymbol{i}} = u_0(\boldsymbol{x_i}) + hu_1(\boldsymbol{x_i}) + \mathcal{O}(h^2) \tag{46}$$

and assuming smooth functions u_k, we derive differential equations from equation (44) by Taylor expansion. Specifically, we find

$$\partial_x u_0 + \partial_y u_0 = 0 \quad \text{in } \Omega \tag{47}$$

and

$$\partial_x u_1 + \partial_y u_1 = \frac{1}{2}\Delta u_0 \quad \text{in } \Omega. \tag{48}$$

Note that the equation for the leading error term u_1 has a source $\Delta u_0/2$ which is large at points where the solution u_0 has a strong curvature and that this error is transported in the advection direction. The error source is related to the diffusive behavior of the scheme and, regardless of the boundary conditions, we see that u_1 will be nonzero in general, so that the consistency order is at most one.

To explain the derivation of the boundary conditions, we consider a point $\boldsymbol{x_i}$ next to the west boundary where (46) reads

$$v_{\boldsymbol{i}} - g\left(\Pi_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x_i})\right) + v_{\boldsymbol{i}} - v_{\boldsymbol{i}-\boldsymbol{e}_2} = 0, \qquad \boldsymbol{i} \in I_w \tag{49}$$

with $I_w = \{\boldsymbol{i} \in I : \boldsymbol{i} - \boldsymbol{e}_1 \notin I,\, \boldsymbol{i} - \boldsymbol{e}_2 \in I\}$. To extract information about the coefficients u_0, u_1, we insert the expansion (46) into (49) and expand around the boundary point

$$\bar{\boldsymbol{x}}_{\boldsymbol{i}} = \Pi_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x_i}) = \boldsymbol{x_i} - d_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x_i})\boldsymbol{e}_1.$$

This leads to

$$\begin{aligned} u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) - g(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) + hu_1(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) + (d_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x_i}) - h)\partial_x u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) \\ = -h(\partial_x u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) + \partial_y u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}})) + \mathcal{O}(h^2). \end{aligned} \tag{50}$$

In view of (47), we first note that the bracket on the right hand side vanishes. Secondly, the distance $d_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x_i})$ between the point $\boldsymbol{x_i}$ and Γ along the line in direction $-\boldsymbol{e}_1$ is of order h. We remark that $d_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x_i})$ can also be expressed

in terms of the boundary point $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$. A straight forward calculation shows that with

$$p(G, \boldsymbol{e}, \boldsymbol{x}) = 1 - \left(\frac{(\boldsymbol{x} - \boldsymbol{\alpha}) \cdot \boldsymbol{e}}{h} - \left\lfloor \frac{(\boldsymbol{x} - \boldsymbol{\alpha}) \cdot \boldsymbol{e}}{h} \right\rfloor \right), \tag{51}$$

we have the relation

$$d_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x}_{\boldsymbol{i}}) = p\left(G, \boldsymbol{e}_1, \Pi_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x}_{\boldsymbol{i}})\right) h.$$

Thus (50) can be rewritten as

$$u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) - g(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) + h u_1(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) + (p(G, \boldsymbol{e}_1, \bar{\boldsymbol{x}}_{\boldsymbol{i}}) - 1) h \partial_x u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) = \mathcal{O}(h^2). \tag{52}$$

Given $\bar{\boldsymbol{x}} \in \Gamma_w$ and a grid sequence $G_n = (h_n, \boldsymbol{\alpha}_n)$ with $h_n \to 0$, we can find a corresponding sequence of indices $\boldsymbol{i}_n \in I_w(G_n)$ such that $\Pi_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x}_{\boldsymbol{i}_n}) \to \bar{\boldsymbol{x}}$ for $n \to \infty$. In this way, we deduce from (52) that

$$u_0(\bar{\boldsymbol{x}}) = g(\bar{\boldsymbol{x}}), \qquad \bar{\boldsymbol{x}} \in \Gamma_w.$$

With the particular choice $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_{\boldsymbol{i}}$, we can reduce (52) to

$$u_1(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) + (p(G, \boldsymbol{e}_1, \bar{\boldsymbol{x}}_{\boldsymbol{i}}) - 1) \partial_x u_0(\bar{\boldsymbol{x}}_{\boldsymbol{i}}) = \mathcal{O}(h). \tag{53}$$

As in our previous example, we see that for general grid sequences G_n, the term involving p need not converge so that a contradiction arises from our assumption that u_1 only depends on position. As before, we go over to the more general expansion

$$v_{\boldsymbol{i}} = u_0(\boldsymbol{x}_{\boldsymbol{i}}) + h u_1(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{m}_G) + \mathcal{O}(h^2)$$

Then, (53) can be satisfied if we set

$$\begin{aligned}
\boldsymbol{m}_G(\bar{\boldsymbol{x}}) &= (1 - p(G, \boldsymbol{e}_1, \bar{\boldsymbol{x}}))\boldsymbol{e}_1, \qquad \bar{\boldsymbol{x}} \in \Gamma_w \\
\boldsymbol{m}_G(\bar{\boldsymbol{x}}) &= (1 - p(G, \boldsymbol{e}_2, \bar{\boldsymbol{x}}))\boldsymbol{e}_2, \qquad \bar{\boldsymbol{x}} \in \Gamma_s \\
\boldsymbol{m}_G(0) &= ((1 - p(G, \boldsymbol{e}_1, 0))\boldsymbol{e}_1 + (1 - p(G, \boldsymbol{e}_2, 0)\boldsymbol{e}_2)/2,
\end{aligned}$$

and

$$u_1(\bar{\boldsymbol{x}}, \boldsymbol{m}_G) = \boldsymbol{m}_G(\bar{\boldsymbol{x}}) \cdot \nabla u_0(\bar{\boldsymbol{x}}), \qquad \bar{\boldsymbol{x}} \in \Gamma. \tag{54}$$

To check the prediction of the asymptotic analysis, we first consider the example with the source $\tilde{g}(x, y) = \sin(2\pi(y - x))$ on a grid $G(h, \boldsymbol{\alpha})$ with $\alpha = (h, h)$, so that

$$p(G, \boldsymbol{e}_1, \bar{\boldsymbol{x}}) = 1, \quad \bar{\boldsymbol{x}} \in \Gamma_w, \qquad p(G, \boldsymbol{e}_2, \bar{\boldsymbol{x}}) = 1, \quad \bar{\boldsymbol{x}} \in \Gamma_s.$$

In this case, $\boldsymbol{m}_G$ vanishes, and $\tilde{u}_1$ is determined by (48) with zero boundary values, giving rise to

$$\tilde{u}_1(x, y, 0) = -4\pi^2 \sin(2\pi(y - x)) \min(x, y).$$

Note that $\tilde{u}_1$ is continuously differentiable but not twice differentiable which affects the next order term u_2. In the left part of figure 8, we can see that $\sup_{\boldsymbol{i}\in I} |\tilde{v}_{\boldsymbol{i}} - \tilde{u}(\boldsymbol{x}_{\boldsymbol{i}}) - h\tilde{u}_1(\boldsymbol{x}_{\boldsymbol{i}}, 0)|$ is numerically of order 1.85 in h which supports that $\tilde{u}_1$ is the correct first order term. The right part in figure 8 shows ten

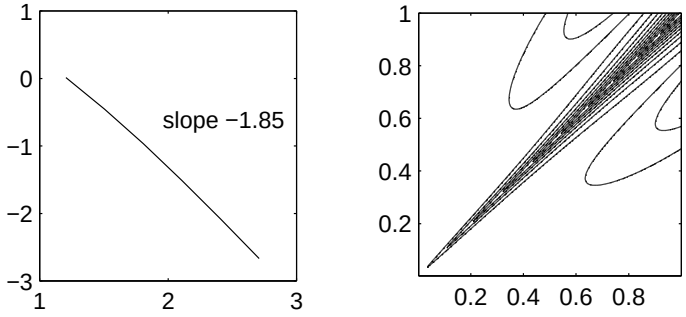

Fig. 8. Double logarithmic plot of the maximal deviation between grid function and first order expansion versus $1/h$ (left) and numerical derivative of the expansion's remainder (right).

equispaced isolines of the numerical derivative of the remainder $(\tilde{v}_{\boldsymbol{i}} - \tilde{u}(\boldsymbol{x}_{\boldsymbol{i}}) - h\tilde{u}_1(\boldsymbol{x}_{\boldsymbol{i}}, 0))/h^2$ which is, in leading order, identical to u_2. The concentration of the isolines along the diagonal is a consequence of the restricted smoothness of $\tilde{u}_1$. In fact, if we expand up to second order, we see that $\tilde{u}_2$ also satisfies an advection equation with source given by $\Delta\tilde{u}_1/2$ which has a jump across the diagonal.

To study the implications of non-differentiability of the coefficients more carefully, let us turn to the example $\hat{g}(x, y) = y - x$ which we consider on a grid with $\alpha = h(p_1, p_2)$ where $p_1, p_2 \in (0, 1]$. Now

$$p(G, \boldsymbol{e}_1, \bar{\boldsymbol{x}}) = p_1, \quad \bar{\boldsymbol{x}} \in \Gamma_w, \qquad p(G, \boldsymbol{e}_2, \bar{\boldsymbol{x}}) = p_2, \quad \bar{\boldsymbol{x}} \in \Gamma_s$$

and since $\Delta\hat{u}_0 = \Delta(y - x) = 0$, equation (48) together with (54) yields

$$\hat{u}_1(\boldsymbol{x}, \boldsymbol{m}_G) = \begin{cases} p_1 - 1 & x < y \\ p_1 - p_2 & x = y \\ 1 - p_2 & x > y \end{cases} \tag{55}$$

The first observation is that, in the particular situation $p_1 = p_2 = 1$, the error term $\hat{u}_1$ vanishes. In fact, one can easily check that the upwind scheme is exact in that case, i.e. $v_{\boldsymbol{i}} = \hat{u}_0(\boldsymbol{x}_{\boldsymbol{i}})$. On a square domain, one would choose $p_1 = p_2 = 1$, of course. But in more general domains, the distance of the grid points to the boundary cannot always be equal to h – not even in axis parallel geometries, as indicated in figure 9. In the more general case $p_1, p_2 \in (0, 1)$, $\hat{u}_1$ is a piecewise constant function with a jump along the diagonal. Since smoothness of the coefficients is required in the derivation, (48) may be a poor description of the scheme's behavior around $x = y$. Indeed, a numerical

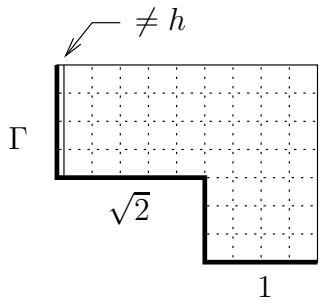

Fig. 9. Even in axis parallel geometries, the distance of grid lines to the boundary cannot always be chosen equal to h.

test shows the well known property of the upwind method to smear jump discontinuities (see figure 10). Nevertheless, the constant states are correctly predicted by (55) and the region with the smooth transition between the states is confined to a region around the diagonal which disappears in the limit $h \to 0$. In other words, the scheme produces a solution with an *interior layer*.

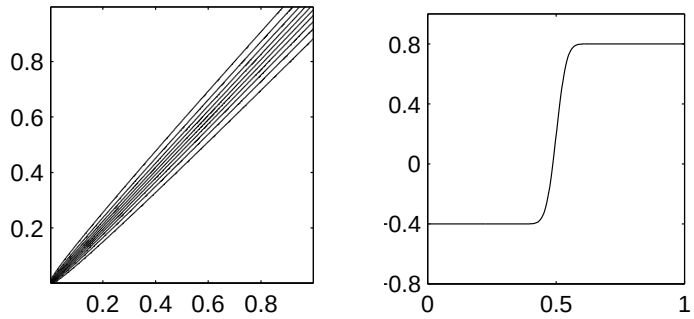

Fig. 10. Ten isolines of the error (left) and scaled error $(\tilde{v}_{\boldsymbol{i}} - \tilde{u}(\boldsymbol{x}_{\boldsymbol{i}}))/h$ of the upwind method with $p_1 = 0.6$ and $p_2 = 0.2$ along the line $(x, 1-x)$ (right).

The classical approach in asymptotic analysis to investigate such layers is a matched asymptotic expansion [24]. The idea is to use the result of a regular *outer* expansion of the form (46) in most of the domain and to employ a so called *inner* expansion around the location of the layer by changing the length scale appropriate to the layer's thickness. Afterwards, the two expansions are matched by comparing them in an intermediate length scale.

In our case, we introduce the coordinate system $t = y + x$, $s = (y - x)/\sqrt{h}$ where t and s are proportional to the distance along and perpendicular to the diagonal. The coordinate transform is summarized with the matrix

$$T_h = \begin{pmatrix} 1 & 1 \\ -h^{-\frac{1}{2}} & h^{-\frac{1}{2}} \end{pmatrix}.$$

We insert the ansatz

$$v_{\boldsymbol{i}} = w_0(T_h \boldsymbol{x}_{\boldsymbol{i}}) + \sqrt{h} w_{1/2}(T_h \boldsymbol{x}_{\boldsymbol{i}}) + h w_1(T_h \boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{m}_G) + \dots \tag{56}$$

into the algorithm (44) and expand around $T_h \boldsymbol{x}_{\boldsymbol{i}}$. For example,

$$w(T_h \boldsymbol{x}_{\boldsymbol{i}-\boldsymbol{e}_k}) = w(T_h \boldsymbol{x}_{\boldsymbol{i}} - h T_h \boldsymbol{e}_k) = w(T_h \boldsymbol{x}_{\boldsymbol{i}}) - h \left[(T_h \boldsymbol{e}_k) \cdot \nabla \right] w(T_h \boldsymbol{x}_{\boldsymbol{i}}) + \dots$$

where the higher order terms involve powers of the differential operators

$$h(T_h \boldsymbol{e}_1) \cdot \nabla = h\partial_t - \sqrt{h}\partial_s, \qquad h(T_h \boldsymbol{e}_2) \cdot \nabla = h\partial_t + \sqrt{h}\partial_s.$$

Given $\bar{\boldsymbol{z}} = (\bar{t}, \bar{s}) \in \mathbb{R}^+ \times \mathbb{R}$ and a general grid sequence $G_n = (h_n, \boldsymbol{\alpha}_n)$ with $h_n \to 0$, we evaluate the expansion at the points with index

$$\boldsymbol{i}_n = \left\lfloor \frac{1}{h_n} T_{h_n}^{-1} \bar{\boldsymbol{z}} - \boldsymbol{\alpha}_n \right\rfloor$$

where the bracket $\lfloor \cdot \rfloor$ is applied component wise. This assures that the scaled coordinates $T_{h_n} \boldsymbol{x}_{\boldsymbol{i}_n}(G_n)$ converge to $\bar{\boldsymbol{z}}$ for $n \to \infty$. In this way, we successively derive the equations

$$\partial_t w_0 = \frac{1}{2}\partial_s^2 w_0, \qquad \partial_t w_{1/2} = \frac{1}{2}\partial_s^2 w_{1/2}, \qquad \text{on } \mathbb{R}^+ \times \mathbb{R} \tag{57}$$

and

$$\partial_t w_1 = \frac{1}{2}\partial_s^2 w_1 + \left(\frac{1}{2}\partial_t^2 - \frac{1}{2}\partial_t \partial_s^2 + \frac{1}{24}\partial_s^4 \right) w_0, \qquad \text{on } \mathbb{R}^+ \times \mathbb{R} \tag{58}$$

which describe the diffusive behavior of the upwind method. The right hand side of equation (58) can be simplified since

$$\frac{1}{2}\partial_t^2 - \frac{1}{2}\partial_t \partial_s^2 + \frac{1}{24}\partial_s^4 = \frac{1}{2}\partial_t \left(\partial_t - \frac{1}{2}\partial_s^2 \right)^2 - \frac{1}{12}\partial_s^4$$

so that

$$\partial_t w_1 = \frac{1}{2}\partial_s^2 w_1 - \frac{1}{12}\partial_s^4 w_0, \qquad \text{on } \mathbb{R}^+ \times \mathbb{R} \tag{59}$$

In order to calculate $w_0, w_{1/2}$, and w_1 we need to determine the values at $t = 0$. Note, however, that $t = 0$ corresponds to the lower left corner of the unit square and that the layer coordinates $(t, s) \in \mathbb{R}^+ \times \mathbb{R}$ are not appropriate close to that point because the s-coordinate is clearly restricted. We therefore investigate the behavior of the scheme in the lower left corner with the scaling $\bar{\boldsymbol{x}} = \boldsymbol{x}/\sqrt{h}$ and the expansion

$$v_{\boldsymbol{i}} = c_0 \left(\frac{\boldsymbol{x}_{\boldsymbol{i}}}{\sqrt{h}} \right) + \sqrt{h} c_{1/2} \left(\frac{\boldsymbol{x}_{\boldsymbol{i}}}{\sqrt{h}} \right) + h c_1 \left(\frac{\boldsymbol{x}_{\boldsymbol{i}}}{\sqrt{h}}, \boldsymbol{m}_G \right) + \dots . \tag{60}$$

In figure 11, the different expansion regions in the computational domain are indicated: region I is dominated by the regular outer expansion, region II is described by the inner expansion, and region III is required to find initial values for the diffusion equations (57), (59). The plan is to determine the coefficients c_* and then match (60) and (56) with the assumption that (60)

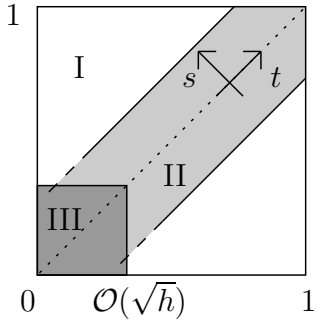

Fig. 11. Different domains of expansion

for large values of $\bar{t} = (x+y)/\sqrt{h}$ should behave like (56) with small values for t.

Since x and y coordinates in (60) are both scaled with the same factor, the expansion parallels the outer expansion (46). The only difference is that, in the new coordinates $\bar{\boldsymbol{x}} = \boldsymbol{x}/\sqrt{h}$, the grid spacing is $\bar{h} = \sqrt{h}$ and because of the focused view, the boundary values look like $g(0) + \sqrt{h}\bar{y}\partial_y g(0) + h\bar{y}^2\partial_y^2 g(0)/2 + \ldots$ on the west boundary and $g(0) + \sqrt{h}\bar{x}\partial_x g(0) + h\bar{x}^2\partial_x^2 g(0)/2 + \ldots$ on the south boundary. Specifically, we find on the quarter plane $Q = \mathbb{R}^+ \times \mathbb{R}^+$

$$\partial_{\bar{x}} c_0 + \partial_{\bar{y}} c_0 = 0 \quad \text{in } Q, \qquad c_0 = g(0) \quad \text{on } \partial Q$$

so that $c_0(\bar{\boldsymbol{x}}) = g(0)$. In order $\sqrt{h}$, we obtain

$$\partial_{\bar{x}} c_{1/2} + \partial_{\bar{y}} c_{1/2} = 0 \quad \text{in } Q, \qquad c_{1/2}(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}} \cdot \nabla g(0) \quad \text{on } \partial Q \tag{61}$$

and, in order h,

$$\begin{aligned} \partial_{\bar{x}} c_1 + \partial_{\bar{y}} c_1 &= \frac{1}{2}\Delta c_{1/2} \quad \text{in } Q, \\ c_1(\bar{\boldsymbol{x}}, \boldsymbol{m}_G) &= \frac{1}{2}(\bar{x}^2\partial_{\bar{x}}^2 + \bar{y}^2\partial_{\bar{y}}^2) g(0) + \boldsymbol{m}_G(\bar{\boldsymbol{x}}) \cdot \nabla c_{1/2}(\bar{\boldsymbol{x}}) \quad \text{on } \partial Q \end{aligned} \tag{62}$$

As in the outer expansion, we face the problem that $c_{1/2}$ and c_1 are, in general, not smooth along the diagonal. This leads again to an interior layer which has a typical width $\bar{h}^{1/2} = h^{1/4}$ in the corner coordinates, respectively $h^{3/4}$ in our base coordinates. This narrowing of the layer from $h^{1/2}$ down to $h^{3/4}$ can be seen in figure 10. In fact, a closer look into the corner with the scaling $\boldsymbol{x}/h^\alpha$, $1/2 < \alpha < 1$ shows an even thinner layer of typical size $h^{(1+\alpha)/2}$. An inner expansion around the diagonal reveals that the grid function is again described by diffusion equations but we can skip the details because our aim is only to determine the initial values for the coefficients w_* in region II. To this end, we evaluate w_* at $(t,s) = (h^\alpha, s)$ with $\alpha < 1/2$ and c_* at the corresponding coordinate $(\bar{t}, \bar{s}) = (h^{\alpha-1/2}, s)$ where we set $\bar{t} = \bar{y} + \bar{x}$, $\bar{s} = \bar{y} - \bar{x}$. Being governed by diffusion equations, we find that the interior layer has a typical width of $h^{1/4}\sqrt{\bar{t}}$ (in the corner coordinates) so that, as $\bar{t} = h^{\alpha-1/2}$ increases

with $h \to 0$, the thickness of the layer actually shrinks like $h^{\alpha/2}$. Given some $s \neq 0$, we therefore conclude that $(h^{\alpha-1/2}, s)$, for h small enough, corresponds to a point in the quarter plane outside the interior layer. For our purpose, we can therefore neglect a detailed investigation of the interior layer in the corner coordinates. For $\bar{s} \neq 0$ and $\bar{t} > \bar{s}$, the coefficients are given by

$$c_0(\bar{\boldsymbol{x}}) = g(0), \qquad c_{1/2}(\bar{\boldsymbol{x}}) = c_{1/2}^{\infty}(\bar{s}) = \begin{cases} \bar{s}\partial_y g(0) & \bar{s} > 0 \\ -\bar{s}\partial_x g(0) & \bar{s} < 0 \end{cases}$$

and with a grid $G = (h, (hp_1, hp_2))$

$$c_1(\bar{\boldsymbol{x}}) = c_1^{\infty}(\bar{s}) = \begin{cases} (p_1 - 1)\partial_y g(0) + \frac{1}{2}\bar{s}^2\partial_y^2 g(0) & \bar{s} > 0 \\ (1 - p_2)\partial_x g(0) + \frac{1}{2}\bar{s}^2\partial_x^2 g(0) & \bar{s} < 0 \end{cases}$$

Since $c_0, c_{1/2}$, and c_1 are $\bar{t}$-independent, the asymptotic coincidence on the intermediate scale

$$\begin{aligned} &w_0(h^\alpha, s) + \sqrt{h} w_{1/2}(h^\alpha, s) + h w_1(h^\alpha, s, \boldsymbol{m}_G) \\ &\quad = c_0(h^{\alpha-1/2}, s) + \sqrt{h} c_{1/2}(h^{\alpha-1/2}, s) + h c_1(h^{\alpha-1/2}, s, \boldsymbol{m}_G) + \mathcal{O}(h^{3/2}) \end{aligned}$$

implies

$$w_0(0, s) = g(0), \qquad w_{1/2}(0, s) = c_{1/2}^{\infty}(s), \qquad w_1(0, s) = c_1^{\infty}(s)$$

For our example with the boundary value $\hat{g}(x, y) = y - x$, this eventually leads to

$$\hat{w}_0(t, s) = 0, \quad \hat{w}_{1/2}(t, s) = s, \quad \hat{w}_1(t, s) = 1 - p_2 + \frac{p_1 + p_2 - 2}{2}\left(1 + \operatorname{erf}\left(\frac{s}{\sqrt{2t}}\right)\right)$$

and in the case $\check{g}(x, y) = \sin(x)$, we find

$$\check{w}_0(t, s) = 0, \qquad \check{w}_{1/2}(t, s) = \frac{s}{2}\left(\operatorname{erf}\left(\frac{s}{\sqrt{2t}}\right) - 1\right) + \sqrt{\frac{t}{2\pi}}\exp\left(-\frac{s^2}{2t}\right).$$

Graphical representations of the expansions are given in figure 12. A combined expansion in region I and II is obtained by adding outer and inner expansion and subtracting the (wrong) behavior of the outer expansion inside the layer (see [24]). More precisely, the terms to be subtracted are obtained by evaluating the outer expansion at points $T_h^{-1}(t, s) = (t - \sqrt{h}s, t + \sqrt{h}s)/2$ close to the diagonal and expanding in $\sqrt{h}$. Note that $T_h^{-1}(t, s)$ approaches the diagonal point $\boldsymbol{x} = (t, t)/2$ from above if $s > 0$ and from below in the case $s < 0$. Since u_0 is known explicitly and is smooth away from the diagonal, we can simplify the resulting expressions and eventually find for the case $\hat{g}(x, y) = y - x$

$$\hat{v}_{\boldsymbol{i}} = \hat{u}_0(\boldsymbol{x}_{\boldsymbol{i}}) + h\hat{w}_1(T_h \boldsymbol{x}_{\boldsymbol{i}}) + \dots \tag{63}$$

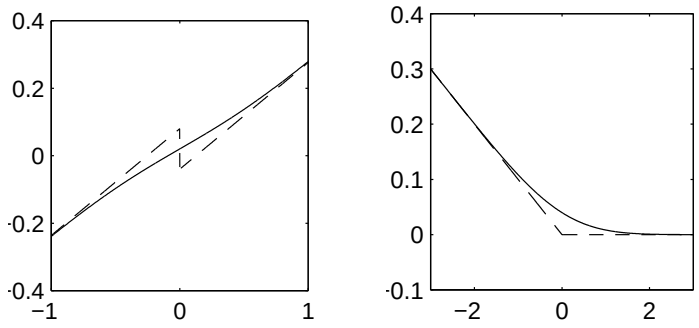

Fig. 12. Typical structure of the interior layer for the boundary value $\hat{g}(x,y) = y - x$ (left) and $\check{g}(x,y) = \sin(x)$ (right). The expansions are shown for fixed t and s varying along the horizontal axis. The dashed line represents the outer expansion.

and for $\check{g}(x,y) = \sin(x)$

$$\check{v}_{\boldsymbol{i}} = \check{u}_0(\boldsymbol{x}_{\boldsymbol{i}}) + \sqrt{h}(\check{w}_1(T_h\boldsymbol{x}_{\boldsymbol{i}}) - \check{c}^{\infty}_{1/2}(T_h\boldsymbol{x}_{\boldsymbol{i}})) + \dots. \tag{64}$$

Note that the matched expansions will not be good approximations in the lower left corner (region III in figure 11) because we have neglected the narrowing of the interior layer in that region. In fact, if we compare left and right hand side of (63), we find a big discrepancy in the corner (see figure 13). However, along the diagonal from the upper left to the lower right corner which is sufficiently far from region III, the prediction is quite good. In the right part of figure 13 we can see that the difference between $\hat{v}_{\boldsymbol{i}}$ and the truncated expansion $\hat{u}_0(\boldsymbol{x}_{\boldsymbol{i}}) + h\hat{w}_1(T_h\boldsymbol{x}_{\boldsymbol{i}})$ is of order $h^{3/2}$ which is exactly the truncation order of our inner expansion. Similarly, we find that our expansion (64) coincides with the numerical result in the considered order: the difference $|\check{v}_{\boldsymbol{i}} - \check{u}_0(\boldsymbol{x}_{\boldsymbol{i}})|$ converges essentially like $\sqrt{h}$ (see figure 14) and the correction of $\check{u}_0(\boldsymbol{x}_{\boldsymbol{i}})$ given by the right hand side of (64) differs from $\check{v}_{\boldsymbol{i}}$ only in order h (figure 14) which is the truncation order in the outer expansion.

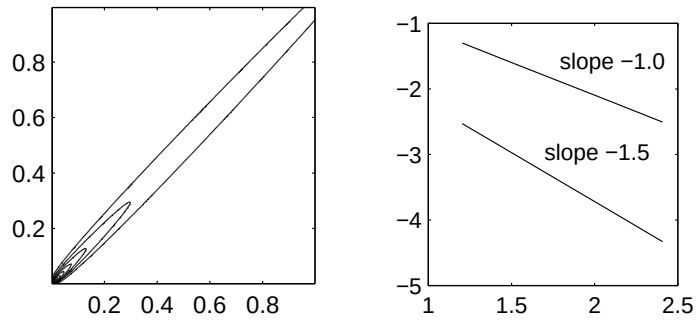

Fig. 13. Left: the difference between expansion and upwind result for $\hat{g}$ is maximal in the lower left corner. Right: along the diagonal $(x, 1-x)$, the numerical convergence order of the upwind method is 1.0 and the difference between numerical result and expansion (63) is of order 1.5.

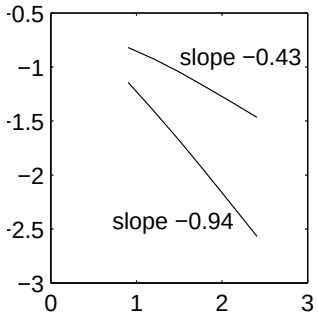

Fig. 14. Maximal error versus $1/h$: the difference between $\check{v}_{\boldsymbol{i}}$ and the exact solution $\check{u}_0(\boldsymbol{x}_{\boldsymbol{i}})$ is of order 0.43 and left and right hand side of (64) differ by a term of order 0.94.

2.7 General geometries

In the previous examples, we have seen that the expansion coefficients u_k can become grid dependent. The order from which on that happens depends on the structure of the boundary algorithm but the reason is always a change of the scheme's length scale: inside the computational domain, the relevant points have fixed grid distance h but at the boundary, the terminal grid points $\boldsymbol{x}_{\boldsymbol{i}}$ can have a distance $d_{\Gamma,\boldsymbol{e}}(\boldsymbol{x}_{\boldsymbol{i}}) < h$ to the boundary. In axis parallel geometries like the one in figure 9, the distance $d_{\Gamma,\boldsymbol{e}}$ is constant along each edge and gives rise to piecewise constant boundary values for the error terms (symbolized by $\boldsymbol{m}_G$ in our example). However, if we go over from axis parallel to more general domains, we observe a drastic change in this behavior. To illustrate the problem, let us consider again the hyperbolic equation (42) with the same upwind discretization but on a domain Ω where the west boundary Γ_w has a general angle $0 < \phi < \pi/2$ with the vertical axis (see figure 15). With the same

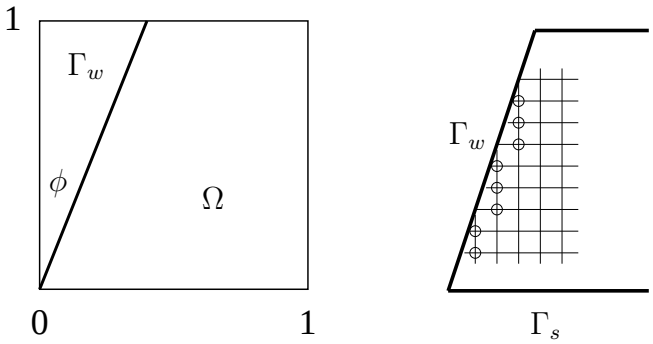

Fig. 15. Domain with non-axis parallel boundaries (left) and details of the regular grid close to Γ_w (right). Terminal grid points are marked with a circle.

analysis as in section 2.6, we also arrive at equation (53) which determines the boundary value for u_1. Using a grid $G = (h, \boldsymbol{\alpha})$, $\alpha = (h, h)$, the boundary

nodes $\bar{\boldsymbol{x}}_{\boldsymbol{i}} = \Pi_{\Gamma,-\boldsymbol{e}_1}(\boldsymbol{x}_{\boldsymbol{i}})$ with $\boldsymbol{i} \in I_w$ are from the set

$$\{\boldsymbol{z}_j = j\,h\,(\tan\phi, 1) : 1 \le j \le \lfloor 1/h \rfloor\}$$

and have a relative distance $h/\cos\phi$. According to (51), the function p attains values

$$p(G, \boldsymbol{e}_1, \boldsymbol{z}_j) = 1 - (j\tan\phi - \lfloor j\tan\phi \rfloor).$$

In figure 16, the values of p are plotted at the boundary nodes on the west boundary Γ_w for different values of ϕ. The typical increment between two

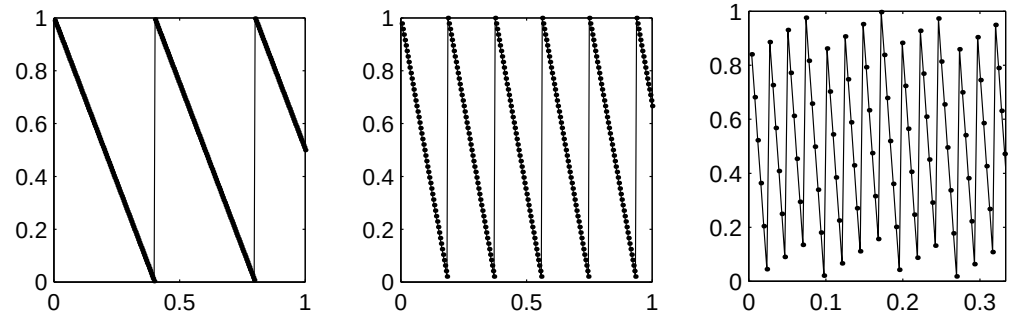

Fig. 16. Function $p(G, \boldsymbol{e}_1, \boldsymbol{z}_j)$ along Γ_w for different values of the angle ϕ. Left: $\tan\phi = 5h/2$, middle: $\tan\phi = \sqrt{h}/3$, right: $\tan\phi = 1/(2\pi)$ with only one third of the boundary shown. The grid spacing is $h = 2^{-8}$.

function values $p(G, \boldsymbol{e}_1, \boldsymbol{z}_j)$ and $p(G, \boldsymbol{e}_1, \boldsymbol{z}_{j+1})$ is $\tan\phi$ unless $j\tan\phi < n \le (j+1)\tan\phi$ for some $n \in \mathbb{N}$ in which case the jump can be even larger. Hence, the boundary value has a typical gradient of $\tan\phi/(h/\cos\phi) = \sin\phi/h$ with possible stronger jumps in approximate distance $h/\sin\phi$. If $\tan\phi$ is less than a small multiple of h, we conclude that the boundary value for u_1 is piecewise smooth (slope of order one) with a few jumps along the boundary. These jumps in the boundary data give rise to inner layers which can be predicted with an asymptotic analysis similar to our example in section 2.6. For the boundary data $\hat{g}(x, y) = y - x$, the difference between the upwind solution $\hat{v}_{\boldsymbol{i}}$ and the regular expansion $\hat{u}_0 + h\hat{u}_1$ is shown in figure 17 where $\hat{u}_0(x, y) = y - x$ and $\hat{u}_1$ is determined from (48) and (54). Since the typical width of the inner layer is $\mathcal{O}(\sqrt{h})$, a similar analysis would even be possible if $\tan\phi$ reaches a small fraction of $\sqrt{h}$ because the inner layers are not interacting in that case (middle picture in figure 17). However, the slope of the boundary data is already of size $1/\sqrt{h}$ so that the basic assumption for the regular outer expansion $\nabla u_1 = \mathcal{O}(1)$ is violated. Hence, a modified expansion which takes care of the strong slope in u_1 is appropriate, for example $v_{\boldsymbol{i}} = u_0(\boldsymbol{x}_{\boldsymbol{i}}) + hu_1(\boldsymbol{x}_{\boldsymbol{i}}/\sqrt{h}) + h^{3/2}u_2(\boldsymbol{x}_{\boldsymbol{i}}/\sqrt{h}) + \cdots$. Finally, for large angles $\tan\phi = \mathcal{O}(1)$, the period of the boundary data is of order h with a slope of order $1/h$ so that the relevant scale is h itself. Since on this scale the grid points never become dense (they remain at distance one), there is no hope for a continuous description in terms of partial differential equations. In our example, it turns out that the predicted coefficient $\hat{u}_1$ is useless because the strongly oscillating boundary values are averaged out in

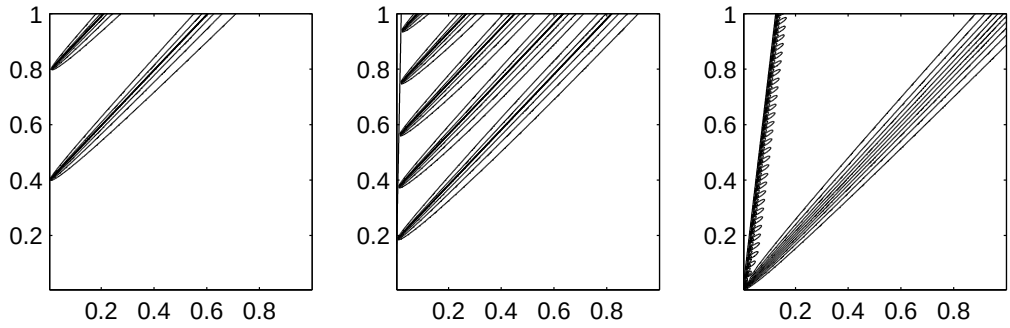

Fig. 17. Difference between numerical solution $\hat{v}_{\boldsymbol{i}}$ and outer expansion $\hat{u}_0 + h\hat{u}_1$ for several boundary angles ϕ. Left: $\tan\phi = 5h/2$, middle: $\tan\phi = \sqrt{h}/3$, right: $\tan\phi = 1/8$ only $\hat{v}_{\boldsymbol{i}} - \hat{u}_0$ is shown because $\hat{u}_1$ is meaningless. The grid spacing is $h = 2^{-8}$.

a thin boundary layer of width $\mathcal{O}(h)$ and a constant instead of an oscillating error is transported into the domain (see figure 17).

Since the reason for the oscillation is obviously related to the boundary condition for u_1, we can apply a direct correction step to move the oscillatory behavior to higher orders. We modify the original scheme by replacing g in (45) with $g - h\boldsymbol{m}_G \cdot \nabla u_0$. Of course, ∇u_0 is generally unknown and we approximate it with finite differences. For example, on the west boundary where $\boldsymbol{m}_G \cdot \nabla u_0 = (1-p)\partial_x u_0$, we use

$$g_h(\boldsymbol{x}) = g(\boldsymbol{x}) - h(1-p)\frac{u(\boldsymbol{x} + (p+1)h\boldsymbol{e}_1) - u(\boldsymbol{x} + ph\boldsymbol{e}_1)}{h}, \quad p = p(G, \boldsymbol{e}_1, \boldsymbol{x})$$

as new boundary value (note that $\boldsymbol{x}+(p+1)h\boldsymbol{e}_1$ and $\boldsymbol{x}+ph\boldsymbol{e}_1$ are grid points). With this modified scheme, we obtain the exact solution of the example with $\hat{g}(x,y) = y - x$. In the case $\tilde{g}(x,y) = \sin(2\pi(y-x))$, the boundary error is successfully removed as shown in figure 18.

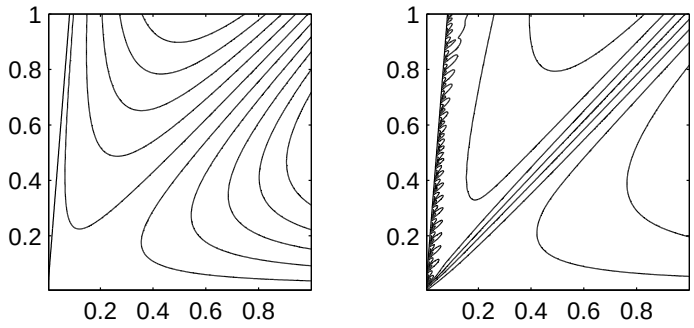

Fig. 18. Numerical error of the upwind scheme with (left) and without direct correction (right). Ten isolines of the error are shown for the case $\tan\phi = 1/12$ and $h = 2^{-8}$.

3 Conclusion

We have shown that several tools of asymptotic analysis can successfully be applied to finite difference schemes: they predict and analyze the behavior very precisely. The gained knowledge can be used to improve numerical methods for example with the direct correction technique. In future contributions, we will apply the method to lattice Boltzmann schemes for which even consistency is often an unresolved problem due to a lack of analytical tools. With the generally applicable approach presented in this paper, this gap has been closed.

Acknowledgement

This work has been supported by the *Deutsche Forschungsgemeinschaft* DFG in the project *Analyse von Lattice Boltzmann Methoden Ju440/1-1.*

References

[1] S. Chen, G. D. Doolen, Lattice Boltzmann method for fluid flows, Annu. Rev. Fluid Mech. 30 (1998) 329–364.

[2] G. Marchuk, V. Shaidurov, Difference methods and their extrapolations, Springer, 1983.

[3] O. Widlund, Some recent applications of asymptotic error expansions to finite-difference schemes, Proc. R. Soc. Lond. Ser. A 323 (1971) 167–177.

[4] L. Richardson, The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam, Phil. Trans, A 210 (1910) 307–357.

[5] L. Richardson, The deferred approach to the limit, Phil. Trans, A 226 (1927) 299–349.

[6] V. Pereyra, On improving an approximate solution of a functional equation by deferred corrections, Numer. Math. 8 (1966) 376–391.

[7] V. Pereyra, Iterated deferred corrections for nonlinear boundary value problems, Numer. Math. 11 (1968) 111–125.

[8] E. Hairer, S. P. Nørsett, G. Wanner, Solving ordinary differential equations. I: Nonstiff problems. 2. rev. ed., Springer, 1993.

[9] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration, Springer, 2002.

[10] K. Böhmer, Asymptotic expansions for the discretization error in poisson's equation on general domains, in: W. Schempp (Ed.), Multivariate approximation theory, Proc. Conf. math. Res. Inst., Oberwolfach 1979, Vol. 51 of ISNM, 1979, pp. 30–45.

[11] A. Chorin, On the convergence of discrete approximations to the navier-stokes equations, Math. Comput. 23 (1969) 341–353.

[12] G. Strang, Accurate partial difference methods. II: Non-linear problems, Numer. Math. 6 (1964) 37–46.

[13] F. Villatoro, J. Ramos, On the method of modified equations. I: Asymptotic analysis of the Euler forward difference method, Appl. Math. Comput. 103 (1999) 111–139.

[14] F. Villatoro, J. Ramos, On the method of modified equations. II: Numerical techniques based on the equivalent equation for the Euler forward difference method, Appl. Math. Comput. 103 (1999) 141–177.

[15] F. Villatoro, J. Ramos, On the method of modified equations. III: Numerical techniques based on the second equivalent equation for the Euler forward difference method, Appl. Math. Comput. 103 (1999) 179–212.

[16] F. Villatoro, J. Ramos, On the method of modified equations. IV: Numerical techniques based on the modified equation for the Euler forward difference method, Appl. Math. Comput. 103 (1999) 213–240.

[17] F. Villatoro, J. Ramos, On the method of modified equations. V: Asymptotics analysis of direct-correction and asymptotic successive-correction techniques for the implicit midpoint method, Appl. Math. Comput. 103 (1999) 241–285.

[18] F. Villatoro, J. Ramos, On the method of modified equations. VI: Asymptotic analysis of and asymptotic successive-corrections techniques for two-point, boundary-value problems in ODE's, Appl. Math. Comput. 105 (1999) 137–171.

[19] F. Hoppensteadt, W. Miranker, Multitime methods for systems of difference equations, Studies Appl. Math. 56 (1977) 273–289.

[20] R. Warming, B. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods, J. Comput. Phys. 14 (1974) 159–179.

[21] D. Griffiths, J. Sanz-Serna, On the scope of the method of modified equations, SIAM J. Sci. Stat. Comput. 7 (1986) 994–1008.

[22] S.-C. Chang, A critical analysis of the modified equation technique of warming and hyett, J. Comput. Phys. 86 (1990) 107–126.

[23] P. Deuflhard, F. Bornemann, Scientific computing with ordinary differential equations, Springer, 2002.

[24] M. H. Holmes, Introduction to perturbation methods, Springer, 1995.

[25] A. Samarsky, Theorie der Differenzenverfahren, Akademische Verlagsgesellschaft Leipzig, 1984.

[26] G. Shortley, R. Weller, The numerical solution of Laplace's equation, J. Appl. Phys. 9 (1938) 334–344.