# Linear schemes with several degrees of freedom for the transport equation and the long-time simulation accuracy

P.A. Bakhvalov[a,*], M.D. Surnachev[a]

[a] *Keldysh Institute of Applied Mathematics, 4, Miusskaya Sq., Moscow, 125047, Russia*

**Abstract**

We consider linear schemes with several degrees of freedom for the transport equation on uniform meshes. For these schemes the solution error is $O(h^p + th^q)$, where $p$ is equal to or greater by one than the order of the truncation error and $q \geqslant p$. We prove the existence of a mapping of smooth functions on the mesh space providing the $q$-th order of the truncation error and deviating from the standard mapping ($L_2$-projection for example) by $O(h^p)$. In 1D case this mapping can be found in the class of local mappings. In more dimensions the existence of a local mapping with such properties is guaranteed only under additional assumptions.

*Keywords:* consistency and accuracy, superconvergence, long-time simulation accuracy, Fourier analysis

## 1. Introduction

The simplest approach to establish an estimate for the solution error of a numerical method is the analysis of the truncation error. A stable scheme possessing the truncation error of order $P_A$ gives the solution with the error of the same order. But this estimate can be very far from optimal. For instance, the DG($k$) method (discontinuous Galerkin method based on the polynomials of order $k$) on the uniform meshes has the truncation error of order $\max\{k, 1\}$, while the numerical solution of the Cauchy problem for the model transport equation $\partial v/\partial t + \partial v/\partial x = 0$ obtained by this scheme possesses an error estimate of the form

$$\|\varepsilon_h(t)\|_2 \leqslant C_1 h^{k+1} + C_2 h^{2k+1} t. \tag{1.1}$$

The comparison of the solution error by the 5-th order finite-difference scheme and the DG(4) method of the same order of accuracy is presented in Fig. 1. The estimate (1.1) follows directly from the results of [1] based on negative norm estimates which essentially use the finite-element nature of the DG scheme. We will consider another two methods that can be used to obtain such estimates.

The spectral analysis is a classical approach to study linear difference schemes (both semi-discrete and fully discrete) with constant coefficients on uniform meshes under periodic boundary conditions. In the case of 1 DOF per cell a wide range of its applications can be found in [2]. This method was successfully applied also for the schemes with several DOFs. For the DG method it was used in [3, 4, 5, 6], for the spectral difference method in [7, 8], for the flux reconstruction method (the class including the first two) in [9, 10], etc. Although the main application of the spectral analysis is to ensure stability, it can be applied for the accuracy analysis also. Lowrie [3] used it to understand the enhanced accuracy of DG($k$) for $k = 1, 2, 3$, but details were not presented. In [5] the estimate (1.1) was proved for DG(1). In [6] this was proved for DG(2) and DG(3) using symbolic computations. To find the spectrum, one needs to find the roots of the polynomial of order $m = k + 1$ with the coefficients depending on the wave number. If $m > 4$, it is not generally solvable by radicals. Thus the authors of [6] could not extend their analysis to $k > 3$. Note also the papers [11, 12] where the spectral analysis was combined with the finite-element technique to prove $\|\varepsilon_h(t)\|_2 \leqslant C_1 h^{k+1} + C_2 h^{k+3/2} t$ for DG($k$), $k \geqslant 1$, on non-uniform meshes.

*Corresponding author

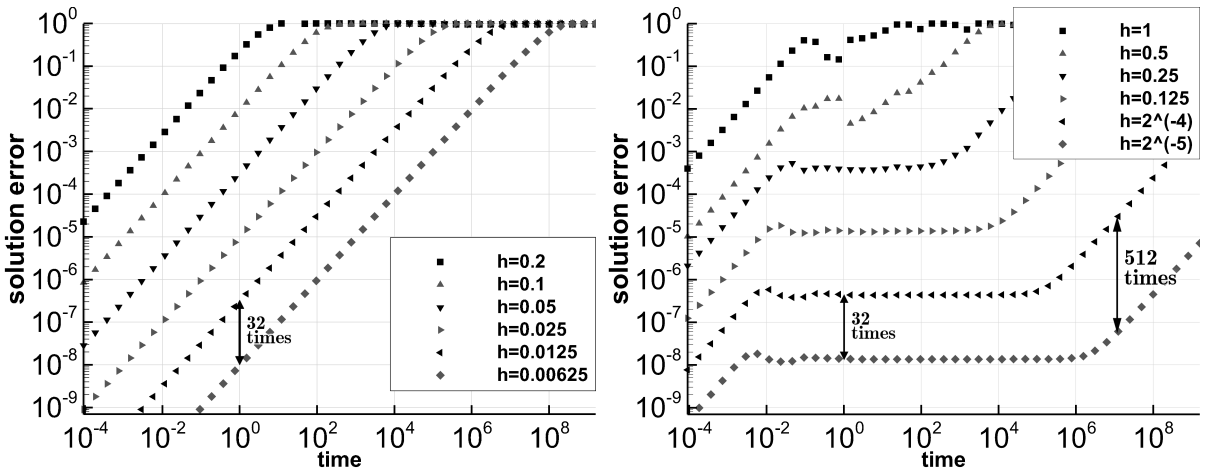



Figure 1: The $L_2$ norm of the solution error of $\partial v/\partial t + \partial v/\partial x = 0$, $v_0 = \sin(2\pi x)$ obtained by the 5th order finite-difference scheme (left) and the DG(4) method (right)

The error estimate (1.1) for DG($k$), $k \in \mathbb{N}$, was obtained by another approach which we call the method of auxiliary mapping. Let $\Pi_h$ be an operator prescribing the initial data for the discrete (or semidiscrete) problem, i. e. taking each $f$ to a mesh function $f_h$. In the case of DG($k$), a mesh function is a piecewise polynomial of order $k$, and the operator $\Pi_h$ is the $L_2$-projection onto the space of mesh functions. The idea of the method is to introduce a new mapping $\tilde{\Pi}_h$ such that for a sufficiently smooth function $f$ there holds

(A1) $\|\tilde{\Pi}_h f - \Pi_h f\|_2 \leqslant C_0 h^P$;
(A2) the scheme possesses the truncation error of order $Q \geqslant P$ in the sense of $\tilde{\Pi}_h$.

By stability these properties guarantee (see Proposition 4.2) that in the sense of $\Pi_h$ the solution error possesses an estimate of the form

$$\|\varepsilon_h(t)\|_2 \leqslant C_1 h^P + C_2 h^Q t, \quad Q \geqslant P > 0. \tag{1.2}$$

For the standard Galerkin method for elliptic and parabolic problems, the auxiliary mapping is usually chosen as the Ritz projection.

The first application of the method of auxiliary mapping to the DG scheme for the transport equation is also due to Lowrie for $k = 1$ (see formulas (5.81) and (5.82) in [3]). In [13] the estimate (1.2) was proved for an arbitrary $k$ and $P = Q = k+1$ using the Gauss–Radau projection. In [14] the estimate $\|\varepsilon_h(t)\|_2 \leqslant C_1 h^{k+1} + C_2 h^{k+2} t^2$ was proved. Finally, the optimal result (1.1) was obtained in [15]. On a uniform mesh the auxiliary mapping $\Pi_h^{(P,Q)}$ introduced in [15] takes each $f \in W_{2,loc}^{Q+1}(\mathbb{R})$ to the mesh function $\Pi_h^{(P,Q)} f$ defined on each cell $\eta \in \mathbb{Z}$ by

$$\left(\Pi_h^{(P,Q)} f\right)_\eta = (\Pi_h f)_\eta + \sum_{m=P}^{Q} h^m \left(\mathcal{P}_h^{(m)} \left(\frac{d^m f}{dx^m}\right)\right)_\eta, \tag{1.3}$$

where $(\Pi_h f)_\eta$ and $(\Pi_h^{(P,Q)} f)_\eta$ are the polynomials at cell $\eta$, and $\mathcal{P}_h^{(m)}$ are some linear mappings to the mesh space. The difference between the modified mapping and the original mapping (the last term in (1.3)) is called a correction function (in [15]) or a corrector (in [16]). Note that estimate (1.1) is valid for arbitrary non-uniform meshes (assuming $h = h_{\max}$), but this is a specific feature of the DG method. Further results in the accuracy of the DG method were obtained in [17] (2D), [18, 19, 20] (variable coefficients), [21] (energy-conserving DG), Runge –Kutta time integration [22] etc.

Estimates of the form (1.2) with some $P$ and $Q$, where $Q$ is greater than the order of the truncation error, are valid not only for the DG method but also for some other schemes with several degrees of freedom per cell. A scheme on

a periodic mesh shows the same behavior if the mesh is refined by scaling (see Fig. 3). In this case we assume that if the mesh is scaled by a factor then $h$ is multiplied by the same factor.

Consider an $L_2$-stable scheme with a local mapping $\Pi_h$ (for the definition of "local" see Section 2.5). The existence of a mapping $\tilde{\Pi}_h$ satisfying (A1) and (A2) for some $P$ and $Q$ proves the estimate (1.2). We are concerned with the following related problems.

1. How to get the optimal values of $P$ and $Q$ in the estimate (1.2)?
2. Does a mapping $\tilde{\Pi}_h$ satisfying (A1) and (A2) for these $P$ and $Q$ exist?
3. How to construct it?

The main results of this paper are presented in three theorems. Theorem 1 gives the optimal value of $P$. Theorem 2 claims that the auxiliary mapping $\tilde{\Pi}_h$ does exist. Theorem 3 states that in 1D and some multidimensional cases it can be found in a form similar to (1.3), that is the original mapping is modified using a local corrector. This leads to an algorithm which reduces the problem of finding the optimal values of $P$ and $Q$ in the estimate (1.2) to solving a linear system. However, in the general multidimensional case the situation is more tricky: generally the required corrector is nonlocal and the optimal value of $Q$ can be non-integer.

The rest of the paper is organized as follows. In *Section 2* we present a mathematical formulation of the problem. In *Section 3* we state the main results. In *Section 4* we recall the Lax – Ryabenkii theorem and present the basics of the method of auxiliary mapping. *Section 5* is auxiliary and contains technical results. In *Section 6* we use the spectral analysis to reduce our problem to a finite-dimensional formulation. In *Section 7* we consider 1D case. While the results of this section can be treated as partial cases of the results from the next section, the 1D structure allows us to use simpler methods and understand the phenomenon of the enhanced accuracy in the long-time simulation more clearly. In *Section 8* for a scheme possessing an estimate of the form (1.2) we prove the existence of an auxiliary mapping such that (A1) and (A2) hold. In *Section 9* we consider two cases when this mapping is local. In *Section 10* for these cases we present algorithms to get the optimal values $P$ and $Q$ in the estimate (1.2). In *Section 11* we demonstrate our method on 1D examples. In *Sections 12 and 13* we construct counterexamples for the 2D transport equation. In particular, we show that a local mapping $\Pi_h^{(P,Q)}$ such that (A1) and (A2) hold generally does not exist. In Conclusion we summarize primary results of the paper.

## 2. Problem formulation

### *2.1. Solution spaces*

Let $d \in \mathbb{N}$ be the space dimension. Let $\boldsymbol{a}_j$, $j = 1, \dots, d$, form a basis in $\mathbb{R}^d$. We consider $h\boldsymbol{a}_j$ as lattice vectors of the mesh. Let $T$ be the map of $\mathbb{R}^d$ to $\mathbb{R}^d$ that takes each $\boldsymbol{\eta}$ to

$$T\boldsymbol{\eta} = \sum_{j=1}^{d} \eta_j \boldsymbol{a}_j. \tag{2.1}$$

If $\boldsymbol{\eta} \in \mathbb{Z}^d$ then $T\boldsymbol{\eta}$ is the offset of the block $\boldsymbol{\eta}$ from the zero block. We have $(T^*\boldsymbol{x})_j = \boldsymbol{x} \cdot \boldsymbol{a}_j$. If $\boldsymbol{a}_j$ coincide with the vectors of the standard basis, $T$ is the identity operator.

We say that $f \in L_{2,loc}(\mathbb{R}^d)$ is periodic with *period* $N_0 \in \mathbb{N}$ if for all $j = 1, \dots, d$ there holds $f(\boldsymbol{r} + N_0\boldsymbol{a}_j) = f(\boldsymbol{r})$ for almost all $\boldsymbol{r} \in \mathbb{R}^d$. Unless specifically stated, we will consider complex-valued functions. Denote by $L_{2,per}(\mathbb{R}^d)$ the linear space of periodic $f \in L_{2,loc}(\mathbb{R}^d)$ equipped with the norm

$$\|f\|^2 = \frac{1}{|\square|} \int\limits_{\square} |f(\boldsymbol{r})|^2 dV,$$

where $\square$ is the parallelepiped generated by the vectors $N_0\boldsymbol{a}_1, \dots, N_0\boldsymbol{a}_d$ and $N_0$ is a period of $f$.

Denote $H^q_{per}(\mathbb{R}^d) = L_{2,per}(\mathbb{R}^d) \cap W^q_{2,loc}(\mathbb{R}^d)$ for $q \geqslant 0$ and $C^q_{per}(\mathbb{R}^d) = L_{2,per}(\mathbb{R}^d) \cap C^q(\mathbb{R}^d)$ for $q \in \mathbb{N} \cup \{0\}$. For $w \in H^q_{per}(\mathbb{R}^d)$ (or $w \in C^q_{per}(\mathbb{R}^d)$), $q \in \mathbb{N} \cup \{0\}$, $r = 0, \dots, q$, denote

$$\|\nabla^r w\|^2 = \sum_{|\boldsymbol{m}|=r} \frac{r!}{\boldsymbol{m}!} \|D^{\boldsymbol{m}} w\|^2, \quad \|\nabla^r w\|^2_\infty = \sum_{|\boldsymbol{m}|=r} \frac{r!}{\boldsymbol{m}!} \max_{\mathbb{R}^d} |D^{\boldsymbol{m}} w|^2, \quad D^{\boldsymbol{m}} = \frac{\partial^{|\boldsymbol{m}|}}{\partial x_1^{m_1} \dots \partial x_d^{m_d}}. \tag{2.2}$$

Here $\boldsymbol{m} = (m_1, \ldots, m_d)$ is a multiindex: $m_j \in \mathbb{N} \cup \{0\}$, $|\boldsymbol{m}| = m_1 + \ldots + m_d$, $\boldsymbol{m}! = m_1! \ldots m_d!$. Obviously, $\|\nabla^r f\| \leqslant \|\nabla^r f\|_\infty$. Denote $\boldsymbol{r}^m = x_1^{m_1} \ldots x_d^{m_d}$ for $\boldsymbol{r} = (x_1, \ldots, x_d)$. The notation $\boldsymbol{l} \leqslant \boldsymbol{m}$ means that for each $j = 1, \ldots, d$ there holds $l_j \leqslant m_j$.

Each $w \in L_{2,per}(\mathbb{R}^d)$ with a period $N_0$ can be represented by the Fourier series

$$w = \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \tilde{w}_{\boldsymbol{k}} \exp\left( i \frac{2\pi}{N_0} \boldsymbol{k} \cdot T^{-1} \boldsymbol{r} \right), \tag{2.3}$$

converging in $L_{2,per}(\mathbb{R}^d)$. If we consider $w$ as a function with period $2N_0$, the coefficients $\tilde{w}_{\boldsymbol{k}}$ will change. To avoid this, we introduce

$$\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^d : T^* \boldsymbol{\alpha}/\pi \in \mathbb{Q}^d\}.$$

Denoting $\boldsymbol{\alpha} = 2\pi (T^*)^{-1} \boldsymbol{k}/N_0$ we rewrite (2.3) as

$$w = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} w_{\boldsymbol{\alpha}} \exp\left( i \boldsymbol{\alpha} \cdot \boldsymbol{r} \right). \tag{2.4}$$

Here and below in such sums we imply that there exists a common multiple of denominators of all $T^* \boldsymbol{\alpha}/(2\pi)$ such that $w_{\boldsymbol{\alpha}} \neq 0$. For $w \in L_{2,per}(\mathbb{R}^d)$, Parseval's identity $\|w\|^2 = \sum |w_{\boldsymbol{\alpha}}|^2$ is valid. The correspondence between $w \in L_{2,per}(\mathbb{R}^d)$ and its Fourier coefficients $w_{\boldsymbol{\alpha}}$ in (2.4) establishes a bijective isometry between $L_{2,per}(\mathbb{R}^d)$ and a dense subspace of $l_2(\mathcal{A}, \mathbb{C})$. The set $\mathcal{A}$ is the frequency domain of this Fourier transform.

For $w \in H^q_{per}(\mathbb{R}^d)$, $q \in \mathbb{R}$, $q \geqslant 0$, denote

$$\|\nabla^q w\|^2 = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 |\boldsymbol{\alpha}|^{2q}. \tag{2.5}$$

For $q \in \mathbb{N} \cup \{0\}$ the definitions (2.2) and (2.5) coincide (see Lemma A.1 in Appendix). We equip $H^q_{per}(\mathbb{R}^d)$, $q \in \mathbb{R}$, $q \geqslant 0$ and $C^q_{per}(\mathbb{R}^d)$, $q \in \mathbb{N} \cup \{0\}$ with the families of norms

$$\|f\|^2_{(q,h)} = \|f\|^2 + h^{2q} \|\nabla^q f\|^2, \quad \|f\|^2_{(q,h,\infty)} = \|f\|^2_\infty + h^{2q} \|\nabla^q f\|^2_\infty.$$

Let $\Omega$ be the unit sphere in $\mathbb{R}^d$. Denote by $\mathring{\Omega}$ a set of vectors $\boldsymbol{e} \in \Omega$ such that $\lambda T^* \boldsymbol{e} \in \mathbb{Q}^d$ for some $\lambda > 0$. We say that $f \in L_{2,loc}(\mathbb{R}^d)$ has a direction $\boldsymbol{e} \in \Omega$ and write $f \in L_{2,loc,\boldsymbol{e}}(\mathbb{R}^d)$ if there exists a function $g \in L_{2,loc}(\mathbb{R})$ such that $f(\boldsymbol{r}) = g(\boldsymbol{r} \cdot \boldsymbol{e})$ for almost all $\boldsymbol{r} \in \mathbb{R}^d$. For $\boldsymbol{e} \in \mathring{\Omega}$ we denote $H^q_{per,\boldsymbol{e}}(\mathbb{R}^d) = L_{2,loc,\boldsymbol{e}}(\mathbb{R}^d) \cap H^q_{per}(\mathbb{R}^d)$, $C^q_{per,\boldsymbol{e}}(\mathbb{R}^d) = L_{2,loc,\boldsymbol{e}}(\mathbb{R}^d) \cap C^q_{per}(\mathbb{R}^d)$. The spaces $H^q_{per,\boldsymbol{e}}(\mathbb{R}^d)$ and $C^q_{per,\boldsymbol{e}}(\mathbb{R}^d)$ contain nonconstant functions (see Lemma A.2 in Appendix).

### *2.2. Mesh function spaces*

Let $M^0$ be a finite set of the degrees of freedom (DOFs) at a mesh block and $M = \mathbb{Z}^d \times M^0$ be the general set of DOFs. For $f \in \mathbb{C}^M$ by $f_{\boldsymbol{\eta}} \in \mathbb{C}^{M^0}$ we denote a part of the vector $f$ in the block $\boldsymbol{\eta} \in \mathbb{Z}^d$. We shall also call $f_{\boldsymbol{\eta}}$ the block component of $f$ at $\boldsymbol{\eta}$. The set of sequences with period $N$ is

$$V^N_{per} = \{f \in \mathbb{C}^M \ : \ \forall \boldsymbol{\eta}, \boldsymbol{\zeta} \in \mathbb{Z}^d \ \ f_{\boldsymbol{\eta} + N\boldsymbol{\zeta}} = f_{\boldsymbol{\eta}}\}.$$

The set of periodic sequences $V_{per} = \bigcup\limits_{N \in \mathbb{N}} V^N_{per}$ is equipped with the scalar product defined for $f \in V^{N(f)}_{per}$, $g \in V^{N(g)}_{per}$ as

$$(f, g) = \frac{1}{N^d} \sum_{\boldsymbol{\eta} = (0, \ldots, N-1)^d} (f_{\boldsymbol{\eta}}, g_{\boldsymbol{\eta}}), \quad N = N(f) N(g).$$

Here $(f_{\boldsymbol{\eta}}, g_{\boldsymbol{\eta}})$ is any scalar product on $\mathbb{C}^{M^0}$. If the function $f$ has a period $N(f)$ then it has the period $nN(f)$ for all $n \in \mathbb{N}$, but the substitution $nN(f)$ for $N(f)$ does not change the value of the scalar product. We equip $\mathbb{C}^{M^0}$ and $V_{per}$ with the norms induced by these scalar products: $\|f_{\boldsymbol{\eta}}\|^2 = (f_{\boldsymbol{\eta}}, f_{\boldsymbol{\eta}})$; $\|f\|^2 = (f, f)$. The space $V_{per}$ is incomplete (see Corollary 6.3).

*2.3. The equation and the schemes*

In this paper we consider the initial value problem for the model linear transport equation

$$\frac{\partial v}{\partial t} + \boldsymbol{\omega} \cdot \nabla v = 0, \quad \boldsymbol{r} \in \mathbb{R}^d, \tag{2.6}$$

$$v(0, \boldsymbol{r}) = v_0(\boldsymbol{r}) \in L_{2,per}(\mathbb{R}^d). \tag{2.7}$$

The transport velocity $\boldsymbol{\omega}$ is constant in space and time.

To approximate (2.6) consider semi-discrete schemes of the form

$$\sum_{\boldsymbol{\zeta} \in \mathcal{S}} Z_{\boldsymbol{\zeta}} \frac{du_{\boldsymbol{\eta}+\boldsymbol{\zeta}}}{dt}(t) + \frac{1}{h} \sum_{\boldsymbol{\zeta} \in \mathcal{S}} L_{\boldsymbol{\zeta}} u_{\boldsymbol{\eta}+\boldsymbol{\zeta}}(t) = 0, \quad \boldsymbol{\eta} \in \mathbb{Z}^d, \quad u_{\boldsymbol{\eta}} \in \mathbb{C}^{M^0}, \tag{2.8}$$

where $\mathcal{S} \subset \mathbb{Z}^d$ is a finite set (the stencil of the scheme), $Z_{\boldsymbol{\zeta}}$ and $L_{\boldsymbol{\zeta}}$ are real-valued matrices. For $\boldsymbol{\zeta} \notin \mathcal{S}$ put $Z_{\boldsymbol{\zeta}} = L_{\boldsymbol{\zeta}} = 0$. Let $Z : \mathbb{C}^M \to \mathbb{C}^M$ and $L : \mathbb{C}^M \to \mathbb{C}^M$ take each $u$ to $Zu$ and $Lu$ such that

$$(Zu)_{\boldsymbol{\eta}} = \sum_{\boldsymbol{\zeta} \in \mathcal{S}} Z_{\boldsymbol{\zeta}} u_{\boldsymbol{\eta}+\boldsymbol{\zeta}}, \quad (Lu)_{\boldsymbol{\eta}} = \sum_{\boldsymbol{\zeta} \in \mathcal{S}} L_{\boldsymbol{\zeta}} u_{\boldsymbol{\eta}+\boldsymbol{\zeta}}. \tag{2.9}$$

With this notation, (2.8) is equivalent to

$$Z\frac{du}{dt} + \frac{1}{h} Lu = 0. \tag{2.10}$$

Obviously, $Z\,V_{per}^N \subseteq V_{per}^N$, $L\,V_{per}^N \subseteq V_{per}^N$, hence, $Z\,V_{per} \subseteq V_{per}$, $L\,V_{per} \subseteq V_{per}$. The operators $L$ and $Z$ are bounded on $V_{per}$ with

$$\|Z\| \leqslant \sum_{\boldsymbol{\zeta} \in \mathcal{S}} \|Z_{\boldsymbol{\zeta}}\|, \quad \|L\| \leqslant \sum_{\boldsymbol{\zeta} \in \mathcal{S}} \|L_{\boldsymbol{\zeta}}\|,$$

where the operator norms on the LHS are induced by the norm on $V_{per}$ and the operator norms on the RHS are induced by the norm on $\mathbb{C}^{M^0}$.

**Lemma 2.1.** *The operator $Z : V_{per} \to V_{per}$ has a bounded inverse if and only if for each $N$ there exists an inverse for the restriction of $Z$ to $V_{per}^N$, with its norm being limited by a constant independent of $N$.*

*Proof.* Let $Z$ be invertible and $\|Z^{-1}\| < \infty$. Then $\operatorname{Ker} Z = \{0\}$. In particular, for each $N$ the restriction of $Z$ to $V_{per}^N$ has zero kernel, and is invertible. Then $Z^{-1}V_{per}^N = V_{per}^N$. Clearly, the norm of the restriction of $Z^{-1}$ to $V_{per}^N$ does not exceed the norm of $Z^{-1}$ on $V_{per}$. To prove the reverse implication note that the norm of $Z^{-1}$ on $V_{per}$ is the supremum of the norms of its restrictions to $V_{per}^N$. □

Further in this paper we assume that for $Z : V_{per} \to V_{per}$ there exists a bounded inverse.

**Lemma 2.2.** *For any initial data $u_0 \in V_{per}$ the ODE system* (2.10) *has a unique solution $u \in C^\infty([0,\infty), V_{per})$ satisfying $u(0) = u_0$. Besides, if $u_0 \in V_{per}^N$, then for any $t > 0$ we have $u(t) \in V_{per}^N$.*

*Proof.* Let $u_0 \in V_{per}^N$. Since $Z$ is invertible on $V_{per}^N$ and $Z^{-1}LV_{per}^N \subset V_{per}^N$, (2.10) has a solution $u \in C^\infty([0,\infty), V_{per}^N)$. The uniqueness on $V_{per}$ follows from the boundedness of $Z^{-1}L$ on $V_{per}$. □

Since the operator $Z^{-1}L$ is bounded on $V_{per}$, and $V_{per}^N$ are its invariant subspaces, it generates the uniformly continuous operator group $\exp(-zZ^{-1}L) = \sum_{k=0}^\infty (-zZ^{-1}L)^k/k!$, $z \in \mathbb{C}$, which is analytical on the whole complex plane and also has $V_{per}^N$ as its invariant subspaces. Since $V_{per}^N$ are finite dimensional, on these subspaces $\exp(-Z^{-1}Lz)$ can be represented as the standard matrix exponential. For $u_0 \in V_{per}$ the function $u(t) = \exp(-Z^{-1}Lt/h)u_0$ solves (2.10) with the initial data $u(0) = u_0$.

A scheme (2.8) is called *stable* on $V_{per}$ with the stability constant $K$, if for all $h$ and each $u \in C^\infty([0,\infty), V_{per})$ satisfying (2.8) and each $t \geqslant 0$ there holds $\|u(t)\| \leqslant K\|u(0)\|$. In other words, a scheme (2.8) is stable with stability

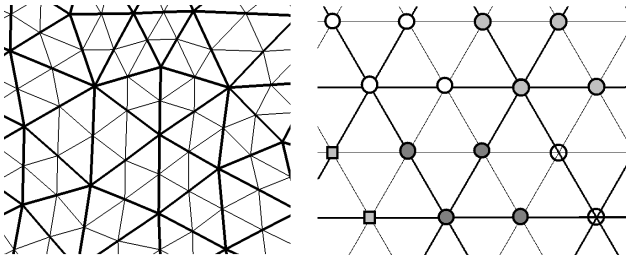


Figure 2: Fragments of a mesh with spectral elements. Left: unstructured mesh. Right: translationally-invariant mesh. In the right figure, nodes from different mesh blocks are marked by different symbols

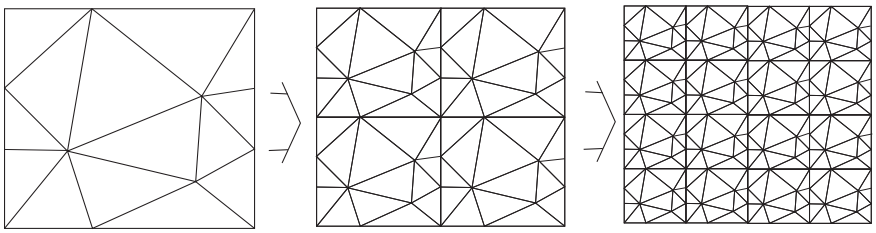


Figure 3: Block refinement of a mesh

constant $K$ iff $\sup\limits_{t,h>0} \|\exp(-Z^{-1}Lt/h)\| = \sup\limits_{\nu\geqslant 0} \|\exp(-Z^{-1}L\nu)\| \leqslant K$.

The most simple schemes of the form (2.8) are the finite-difference schemes on the uniform meshes. In this case $|M^0| = 1$, the order of accuracy coincides with the order of the truncation error and there is no enhanced accuracy in the long-time simulation (see Proposition 9.9). The schemes of the form (2.8) such that $|M^0| > 1$ arise in the following situations.

- Schemes with several DOFs per cell on the uniform meshes: discontinuous Galerkin, spectral difference, active flux etc. In this case $\boldsymbol{a}_j$ are the linearly independent vectors of the mesh edges and $M^0$ is the set of the DOFs in a cell. We assume that the ratio of the mesh steps along different directions remains constant under mesh refinement.

- The use of spectral elements. For example, in the flux correction method [24] one needs to compute the gradients at nodes of an unstructured mesh with the second order of the truncation error. It is convenient to use an unstructured mesh obtained by refining a base mesh (see Fig. 2, left). Elements of the base mesh are called spectral elements. On each spectral element one can construct the 2-nd order interpolation polynomial based on the nodal values of the mesh function. A gradient of the mesh function at a node can be defined as a weighted average of the gradients of interpolation polynomials at all the spectral elements containing this node. The resulting scheme on the translationally-invariant meshes has the form (2.8) with $|M^0| = 4$ (see Fig. 2, right) since the gradients are defined differently in the nodes on the base mesh and in the nodes at edges of the spectral elements. We say that the mesh is *translationally-invariant* if it is invariant with respect to the translation by the vector of any mesh edge.

- Finite-difference or finite-volume schemes if the block refinement is used (see Fig. 3). In this case the whole mesh is uniform-block, i. e. space is tessellated by blocks of an unstructured mesh. Blocks can be indexed by $\eta_1, \ldots, \eta_d$, as if they were cells of the uniform mesh. Vectors $\boldsymbol{a}_j$, $j = 1, \ldots, d$ are the offsets of blocks adjacent to a reference block and $|M^0|$ is the number of DOFs per block.

- Combination of the cases mentioned above. For example, the DG method on a simplicial translationally-invariant mesh.

*2.4. Homogeneous mappings to the mesh space*

Throughout this paper we assume that $1/h \in \mathbb{N}$. In order to transform data to the mesh space (prescribe the initial data for the difference problem etc.) one needs a mapping. Let $\mathcal{H} = H^q_{per}(\mathbb{R}^d)$ or $\mathcal{H} = C^q_{per}(\mathbb{R}^d)$. Consider a family of linear mappings $\Pi_h : \mathcal{H} \to V_{per}$, $1/h \in \mathbb{N}$, with the following properties:

(P1) for each $f, g \in \mathcal{H}$ such that almost everywhere $g(\boldsymbol{r}) = f(\boldsymbol{r} + hT\boldsymbol{\zeta})$ there holds $(\Pi_h f)_{\boldsymbol{\eta}+\boldsymbol{\zeta},\xi} = (\Pi_h g)_{\boldsymbol{\eta},\xi}$ for each $\boldsymbol{\eta}, \boldsymbol{\zeta} \in \mathbb{Z}^d$, $\xi \in M^0$;

(P2) for each $f, g \in \mathcal{H}$ such that almost everywhere $g(\boldsymbol{r}) = f(\boldsymbol{r}/h)$ there holds $(\Pi_h g)_{\boldsymbol{\eta},\xi} = (\Pi_1 f)_{\boldsymbol{\eta},\xi}$;

(P3) if $f \in \mathcal{H}$ is real-valued then so is $\Pi_h f$.

Each $\Pi_h$ from this family we call a *homogeneous mapping*. By (P2), a homogeneous mapping $\Pi_h$ uniquely defines the family it belongs to.

A homogeneous mapping $\Pi_h$ maps a function $f \in \mathcal{H}$ with a period $N_0$ to a sequence $\Pi_h f \in V_{per}$ with the period $N_0/h$. This follows directly from the definition. A homogeneous mapping $\Pi_h$ of $H^r_{per}(\mathbb{R}^d)$ to $V_{per}$ is *bounded* if $\|\Pi_h\|_{(r,h)} := \sup \|\Pi_h f\|/\|f\|_{(r,h)} < \infty$. A homogeneous mapping $\Pi_h$ of $C^r_{per}(\mathbb{R}^d)$ to $V_{per}$ is *bounded* if $\|\Pi_h\|_{(r,h,\infty)} := \sup \|\Pi_h f\|/\|f\|_{(r,h,\infty)} < \infty$. By construction, $\|\Pi_h\|_{(r,h)}$ and $\|\Pi_h\|_{(r,h,\infty)}$ do not depend on $h$.

*2.5. Local mappings to the mesh space*

Consider a family of mappings $\Pi_h$, $1/h \in \mathbb{N}$, from $L_{2,loc}(\mathbb{R}^d)$ to $\mathbb{C}^M$ of the form

$$(\Pi_h f)_{\boldsymbol{\eta},\xi} = \langle \mu_\xi, f(h(\,\cdot\, + T\boldsymbol{\eta}))\rangle \tag{2.11}$$

where $\boldsymbol{\eta} \in \mathbb{Z}^d$ is the index of a mesh block, $\xi \in M^0$ is the index of a variable inside the block, $\mu_\xi \in (W^q_2(G))^*$ or $(C^q(G))^*$ for some $q \in \mathbb{N} \cup \{0\}$ and a bounded domain $G$, $\Pi_h f$ is real-valued for real-valued $f$, and $\langle \mu_\xi, 1\rangle \neq 0$ for at least one $\xi \in M^0$. Each $\Pi_h$ from this family we call a *local mapping* and $\mu_\xi$ we call its kernel. Particularly, if $\mu_\xi$ is a measure then

$$(\Pi_h f)_{\boldsymbol{\eta},\xi} = \int\limits_G f(h(\boldsymbol{r} + T\boldsymbol{\eta}))d\mu_\xi = \int\limits_G f\left(h\boldsymbol{r} + h\sum_{j=1}^d \eta_j \boldsymbol{a}_j\right)d\mu_\xi.$$

If $\mu_\xi \in L_2(G)$, then $\Pi_h$ maps $L_{2,loc}(\mathbb{R}^d)$ to $\mathbb{C}^M$. For example, the standard $L_2$-projection onto a space of functions, which are polynomials on each mesh cell, can be represented as a local mapping. If $\mu_\xi \in (W^q_2(G))^*$ for some $q \in \mathbb{N} \cup \{0\}$, then $\Pi_h$ maps $H^q_{loc}(\mathbb{R}^d)$ to $\mathbb{C}^M$. If $\mu_\xi \in (C^q(G))^*$ for some $q \in \mathbb{N} \cup \{0\}$, then $\Pi_h$ maps $C^q(\mathbb{R}^d)$ to $\mathbb{C}^M$. An example is the pointwise mapping used in finite-difference schemes, namely, $(\Pi_h f)_{\boldsymbol{\eta},\xi} = f(h(\boldsymbol{\rho}_\xi + T\boldsymbol{\eta}))$ (i. e. $\mu_\xi = \delta(\boldsymbol{r} - \boldsymbol{\rho}_\xi)$) where $\boldsymbol{\rho}_\xi$ are collocation points.

**Lemma 2.3.** *A local mapping $\Pi_h$ with $\mu_\xi \in (W^q_2(G))^*$ is a bounded homogeneous mapping of $H^q_{per}(\mathbb{R}^d)$ to $V_{per}$. A local mapping $\Pi'_h$ with $\mu_\xi \in (C^q(G))^*$ is a bounded homogeneous mapping of $C^q_{per}(\mathbb{R}^d)$ to $V_{per}$.*

*Proof.* Let $N_0$ be a period of $f$. Then for each $\boldsymbol{\eta}, \boldsymbol{\zeta} \in \mathbb{Z}^d$, $\xi \in M^0$ there holds

$$(\Pi_h f)_{\boldsymbol{\eta}+\boldsymbol{\zeta}N_0/h,\xi} = \langle \mu_\xi, f\left(h\left(\,\cdot\, + T\boldsymbol{\eta}\right) + N_0 T\boldsymbol{\zeta}\right)\rangle = \langle \mu_\xi, f\left(h\left(\,\cdot\, + T\boldsymbol{\eta}\right)\right)\rangle = (\Pi_h f)_{\boldsymbol{\eta},\xi},$$

and the same for $\Pi'_h$. Thus $\Pi_h H^q_{per}(\mathbb{R}^d) \subseteq V_{per}$, $\Pi'_h C^q_{per}(\mathbb{R}^d) \subseteq V_{per}$. The properties (P1) and (P2) are obvious. The boundedness of $\Pi'_h$ is obvious; the proof of the boundedness of $\Pi_h$ is in Appendix (see Proposition A.4). □

In this paper we will use mappings $\Pi^{(p,q)}_h$ and $\Pi^{(p,q)}_{h,\boldsymbol{e}}$, $p, q \in \mathbb{N}$, given by

$$\left(\Pi^{(p,q)}_h f\right)_{\boldsymbol{\eta}} = (\Pi_h f)_{\boldsymbol{\eta}} + \sum_{p \leqslant |\boldsymbol{m}| \leqslant q} h^{|\boldsymbol{m}|} \mathfrak{C}^{(\boldsymbol{m})} \left(\mathcal{P}_h D^{\boldsymbol{m}} f\right)_{\boldsymbol{\eta}}, \tag{2.12}$$

$$\left(\Pi^{(p,q)}_{h,\boldsymbol{e}} f\right)_{\boldsymbol{\eta}} = (\Pi_h f)_{\boldsymbol{\eta}} + \sum_{m=p}^{q} h^m \mathfrak{C}^{(m)}_{\boldsymbol{e}} \left(\mathcal{P}_h \frac{\partial^m f}{\partial \boldsymbol{e}^m}\right)_{\boldsymbol{\eta}}, \tag{2.13}$$

where $\Pi_h$ and $\mathcal{P}_h$ are some local mappings, the kernel $\hat{\mu}_\xi$ of $\mathcal{P}_h$ satisfies $\langle \hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$, $\mathfrak{C}^{(\boldsymbol{m})}$ and $\mathfrak{C}_{\boldsymbol{e}}^{(m)}$ are diagonal square real-valued matrices of size $|M^0|$, and $\boldsymbol{e}$ is a unit vector. It is easy to see that $\Pi_h^{(p,q)}$ and $\Pi_{h,\boldsymbol{e}}^{(p,q)}$ are local mappings. If the kernels of $\Pi_h$ and $\mathcal{P}_h$ belong to $L_2(G)$ then $\Pi_h^{(p,q)}$ maps $W^q_{2,loc}(\mathbb{R}^d)$ to $\mathbb{C}^M$ and $H^q_{per}(\mathbb{R}^d)$ to $V_{per}$. If the kernels of $\Pi_h$ and $\mathcal{P}_h$ belong to $(C(G))^*$ then $\Pi_h^{(p,q)}$ maps $C^q(\mathbb{R}^d)$ to $\mathbb{C}^M$ and $C^q_{per}(\mathbb{R}^d)$ to $V_{per}$.

### *2.6. The truncation error and the solution error*

Suppose $\boldsymbol{\Pi}$ maps a subspace of $H^1_{loc}(\mathbb{R}^d)$ to $\mathbb{C}^M$. The *truncation error* on the function $f$ in the sense of $\boldsymbol{\Pi}$ is the vector $\epsilon_h(f, \boldsymbol{\Pi}) \in \mathbb{C}^M$ defined as

$$\epsilon_h(f, \boldsymbol{\Pi}) = -Z\boldsymbol{\Pi}(\boldsymbol{\omega} \cdot \nabla f) + \frac{1}{h} L\boldsymbol{\Pi} f. \tag{2.14}$$

Let $\Pi_h$ be a homogeneous mapping of $f \in H^q_{per}(\mathbb{R}^d)$ (or $C^q_{per}(\mathbb{R}^d)$) to $V_{per}$. Then for each $f$ with a period $N_0$ and each $h$ there holds $\epsilon_h(f, \Pi_h) \in V_{per}^{(N_0/h)}$.

**Definition 1.** We say that the scheme possesses *the truncation error of order* $P_A$ in the sense of $\Pi_h$ if for each $\boldsymbol{\alpha} \in \mathcal{A}$ there holds $\|\epsilon_h(e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant C(\boldsymbol{\alpha}) h^{P_A}$. We say that the scheme possesses the truncation error of order $P_A$ in the sense of $\Pi_h$ on the direction $\boldsymbol{e} \in \mathring{\Omega}$ if for each $\alpha \in \mathbb{R}$ such that $\alpha \boldsymbol{e} \in \mathcal{A}$ there holds $\|\epsilon_h(e^{i\alpha\boldsymbol{e}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant C(\alpha) h^{P_A}$. We say that the optimal order of the truncation error is $P_A$ if the scheme possesses the truncation error of order $P_A$ and does not possess the truncation error of order $P_A + \delta$ for each $\delta > 0$.

Later (see Corollary 6.21) we will show that for a local mapping $\Pi_h$ with kernel $\mu \in (W^q_2(G))^*$ the scheme possesses the truncation error of order $P_A \in \mathbb{N}$ iff for some $c_1, c_2 > 0$ there holds

$$\|\epsilon_h(v_0, \Pi_h)\| \leqslant c_1 \|\nabla^{P_A+1} v_0\| h^{P_A} + c_2 \|\nabla^{\max\{P_A,q\}+1} v_0\| \, h^{\max\{P_A,q\}},$$

for each $v_0 \in H^{\max\{P_A,q\}+1}_{per}(\mathbb{R}^d)$. A similar property holds if $\mu \in (C^q(G))^*$.

**Definition 2.** Suppose $\boldsymbol{\Pi}$ maps $H^r_{per}(\mathbb{R}^d)$ or $C^r_{per}(\mathbb{R}^d)$ to $V_{per}$ for some $r$. The *solution error* in the sense of $\boldsymbol{\Pi}$ with the initial data $v_0$ is the vector

$$\varepsilon_h(t, v_0, \boldsymbol{\Pi}) = u(t) - \boldsymbol{\Pi} v(t, \,\cdot\,) \in V_{per}, \tag{2.15}$$

where $u(t) = \exp(-Z^{-1}Lt/h)\,\boldsymbol{\Pi} v_0$ is the solution of (2.10) with the initial data $u(0) = \boldsymbol{\Pi} v_0$, and $v(t, \boldsymbol{r}) = v_0(\boldsymbol{r} - \boldsymbol{\omega} t)$.

If $\Pi_h$ is a local mapping with $\mu_\xi \in (W^q_2(G))^*$ (or $(C^q(G))^*$), then for each $f \in H^q_{per}(\mathbb{R}^d)$ (or $C^q_{per}(\mathbb{R}^d)$) with a period $N_0$ and each $h$ and $t \geqslant 0$ there holds $\varepsilon_h(t, f, \Pi_h) \in V_{per}^{(N_0/h)}$. Note that by definition we have $\varepsilon_h(0, v_0, \boldsymbol{\Pi}) \equiv 0$, so in contrast to the concepts conventional in the finite-element analysis we do not consider the error of the initial data mapping to be a part of the solution error.

### *2.7. Long-time simulation accuracy*

We introduce two different definitions of the scheme order in the long-time simulation.

**Definition 3.** Consider a scheme of the form (2.8) and a homogeneous mapping $\Pi_h$ of $\mathcal{H} \subseteq L_{2,per}(\mathbb{R}^d)$ to $V_{per}$. Let $P$, $Q$ satisfy $0 < P \leqslant Q \leqslant \infty$. Suppose for each $v_0 \in \mathcal{H}$ there exist non-negative constants $c_1(v_0)$, $c_2(v_0)$ such that for each $h$, and each $t \geqslant 0$ the scheme possesses the error estimate

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant c_1(v_0) h^P + c_2(v_0) t h^Q \tag{2.16}$$

(here we assume $h^\infty = 0$). Then we say that in the sense of $\Pi_h$ on $\mathcal{H}$ the scheme possesses *the formal order of accuracy* $P$ and *the long-time simulation order* $Q$ *in the weak sense*.

**Definition 4.** Let $\mathcal{H}$ be one of the following spaces: $H^R_{per}(\mathbb{R}^d)$ for some $R \geqslant 0$, $C^R_{per}(\mathbb{R}^d)$ for some $R \in \mathbb{N} \cup \{0\}$, $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ for some $R \geqslant 0$ and $\boldsymbol{e} \in \mathring{\Omega}$, $C^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ for some $R \in \mathbb{N} \cup \{0\}$ and $\boldsymbol{e} \in \mathring{\Omega}$. Consider a scheme of the form (2.8) and a homogeneous mapping $\Pi_h$ with $\mathcal{H}$ in the domain of $\Pi_h$.
**i.** Let $P$ and $Q$ be real numbers satisfying $0 < P \leqslant Q < \infty$ and $Q + 1 \leqslant R$. We say that the scheme possesses *the formal order of accuracy* $P$ and *the long-time simulation order* $Q$ in the sense of $\Pi_h$ on $\mathcal{H}$ if for each initial data $v_0 \in \mathcal{H}$, each $h$, and each $t \geqslant 0$ the scheme possesses the error estimate

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 h^P \|\nabla^P v_0\| + C_2 t h^Q \|\nabla^{Q+1} v_0\| + C_3 h^R \|\nabla^R v_0\|_* \tag{2.17}$$

where $C_1$, $C_2$, $C_3$ are non-negative constants independent of $v_0$, $h$, and $t$. Here $* \equiv \infty$ for $\mathcal{H} = C^R_{per}(\mathbb{R}^d)$ or $C^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ and dropped otherwise[1].
**ii.** Let $P$ be a real number satisfying $0 < P \leqslant R$. We say that the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q = \infty$ in the sense of $\Pi_h$ on $\mathcal{H}$ if for each initial data $v_0 \in \mathcal{H}$, each $h$, and each $t \geqslant 0$ the scheme possesses the error estimate

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 h^P \|\nabla^P v_0\| + C_3 h^R \|\nabla^R v_0\|_* \tag{2.18}$$

with nonnegative constants $C_1$ and $C_3$ independent of $v_0$, $h$, and $t$, and $*$ treated as in (2.17).
**iii.** If for each initial data $v_0 \in \mathcal{H}$ there holds $\varepsilon_h(t, v_0, \Pi_h) \equiv 0$ then we say that the scheme possesses the formal order of accuracy $P = \infty$ and the long-time simulation order $Q = \infty$ on $\mathcal{H}$.

Obviously, if the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ then it possesses the same orders in the weak sense. Unless specifically stated, speaking about the formal order of accuracy and the long-time simulation order we imply the use of Definition 4.

For $y \in \mathbb{R}$ denote by $\lfloor y \rfloor$ the largest integer that is less than or equal to $y$ and by $\lceil y \rceil$ the lowest integer that is greater than or equal to $y$. The formal order of accuracy and the long-time simulation order given by Definition 4 have the following properties, which are intuintively clear.

**Lemma 2.4.** *Let $\Pi_h$ be a local mapping with $\mu_\xi \in (W^s_2(G))^*$ (or $(C^s(G))^*$). Let the scheme* (2.8) *be stable and possess the formal order of accuracy $P$ and the long-time simulation order $Q$ on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r \rceil}_{per}(\mathbb{R}^d)$), $r \geqslant \max\{Q+1, s\}$ (or $r \geqslant \max\{P, s\}$ if $Q = \infty$). Then for each $p$ and $q$ such that $0 < p \leqslant P$ and $p \leqslant q \leqslant Q$ the scheme possesses the formal order $p$ and the long-time simulation order $q$ on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r \rceil}_{per}(\mathbb{R}^d)$). If $q < Q = \infty$ we additionally assume $q \leqslant r - 1$. The same holds for the definition in the weak sense.*

In the sense of Definition 3 this is obvious; in the sense of Definition 4 this will be proved below (see Corollary 6.18). So we can say that the scheme *possesses the order of accuracy* $P$, $0 < P \leqslant \infty$, if the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $P$ on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r \rceil}_{per}(\mathbb{R}^d)$) with $r \geqslant P$.

**Lemma 2.5.** *Let $P > 0$, $Q \geqslant P$, $R \geqslant Q + 1$. Let $\Pi_h$ be a homogeneous mapping of $H^R_{per}(\mathbb{R}^d)$. Let $r \geqslant R$. The scheme* (2.8) *possesses the formal order $P$ and the long-time simulation order $Q$ on $H^R_{per}(\mathbb{R}^d)$ iff it possesses these orders on $H^r_{per}(\mathbb{R}^d)$.*

**Lemma 2.6.** *Let $P > 0$, $Q \geqslant P$, $R \geqslant Q + 1$, $R \in \mathbb{N}$. Let $\Pi_h$ be a local mapping with $\mu_\xi \in (W^R_2(G))^*$. The scheme* (2.8) *possesses the formal order $P$ and the long-time simulation order $Q$ on $H^R_{per}(\mathbb{R}^d)$ iff it possesses the formal order $P$ and the long-time simulation order $Q$ on $C^R_{per}(\mathbb{R}^d)$.*

These lemmas follow from Theorem 6.17. Thus we will drop the class of functions on which the scheme possesses these orders, unless this leads to ambiguity.

Definitions 3 and 4 introduce the concepts of the formal order and the long-time simulation order simultaneously. Now we introduce the concept of the optimal values in such a way that one should first define the optimal value of the formal order of accuracy and then find the optimal value of the long-time simulation order.

[1] If $P, Q \in \mathbb{N}$, then $\|\nabla^P v_0\|$ and $\|\nabla^{Q+1} v_0\|$ can be replaced by $\|\nabla^P v_0\|_*$ and $\|\nabla^{Q+1} v_0\|_*$, which leads to an equivalent definition.

**Definition 5.** Let $\Pi_h$ be a homogeneous mapping. Let the scheme (2.8) possess the formal order $P$ and the long-time simulation order $Q$ in the sense of $\Pi_h$ on $H^R_{per}(\mathbb{R}^d)$ (or $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$). If for each $P' > P$, $Q' \geqslant P'$ and for each $P' = P$, $Q' > Q$ the scheme does not possess the formal order $P'$ and the long-time simulation order $Q'$ in the sense of $\Pi_h$ on any $H^r_{per}(\mathbb{R}^d)$ (or $H^r_{per,\boldsymbol{e}}(\mathbb{R}^d)$), then we call the values $P$ and $Q$ *optimal*. The optimal values for homogeneous mappings of $C^q_{per}(\mathbb{R}^d)$ and the optimal values in the weak sense are defined in the same way.

## 3. Main results

For $n \in \mathbb{N}$, let $\mathring{\Omega}_n$ be a set of vectors $\{\boldsymbol{e}_k \in \mathring{\Omega}, k = 1, \dots, C^{d-1}_{n+d-1}\}$ such that $\{(\boldsymbol{e}_k \cdot \boldsymbol{r})^n\}$ form a basis in the space of homogeneous polynomials of order $n$. For the existence of this set see Lemma A.6. Denote $L(0) = \sum_{\boldsymbol{\zeta}} L_{\boldsymbol{\zeta}}$.

**Theorem 1.** *Consider a scheme of the form* (2.8)*, stable with a constant $K$, and local mappings $\Pi_h$, $\mathcal{P}_h$ with kernels $\mu_\xi, \hat{\mu}_\xi \in (W^q_2(G))^*$ (or $(C^q(G))^*$). Let $P_A$ and $P$ be the optimal orders of the truncation error and accuracy, correspondingly, in the sense of $\Pi_h$. Let $\hat{\varkappa} = \min_{\xi \in M^0} |\langle \hat{\mu}_\xi, 1\rangle| > 0$ and $R = \max\{q, P\} + 1$. Then the following holds.*

1. *$P_A$ and $P$ are integers.*
2. *Either $P = P_A$ or $P = P_A + 1$.*
3. *If $P = P_A + 1$ then there exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}| = P$, such that the scheme possesses the truncation error of order $P$ in the sense of $\Pi_h^{(P,P)}$ given by* (2.12)*. Moreover, $\|\mathfrak{C}^{(\boldsymbol{m})}\| \leqslant \tilde{\delta} C_1$, where $C_1$ is the constant in estimate* (2.17) *and $\tilde{\delta}$ depends only on $K$, $\hat{\varkappa}$, $P$, $|M^0|$, and the norm on $\mathbb{C}^{M^0}$.*
4. *If $P = P_A$ then there exists no set of matrices $\{\mathfrak{C}^{(\boldsymbol{m})}, |\boldsymbol{m}| = P_A + 1\}$, such that the scheme possesses the truncation error of order $P_A + 1$ in the sense of $\Pi_h^{(P_A+1,P_A+1)}$ given by* (2.12)*.*
5. *If $L(0) = 0$ then $P = P_A$.*
6. *If the scheme possesses the formal order of accuracy $P_A+1$ on $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ (or $C^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$) for each $\boldsymbol{e} \in \mathring{\Omega}_{P_A+1}$, then it possesses the formal order of accuracy $P = P_A + 1$ on $H^R_{per}(\mathbb{R}^d)$ (or $C^R_{per}(\mathbb{R}^d)$).*
7. *$P$ coincides with the optimal order of accuracy in the weak sense.*

**Theorem 2.** *Let $\Pi_h$ be a local mapping with $\mu_\xi \in (W^q_2(G))^*$ or $(C^q(G))^*$. Let $P, Q > 0$, $r \geqslant \max\{P, q\}$, and $R = \max\{Q + 1, r\}$. Let the scheme* (2.8) *be stable and possess the error estimate* (2.17) *on $H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$). Then there exists a homogeneous mapping $\tilde{\Pi}_h : L_{2,per}(\mathbb{R}^d) \to V_{per}$ such that*

$$\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant C(h^P \|\nabla^P f\| + h^r \|\nabla^r f\|_*), \quad \|\epsilon_h(f, \tilde{\Pi}_h)\| \leqslant C h^Q \|\nabla^{Q+1} f\|$$

*for each $h$ and $f \in H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$), where $C$ does not depend on $h$ and $f$, and $\|\cdot\|_*$ means either $\|\cdot\|$ or $\|\cdot\|_\infty$ depending on the case.*

Note that the mapping $\tilde{\Pi}_h$ given by this theorem is generally *not* local.

In order to formulate the third result, we need two more definitions.

**Definition 6.** We say that a scheme of the form (2.8) is *quasi-1D* if the stencil $\mathcal{S}$ of the scheme belongs to a 1D subset of $\mathbb{Z}^d$, i. e. there exists $\boldsymbol{\eta} \in \mathbb{Z}^d$ such that $\mathcal{S} \subset \{m\boldsymbol{\eta}, m \in \mathbb{Z}\}$.

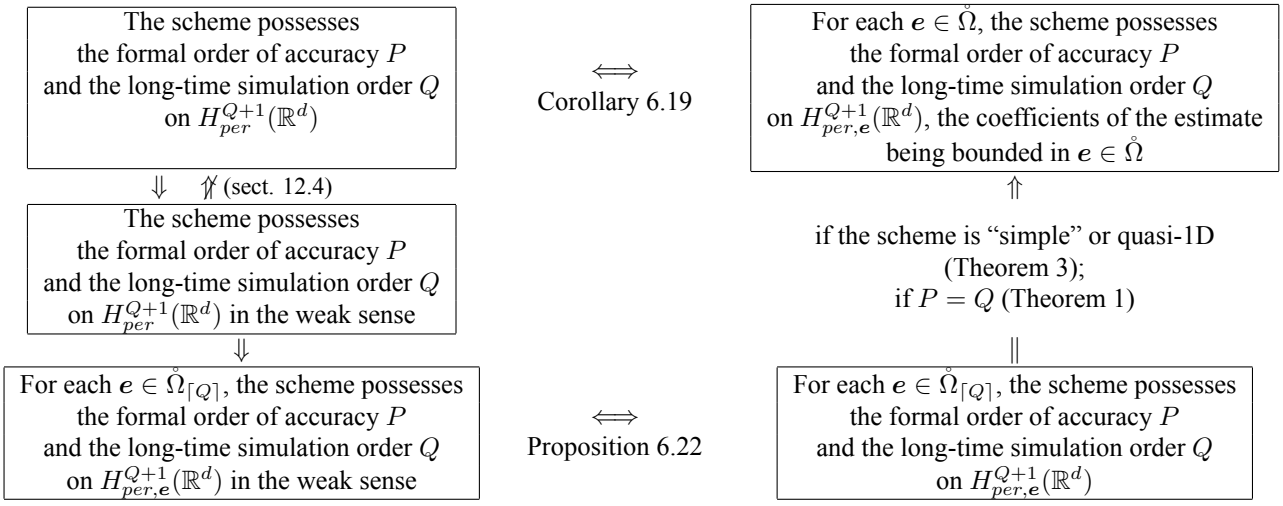
In 1D case, each scheme is quasi-1D.

**Definition 7.** We say that a scheme of the form (2.8) is *simple* if the matrix $L(0) \equiv \sum_{\boldsymbol{\zeta}} L_{\boldsymbol{\zeta}}$ has rank $|M^0| - 1$.

Note that if $L(0)$ has rank $|M^0|$ then the scheme does not preserve a constant solution, thus its order of the truncation error is equal to $P_A = -1$, then $P \leqslant 0$ by Theorem 1, i. e. there is no solution convergence.

**Theorem 3.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with kernels $\mu_\xi, \hat{\mu}_\xi \in (W^r_2(G))^*$ (or $(C^r(G))^*$). Let $\hat{\varkappa} = \min_{\xi \in M^0} |\langle \hat{\mu}_\xi, 1\rangle| > 0$. Let the scheme* (2.8) *be 1) stable and 2) either quasi-1D or simple. Let $P_A$, $P$, and*

Table 1: Relations between the definitions




*$Q$ be the optimal values of the order of the truncation error, the formal order of accuracy, and the long-time simulation order, correspondingly. Then the following holds.*

1. $Q \in \mathbb{N} \cup \{0\}$.
2. *There exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$, $P \leqslant |\boldsymbol{m}| \leqslant Q$, such that the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_h^{(P,Q)}$ given by* (2.12)*.*
3. *Let $p, q > 0$. If for each $\boldsymbol{e} \in \mathring{\Omega}_{\lceil q \rceil}$ the scheme possesses the formal order of accuracy $p$ and the long-time simulation order $q$ on $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ (or $C^{\lceil R \rceil}_{per,\boldsymbol{e}}(\mathbb{R}^d)$), where $R \geqslant \max\{r, p, q+1\}$, then it possesses the formal order of accuracy $p$ and the long-time simulation order $q$ on $H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R \rceil}_{per}(\mathbb{R}^d)$). Particularly, if $p = P$, then $Q \geqslant q$.*
4. *$Q$ coincides with the optimal value of the long-time simulation order in the weak sense.*
5. *If the scheme is simple, then either $Q \geqslant P = P_A + 1$ or $Q = P = P_A$.*

If the scheme is neither quasi-1D nor simple, then the statements 1–4 of Theorem 3 may be wrong, see counter-examples to the statements 1 and 2 in Section 12.4, to the statement 3 in Sections 12.2 and 12.3, to the statement 4 in Section 12.1. If the scheme is not simple, then the statement 5 of Theorem 3 may also be wrong even in 1D case, see Section 11.3.

The optimal value of the formal order of accuracy is the same for Definitions 3 and 4, but this generally does not hold for the long-time simulation order. Relations between the definitions are shown in Table 1. For each implication, a reference to a corresponding statement is given unless it is obvious. For an implication that does not hold, a reference to a counter-example is given.

Basing on Theorems 1 and 3 we construct algorithms giving the optimal values of the formal order of accuracy (for each stable scheme) and of the long-time simulation order (for simple and quasi-1D schemes), see Section 10. Theorem 2 is used below to establish Theorem 8.5 which states that a scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ if and only if a specific function is bounded at zero.

## 4. The method of auxiliary mapping

**Theorem 4.1** (Lax – Ryabenkii)**.** *Let $\Pi_h$ be a map of $H^{Q+1}_{per}(\mathbb{R}^d)$ to $V_{per}$. Let $K$ be the stability constant of the scheme* (2.8)*. Let for each $f \in H^{Q+1}_{per}(\mathbb{R}^d)$ there hold*

$$\|\epsilon(f, \Pi_h)\| \leqslant Ch^Q \|\nabla^{Q+1} f\|. \tag{4.1}$$

*Then for each $v_0$ there holds*

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant CKth^Q \|Z^{-1}\| \|\nabla^{Q+1} v_0\|. \tag{4.2}$$

*Proof.* Put $v(t, \boldsymbol{r}) = v_0(\boldsymbol{r} - \boldsymbol{\omega} t)$. Taking the time derivative of (2.15), multiplying by $Z$ and using (2.10) we get

$$Z \frac{d}{dt} \varepsilon_h(t, v_0, \Pi_h) = -\frac{1}{h} L u(t) + Z \Pi_h (\boldsymbol{\omega} \cdot \nabla) v(t, \cdot) =$$
$$= -\frac{1}{h} L \left( \varepsilon_h(t, v_0, \Pi_h) + \Pi_h v(t, \cdot) \right) + Z \Pi_h (\boldsymbol{\omega} \cdot \nabla) v(t, \cdot).$$

By (2.14) this is equivalent to

$$Z \frac{d}{dt} \varepsilon_h(t, v_0, \Pi_h) + \frac{1}{h} L \varepsilon_h(t, v_0, \Pi_h) = -\epsilon_h(v(t, \cdot), \Pi_h).$$

Since $\varepsilon_h(0, v_0, \Pi_h) = 0$ (see (2.15)),

$$\varepsilon_h(t, v_0, \Pi_h) = -\int_0^t \exp\left( -Z^{-1} L \frac{t - \tau}{h} \right) Z^{-1} \epsilon_h(v(\tau, \cdot), \Pi_h) \, d\tau.$$

By stability $\| \exp(-\nu Z^{-1} L) \| \leqslant K$ for all $\nu \geqslant 0$, so using (4.1) we have

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant Kt \max_{0 \leqslant \tau \leqslant t} \|Z^{-1} \epsilon_h(v(\tau, \cdot), \Pi_h)\| \leqslant CKth^Q \|Z^{-1}\| \cdot \max_{0 \leqslant \tau \leqslant t} \|\nabla^{Q+1} v(\tau, \cdot)\|.$$

It remains to note that $\|\nabla^{Q+1} v(\tau, \cdot)\| = \|\nabla^{Q+1} v_0\|$ for all $\tau$. □

The following proposition describes the method of auxiliary mapping, which is the cornerstone of this paper.

**Proposition 4.2.** *Consider a stable scheme* (2.8) *and homogeneous mappings $\Pi_h$ and $\tilde{\Pi}_h$ from $L_{2,per}(\mathbb{R}^d)$ to $V_{per}$. Suppose for some $P$ and $Q$ for each $f \in H_{per}^{\max\{P, Q+1\}}$ there holds*

$$\|(\Pi_h - \tilde{\Pi}_h) f\| \leqslant C_1 h^P \|\nabla^P f\| \quad \text{and} \quad \|\epsilon(f, \tilde{\Pi}_h)\| \leqslant C_2 h^Q \|\nabla^{Q+1} f\|.$$

*Then the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ in the sense of $\Pi_h$.*

*Proof.* Let $u(t)$ and $\tilde{u}(t)$ be the solutions of (2.8) with the initial data $u(0) = \Pi_h v_0$ and $\tilde{u}(0) = \tilde{\Pi}_h v_0$, correspondingly. Denote by $v$ the exact solution $v(t, \boldsymbol{r}) = v_0(\boldsymbol{r} - \boldsymbol{\omega} t)$. Then

$$\|\varepsilon_h(t, v_0, \Pi_h)\| = \|u(t) - \Pi_h v(t, \cdot)\| \leqslant$$
$$\leqslant \|u(t) - \tilde{u}(t)\| + \|\tilde{u}(t) - \tilde{\Pi}_h v(t, \cdot)\| + \|\tilde{\Pi}_h v(t, \cdot) - \Pi_h v(t, \cdot)\|.$$

By stability there holds

$$\|u(t) - \tilde{u}(t)\| \leqslant K \|u(0) - \tilde{u}(0)\| = K \|\tilde{\Pi}_h v_0 - \Pi_h v_0\|,$$

where $K$ is the stability constant of the scheme. Note that $\|v(t, \cdot)\| \equiv \|v_0\|$. Then we get

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant \|\varepsilon_h(t, v_0, \tilde{\Pi}_h)\| + (K + 1) C_1 h^P \|\nabla^P v_0\|. \tag{4.3}$$

By Theorem 4.1 we get the desired estimate (2.17) or (2.18). □

## 5. Preliminaries

### *5.1. Some properties of the functional spaces*

Recall that for $n \in \mathbb{N}$, by $\mathring{\Omega}_n$ we denote a set of vectors $\{\boldsymbol{e}_k \in \mathring{\Omega}, k = 1, \dots, C^{d-1}_{n+d-1}\}$ such that the polynomials $(\boldsymbol{e}_k \cdot \boldsymbol{r})^n$ form a basis in the space of homogeneous polynomials of order $n$ (see Lemma A.6).

**Lemma 5.1.** *Let $\boldsymbol{m}$ be a multiindex, $\mathring{\Omega}_{|\boldsymbol{m}|} = \{\boldsymbol{e}_k\}$. Then there exists $\gamma_k^{(\boldsymbol{m})} \in \mathbb{R}$, $k = 1, \dots, C^{d-1}_{|\boldsymbol{m}|+d-1}$, such that*

$$D^{\boldsymbol{m}} = \sum_{k=1}^{C^{d-1}_{|\boldsymbol{m}|+d-1}} \gamma_k^{(\boldsymbol{m})} (\boldsymbol{e}_k \cdot \nabla)^{|\boldsymbol{m}|}. \tag{5.1}$$

*Proof.* Since the polynomials $(\boldsymbol{e}_k \cdot \boldsymbol{r})^n$ form a basis in the space of homogeneous polynomials of order $n$, there exists a set of coefficients $\gamma_k^{(\boldsymbol{m})}$ such that

$$\boldsymbol{r}^{\boldsymbol{m}} = \sum_{k=1}^{C^{d-1}_{|\boldsymbol{m}|+d-1}} \gamma_k^{(\boldsymbol{m})} (\boldsymbol{e}_k \cdot \boldsymbol{r})^{|\boldsymbol{m}|}.$$

Since this is a polynomial identity, it holds with the substitution of $\nabla$ for $\boldsymbol{r}$. Thus we get (5.1). □

**Lemma 5.2.** *Let $G$ be a neighborhood of zero in $\mathbb{R}^d$, $f \in C^p(G)$, $p \in \mathbb{N}$ and $|f(\boldsymbol{x})| \leqslant c_f |\boldsymbol{x}|^p$ in $G$. Then $|D^{\boldsymbol{m}} f(0)| \leqslant c_f c_p$ for each $|\boldsymbol{m}| = p$, where $c_p$ does not depend on $f$.*

*Proof.* For the 1D case, this follows from the Taylor expansion with remainder term in Peano form with $c_p = p!$. Thus for each direction $\boldsymbol{e} \in \Omega$ we have $|(\boldsymbol{e} \cdot \nabla)^p f| \leqslant c_f p!$, and it remains to use Lemma 5.1. □

**Lemma 5.3.** *Let $f(\boldsymbol{\phi})$ be a holomorphic function at $\boldsymbol{\phi} = 0$, and $n, m \in \mathbb{N} \cup \{0\}$, $m \geqslant n$. Suppose for each $\boldsymbol{e} \in \mathring{\Omega}_m$ there exists $c_{\boldsymbol{e}} \geqslant 0$ such that $|f(\boldsymbol{e}\psi)| \leqslant c_{\boldsymbol{e}} |\psi|^{n+1}$ in a neighborhood of $\psi = 0$. Then there exists $\tilde{c}$ such that $|f(\boldsymbol{\phi})| \leqslant \tilde{c} |\boldsymbol{\phi}|^{n+1}$ in a neighborhood of $\boldsymbol{\phi} = 0$.*

*Proof.* By Lemma A.5, the system $\{(\boldsymbol{e} \cdot \boldsymbol{r})^n, \boldsymbol{e} \in \mathring{\Omega}_m\}$ is complete in the space of homogeneous polynomials of order $n$. Let $g(\boldsymbol{e}, \boldsymbol{\phi}) = \partial^n f / \partial \boldsymbol{e}^n (\boldsymbol{\phi})$. As a function of $\boldsymbol{e}$, $g$ is a homogeneous polynomial of order $n$. By assumption $g(\boldsymbol{e}, 0) = 0$ for $\boldsymbol{e} \in \mathring{\Omega}_m$. So we have $g(\boldsymbol{e}, 0) = 0$ for each $\boldsymbol{e} \in \Omega$. By Lemma 5.1 for each $\boldsymbol{m}$, $|\boldsymbol{m}| = n$, we have $D^{\boldsymbol{m}} f(\boldsymbol{\phi}) = 0$. Then the statement of the Lemma is obvious. □

**Lemma 5.4.** *For each $p \leqslant m \leqslant q$ and each $w \in H^q_{per}(\mathbb{R}^d)$ there holds*

$$h^{2m} \|\nabla^m w\|^2 \leqslant h^{2p} \|\nabla^p w\|^2 + h^{2q} \|\nabla^{2q} w\|^2. \tag{5.2}$$

*Proof.* For $p = m = q$ this is obvious so assume $q > p$. Let $w_{\boldsymbol{\alpha}}$ be Fourier coefficients of $w$. By (2.5),

$$h^{2m} \|\nabla^m w\|^2 = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 |h\boldsymbol{\alpha}|^{2m} \leqslant$$

$$\leqslant \sum_{\boldsymbol{\alpha} \in \mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 \left( \frac{q-m}{q-p} |h\boldsymbol{\alpha}|^{2p} + \frac{m-p}{q-p} |h\boldsymbol{\alpha}|^{2q} \right) \leqslant h^{2p} \|\nabla^p w\|^2 + h^{2q} \|\nabla^q w\|^2,$$

where we have used the Young inequality for products. □

**Lemma 5.5.** *For each $p, q \in \mathbb{N} \cup \{0\}$ there exists $C(p, q)$ such that for each $m = p, \dots, q$ and $f \in C^q_{per}(\mathbb{R}^d)$ there holds*

$$h^m \|\nabla^m f\|_\infty \leqslant C(p, q) \left( h^p \|\nabla^p f\|_\infty + h^q \|\nabla^q f\|_\infty \right).$$

*Proof.* By the coordinate transformation we can assume without loss that $h = 1$. For 1D case this follows from the Landau – Kolmogorov inequality [25]. For each $\boldsymbol{e} \in \Omega$ this yields

$$\left\| \frac{\partial^m f}{\partial \boldsymbol{e}^m} \right\|_\infty \leqslant C(p,q) \left( \left\| \frac{\partial^p f}{\partial \boldsymbol{e}^p} \right\|_\infty + \left\| \frac{\partial^q f}{\partial \boldsymbol{e}^q} \right\|_\infty \right).$$

Then by Lemma 5.1 we have

$$\|\nabla^m f\|_\infty^2 = \sum_{|\boldsymbol{m}|=m} \frac{m!}{\boldsymbol{m}!} \max_{\mathbb{R}^d} |D^{\boldsymbol{m}} f|^2 \leqslant \sum_{|\boldsymbol{m}|=m} \frac{m!}{\boldsymbol{m}!} \gamma^2 \sup_{\boldsymbol{e}\in\Omega} \left\| \frac{\partial^m f}{\partial \boldsymbol{e}^m} \right\|_\infty^2 \leqslant$$

$$\leqslant \sum_{|\boldsymbol{m}|=m} \frac{m!}{\boldsymbol{m}!} \gamma^2 C^2(p,q) \sup_{\boldsymbol{e}\in\Omega} \left( \left\| \frac{\partial^p f}{\partial \boldsymbol{e}^p} \right\|_\infty + \left\| \frac{\partial^q f}{\partial \boldsymbol{e}^q} \right\|_\infty \right)^2 \leqslant \tilde{C}(p,q) \left( \|\nabla^p f\|_\infty + \|\nabla^q f\|_\infty \right)^2,$$

where $\gamma = \max\limits_{|\boldsymbol{m}|=m} \sum\limits_k |\gamma_k^{(\boldsymbol{m})}|$ and $\gamma_k^{(\boldsymbol{m})}$ are given by Lemma 5.1. □

**Lemma 5.6.** *Let $0 \leqslant r \leqslant q < \infty$. Then for each $f \in H^q_{per}(\mathbb{R}^d)$ there holds $\|f\|_{(r,h)} \leqslant 2\|f\|_{(q,h)}$. For $r, q \in \mathbb{N} \cup \{0\}$, $r \leqslant q$, there exists $C(p,q)$ such that for each $f \in C^q_{per}(\mathbb{R}^d)$ there holds $\|f\|_{(r,h,\infty)} \leqslant C(p,q)\|f\|_{(q,h,\infty)}$.*

The proof is obvious by Lemma 5.4 and Lemma 5.5.

**Lemma 5.7.** *For each $v_0 \in H^{r+1}_{per}(\mathbb{R}^d)$ there holds*

$$\|\nabla^r(\boldsymbol{\omega}\cdot\nabla v_0)\| \leqslant |\boldsymbol{\omega}|\,\|\nabla^{r+1} v_0\|, \quad h\|\boldsymbol{\omega}\cdot\nabla v_0\|_{(r,h)} \leqslant \sqrt{2}|\boldsymbol{\omega}|\,\|v_0\|_{(r+1,h)}.$$

*There exists $c_r$ such that for each $v_0 \in C^{r+1}_{per}(\mathbb{R}^d)$ there holds*

$$\|\nabla^r(\boldsymbol{\omega}\cdot\nabla v_0)\|_\infty \leqslant |\boldsymbol{\omega}|\,\|\nabla^{r+1} v_0\|_\infty, \quad h\|\boldsymbol{\omega}\cdot\nabla v_0\|_{(r,h,\infty)} \leqslant c_r|\boldsymbol{\omega}|\,\|v_0\|_{(r+1,h,\infty)}.$$

*Proof.* For $v_0 \in H^{r+1}_{per}(\mathbb{R}^d)$, $r \geqslant 0$, we have

$$\|\nabla^r(\boldsymbol{\omega}\cdot\nabla v_0)\|^2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} |f_{\boldsymbol{\alpha}}|^2 (\boldsymbol{\omega}\cdot\boldsymbol{\alpha})^2 |\boldsymbol{\alpha}|^{2r} \leqslant \sum_{\boldsymbol{\alpha}\in\mathcal{A}} |f_{\boldsymbol{\alpha}}|^2 |\boldsymbol{\omega}|^2 |\boldsymbol{\alpha}|^{2r+2} = |\boldsymbol{\omega}|^2\,\|\nabla^{r+1} v_0\|^2.$$

Then, by Lemma 5.4,

$$h^2\|\boldsymbol{\omega}\cdot\nabla v_0\|^2_{(r,h)} = h^2\|\boldsymbol{\omega}\cdot\nabla v_0\|^2 + h^{2r+2}\|\nabla^r(\boldsymbol{\omega}\cdot\nabla v_0)\|^2 \leqslant h^2|\boldsymbol{\omega}|^2\|\nabla v_0\|^2 + h^{2r+2}|\boldsymbol{\omega}|^2\|\nabla^{r+1} v_0\|^2 \leqslant$$

$$\leqslant |\boldsymbol{\omega}|^2 \left( \|v_0\|^2 + 2h^{2r+2}\|\nabla^{r+1} v_0\|^2 \right) \leqslant 2|\boldsymbol{\omega}|^2\,\|v_0\|^2_{(r+1,h)}.$$

For $v_0 \in C^{r+1}_{per}(\mathbb{R}^d)$, $r \in \mathbb{N} \cup \{0\}$, we have

$$\|\nabla^r(\boldsymbol{\omega}\cdot\nabla v_0)\|^2_\infty = \sum_{|\boldsymbol{m}|=r} \frac{r!}{\boldsymbol{m}!} \|D^{\boldsymbol{m}}(\boldsymbol{\omega}\cdot\nabla v_0)\|^2_\infty =$$

$$= \sum_{|\boldsymbol{m}|=r} \frac{r!}{\boldsymbol{m}!} \left\| \sum_{|\boldsymbol{l}|=1} \boldsymbol{\omega}^{\boldsymbol{l}} D^{\boldsymbol{m}+\boldsymbol{l}} v_0 \right\|^2_\infty \leqslant \sum_{|\boldsymbol{m}|=r} \frac{r!}{\boldsymbol{m}!} \left( \sum_{|\boldsymbol{l}|=1} |\boldsymbol{\omega}^{\boldsymbol{l}}| \left\| D^{\boldsymbol{m}+\boldsymbol{l}} v_0 \right\|_\infty \right)^2 \leqslant |\boldsymbol{\omega}|^2 \sum_{|\boldsymbol{m}|=r} \frac{r!}{\boldsymbol{m}!} \sum_{|\boldsymbol{l}|=1} \left\| D^{\boldsymbol{m}+\boldsymbol{l}} v_0 \right\|^2_\infty.$$

Denote $\boldsymbol{n} = \boldsymbol{m} + \boldsymbol{l}$, then we continue the chain:

$$\|\nabla^r(\boldsymbol{\omega}\cdot\nabla v_0)\|^2_\infty \leqslant |\boldsymbol{\omega}|^2 \sum_{|\boldsymbol{n}|=r+1} \sum_{|\boldsymbol{l}|=1, \boldsymbol{l}\leqslant\boldsymbol{n}} \frac{(r+1)!}{(\boldsymbol{n}-\boldsymbol{l})!\,(r+1)} \|D^{\boldsymbol{n}} v_0\|^2_\infty =$$

$$= |\boldsymbol{\omega}|^2 \sum_{|\boldsymbol{n}|=r+1} \frac{(r+1)!}{\boldsymbol{n}!} \|D^{\boldsymbol{n}} v_0\|_\infty^2 \sum_{|\boldsymbol{l}|=1, \boldsymbol{l} \leqslant \boldsymbol{n}} \frac{\boldsymbol{n}^{\boldsymbol{l}}}{|\boldsymbol{n}|} = |\boldsymbol{\omega}|^2 \sum_{|\boldsymbol{n}|=r+1} \frac{(r+1)!}{\boldsymbol{n}!} \|D^{\boldsymbol{n}} v_0\|_\infty^2 = |\boldsymbol{\omega}|^2 \|\nabla^{r+1} v_0\|_\infty^2.$$

Then, by Lemma 5.5,

$$h^2 \|\boldsymbol{\omega} \cdot \nabla v_0\|_{(r,h,\infty)}^2 = h^2 \|\boldsymbol{\omega} \cdot \nabla v_0\|_\infty^2 + h^{2r+2} \|\nabla^r (\boldsymbol{\omega} \cdot \nabla v_0)\|_\infty^2 \leqslant h^2 |\boldsymbol{\omega}|^2 \|\nabla v_0\|_\infty^2 + h^{2r+2} |\boldsymbol{\omega}|^2 \|\nabla^{r+1} v_0\|_\infty^2 \leqslant$$
$$\leqslant |\boldsymbol{\omega}|^2 \left( C(1, r+1) \|v_0\|^2 + (1 + C(1, r+1)) h^{2r+2} \|\nabla^{r+1} v_0\|^2 \right) \leqslant (1 + C(1, r+1)) |\boldsymbol{\omega}|^2 \|v_0\|_{(r+1,h,\infty)}^2.$$

□

*5.2. Approximation of local mappings*

In this subsection we prove that any local mapping can be approximated by a local mapping with a continuous kernel.

**Lemma 5.8.** *Let $s \in \mathbb{N}$. Then there exists $\varkappa^{(s)} \in C^s(\{|\boldsymbol{x}| \leqslant 1\})$ such that the following conditions hold.*

*(i) $\varkappa^{(s)}$ and its derivatives up to the order $s$ are equal to zero on $|\boldsymbol{x}| = 1$;*

*(ii) there holds*

$$\int\limits_{|\boldsymbol{x}|<1} \varkappa^{(s)}(\boldsymbol{x}) d\boldsymbol{x} = 1; \tag{5.3}$$

*(iii) for each $\sigma = 0, \dots, s+1$, each $f \in C_{per}^\sigma(\mathbb{R}^d)$, and each $\boldsymbol{y} \in \mathbb{R}^d$ there holds*

$$\left| \int\limits_{|\boldsymbol{x}|<1} \varkappa^{(s)}(\boldsymbol{x}) f(\boldsymbol{y} + h\boldsymbol{x}) d\boldsymbol{x} - f(\boldsymbol{y}) \right| \leqslant C(d, s) h^\sigma \|\nabla^\sigma f\|_\infty. \tag{5.4}$$

*Proof.* Put $\varkappa^{(s)}(\boldsymbol{x}) = |\boldsymbol{x}|^{2s}(1 - |\boldsymbol{x}|)^s g(|\boldsymbol{x}|)$, where $g(r)$ is a smooth function on $[0, 1]$ such that

$$|\Omega| \int\limits_0^1 r^{2s+d-1} (1-r)^s g(r) dr = 1, \tag{5.5}$$

$$\int\limits_0^1 r^{2s+d-1+m} (1-r)^s g(r) dr = 0, \quad m = 1, \dots, s. \tag{5.6}$$

The factor $|\boldsymbol{x}|^{2s}$ guarantees the existence and continuity of the derivatives of $\varkappa^{(s)}(\boldsymbol{x})$ up to the order $s$ at $r = 0$, and the factor $(1 - |\boldsymbol{x}|)^s$ makes the boundary conditions satisfied. The condition (5.5) yields (5.3). Now we show (5.4). For $\sigma = 0$ inequality (5.4) is obvious, so assume $\sigma > 0$. For each $\boldsymbol{y}$ and $\boldsymbol{x}$ there holds

$$\left| f(\boldsymbol{y} + h\boldsymbol{x}) - f(\boldsymbol{y}) - \sum_{k=1}^{\sigma-1} \frac{1}{k!} \frac{\partial^k f(\boldsymbol{y})}{\partial \boldsymbol{e}^k} h^k |\boldsymbol{x}|^k \right| \leqslant \|\nabla^\sigma f\|_\infty \frac{h^\sigma |\boldsymbol{x}|^\sigma}{\sigma!},$$

where $\boldsymbol{e}$ is the unit vector along $\boldsymbol{x}$. Hence, using (5.3) we get

$$\left| \int \varkappa^{(s)}(\boldsymbol{x}) f(\boldsymbol{y} + h\boldsymbol{x}) d\boldsymbol{x} - f(\boldsymbol{y}) \right| = \left| \int \varkappa^{(s)}(\boldsymbol{x}) (f(\boldsymbol{y} + h\boldsymbol{x}) - f(\boldsymbol{y})) d\boldsymbol{x} \right| \leqslant$$
$$\leqslant \left| \int \varkappa^{(s)}(\boldsymbol{x}) \left[ \sum_{k=1}^{\sigma-1} \frac{1}{k!} \frac{\partial^k f(\boldsymbol{y})}{\partial \boldsymbol{e}^k} h^k |\boldsymbol{x}|^k \right] d\boldsymbol{x} \right| + \|\nabla^\sigma f\|_\infty \frac{h^\sigma}{\sigma!} \int |\varkappa^{(s)}(\boldsymbol{x})| d\boldsymbol{x}$$

(all integrals are over $|\boldsymbol{x}| < 1$). We claim that the first term in the right-hand size is equal to zero. Indeed, put $\boldsymbol{x} = r\boldsymbol{\gamma}$, where $0 < r < 1$ and $\boldsymbol{\gamma}$ varies over $\Omega$. Then for $k = 1, \dots, \sigma - 1$ we have

$$\int\limits_{|\boldsymbol{x}|<1} \varkappa^{(s)}(\boldsymbol{x}) \frac{\partial^k f(\boldsymbol{y})}{\partial \boldsymbol{e}^k} |\boldsymbol{x}|^k d\boldsymbol{x} = \int\limits_0^1 r^{2s} (1-r)^s r^{d-1} r^k g(r) dr \int\limits_\Omega \frac{\partial^k f(\boldsymbol{y})}{\partial \boldsymbol{e}_{\boldsymbol{\gamma}}^k} d\boldsymbol{\gamma},$$

that is equal to zero by (5.6). Thus the inequality (5.4) is proved. □

**Lemma 5.9.** *Let $\tilde{\Pi}_h$ be a local mapping with kernel $\tilde{\mu}_\xi \in (C^q(G))^*$, $q \in \mathbb{N} \cup \{0\}$. Let $s \geqslant q$, $s \in \mathbb{N}$, and $\varkappa^{(s)}$ be given by Lemma 5.8. Then the convolution $\mu_\xi = \tilde{\mu}_\xi * \varkappa^{(s)}$ is a continuous function with support in $G + B_1(0)$, and the local mapping $\Pi_h$ with kernel $\mu_\xi$ satisfies*

$$\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant Ch^r \|\nabla^r f\|_\infty$$

*for each $r = q, \dots, s$ and all $f \in C^r_{per}(\mathbb{R}^d)$, where $C$ depends only on $d$, $s$, $\tilde{\mu}_\xi$, and the norm on $\mathbb{C}^{M^0}$.*

*Proof.* For each $\xi \in M^0$, each $h$, and $\boldsymbol{\eta} \in \mathbb{Z}^d$ we have

$$(\tilde{\Pi}_h f)_{\boldsymbol{\eta},\xi} - (\Pi_h f)_{\boldsymbol{\eta},\xi} = \langle \tilde{\mu}_\xi - \mu_\xi, f(h(\,\cdot\, + T\boldsymbol{\eta}))\rangle = \langle \tilde{\mu}_\xi, \Phi[f]\rangle,$$

where $\Phi[f]$ is given by

$$\Phi[f](\boldsymbol{r}) = \int\limits_{|\boldsymbol{x}|<1} \varkappa^{(s)}(\boldsymbol{x}) f(h(\boldsymbol{r} + T\boldsymbol{\eta} + \boldsymbol{x}))d\boldsymbol{x} - f(h(\boldsymbol{r} + T\boldsymbol{\eta})).$$

For $|\boldsymbol{m}| \leqslant q$ we have $D^{\boldsymbol{m}}\Phi[f] = h^{|\boldsymbol{m}|}\Phi[D^{\boldsymbol{m}} f]$. Then by (5.4) with $\sigma = r - |\boldsymbol{m}|$ we get

$$\sup_{\boldsymbol{r}\in\mathbb{R}^d} |D^{\boldsymbol{m}}\Phi[f](\boldsymbol{r})| \leqslant h^{|\boldsymbol{m}|} C(d,s) h^{r-|\boldsymbol{m}|} \|\nabla^{r-|\boldsymbol{m}|}(D^{\boldsymbol{m}} f)\|_\infty \leqslant h^r C(d,s)\|\nabla^r f\|_\infty.$$

Equip $C^q(G)$ with the norm

$$\|g\|_{C^q(G)} = \max_{|\boldsymbol{m}|\leqslant q} \max_{\boldsymbol{r}\in G} |D^{\boldsymbol{m}} g|.$$

Then $\|\Phi[f]\|_{C^q(G)} \leqslant h^r C(d,s)\|\nabla^r f\|_\infty$ and

$$\left|(\tilde{\Pi}_h f)_{\boldsymbol{\eta},\xi} - (\Pi_h f)_{\boldsymbol{\eta},\xi}\right| \leqslant \|\tilde{\mu}_\xi\| h^r C(d,s)\|\nabla^r f\|_\infty.$$

Then the inequality to prove is by the definition of the norm on $V_{per}$. □

*5.3. Properties of the matrix exponential*

In this subsection we consider a matrix norm induced by a vector norm.

**Lemma 5.10.** *If $A(\boldsymbol{\phi})$ is a holomorphic matrix function at $\boldsymbol{\phi} = \boldsymbol{\phi}_0$, then so is* $\exp(A(\boldsymbol{\phi}))$.

This follows from the representation

$$\exp(A(\boldsymbol{\phi})) = \frac{1}{2\pi i}\int\limits_\gamma \frac{e^z}{zI - A(\boldsymbol{\phi})}dz,$$

where $I$ is the identity matrix and $\gamma$ is any closed contour such that all eigenvalues of $A(\boldsymbol{\phi}_0)$ lay inside $\gamma$.

**Lemma 5.11.** *For each square matrix $Y$ there holds*

$$\|e^Y - I\| \leqslant (\|e^Y\| + e - 1)\min\{1, \|Y\|\}.$$

*Proof.* If $\|Y\| \geqslant 1$ then

$$\|e^Y - I\| \leqslant \|e^Y\| + 1 \leqslant \|e^Y\| + e - 1.$$

If $\|Y\| \leqslant 1$ then

$$\|e^Y - I\| = \left\| Y \sum_{k=1}^\infty \frac{Y^{k-1}}{k!}\right\| \leqslant \|Y\| \sum_{k=1}^\infty \frac{1}{k!} = (e-1)\|Y\| \leqslant (\|e^Y\| + e - 1)\|Y\|.$$

□

**Lemma 5.12.** *Suppose* $\|A\| \leqslant 1$ *and*

$$f(A) = \sum_{k=1}^{\infty} \frac{A^{k-1}}{k!}. \tag{5.7}$$

*Then* $\|(f(A))^{-1}\| \leqslant 4$.

*Proof.* First,

$$\|f(A) - I\| = \left\| \sum_{k=2}^{\infty} \frac{A^{k-1}}{k!} \right\| \leqslant \sum_{k=2}^{\infty} \frac{\|A\|^{k-1}}{k!} \leqslant \sum_{k=2}^{\infty} \frac{1}{k!} = e - 2.$$

Hence

$$\|(f(A))^{-1}\| = \|(I + (f(A) - I))^{-1}\| \leqslant \frac{1}{1 - \|f(A) - I\|} \leqslant \frac{1}{1 - (e-2)} \leqslant 4.$$

□

**Lemma 5.13.** *Let* $A$ *be non-degenerate and* $\|A\| \leqslant 1$. *Then* $\|(e^A - I)^{-1}\| \leqslant 4\|A^{-1}\|$.

*Proof.* For $f(A)$ given by (5.7) we have $e^A - I = Af(A)$. Then

$$\|(e^A - I)^{-1}\| \leqslant \|(f(A))^{-1}\| \, \|A^{-1}\| \leqslant 4\|A^{-1}\|.$$

□

*5.4. Transformation of a matrix to a block-diagonal form*

Let $\mathbb{C}^{n\times n}$ be the space of complex matrices of size $n$. Denote by $\mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$ the set of functions from $\mathbb{C}^d$ to $\mathbb{C}^{n\times n}$ holomorphic at $\boldsymbol{\phi} = 0$. Let $\sigma(Y)$ be the spectrum of a matrix $Y$ and $\varkappa(Y)$ be its condition number, i. e. $\varkappa(Y) = \|Y\| \, \|Y^{-1}\|$.

To proceed, we need to recall some facts from the perturbation theory.

**Lemma 5.14** ([26], §1.5.3)**.** *Let* $A \in \mathbb{C}^{n\times n}$, $\lambda \in \sigma(A)$. *Let* $0 < \rho < \operatorname{dist}(\lambda, \sigma(A) \setminus \{\lambda\})$ *(the distance to the empty set is assumed to be* $+\infty$*). Then the matrix*

$$P_\lambda(A) = \frac{1}{2\pi i} \oint\limits_{|z-\lambda|=\rho} (zI - A)^{-1} dz \tag{5.8}$$

*is a projection onto the algebraic eigenspace corresponding to* $\lambda$. *Besides, for* $\lambda, \mu \in \sigma(A)$ *there holds* $P_\lambda(A)P_\mu(A) = \delta_{\lambda\mu}P_\lambda(A)$.

**Lemma 5.15** ([26], §2.5.1)**.** *Let* $A(\boldsymbol{\phi}) \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$. *Then the eigenvalues* $\lambda_j(\boldsymbol{\phi})$, $j = 1, \ldots, n$, *can be reordered so that* $\lambda_j(\boldsymbol{\phi}) \to \lambda_j(0)$ *as* $\boldsymbol{\phi} \to 0$.

For $A(\boldsymbol{\phi}) \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$ and $\lambda \in \sigma(A(0))$ the set of eigenvalues $\lambda_j(\boldsymbol{\phi})$ of the matrix $A(\boldsymbol{\phi})$ such that $\lambda_j(\boldsymbol{\phi}) \to \lambda$ is referred to as $\lambda$-*group*, and the sum of the corresponding algebraic eigenspaces is the *total eigenspace* of the $\lambda$-group. The operator $\mathcal{P}_\lambda(\boldsymbol{\phi}) = \sum P_{\lambda_j(\boldsymbol{\phi})}(A(\boldsymbol{\phi}))$, where the sum is over the $\lambda$-group, is the *total eigenprojection* of the $\lambda$-group. It is a projection onto the total eigenspace of the $\lambda$-group (see [26], §2.2.1). The sum of all total eigenprojections is the identity operator and $\mathcal{P}_\lambda(\boldsymbol{\phi})\mathcal{P}_\mu(\boldsymbol{\phi}) = \delta_{\lambda\mu}\mathcal{P}_\lambda(\boldsymbol{\phi})$.

**Lemma 5.16.** *Let* $A(\boldsymbol{\phi}) \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$ *and* $\lambda \in \sigma(A(0))$. *Then* $\mathcal{P}_\lambda(\boldsymbol{\phi}) \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$.

*Proof.* Let $\rho = \frac{1}{2}\operatorname{dist}(\lambda, \sigma(A(0)) \setminus \{\lambda\})$. By continuity in a neighborhood of $\boldsymbol{\phi} = 0$ there holds $|\lambda_j(\boldsymbol{\phi}) - \lambda| < \rho$ for each $\lambda_j(\boldsymbol{\phi})$ belonging to the $\lambda$-group corresponding to $\lambda$ and $|\lambda_j(\boldsymbol{\phi}) - \lambda| > \rho$ otherwise. Thus the total eigenprojection of this $\lambda$-group is

$$\mathcal{P}_\lambda(\boldsymbol{\phi}) = \frac{1}{2\pi i} \oint\limits_{|z-\lambda|=\rho} (zI - A(\boldsymbol{\phi}))^{-1} dz. \tag{5.9}$$

Since the integrand is analytical with respect to $\boldsymbol{\phi}$ and $z$ over the integral path, we have $\mathcal{P}_\lambda(\boldsymbol{\phi}) \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$. □

**Lemma 5.17.** *For any $n \in \mathbb{N}$, $K \geqslant 1$ there exists $\hat{C} = \hat{C}(n,K)$ such that for any $B \in \mathbb{C}^{n\times n}$ satisfying $\sup\limits_{\nu\geqslant 0} \|e^{\nu B}\| \leqslant K$ there holds*

$$\frac{|\lambda(B)|_{\max}}{|\lambda(B)|_{\min}} \leqslant \varkappa(B) \leqslant \hat{C}\frac{|\lambda(B)|_{\max}}{|\lambda(B)|_{\min}},$$

*where $|\lambda(B)|_{\max}$ and $|\lambda(B)|_{\min}$ are the maximal and minimal module of the eigenvalues of $B$.*

*Proof.* The first inequality follows from the fact that any norm of the matrix induced by a vector norm is greater than or equal to its spectral radius. The proof of the second inequality is based on the Kreiss matrix theorem [27], see Lemma 6 and Lemma 8 in [28]. □

**Theorem 5.18** ([28])**.** *For any $n \in \mathbb{N}$, $K \geqslant 1$ there exists $C = C(n,K)$ such that for any $B \in \mathbb{C}^{n\times n}$ satisfying $\sup\limits_{\nu\geqslant 0} \|e^{\nu B}\| \leqslant K$ there exists a matrix $U$ such that $\varkappa(U) \leqslant C$, the matrix $M = U^{-1}BU$ is block-diagonal, and each block $M_j$ of $M$ satisfies either $M_j = 0$ or $\varkappa(M_j) \leqslant C$.*

Without loss we assume the spectra of $M_j$ do not intersect. Indeed, if two blocks $M_j$ have common eigenvalues then by Lemma 5.17 they can be united into one block with the condition number not greater than the product of the conditional numbers of the original blocks multiplied by $\hat{C}$.

**Theorem 5.19.** *Let $A \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$ and $\sup\limits_{\nu\geqslant 0} \|\exp(\nu A(0))\| \leqslant K$. Let $A(0)$ have zero eigenvalue of algebraic multiplicity $n_0$. Then in a neighborhood of $\boldsymbol{\phi} = 0$ there holds $A(\boldsymbol{\phi}) = S(\boldsymbol{\phi})M(\boldsymbol{\phi})S^{-1}(\boldsymbol{\phi})$ where $S, M, S^{-1} \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$,*

$$M(\boldsymbol{\phi}) = \begin{pmatrix} M^{(1)}(\boldsymbol{\phi}) & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & M^{(m)}(\boldsymbol{\phi}) & 0 \\ 0 & \dots & 0 & M^{(0)}(\boldsymbol{\phi}) \end{pmatrix}, \tag{5.10}$$

*$M^{(0)}(\boldsymbol{\phi})$ is a $n_0 \times n_0$-matrix, $M^{(0)}(0) = 0$, $\varkappa(M^{(j)}(0)) \leqslant \delta$ for $j \neq 0$, $\|S(0)\| \leqslant \delta$, $\|S^{-1}(0)\| \leqslant \delta$, and $\delta$ depends on $n$ and $K$ only.*

*Proof.* By Theorem 5.18 there exists a matrix $Y$ such that $A(0) = YBY^{-1}$, where $B$ is a block-diagonal matrix with blocks $B_j$ of size $n_j$, $\varkappa(Y) \leqslant C(n,K)$, and $\varkappa(B_j) \leqslant C(n,K)$. Denote $B(\boldsymbol{\phi}) = Y^{-1}A(\boldsymbol{\phi})Y$.

Denote by $P_j(\boldsymbol{\phi})$ the sum of the total eigenprojections of the $\lambda$-groups for the matrix function $B(\boldsymbol{\phi})$ corresponding to $\sigma(B_j)$. By construction

$$P_j(0) = \begin{pmatrix} 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & I & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix},$$

where $I$ is the identity matrix of size $n_j$ and its position corresponds to the position of $B_j$ in the matrix $B$. By continuity the submatrix of $P_j(\boldsymbol{\phi})$ taking the same position as $B_j$ is non-degenerate in a neighborhood of $\boldsymbol{\phi} = 0$. Therefore, for a given $j$ the columns of $P_j(\boldsymbol{\phi})$ corresponding to the block $B_j$ form a basis in the sum of algebraic eigenspaces of $B(\boldsymbol{\phi})$ corresponding to $\sigma(B_j)$.

Introduce the matrix function $\tilde{S}(\boldsymbol{\phi})$ composed from these basis columns of all $P_j(\boldsymbol{\phi})$. By construction the matrix $M(\boldsymbol{\phi}) = \tilde{S}^{-1}(\boldsymbol{\phi})B(\boldsymbol{\phi})\tilde{S}(\boldsymbol{\phi})$ has the block-diagonal form (5.10), where $M^{(j)}(0) = B_j$ and $\tilde{S}(0) = I$. The matrix $\tilde{S}(\boldsymbol{\phi})$ is analytical in a neighborhood of $\boldsymbol{\phi} = 0$ since all of its columns are analytical by Lemma 5.16.

By continuity $\det\tilde{S}(\boldsymbol{\phi}) \neq 0$ in a neighborhood of $\boldsymbol{\phi} = 0$. The elements of $\tilde{S}^{-1}(\boldsymbol{\phi})$ are products of elements of $\tilde{S}(\boldsymbol{\phi})$ divided by $\det\tilde{S}(\boldsymbol{\phi})$, so $\tilde{S}^{-1}(\boldsymbol{\phi})$ is also holomorphic. Hence so is $M(\boldsymbol{\phi})$. Thus we get representation (5.10) with $S(\boldsymbol{\phi}) = Y\tilde{S}(\boldsymbol{\phi})$. □

## 6. Spectral analysis

The purpose of this section is to show that the error estimate for an arbitrary smooth solution is equivalent to the error estimate for a single complex exponent. This allows us to reduce our further analysis to a single wave and thus to a finite-dimensional analysis.

### *6.1. The basics*

Recall that $\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^d : T^*\boldsymbol{\alpha}/\pi \in \mathbb{Q}^d\}$. For $\beta > 0$ denote $\mathcal{A}_\beta = \{\boldsymbol{\phi} \in \mathcal{A} : T^*\boldsymbol{\phi} \in [-\beta, \beta)^d\}$. Let $l_2(\mathcal{A}_\pi, \mathbb{C}^{M^0})$ be the space of maps $w$ of $\mathcal{A}_\pi$ to $\mathbb{C}^{M^0}$ such that $\sum_{\boldsymbol{\phi}} \|w_{\boldsymbol{\phi}}\|^2 < \infty$ with the scalar product $(w, w') = \sum_{\boldsymbol{\phi}}(w_{\boldsymbol{\phi}}, w'_{\boldsymbol{\phi}})$. Let $V_F \subset l_2(\mathcal{A}_\pi, \mathbb{C}^{M^0})$ be the space of maps $w$ such that $w(\boldsymbol{\phi}) = 0$ for all but a finite number $\boldsymbol{\phi} \in \mathcal{A}_\pi$. Clearly $l_2(\mathcal{A}_\pi, \mathbb{C}^{M^0})$ is a Hilbert space, and $V_F$ is its dense subspace. Also denote $U = (T^*)^{-1}$.

Let $F : V_{per} \to V_F$ be the mapping taking each $f \in V_{per}^N$ to

$$F[f](\boldsymbol{\phi}) = \frac{1}{N^d} \sum_{\boldsymbol{\eta} \in \{0,\dots,N-1\}^d} f_{\boldsymbol{\eta}} \exp(-i\boldsymbol{\phi} \cdot T\boldsymbol{\eta}) \quad \text{for} \quad \boldsymbol{\phi} = 2\pi U \frac{\boldsymbol{k}}{N}, \quad \boldsymbol{k} \in \mathbb{Z}^d \cap [-N/2, N/2)^d, \tag{6.1}$$

and $F[f](\boldsymbol{\phi}) = 0$ otherwise. Note that $f \in V_{per}^N$ also belongs to $f \in V_{per}^{nN}$ for each $n \in \mathbb{N}$, but this does not affect the value $F[f]$. It is easy to see that $F$ is a linear mapping. This mapping is invertible. Indeed, for $w \in V_F$ let $F^{-1}[w]$ be the sequence with components

$$(F^{-1}[w])_{\boldsymbol{\eta}} = \sum_{\boldsymbol{\phi} \in \mathcal{A}_\pi} w(\boldsymbol{\phi}) \exp(i\boldsymbol{\phi} \cdot T\boldsymbol{\eta}). \tag{6.2}$$

Obviously, $F^{-1}[w] \in V_{per}^N$, where $N$ is the product of all denominators of $\boldsymbol{\phi}$ such that $w(\boldsymbol{\phi}) \neq 0$ (by definition, the number of such $\boldsymbol{\phi}$ is finite). It is easy to see that $F^{-1}[F[f]] = f$ and $F[F^{-1}[w]] = w$.

Denote

$$\delta_{\boldsymbol{\phi}_1,\boldsymbol{\phi}_2} = \begin{cases} 1, & \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2; \\ 0, & otherwise; \end{cases} \qquad \delta^{\text{mod}}_{\boldsymbol{\phi}_1,\boldsymbol{\phi}_2} = \sum_{\boldsymbol{c} \in \mathbb{Z}^d} \delta_{\boldsymbol{\phi}_1,\boldsymbol{\phi}_2 + 2\pi U\boldsymbol{c}} = \begin{cases} 1, & \exists \boldsymbol{c} \in \mathbb{Z}^d : \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 + 2\pi U\boldsymbol{c}; \\ 0, & otherwise. \end{cases}$$

**Lemma 6.1.** *If $f \in V_{per}$ has components $f_{\boldsymbol{\eta}} = y \exp(i\boldsymbol{\phi} \cdot T\boldsymbol{\eta})$ with some $\boldsymbol{\phi} \in \mathcal{A}$, $y \in \mathbb{C}^{M^0}$, then*

$$F[f](\boldsymbol{\phi}') = y\delta^{\text{mod}}_{\boldsymbol{\phi},\boldsymbol{\phi}'} \tag{6.3}$$

*for each $\boldsymbol{\phi}' \in \mathcal{A}_\pi$.*

The proof is obvious.

**Lemma 6.2.** *For $f, g \in V_{per}$ there holds $(F[f], F[g]) = (f, g)$ and thus $F$ is an isometry: $\|F[f]\| = \|f\|$.*

*Proof.* Let $f \in V_{per}^{N_1}$, $g \in V_{per}^{N_2}$, then $f, g \in V_{per}^N$ with $N = N_1N_2$. We have

$$(F[f], F[g]) = \sum_{\boldsymbol{k} \in \mathbb{Z}^d \cap [-N/2, N/2)^d} (F[f](2\pi U\boldsymbol{k}/N), F[g](2\pi U\boldsymbol{k}/N)) =$$

$$= \frac{1}{N^{2d}} \sum_{\boldsymbol{\eta} \in \{0,\dots,N-1\}^d} \sum_{\boldsymbol{\zeta} \in \{0,\dots,N-1\}^d} (f_{\boldsymbol{\eta}}, g_{\boldsymbol{\zeta}}) \sum_{\boldsymbol{k} \in \mathbb{Z}^d \cap [-N/2, N/2)^d} \exp\left(2\pi i \frac{\boldsymbol{k} \cdot (\boldsymbol{\eta} - \boldsymbol{\zeta})}{N}\right).$$

The sum over $\boldsymbol{k}$ is equal to zero for $\boldsymbol{\eta} \neq \boldsymbol{\zeta}$ and to $N^d$ for $\boldsymbol{\eta} = \boldsymbol{\zeta}$, thus

$$(F[f], F[g]) = \frac{1}{N^d} \sum_{\boldsymbol{\eta} \in (0,\dots,N-1)^d} (f_{\boldsymbol{\eta}}, g_{\boldsymbol{\eta}}) = (f, g).$$

□

**Corollary 6.3.** *The space $V_{per}$ is incomplete.*

*Proof.* Since $F$ is a bijective isometry from $V_{per}$ to $V_F$, and the latter is incomplete, so is the former. □

For each $\boldsymbol{\phi} \in \mathbb{C}^d$ put by definition

$$Z(\boldsymbol{\phi}) = \sum_{\boldsymbol{\eta} \in \mathcal{S} \subset \mathbb{Z}^d} Z_{\boldsymbol{\eta}} \exp\left(i\boldsymbol{\phi} \cdot T\boldsymbol{\eta}\right), \quad L(\boldsymbol{\phi}) = \sum_{\boldsymbol{\eta} \in \mathcal{S} \subset \mathbb{Z}^d} L_{\boldsymbol{\eta}} \exp\left(i\boldsymbol{\phi} \cdot T\boldsymbol{\eta}\right). \tag{6.4}$$

Functions $Z(\boldsymbol{\phi})$ and $L(\boldsymbol{\phi})$ are periodic with the periodic cell $(T^*)^{-1}[-\pi, \pi)^d$.

**Lemma 6.4.** *Let $Z$ and $L$ be given by* (2.9)*. Then for $f \in V_{per}$ there holds*

$$F[Zf](\boldsymbol{\phi}) = Z(\boldsymbol{\phi})F[f](\boldsymbol{\phi}), \quad F[Lf](\boldsymbol{\phi}) = L(\boldsymbol{\phi})F[f](\boldsymbol{\phi}), \tag{6.5}$$

$$F[\exp(-Z^{-1}Lt/h)f](\boldsymbol{\phi}) = \exp(-Z^{-1}(\boldsymbol{\phi})L(\boldsymbol{\phi})t/h)F[f](\boldsymbol{\phi}). \tag{6.6}$$

*Proof.* By linearity it is enough to check this for $f_{\boldsymbol{\eta}} = \exp(i\boldsymbol{\phi} \cdot T\boldsymbol{\eta})y$, with some $\boldsymbol{\phi} \in \mathcal{A}_\pi$, $y \in \mathbb{C}^{M^0}$. Then

$$(Zf)_{\boldsymbol{\eta}} = \sum_{\boldsymbol{\zeta} \in \mathcal{S}} Z_{\boldsymbol{\zeta}} \exp(i\boldsymbol{\phi} \cdot T(\boldsymbol{\eta} + \boldsymbol{\zeta}))y = \exp(i\boldsymbol{\phi} \cdot T\boldsymbol{\eta})Z(\boldsymbol{\phi})y,$$

so $F[Zf](\boldsymbol{\phi}') = Z(\boldsymbol{\phi})y\delta_{\boldsymbol{\phi},\boldsymbol{\phi}'} = Z(\boldsymbol{\phi})F[f](\boldsymbol{\phi}')$. The proof of the second identity is similar. To prove (6.6), note that $u(t) = \exp(-Z^{-1}Lt/h)f$ is the solution of (2.10) with the initial data $f$. Applying $F$ to (2.10) and using (6.5) we get

$$Z(\boldsymbol{\phi})\frac{d}{dt}F[u(t)](\boldsymbol{\phi}) + \frac{1}{h}L(\boldsymbol{\phi})F[u(t)](\boldsymbol{\phi}) = 0, \quad F[u(0)](\boldsymbol{\phi}) = F[f](\boldsymbol{\phi}).$$

Solving this equation we get (6.6). □

*6.2. Mappings to the mesh space*

For $\boldsymbol{\phi} \in \mathbb{R}^d$ there holds $e^{i\boldsymbol{\phi}\cdot\boldsymbol{r}} \in L_{2,per}(\mathbb{R}^d)$ iff $\boldsymbol{\phi} \in \mathcal{A}$. For a homogeneous mapping $\Pi_h$ and $\boldsymbol{\phi} \in \mathcal{A}$ put by definition

$$v(\boldsymbol{\phi}, \Pi_h) = \left(\Pi_1 e^{i\boldsymbol{\phi}\cdot\boldsymbol{r}}\right)_0, \tag{6.7}$$

where the subscript 0 means the block component at $\boldsymbol{\eta} = 0$. If $\Pi_h$ is a local mapping, then (6.7) defines $v(\boldsymbol{\phi}, \Pi_h)$ for each $\boldsymbol{\phi} \in \mathbb{C}^d$, and $v(\boldsymbol{\phi}, \Pi_h)$ is a holomorphic function.

**Lemma 6.5.** *Let $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathcal{A}$, $\boldsymbol{\phi}' \in \mathcal{A}_\pi$, and $\Pi_h$ be a homogeneous mapping. Then for $\boldsymbol{\alpha} = \boldsymbol{\phi}/h$, $\boldsymbol{\beta} = \boldsymbol{\psi}/h$ there holds $\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}} \in V_{per}$ and*

$$F[\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}](\boldsymbol{\phi}') = \delta^{\text{mod}}_{\boldsymbol{\phi},\boldsymbol{\phi}'}\, v(\boldsymbol{\phi}, \Pi_h), \tag{6.8}$$

$$(\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h e^{i\boldsymbol{\beta}\cdot\boldsymbol{r}}) = \delta^{\text{mod}}_{\boldsymbol{\phi},\boldsymbol{\psi}}\, (v(\boldsymbol{\phi}, \Pi_h), v(\boldsymbol{\psi}, \Pi_h)). \tag{6.9}$$

*Proof.* By definition,

$$(\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}})_{\boldsymbol{\eta}} = (\Pi_h e^{i\boldsymbol{\alpha}\cdot(\boldsymbol{r}+hT\boldsymbol{\eta})})_0 = e^{i\boldsymbol{\alpha}\cdot hT\boldsymbol{\eta}}(\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}})_0 = e^{i\boldsymbol{\phi}\cdot T\boldsymbol{\eta}}\, v(\boldsymbol{\phi}, \Pi_h).$$

Applying Lemma 6.1 we get (6.8). By Lemma 6.2

$$(\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h e^{i\boldsymbol{\beta}\cdot\boldsymbol{r}}) = \sum_{\boldsymbol{\psi}' \in \mathcal{A}_\pi} (F[\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}](\boldsymbol{\psi}'), F[\Pi_h e^{i\boldsymbol{\beta}\cdot\boldsymbol{r}}](\boldsymbol{\psi}')) =$$

$$= \sum_{\boldsymbol{\psi}' \in \mathcal{A}_\pi} \delta^{\text{mod}}_{\boldsymbol{\phi},\boldsymbol{\psi}'}\delta^{\text{mod}}_{\boldsymbol{\psi},\boldsymbol{\psi}'}(v(\boldsymbol{\phi}, \Pi_h), v(\boldsymbol{\psi}, \Pi_h)) = \delta^{\text{mod}}_{\boldsymbol{\phi},\boldsymbol{\psi}}\, (v(\boldsymbol{\phi}, \Pi_h), v(\boldsymbol{\psi}, \Pi_h)).$$

□

**Lemma 6.6.** *Let $G \subset \mathcal{A}_\pi$ be symmetric with respect to the origin. Let $v$ be a map of $G$ to $\mathbb{C}^{M^0}$ such that $v(-\boldsymbol{\phi}) = \overline{v(\boldsymbol{\phi})}$. Then there exists a homogeneous mapping $\Pi_h$ of $L_{2,per}(\mathbb{R}^d)$ to $V_{per}$ such that $v(\boldsymbol{\phi}) = v(\boldsymbol{\phi}, \Pi_h)$ for $\boldsymbol{\phi} \in G$ and $\|\Pi_h\| \leqslant \sup\limits_G \|v(\boldsymbol{\phi})\|$.*

*Proof.* Extend $v$ to $\mathcal{A}_\pi \setminus G$ by zero. Let $f \in L_{2,per}(\mathbb{R}^d)$ have Fourier series $\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}} f_{\boldsymbol{\alpha}} \exp(i\boldsymbol{\alpha}\cdot\boldsymbol{r})$. Note that there is only finitely many nonzero $f_{\boldsymbol{\alpha}}$ in any bounded part of the frequency domain. Define $w_f \in V_F$ by $w_f(\boldsymbol{\phi}) = v(\boldsymbol{\phi}) f_{\boldsymbol{\phi}/h}$. Let $\Pi_h$ is a linear mapping of $L_{2,per}(\mathbb{R}^d)$ to $V_{per}$ defined by $\Pi_h f = F^{-1}[w_f]$. Since

$$\|\Pi_h f\|^2 = \|F^{-1}[w_f]\|^2 = \|w_f\|^2 = \sum_{\boldsymbol{\phi}\in G} \|v(\boldsymbol{\phi}) f_{\boldsymbol{\phi}/h}\|^2 \leqslant \sup_G \|v(\boldsymbol{\phi})\|^2 \sum_{\boldsymbol{\phi}\in\mathcal{A}} |f_{\boldsymbol{\phi}}|^2 = \sup_G \|v(\boldsymbol{\phi})\|^2 \|f\|^2,$$

there holds $\|\Pi_h\| \leqslant \sup\limits_G \|v(\boldsymbol{\phi})\|$.

Now we check that $\Pi_h$ is a homogeneous mapping. To prove (P3), consider a real-valued function $f \in L_{2,per}(\mathbb{R}^d)$. Then $f_{-\boldsymbol{\alpha}} = \overline{f_{\boldsymbol{\alpha}}}$ and $w_f(-\boldsymbol{\phi}) = \overline{w_f(\boldsymbol{\phi})}$. From (6.2), $F^{-1}[w_f]$ is real-valued, i. e. (P3) is proved. To prove (P1) and (P2), it is enough to consider a single exponent $f(\boldsymbol{r}) = \exp(i\boldsymbol{\alpha}\cdot\boldsymbol{r})$, $\boldsymbol{\alpha} \in \mathcal{A}$. Clearly,

$$(\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}})_{\boldsymbol{\eta}} = \begin{cases} v(\boldsymbol{\alpha}h)\exp(i\boldsymbol{\alpha}h\cdot T\boldsymbol{\eta}), & \boldsymbol{\alpha}h \in G, \\ 0, & \boldsymbol{\alpha}h \notin G. \end{cases} \tag{6.10}$$

For

$$g(\boldsymbol{r}) = f(\boldsymbol{r} + hT\boldsymbol{\zeta}) = \exp(i\boldsymbol{\alpha}\cdot hT\boldsymbol{\zeta}) f(\boldsymbol{r}),$$

we have

$$(\Pi_h g)_{\boldsymbol{\eta}} = \exp(i\boldsymbol{\alpha}\cdot hT\boldsymbol{\zeta}) \exp(ih\boldsymbol{\alpha}\cdot T\boldsymbol{\eta}) v(h\boldsymbol{\alpha}) = \exp(ih\boldsymbol{\alpha}\cdot T(\boldsymbol{\eta}+\boldsymbol{\zeta})) v(h\boldsymbol{\alpha}) = (\Pi_h f)_{\boldsymbol{\eta}+\boldsymbol{\zeta}}.$$

The property (P1) is proved. For $g(\boldsymbol{r}) = f(\boldsymbol{r}/h) = \exp(i\boldsymbol{\alpha}\cdot\boldsymbol{r}/h)$ we have

$$(\Pi_h g)_{\boldsymbol{\eta}} = \exp(ih\boldsymbol{\alpha}\cdot T\boldsymbol{\eta}/h) v(h\boldsymbol{\alpha}/h) = (\Pi_1 f)_{\boldsymbol{\eta}}.$$

The property (P2) is proved.

The property $v(\boldsymbol{\phi}, \Pi_h) = v(\boldsymbol{\phi})$ follows directly from (6.10). □

**Lemma 6.7.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with the kernels $\mu_\xi$ and $\hat{\mu}_\xi$, correspondingly, and $\langle\hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$. Suppose $W(\boldsymbol{\phi})$ maps a neighborhood of zero in $\mathbb{R}^d$ to $\mathbb{C}^{M^0}$, has $q+1$ continuous derivatives at $\boldsymbol{\phi} = 0$, satisfies $W(-\boldsymbol{\phi}) = \overline{W(\boldsymbol{\phi})}$, and $\|v(\boldsymbol{\phi}, \Pi_h) - W(\boldsymbol{\phi})\| = O(|\boldsymbol{\phi}|^p)$ as $\boldsymbol{\phi} \to 0$, where $p, q \in \mathbb{N}$. Then there exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$, $p \leqslant |\boldsymbol{m}| \leqslant q$, such that for the mapping $\Pi_h^{(p,q)}$ given by* (2.12) *there holds*

$$v(\boldsymbol{\phi}, \Pi_h^{(p,q)}) = W(\boldsymbol{\phi}) + O(|\boldsymbol{\phi}|^{q+1}). \tag{6.11}$$

*Proof.* Let the diagonal matrix $\mathfrak{C}(\boldsymbol{\phi})$ be defined in a neighborhood of $\boldsymbol{\phi} = 0$ by

$$W(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h) = \mathfrak{C}(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \mathcal{P}_h). \tag{6.12}$$

Since $(v(\boldsymbol{\phi}, \mathcal{P}_h))_\xi \to \langle\hat{\mu}_\xi, 1\rangle \neq 0$ as $\boldsymbol{\phi} \to 0$, then in a neighborhood of $\boldsymbol{\phi} = 0$ the matrix function $\mathfrak{C}(\boldsymbol{\phi})$ is well-defined and $\|\mathfrak{C}(\boldsymbol{\phi})\| = O(|\boldsymbol{\phi}|^p)$. Put

$$\mathfrak{C}^{(\boldsymbol{m})} = \frac{1}{\boldsymbol{m}!} \frac{1}{i^{|\boldsymbol{m}|}} D^{\boldsymbol{m}} \mathfrak{C}(\boldsymbol{\phi})|_{\boldsymbol{\phi}=0}. \tag{6.13}$$

Taking complex conjugation from (6.12) we get

$$W(-\boldsymbol{\phi}) - v(-\boldsymbol{\phi}, \Pi_h) = \overline{\mathfrak{C}(\boldsymbol{\phi})} v(-\boldsymbol{\phi}, \mathcal{P}_h),$$

thus $\overline{\mathfrak{C}(\boldsymbol{\phi})} = \mathfrak{C}(-\boldsymbol{\phi})$ and

$$\overline{\mathfrak{C}^{(\boldsymbol{m})}} = \frac{1}{\boldsymbol{m}!}\frac{1}{(-i)^{|\boldsymbol{m}|}}\, D^{\boldsymbol{m}}(\overline{\mathfrak{C}(\boldsymbol{\phi})})\Big|_{\boldsymbol{\phi}=0} = \frac{1}{\boldsymbol{m}!}\frac{1}{(-i)^{|\boldsymbol{m}|}}\, D^{\boldsymbol{m}}(\mathfrak{C}(-\boldsymbol{\phi}))|_{\boldsymbol{\phi}=0} = \mathfrak{C}^{(\boldsymbol{m})},$$

so $\mathfrak{C}^{(\boldsymbol{m})}$ is real-valued. By definition

$$v(\boldsymbol{\phi}, \Pi_h^{(p,q)}) = v(\boldsymbol{\phi}, \Pi_h) + \left[\sum_{p\leqslant|\boldsymbol{m}|\leqslant q} \mathfrak{C}^{(\boldsymbol{m})}(i\boldsymbol{\phi})^{\boldsymbol{m}}\right] v(\boldsymbol{\phi}, \mathcal{P}_h).$$

Since $\|\mathfrak{C}(\boldsymbol{\phi})\| = O(|\boldsymbol{\phi}|^p)$, there holds $\mathfrak{C}^{(\boldsymbol{m})} = 0$ whenever $|\boldsymbol{m}| < p$. Thus the sum inside the brackets is the Taylor polynomial of the function $\mathfrak{C}(\boldsymbol{\phi})$ of order $q$ and

$$v(\boldsymbol{\phi}, \Pi_h^{(p,q)}) = v(\boldsymbol{\phi}, \Pi_h) + (\mathfrak{C}(\boldsymbol{\phi}) + O(|\boldsymbol{\phi}|^{q+1}))v(\boldsymbol{\phi}, \mathcal{P}_h).$$

Combining this with (6.12) we get (6.11). □

**Lemma 6.8.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with the kernels $\mu_\xi$ and $\hat{\mu}_\xi$, correspondingly, and $\langle\hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$. Let $\boldsymbol{e} \in \mathring{\Omega}$. Let $W(\psi)$ have $q+1$ continuous derivatives at $\psi = 0$, $W(-\psi) = \overline{W(\psi)}$, and $\|v(\psi\boldsymbol{e}, \Pi_h) - W(\psi)\| = O(|\psi|^p)$ as $\psi \to 0$, where $p, q \in \mathbb{N}$. Then there exist real-valued diagonal matrices $\mathfrak{C}_{\boldsymbol{e}}^{(m)}$, $m = p, \dots, q$, such that for the mapping $\Pi_{h,\boldsymbol{e}}^{(p,q)}$ given by* (2.13) *there holds*

$$v(\psi\boldsymbol{e}, \Pi_{h,\boldsymbol{e}}^{(p,q)}) = w(\psi) + O(|\psi|^{q+1}).$$

The proof is similar to the previous Lemma.

*6.3. Truncation error, stability, and accuracy*

We need the following auxiliary result.

**Lemma 6.9.** *Let $\Pi_h$ be a homogeneous mapping of $H^q_{per}(\mathbb{R}^d)$ (or $C^q_{per}(\mathbb{R}^d)$) to $V_{per}$. Then the operator $F_h^1$ defined as $F_h^1 f = h\epsilon_h(f, \Pi_h)$ is a homogeneous mapping of $H^{q+1}_{per}(\mathbb{R}^d)$ (or $C^{q+1}_{per}(\mathbb{R}^d)$). For each $t \geqslant 0$ the operator $F_h^2$ defined as $F_h^2 f = \varepsilon_h(t, f, \Pi_h)$ is a homogeneous mapping of $H^q_{per}(\mathbb{R}^d)$ (or $C^q_{per}(\mathbb{R}^d)$) to $V_{per}$. Besides,*

$$\|F_h^1\|_{(q+1,h,*)} \leqslant \|\Pi_h\|_{(q,h,*)}(\|Z\||\boldsymbol{\omega}| + \|L\|), \quad \|F_h^2\|_{(q,h,*)} \leqslant (K+1)\|\Pi_h\|_{(q,h,*)},$$

*where $K$ is the stability constant of the scheme, and $*$ means either nothing or infinity depending on the case.*

*Proof.* By the definitions of $\epsilon_h(f, \Pi_h)$ (see (2.14)) and $\varepsilon_h(t, f, \Pi_h)$ (see (2.15)) it immediately follows that the mappings $F_h^1$ and $F_h^2$ are homogeneous.

For each $v_0 \in H^{q+1}_{per}(\mathbb{R}^d)$ (or $C^{q+1}_{per}(\mathbb{R}^d)$) by Lemma 5.7 there holds

$$\|F_h^1 v_0\| = \|h\epsilon_h(v_0, \Pi_h)\| = \|-hZ\Pi_h(\boldsymbol{\omega}\cdot\nabla v_0) + L\Pi_h v_0\| \leqslant (\|Z\||\boldsymbol{\omega}| + \|L\|)\|\Pi_h\|_{(q,h,*)}\|v_0\|_{(q+1,h,*)}.$$

Now let $v_0 \in H^q_{per}(\mathbb{R}^d)$ (or $C^q_{per}(\mathbb{R}^d)$), $v(t, \boldsymbol{r}) = v_0(\boldsymbol{r} - \boldsymbol{\omega}t)$, and $u(t)$ be the solution of (2.8) with the initial data $u(0) = \Pi_h v_0$. Using $\|v_0\| = \|v(t, \cdot)\|$ we get

$$\|F_h^2 v_0\| = \|\varepsilon_h(t, v_0, \Pi_h)\| = \|u(t) - \Pi_h v(t, \cdot)\| \leqslant K\|\Pi_h v_0\| + \|\Pi_h v(t, \cdot)\| \leqslant (K+1)\,\|\Pi_h\|_{(q,h,*)}\,\|v_0\|_{(q,h,*)},$$

where we have used the stability condition. □

By definition, put

$$A(\boldsymbol{\phi}) = -Z^{-1}(\boldsymbol{\phi})L(\boldsymbol{\phi}) + i\boldsymbol{\omega}\cdot\boldsymbol{\phi} I, \tag{6.14}$$

$$\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = A(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \Pi_h) \tag{6.15}$$

$$\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h) = \left(e^{\nu A(\boldsymbol{\phi})} - I\right) v(\boldsymbol{\phi}, \Pi_h). \tag{6.16}$$

where $I$ is the identity matrix of size $|M^0|$, $Z(\boldsymbol{\phi})$ and $L(\boldsymbol{\phi})$ are given by (6.4), and $v(\boldsymbol{\phi}, \Pi_h)$ is given by (6.7).

**Lemma 6.10.** *Let $\Pi_h$ be a homogeneous mapping, $\boldsymbol{\alpha} \in \mathcal{A}$, $\boldsymbol{\phi} = \boldsymbol{\alpha} h$. Then there holds*

$$\|\epsilon_h(e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| = \frac{1}{h}\|Z(\boldsymbol{\phi})\, \hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\|, \tag{6.17}$$

$$\|\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| = \|\hat{\varepsilon}(\boldsymbol{\phi}, t/h, \Pi_h)\|. \tag{6.18}$$

*Proof.* First we obtain the spectral representation of the truncation error. By (6.5) and (6.8), for $\boldsymbol{\phi}' \in \mathcal{A}_\pi$,

$$\begin{gathered} F[\epsilon_h(e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)](\boldsymbol{\phi}') = \left(-(\boldsymbol{\omega}\cdot i\boldsymbol{\alpha})Z(\boldsymbol{\phi}') + \frac{1}{h}L(\boldsymbol{\phi}')\right) F[\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}](\boldsymbol{\phi}') = \\ = -\frac{1}{h}Z(\boldsymbol{\phi}')\left((\boldsymbol{\omega}\cdot i\boldsymbol{\alpha}h)I - Z^{-1}(\boldsymbol{\phi}')L(\boldsymbol{\phi}')\right)\delta^{\mathrm{mod}}_{\boldsymbol{\phi},\boldsymbol{\phi}'}\, v(\boldsymbol{\phi}, \Pi_h) = -\frac{1}{h}Z(\boldsymbol{\phi})\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\delta^{\mathrm{mod}}_{\boldsymbol{\phi},\boldsymbol{\phi}'}. \end{gathered} \tag{6.19}$$

The last identity is due to the $2\pi U\mathbb{Z}^d$-periodicity of $L$ and $Z$. Using Lemma 6.2 we get (6.17).

Now we move to the solution error. By definition,

$$\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h) = u(t) - e^{-i\boldsymbol{\alpha}\cdot\boldsymbol{\omega}t}\, \Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}},$$

where $u(t) = \exp(-Z^{-1}Lt/h)\, \Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$ (see (2.15)). By (6.6), for $\boldsymbol{\phi}' \in \mathcal{A}_\pi$ there holds

$$F[u(t)](\boldsymbol{\phi}') = \exp\left(-Z^{-1}(\boldsymbol{\phi}')L(\boldsymbol{\phi}')\frac{t}{h}\right) F[\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}](\boldsymbol{\phi}').$$

Thus by (6.8) we get

$$F[\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)](\boldsymbol{\phi}') = \left[\exp\left(-Z^{-1}(\boldsymbol{\phi}')L(\boldsymbol{\phi}')\frac{t}{h}\right) - \exp\left(-i\boldsymbol{\alpha}\cdot\boldsymbol{\omega}t\right)\right] \delta^{\mathrm{mod}}_{\boldsymbol{\phi},\boldsymbol{\phi}'}\, v(\boldsymbol{\phi}, \Pi_h).$$

By the $2\pi U\mathbb{Z}^d$-periodicity of $Z(\boldsymbol{\phi})$ and $L(\boldsymbol{\phi})$ we replace $\boldsymbol{\phi}'$ by $\boldsymbol{\phi}$ in their arguments and finally obtain

$$F[\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)](\boldsymbol{\phi}') = \exp\left(-i\boldsymbol{\alpha}\cdot\boldsymbol{\omega}t\right) \hat{\varepsilon}(\boldsymbol{\phi}, t/h, \Pi_h)\delta^{\mathrm{mod}}_{\boldsymbol{\phi},\boldsymbol{\phi}'}.$$

Now Lemma 6.2 yields (6.18). □

**Lemma 6.11.** *Consider a scheme of the form* (2.8) *and a homogeneous mapping $\Pi_h$. Let the error estimate* (2.17) *hold. Then for $\boldsymbol{\phi} \in \mathcal{A}$ there holds*

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant C_1|\boldsymbol{\phi}|^P + C_2\nu|\boldsymbol{\phi}|^{Q+1} + C_3|\boldsymbol{\phi}|^R, \tag{6.20}$$

*where $C_1$, $C_2$, $C_3$ are the same as in* (2.17)*. If $\Pi_h$ is a local mapping, then* (6.20) *holds for each $\boldsymbol{\phi} \in \mathbb{R}^d$.*

*Proof.* Using (2.17) for $v_0 = e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$, $\boldsymbol{\alpha} = \boldsymbol{\phi}/h$, we get

$$\|\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant C_1 h^P|\boldsymbol{\alpha}|^P + C_2 t h^Q|\boldsymbol{\alpha}|^{Q+1} + C_3 h^R|\boldsymbol{\alpha}|^R,$$

and it remains to use Lemma 6.10. If $\Pi_h$ is a local mapping, then (6.20) extends to $\boldsymbol{\phi} \in \mathbb{R}^d$ by continuity. □

395 **Lemma 6.12.** *The optimal value of the order of the truncation error in the sense of a local mapping is an integer.*

*Proof.* By Lemma 6.10 this order is less by one then the order of smallness of $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)$ as $\boldsymbol{\phi}$ tends to zero, which is integer since $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)$ is a holomorphic function of $\boldsymbol{\phi}$. □

**Lemma 6.13.** *A scheme* (2.8) *is stable with the constant $K$ if and only if for each $\boldsymbol{\phi} \in \mathcal{A}$ and each $\nu > 0$ there holds*

$$\|\exp(A(\boldsymbol{\phi})\nu)\| \leqslant K. \tag{6.21}$$

*Proof.* For an operator $Y : V_F \to V_F$ such that $(Yw)(\boldsymbol{\phi}) = y(\boldsymbol{\phi})w(\boldsymbol{\phi})$ there obviously holds $\|Y\| = \sup_{\boldsymbol{\phi}\in\mathcal{A}_\pi} \|y(\boldsymbol{\phi})\|$. Using this fact for $y(\boldsymbol{\phi}) = \exp(-Z^{-1}(\boldsymbol{\phi})L(\boldsymbol{\phi})t/h)$ and Lemmas 6.2 and 6.4, we get

$$\left\|\exp\left(-Z^{-1}L\frac{t}{h}\right)\right\| = \sup_{\boldsymbol{\phi}\in\mathcal{A}_\pi} \left\|\exp\left(-Z^{-1}(\boldsymbol{\phi})L(\boldsymbol{\phi})\frac{t}{h}\right)\right\|.$$

By continuity and $2\pi U\mathbb{Z}^d$-periodicity of $Z^{-1}(\boldsymbol{\phi})$ and $L(\boldsymbol{\phi})$, and finally by the definition of $A(\boldsymbol{\phi})$ we get

$$\left\|\exp\left(-Z^{-1}L\frac{t}{h}\right)\right\| = \sup_{\boldsymbol{\phi}\in\mathcal{A}} \left\|\exp\left(-Z^{-1}(\boldsymbol{\phi})L(\boldsymbol{\phi})\frac{t}{h}\right)\right\| =$$
$$= \sup_{\boldsymbol{\phi}\in\mathcal{A}} \left\|\exp\left(A(\boldsymbol{\phi})\frac{t}{h}\right)\exp(-i\boldsymbol{\omega}\cdot\boldsymbol{\phi}I)\right\| = \sup_{\boldsymbol{\phi}\in\mathcal{A}} \left\|\exp\left(A(\boldsymbol{\phi})\frac{t}{h}\right)\right\|.$$

Taking the supremum over $t, h > 0$, we get

$$\sup_{t,h>0} \left\|\exp\left(-Z^{-1}L\frac{t}{h}\right)\right\| = \sup_{\nu>0}\sup_{\boldsymbol{\phi}\in\mathcal{A}} \|\exp(A(\boldsymbol{\phi})\nu)\|.$$

It remains to recall the definition of the stability constant of a scheme. □

*6.4. From a sine wave to an arbitrary smooth solution*

400 The section is a toolbox for extending various estimates from single Fourier modes to arbitrary smooth solutions.

**Lemma 6.14.** *Let $F_h$ be a homogeneous mapping of $H^r_{per}(\mathbb{R}^d)$ to $V_{per}$, $r \geqslant 0$. Let $p(s) = \sum_{j=1}^J c_j s^{p_j}$ with some $J \in \mathbb{N}$ and $c_j, p_j \geqslant 0$. Suppose there exists $0 < \beta \leqslant \pi$ such that for each $h$ and each $\boldsymbol{\alpha} \in \mathcal{A}_{\beta/h}$ there holds*

$$\|F_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}\| \leqslant p(|\boldsymbol{\alpha}|).$$

*Then for each $w \in H^{\max\{r,p_1,\dots,p_J\}}_{per}(\mathbb{R}^d)$ there holds*

$$\|F_h w\| \leqslant \sum_{j=1}^J c_j \|\nabla^{p_j} w\| + 2h^r c^r \|F_h\|_{(r,h)} \|\nabla^r w\|, \tag{6.22}$$

*where $c = 1 + \beta^{-1}\|T^*\|_2$ and $\|F_h\|_{(r,h)} = \sup \|F_h f\|/\|f\|_{(r,h)}$.*

*Proof.* Consider a function $w \in H^{\max\{r,p_1,\dots,p_J\}}_{per}(\mathbb{R}^d)$. Since $w$ is periodic it has the Fourier representation (2.4):

$$w = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} w_{\boldsymbol{\alpha}} \exp(i\boldsymbol{\alpha}\cdot\boldsymbol{r}).$$

Recall that $\mathcal{A}_{\beta/h} = \{\boldsymbol{\alpha} \in \mathcal{A} : T^*\boldsymbol{\alpha} \in [-\beta/h, \beta/h)^d\}$. By definition, put

$$S_N(\boldsymbol{r}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\beta/h}} w_{\boldsymbol{\alpha}} \exp(i\boldsymbol{\alpha}\cdot\boldsymbol{r}).$$

Obviously, $F_h w = F_h S_N + F_h(w - S_N)$. By Lemma 6.5 for $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{A}_{\beta/h}$, $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ there holds $(F_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, F_h e^{i\boldsymbol{\alpha}'\cdot\boldsymbol{r}}) = 0$. By assumption and using the Minkowski inequality for sums (the triangle inequality for $l_2$) we get

$$\|F_h S_N\|^2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\beta/h}} |w_{\boldsymbol{\alpha}}|^2 \left\|F_h \exp\left(i\boldsymbol{\alpha}\cdot\boldsymbol{r}\right)\right\|^2 \leqslant \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\beta/h}} |w_{\boldsymbol{\alpha}}|^2 \left(p(|\boldsymbol{\alpha}|)\right)^2 =$$

$$= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\beta/h}} \left(\sum_{j=1}^{J} c_j |w_{\boldsymbol{\alpha}}| \, |\boldsymbol{\alpha}|^{p_j}\right)^2 \leqslant \left(\sum_{j=1}^{J} c_j \left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\beta/h}} |w_{\boldsymbol{\alpha}}|^2 |\boldsymbol{\alpha}|^{2p_j}\right)^{1/2}\right)^2.$$

Extending the limits of the sum to $\boldsymbol{\alpha} \in \mathcal{A}$ we obtain

$$\|F_h S_N\| \leqslant \sum_{j=1}^{J} c_j \left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 |\boldsymbol{\alpha}|^{2p_j}\right)^{1/2} = \sum_{j=1}^{J} c_j \|\nabla^{p_j} w\|. \tag{6.23}$$

Now consider $F_h(w - S_N)$. By the definition of $S_N$ we have

$$w - S_N = \sum_{\boldsymbol{\alpha}\in\mathcal{A}\setminus\mathcal{A}_{\beta/h}} w_{\boldsymbol{\alpha}} \exp\left(i\boldsymbol{\alpha}\cdot\boldsymbol{r}\right).$$

For $\boldsymbol{\alpha} \in \mathcal{A} \setminus \mathcal{A}_{\beta/h}$ we have

$$|\boldsymbol{\alpha}|^{-1} \leqslant \|T^*\|_2 \, |T^*\boldsymbol{\alpha}|^{-1} \leqslant \|T^*\|_2 \, \|T^*\boldsymbol{\alpha}\|_\infty^{-1} \leqslant h\beta^{-1}\|T^*\|_2 \leqslant hc.$$

Thus

$$\|w - S_N\|^2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}\setminus\mathcal{A}_{\beta/h}} |w_{\boldsymbol{\alpha}}|^2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}\setminus\mathcal{A}_{\beta/h}} \frac{|w_{\boldsymbol{\alpha}}|^2 |\boldsymbol{\alpha}|^{2r}}{|\boldsymbol{\alpha}|^{2r}} \leqslant (hc)^{2r} \|\nabla^r w\|^2.$$

Taking into account that $c \geqslant 1$ we obtain

$$\|w - S_N\|_{(r,h)}^2 \leqslant \|\nabla^r w\|^2 \left((hc)^{2r} + h^{2r}\right) \leqslant 2\|\nabla^r w\|^2 (hc)^{2r}$$

and

$$\|F_h(w - S_N)\| \leqslant \sqrt{2}\, \|F\|_{(r,h)} \|\nabla^r w\| h^r c^r.$$

Combining this (6.23) we get (6.22). □

**Theorem 6.15.** *Consider a scheme of the form* (2.8) *and a homogeneous mapping* $\Pi_h$. *[Let* $\boldsymbol{e} \in \mathring{\Omega}$.*] Suppose there exist* $C \geqslant 0$, $P_A \geqslant -1$, *and* $0 < \beta \leqslant \pi$ *such that there holds*

$$\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant C|\boldsymbol{\phi}|^{P_A+1} \tag{6.24}$$

*for each* $|\boldsymbol{\phi}| \in \mathcal{A}_\beta$ *[aligned with* $\boldsymbol{e}$*].*

*1) If* $\Pi_h$ *is a bounded homogeneous mapping of* $H^q_{per}(\mathbb{R}^d)$ *to* $V_{per}$, *then for each* $v_0 \in H^{r+1}_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$ *and* $r \geqslant \max\{P_A, q\}$ *there holds*

$$\|\epsilon_h(v_0, \Pi_h)\| \leqslant Ch^{P_A}\|\nabla^{P_A+1} v_0\| + \tilde{C}h^r \|\nabla^{r+1} v_0\|, \tag{6.25}$$

*where* $\tilde{C} = 4c^{r+1}\|\Pi_h\|_{(q,h)}(\|Z\||\boldsymbol{\omega}| + \|L\|)$, $c = 1 + \beta^{-1}\|T^*\|_2$.

*2) If* $\Pi_h$ *is a local mapping with* $\mu_\xi \in (C^q(G))^*$, $q \in \mathbb{N} \cup \{0\}$, *then there exists* $\tilde{C}$ *such that for each* $v_0 \in C^{r+1}_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$, $r \in \mathbb{N}$, $r \geqslant \max\{P_A, q\}$, *there holds*

$$\|\epsilon_h(v_0, \Pi_h)\| \leqslant Ch^{P_A}\|\nabla^{P_A+1} v_0\| + \tilde{C}h^r \|\nabla^{r+1} v_0\|_\infty.$$

405

*Proof.* If $\boldsymbol{e}$ is specified, let $\mathfrak{T} : L_{2,per}(\mathbb{R}^d) \to L_{2,per}(\mathbb{R}^d)$ be the projection operator taking each $e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$ to $e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$ if $\boldsymbol{\alpha}$ is aligned with $\boldsymbol{e}$ and to zero otherwise. If $\boldsymbol{e}$ is not specified, $\mathfrak{T}$ is the identity operator. Clearly, $\|\mathfrak{T}\| = 1$ and maps $H^{\sigma}_{per}(\mathbb{R}^d)$ to $H^{\sigma}_{per}(\mathbb{R}^d)$ for each $\sigma$.

To prove the first statement, put $F_h v_0 = h\epsilon_h(v_0, \Pi_h)$. By (6.24) and Lemma 6.10 we have $\|F_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}\| \leqslant C|h\boldsymbol{\alpha}|^{P_A+1}$ for each $\boldsymbol{\alpha} \in \mathcal{A}_{\beta/h}$ [aligned with $\boldsymbol{e}$]. By Lemma 6.9 we have $\|F_h\|_{(q+1,h)} \leqslant \|\Pi_h\|_{(q,h)}(\|Z\||\boldsymbol{\omega}| + \|L\|)$. Since $r \geqslant q$ then by Lemma 5.6 we have $\|F_h\|_{(r+1,h)} \leqslant 2\|F_h\|_{(q+1,h)}$. Since $\|\mathfrak{T}\| = 1$ we have

$$\|F_h\mathfrak{T}\|_{(r+1,h)} \leqslant 2\|\Pi_h\|_{(q,h)}(\|Z\||\boldsymbol{\omega}| + \|L\|).$$

Using Theorem 6.14 for the mapping $F_h\mathfrak{T}$ and $p(s) = C(hs)^{P_A+1}$ we get (6.25).

Now we prove the second statement. By Lemma 5.9 there exists a local mapping $\tilde{\Pi}_h$ with $\tilde{\mu}_\xi \in L_2(\tilde{G})$, $\tilde{G} = G + B_1(0)$, such that for $k \in \{r, r+1\}$ and each $f \in C^k_{per}(\mathbb{R}^d)$ there holds $\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant ch^k\|\nabla^k f\|_\infty$. By Lemma 2.3, $\tilde{\Pi}_h$ is a bounded homogeneous mapping of $L_{2,per}(\mathbb{R}^d)$ to $V_{per}$. Then for each $f \in C^{r+1}_{per}(\mathbb{R}^d)$ there holds

$$\begin{aligned} &\|\epsilon_h(f, \tilde{\Pi}_h) - \epsilon_h(f, \Pi_h)\| \leqslant \|Z\|\,\|(\tilde{\Pi}_h - \Pi_h)(\boldsymbol{\omega}\cdot\nabla)f\| + h^{-1}\|L\|\,\|(\tilde{\Pi}_h - \Pi_h)f\| \leqslant \\ &\leqslant ch^r\|Z\|\,\|\nabla^r(\boldsymbol{\omega}\cdot\nabla)f\|_\infty + ch^r\|L\|\,\|\nabla^{r+1}f\|_\infty \leqslant ch^r(\|Z\|\,|\boldsymbol{\omega}| + \|L\|)\|\nabla^{r+1}f\|_\infty. \end{aligned} \tag{6.26}$$

The last inequality in this chain is by Lemma 5.7. Particularly, this holds for $f = e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$, so by the triangle inequality for each $\boldsymbol{\alpha} \in \mathcal{A}_{\beta/h}$ [aligned with $\boldsymbol{e}$] we get

$$\|\epsilon_h(e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \tilde{\Pi}_h)\| \leqslant C|\boldsymbol{\alpha}|^{P_A+1}h^{P_A} + \tilde{c}|\boldsymbol{\alpha}|^{r+1}h^r.$$

Put $F_h v_0 = h\epsilon_h(v_0, \tilde{\Pi}_h)$. Since $\|\mathfrak{T}\| = 1$, by Lemma 5.6 and Lemma 6.9 we have

$$\|F_h\mathfrak{T}\|_{(r+1,h)} \leqslant \|F_h\|_{(r+1,h)} \leqslant \|\tilde{\Pi}_h\|_{(q,h)}(\|Z\||\boldsymbol{\omega}| + \|L\|) \leqslant \|\tilde{\Pi}_h\|(\|Z\||\boldsymbol{\omega}| + \|L\|).$$

Using Lemma 6.14 for the mapping $F_h\mathfrak{T}$ and $p(s) = C(hs)^{P_A+1} + \tilde{c}(hs)^{r+1}$ we get

$$\|\epsilon_h(\mathfrak{T}v_0, \tilde{\Pi}_h)\| \leqslant Ch^{P_A}\|\nabla^{P_A+1}v_0\| + \tilde{C}h^r\|\nabla^{r+1}v_0\|.$$

From (6.26) by the triangle inequality for each $v_0 \in C^{r+1}_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$ we obtain

$$\|\epsilon_h(v_0, \Pi_h)\| \leqslant \|\epsilon_h(v_0, \tilde{\Pi}_h)\| + \hat{c}h^r\|\nabla^{r+1}v_0\|_\infty \leqslant Ch^{P_A}\|\nabla^{P_A+1}v_0\| + \hat{C}h^r\|\nabla^{r+1}v_0\|_\infty.$$

In the last inequality we used the inequality $\|\nabla^{r+1}v_0\| \leqslant \|\nabla^{r+1}v_0\|_\infty$. □

**Lemma 6.16.** *Consider a scheme of the form* (2.8) *and a local mapping* $\Pi_h$ *with* $\mu_\xi \in (W^q_2(G))^*$ *(or* $(C^q(G))^*$*),* $q \in \mathbb{N} \cup \{0\}$. *[Let* $\boldsymbol{e} \in \mathring{\Omega}$*.] Suppose for* $\boldsymbol{\phi} \in \mathcal{A}_\beta$ *[aligned with* $\boldsymbol{e}$*] there holds* $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = 0$. *Then for each* $v_0 \in H^{q+1}_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$ *(or* $C^{q+1}_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$*), there holds* $\epsilon_h(v_0, \Pi_h) = 0$.

*Proof.* By assumption

$$y(\boldsymbol{\phi}) := Z(\boldsymbol{\phi})\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = (-L(\boldsymbol{\phi}) + i\boldsymbol{\omega}\cdot\boldsymbol{\phi}Z(\boldsymbol{\phi}))\,v(\boldsymbol{\phi}, \Pi_h) \equiv 0$$

for $\boldsymbol{\phi} \in \mathcal{A}_\beta$ [aligned with $\boldsymbol{e}$]. If $\boldsymbol{e}$ is not defined, $y$ is an entire function of $\boldsymbol{\phi}$, therefore $y \equiv 0$ on $\mathbb{C}^d$ by the uniqueness theorem for a holomorphic function of several complex variables and thus $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = Z^{-1}(\boldsymbol{\phi})y(\boldsymbol{\phi}) = 0$ for $\boldsymbol{\phi} \in \mathbb{R}^d$. If $\boldsymbol{e}$ is defined, the function $y(\boldsymbol{e}\psi)$ is an entire function of $\psi$, thus $y(\boldsymbol{e}\psi) = 0$ for each $\psi \in \mathbb{C}$ and so $\hat{\epsilon}(\boldsymbol{e}\psi, \Pi_h) = 0$ for $\psi \in \mathbb{R}$. By Lemma 6.10 for each $\boldsymbol{\alpha} \in \mathcal{A}$ [aligned with $\boldsymbol{e}$] there holds $\epsilon_h(e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h) = 0$. Since the linear span of $e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$, $\boldsymbol{\alpha} \in \mathcal{A}$, is dense in $H^{r+1}_{per}(\mathbb{R}^d)$ and $C^{r+1}_{per}(\mathbb{R}^d)$, and the linear span of $e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}$, $\boldsymbol{\alpha} \in \mathcal{A}$ and aligned with $\boldsymbol{e}$, is dense in $H^{r+1}_{per,\boldsymbol{e}}(\mathbb{R}^d)$ and $C^{r+1}_{per,\boldsymbol{e}}(\mathbb{R}^d)$, the statement to prove is by boundedness of the linear operator $\epsilon_h(\,\cdot\,, \Pi_h)$ in the corresponding space (see Lemma 6.9). □

**Theorem 6.17.** *Consider a stable scheme of the form* (2.8) *and a homogeneous mapping* $\Pi_h$. *Let* $P, Q, C_1, C_2 \geqslant 0$. *[Let* $\boldsymbol{e} \in \mathring{\Omega}$.*] Assume for all* $\boldsymbol{\phi} \in \mathcal{A}_\beta$, $\beta \in (0, \pi]$, *[aligned with* $\boldsymbol{e}$*] and all* $\nu \geqslant 0$ *there holds*

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant C_1|\boldsymbol{\phi}|^P + C_2\nu|\boldsymbol{\phi}|^{Q+1}. \tag{6.27}$$

*1) If* $\Pi_h$ *is a bounded homogeneous mapping of* $H^q_{per}(\mathbb{R}^d)$ *to* $V_{per}$ *then there exists* $\tilde{C}$ *such that for each initial data* $v_0 \in H^r_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$, $r \geqslant \max\{Q+1, P, q\}$, *there holds*

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 h^P \|\nabla^P v_0\| + C_2 t h^Q \|\nabla^{Q+1} v_0\| + \tilde{C} h^r \|\nabla^r v_0\|. \tag{6.28}$$

*2) If* $\Pi_h$ *is a local mapping with kernel* $\mu_\xi \in (C^q(G))^*$, $q \in \mathbb{N} \cup \{0\}$, *then then there exists* $\tilde{C}$ *such that for each initial data* $v_0 \in C^r_{per[,\boldsymbol{e}]}(\mathbb{R}^d)$, $r \geqslant \max\{\lceil Q \rceil + 1, \lceil P \rceil, q\}$, *there holds*

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 h^P \|\nabla^P v_0\| + C_2 t h^Q \|\nabla^{Q+1} v_0\| + \tilde{C} h^r \|\nabla^r v_0\|_\infty. \tag{6.29}$$

*For* $Q = +\infty$ *the terms with* $C_2$ *vanish and the expression for* $r$ *is replaced by* $r = \max\{P, q\}$ *(or* $r = \max\{\lceil P \rceil, q\}$*). The constants* $\tilde{C}$ *in 1) and 2) depend only on* $\Pi_h, r, T, \beta, K$, *and the norm on* $\mathbb{C}^{M^0}$.

*Proof.* By (6.18) for each $t > 0$, each $h$, and each $\boldsymbol{\alpha} \in \mathcal{A}_{\beta/h}$ [aligned with $\boldsymbol{e}$] there holds

$$\|\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| = \|\hat{\varepsilon}(h\boldsymbol{\alpha}, t/h, \Pi_h)\| \leqslant C_1 h^P |\boldsymbol{\alpha}|^P + C_2 t h^Q |\boldsymbol{\alpha}|^{Q+1}.$$

Let $\mathfrak{T}$ be the same as in Theorem 6.15.

To prove the first statement, put $F_h v_0 = \varepsilon_h(t, \mathfrak{T} v_0, \Pi_h)$. By Lemma 6.9

$$\|F_h\|_{(r,h)} \leqslant (K+1)\|\Pi_h\|_{(r,h)} \leqslant 2(K+1)\|\Pi_h\|_{(q,h)},$$

where $K$ is the stability constant. Then by Lemma 6.14 with $p(s) = C_1(hs)^P + C_2 t h^Q s^{Q+1}$ we have (6.28).

Now we prove the second statement. By Lemma 5.9 there exists a local mapping $\tilde{\Pi}_h : L_{2,per}(\mathbb{R}^d) \to V_{per}$ with kernel $\tilde{\mu}_\xi \in L_2(\tilde{G})$, $\tilde{G} = G + B_1(0)$, such that for each $f \in C^r_{per}(\mathbb{R}^d)$ there holds $\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant c h^r \|\nabla^r f\|_\infty$ with $c = c(d, r, \Pi_h)$. In particular, $\|\tilde{\Pi}_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}} - \Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}\| \leqslant c h^r |\boldsymbol{\alpha}|^r$. By construction and Lemma A.4, $\|\tilde{\Pi}_h\|$ depends only on $\mu_\xi$ and the norm on $\mathbb{C}^{M^0}$. By the triangle inequality and stability we have

$$\begin{gathered}\|\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \tilde{\Pi}_h)\| \leqslant \|\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| + (K+1)\|(\tilde{\Pi}_h - \Pi_h)e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}\| \leqslant \\ \leqslant C_1|\boldsymbol{\alpha}|^P h^P + C_2|\boldsymbol{\alpha}|^{Q+1} t h^Q + c(K+1)|\boldsymbol{\alpha}|^r h^r.\end{gathered}$$

Put $F_h v_0 = \varepsilon_h(t, \mathfrak{T} v_0, \tilde{\Pi}_h)$. By Lemma 6.9 we have

$$\|F_h\|_{(r,h)} \leqslant (K+1)\|\tilde{\Pi}_h\|_{(r,h)} \leqslant (K+1)\|\tilde{\Pi}_h\|.$$

Then by Lemma 6.14 with $p(s) = C_1(hs)^P + C_2 t h^Q s^{Q+1} + c(K+1)(hs)^r$ we have

$$\|\varepsilon_h(t, v_0, \tilde{\Pi}_h)\| \leqslant C_1 h^P \|\nabla^P v_0\| + C_2 t h^Q \|\nabla^{Q+1} v_0\| + \tilde{c} h^r \|\nabla^r v_0\|,$$

where $\tilde{c}$ depends only on $\Pi_h, r, T, \beta, K$, and the norm on $\mathbb{C}^{M^0}$. By the triangle inequality and stability

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant \|\varepsilon_h(t, v_0, \tilde{\Pi}_h)\| + K\|\tilde{\Pi}_h v_0 - \Pi_h v_0\| + \|\tilde{\Pi}_h v(t, \cdot) - \Pi_h v(t, \cdot)\|$$

where $v(t, \boldsymbol{r}) = v_0(t - \boldsymbol{\omega}\cdot\boldsymbol{r})$. From here, (6.29) easily follows. □

*6.5. Miscellaneous facts*

**Corollary 6.18.** *Let $\Pi_h$ be a bounded homogeneous mapping of $H^s_{per}(\mathbb{R}^d)$ to $V_{per}$ or a local mapping with $\mu_\xi \in (C^s(G))^*$. Let the scheme* (2.8) *be stable and possess the formal order of accuracy $P$ and the long-time simulation order $Q$ on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r\rceil}_{per}(\mathbb{R}^d)$), $r \geqslant \max\{Q+1, s\}$ (or $r \geqslant \max\{P, s\}$ if $Q = \infty$). Then for each $p$ and $q$ such that $0 < p \leqslant P$ and $p \leqslant q \leqslant Q$ the scheme possesses the formal order $p$ and the long-time simulation order $q$ on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r\rceil}_{per}(\mathbb{R}^d)$). If $q < Q = \infty$ we additionally assume $q \leqslant r-1$.*

*Proof.* Let the scheme possess the formal order of accuracy $P$ and the long-time simulation order $Q$. Then by Lemma 6.11 for all $\boldsymbol{\phi} \in \mathbb{R}^d$ and $\nu \geqslant 0$ we have

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant C_1|\boldsymbol{\phi}|^P + C_2\nu|\boldsymbol{\phi}|^{Q+1} + C_3|\boldsymbol{\phi}|^R.$$

Hence, in a neighborhood of $\boldsymbol{\phi} = 0$ we get

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant 2C_1|\boldsymbol{\phi}|^p + C_2\nu|\boldsymbol{\phi}|^{q+1}.$$

By Theorem 6.17 we obtain that the scheme possesses the formal order of accuracy $p$ and the long-time simulation order $q$ on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r\rceil}_{per}(\mathbb{R}^d)$). For $Q = \infty$ the proof remains valid if we drop the terms with $C_2$. □

**Corollary 6.19.** *Consider a stable scheme of the form* (2.8) *and a bounded homogeneous mapping $\Pi_h$ of $H^q_{per}(\mathbb{R}^d)$ (or $C^q_{per}(\mathbb{R}^d)$) to $V_{per}$. Then the following two statements are equivalent.*

(Y1) *The scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$.*

(Y2) *For each $\boldsymbol{e} \in \mathring{\Omega}$ the scheme possesses a error estimate of the form*

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1h^P\|\nabla^P v_0\| + C_2th^Q\|\nabla^{Q+1}v_0\| + C_3h^r\|\nabla^r v_0\|_* \tag{6.30}$$

*for $v_0 \in H^r_{per,\boldsymbol{e}}(\mathbb{R}^d)$, $r \geqslant \max\{Q+1, P, q\}$, (or $v_0 \in C^r_{per,\boldsymbol{e}}(\mathbb{R}^d)$, $r \geqslant \max\{\lceil Q\rceil + 1, \lceil P\rceil, q\}$), where $C_1$, $C_2$, and $C_3$ are independent of $\boldsymbol{e}$, and $\|\cdot\|_*$ means either $\|\cdot\|$ or $\|\cdot\|_\infty$ depending on the case.*

*Proof.* The implication (Y1) to (Y2) is obvious. Conversely, assume (Y2). Taking (6.30) for $v_0 = \exp(i\boldsymbol{\phi}\cdot\boldsymbol{r}/h)$, by (6.18) we get

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant C_1|\boldsymbol{\phi}|^P + C_2\nu|\boldsymbol{\phi}|^{Q+1} + C_3|\boldsymbol{\phi}|^r$$

for each $\boldsymbol{\phi} \in \mathcal{A}$. Since $r \geqslant P$, for $\boldsymbol{\phi} \in \mathcal{A}_\pi$ there holds (6.27) with another $C_1$. It remains to use Theorem 6.17. □

**Lemma 6.20.** *Let $f(x) : \mathbb{R} \to \mathbb{C}^n$ be analytical at $x = 0$ and satisfy $\|f(1/k)\| \leqslant g(1/k)$ for each $k \in \mathbb{N}$ where $g(x) \in C^1(\mathbb{R})$ is a function such that $g(0) = 0$ and $g'(x) > 0$ for $0 < x < 1$. Then there exists $\delta > 0$ such that there holds $\|f(x)\| \leqslant Cg(2x)$ for $0 < x < \delta$, where $C$ depends on the norm on $\mathbb{C}^n$ only.*

*Proof.* We interpret $\mathbb{C}^n$ as $\mathbb{R}^{2n}$ and denote by $f_j$, $j = 1, \dots, 2n$, the components of $f$. Since $f_j$ is analytical then there exists $\delta_j$ such that $f_j'(x) \neq 0$ for $x \in (0, \delta_j)$. Put $\delta = \min\{\min\{\delta_j\}, 1\}/2$. Then $f_j$ is monotone on $(0, \delta)$, $f_j(0) = 0$ and so for any $k \in \mathbb{N}$ we have

$$|f_j(x)| \leqslant \left|f_j(2^{-k})\right| \leqslant cg(2^{-k}) \leqslant cg(2x), \quad x \in (2^{-k-1}, 2^{-k}] \cap (0, \delta).$$

The inequality for $\|f(x)\|$ easily follows. □

**Corollary 6.21.** *Let $\Pi_h$ be a local mapping with kernel $\mu \in (W^q_2(G))^*$ (or $(C^q(G))^*$). If for each $\boldsymbol{\alpha} \in \mathcal{A}$ and for each $h$ there holds $\|\epsilon_h(e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant c(\boldsymbol{\alpha})h^{P_A}$, then there holds*

$$\|\epsilon_h(v_0, \Pi_h)\| \leqslant C_1\|\nabla^{P_A+1}v_0\|h^{P_A} + C_2\|\nabla^{\max\{P_A,q\}+1}v_0\|_*\, h^{\max\{P_A,q\}} \tag{6.31}$$

*with some $C_1, C_2 > 0$, where $\|\cdot\|_* \equiv \|\cdot\|$ if $\mu \in (W^q_2(G))^*$ and $\|\cdot\|_* \equiv \|\cdot\|_\infty$ otherwise.*

*Proof.* Let $\boldsymbol{e} \in \mathring{\Omega}$, then there exists $\lambda > 0$ such that $\lambda\boldsymbol{e} \in \mathcal{A}$. Let $h$ be such that $1/h \in \mathbb{N}$. Then we get

$$\|\hat{\epsilon}(\lambda h\boldsymbol{e}, \Pi_h)\| \leqslant h\|Z^{-1}\| \, \|\epsilon_h(e^{i\lambda\boldsymbol{e}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant c(\boldsymbol{e})\|Z^{-1}\| \, h^{P_A+1}.$$

Here the first inequality is by Lemma 6.10 and the second one is by assumption. For $\boldsymbol{\phi} = \lambda h\boldsymbol{e}$, where $1/h \in \mathbb{N}$, we have

$$\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant \frac{c(\boldsymbol{e})}{|\lambda\boldsymbol{e}|^{P_A+1}}|\boldsymbol{\phi}|^{P_A+1} := \hat{c}(\boldsymbol{e})|\boldsymbol{\phi}|^{P_A+1}.$$

Since $\hat{\epsilon}$ is holomorphic with respect to the first argument by Lemma 6.20 there holds $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant C\hat{c}(\boldsymbol{e})|2\boldsymbol{\phi}|^{P_A+1}$ for each $\boldsymbol{\phi}$ aligned with $\boldsymbol{e}$ in a neighborhood of zero. By Lemma 5.3 there holds $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant \tilde{c}|\boldsymbol{\phi}|^{P_A+1}$ in a neighborhood of $\boldsymbol{\phi} = 0$. It remains to apply Theorem 6.15. □

**Proposition 6.22.** *Consider a scheme of the form* (2.8)*, a local mapping* $\Pi_h$ *with the kernel* $\mu_\xi \in (W_2^q(G))^*$*, and* $\boldsymbol{e} \in \mathring{\Omega}$*. Let* $0 < P \leqslant Q < \infty$ *and* $R = \max\{Q+1, q\}$*. The scheme possesses the formal order of accuracy* $P$ *and the long-time simulation order* $Q$ *on* $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ *iff it possesses the formal order of accuracy* $P$ *and the long-time simulation order* $Q$ *on* $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ *in the weak sense.*

*Proof.* Let the scheme possess the formal order of accuracy $P$ and the long-time simulation order $Q$ on $H^Q_{per,\boldsymbol{e}}(\mathbb{R}^d)$ in the weak sense. Since $\boldsymbol{e} \in \mathring{\Omega}$, there exists $\lambda > 0$ such that $\lambda\boldsymbol{e} \in \mathcal{A}$. For $v_0 = e^{i\lambda\boldsymbol{e}\cdot\boldsymbol{r}}$ Definition 3 gives

$$\|\varepsilon_h(t, e^{i\lambda\boldsymbol{e}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant c_1h^P + c_2th^Q.$$

Here $c_1$ and $c_2$ may depend on $\boldsymbol{e}$ but are independent of $t$ and $h$. By Lemma 6.10 there holds

$$\|\hat{\varepsilon}(\lambda h\boldsymbol{e}, t/h, \Pi_h)\| = \|\varepsilon_h(t, e^{i\lambda\boldsymbol{e}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant c_1h^P + c_2th^Q.$$

For $\boldsymbol{\phi} = \lambda h\boldsymbol{e}$, where $1/h \in \mathbb{N}$, and each $\nu \geqslant 0$ we have

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant |\boldsymbol{\phi}|^P\frac{c_1}{|\lambda\boldsymbol{e}|^P} + \nu|\boldsymbol{\phi}|^{Q+1}\frac{c_2}{|\lambda\boldsymbol{e}|^{Q+1}}.$$

Since $\hat{\varepsilon}$ is holomorphic with respect to the first argument by Lemma 6.20 there holds $\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant \tilde{c}_1|\boldsymbol{\phi}|^P + \tilde{c}_2\nu|\boldsymbol{\phi}|^{Q+1}$ for each $\boldsymbol{\phi}$ aligned with $\boldsymbol{e}$ in a neighborhood of zero. By Theorem 6.17 the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ on $H^Q_{per,\boldsymbol{e}}(\mathbb{R}^d)$. The reverse implication is obvious. □

In particular, in the 1D case for local mappings Definition 3 is equivalent to Definition 4.

## 7. The 1D case

In this section we consider the one-dimensional case. The main result of this section (Theorem 7.7) can be obtained from the quasi-one-dimensional case considered in Section 9.3 (see Lemma 9.20). However, the 1D case allows us to use powerful analytical tools that provide a clearer understanding of the enhanced accuracy in the long-time simulation.

Throughout this section we consider a stable scheme of the form (2.8) and a local mapping $\Pi_h$. In 1D the set of the vectors $\boldsymbol{a}_j$ contains only one vector $\boldsymbol{a}_1$, which has only one component. We assume it to be unit, i. e. the value $h$ coincides with the mesh period, which is the natural definition for the schemes with several DOFs per cell. Hence, $T$ and $U$ are the identity operators on $\mathbb{R}^1$.

### *7.1. Matrix decomposition*

Let $\mathbb{R}^{n\times n}$ and $\mathbb{C}^{n\times n}$ be the spaces of real and complex matrices of size $n$. Denote by $\mathcal{A}(\mathbb{C}, \mathbb{R}^{n\times n})$ the set of functions $A(\phi)$ from $\mathbb{C}$ to $\mathbb{C}^{n\times n}$ such that each element of the matrix $A(\phi)$ is a holomorphic function at $\phi = 0$ and for all $i\phi \in \mathbb{R}$ in a neighborhood of $\phi = 0$ there holds $A(\phi) \in \mathbb{R}^{n\times n}$.

We need the following result.

**Theorem 7.1** ([29]). *Let $A \in \mathcal{A}(\mathbb{C}, \mathbb{R}^{n\times n})$. Suppose there exists $K > 0$ such that for all $\phi \in \mathbb{R}$ and $\nu \geqslant 0$ there holds $\| \exp(\nu A(\phi))\| \leqslant K$. Then in a neighborhood of $\phi = 0$ the matrix $A(\phi)$ can be represented in the form*

$$A(\phi) = S(\phi)M(\phi)S^{-1}(\phi), \tag{7.1}$$

$$M(\phi) = \begin{pmatrix} M_0(\phi) & 0 & \dots & 0 & 0 \\ 0 & \phi M_1(\phi) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \phi^m M_m(\phi) & 0 \\ 0 & 0 & \dots & 0 & M_\infty(\phi) \equiv 0 \end{pmatrix}, \tag{7.2}$$

*where $S, M, S^{-1} \in \mathcal{A}(\mathbb{C}, \mathbb{R}^{n\times n})$, the square matrices $M_k(\phi)$, $k \in \mathbb{N} \cup \{0, \infty\}$, are non-degenerate for $\phi = 0$ (except $k = \infty$), and some of them are absent.*

Throughout this section we denote by $S(\phi)$, $M_j(\phi)$, and $M(\phi)$ the matrices given by Theorem 7.1 for the matrix $A(\phi)$ defined in (6.14).

Denote by $\bar{\aleph}$ the set $j \in \mathbb{N} \cup \{0, \infty\}$ such that the block $\phi^j M_j(\phi)$ is present in the matrix $M(\phi)$.

The function $v(\phi, \Pi_h)$ defined in (6.7) has the form $v(\phi, \Pi_h) = (\Pi_1 e^{i\phi x})_0$. Let $v_j(\phi, \Pi_h)$, $j \in \bar{\aleph}$, be the components of $S^{-1}(\phi)v(\phi, \Pi_h)$ corresponding to the blocks $M_j(\phi)$. Since $v_j(\phi, \Pi_h)$ are holomorphic functions there exist $p_j \in \mathbb{N} \cup \{0\}$ and $c_j \in \mathbb{R} \setminus \{0\}$ such that in a neighbourhood of $\phi = 0$ there holds

$$c_j|\phi|^{p_j} \leqslant \|v_j(\phi, \Pi_h)\| \leqslant 2c_j|\phi|^{p_j}. \tag{7.3}$$

If $v_j(\phi) \equiv 0$ we put by definition $p_j = \infty$ and $c_j = 1$. Put

$$\aleph = \{j \in \mathbb{N} \cup \{0\} : p_j < \infty\} \subseteq \bar{\aleph}.$$

This means that $j \in \bar{\aleph} \setminus \aleph$ iff $j = \infty$ or $p_j = \infty$.

Below we show that the values $p_j$, $j \in \bar{\aleph}$, are responsible for the structure of the numerical error and define the order of the truncation error, the order of accuracy, and the long-time simulation order.

**Lemma 7.2.** *There exists $j \in \bar{\aleph}$ such that $p_j = 0$.*

*Proof.* Assume the converse: $p_j > 0$ for each $j \in \bar{\aleph}$. Then for each $j$ there holds $v_j(0) = 0$, therefore $(\Pi_1 1)_0 = v(0, \Pi_h) = 0$. On the other hand, by definition for at least one $\xi \in M^0$ there holds $(\Pi_1 1)_{0,\xi} = \langle \mu_\xi, 1\rangle \neq 0$. This contradiction proves the lemma. □

*7.2. The structure of the truncation error*

**Lemma 7.3.** *Let $\Pi_h$ be a local mapping with $\mu_\xi \in (W_2^q(G))^*$ or $(C^q(G))^*$. If $\aleph = \varnothing$ then the scheme is exact, i. e. for each $v_0 \in H^{q+1}_{per}(\mathbb{R})$ (or $C^{q+1}_{per}(\mathbb{R})$) there holds $\epsilon_h(v_0, \Pi_h) = 0$.*

*Proof.* Let $j \in \bar{\aleph}$. Since $\aleph = \varnothing$, either $j = \infty$ or $p_j = \infty$. In a neighborhood of $\phi = 0$ there holds $M_j(\phi) = 0$ in the first case and $v_j(\phi, \Pi_h) = 0$ in the second case. Thus for each $j \in \bar{\aleph}$ there holds $M_j(\phi)v_j(\phi, \Pi_h) \equiv 0$ and hence

$$\hat{\epsilon}(\phi, \Pi_h) \equiv A(\phi)v(\phi, \Pi_h) \equiv 0. \tag{7.4}$$

It remains to apply Lemma 6.16. □

**Lemma 7.4.** *Let $\Pi_h$ be a local mapping. If $\aleph \neq \varnothing$, then the scheme possesses the truncation error of order*

$$P_A = \min_{j\in\aleph}\{p_j + j - 1\}. \tag{7.5}$$

*This value is optimal, i. e. the scheme does not possess the truncation error of order $p > P_A$.*

*Proof.* Since the matrices $Z(\phi)$, $S(\phi)$, and $M_j(\phi)$ for each $j$ are holomorphic functions in a neighborhood of $\phi = 0$, for small $|\phi|$ we have

$$\|\hat{\epsilon}(\phi, \Pi_h)\| \leqslant C \sum_{j \in \aleph} |\phi|^j \|v_j(\phi, \Pi_h)\| \leqslant \tilde{C} \max_{j \in \aleph} |\phi|^{j+p_j}. \tag{7.6}$$

By Theorem 6.15 the scheme possesses the truncation error of order $P_A$.

Conversely, let the scheme possess the truncation error of order $p$. Then by (6.17) we have $\|\hat{\epsilon}(\phi, \Pi_h)\| \leqslant C|\phi|^{p+1}$. Since $S(\phi)$ is non-degenerate, for small $|\phi|$ for each $j \in \aleph$ there holds

$$\|\phi^j M_j(\phi) v_j(\phi, \Pi_h)\| \leqslant \tilde{c}|\phi|^{p+1}.$$

Since $M_j(0)$ is non-degenerate, we have $\|M_j^{-1}(\phi)\| \leqslant 2\|M_j^{-1}(0)\|$ in a neighborhood of $\phi = 0$. Hence, $\|v_j(\phi, \Pi_h)\| \leqslant 2\tilde{c}\|M_j^{-1}(0)\| \, |\phi|^{p+1-j}$. By (7.3) we get $p_j \geqslant p + 1 - j$. Since this holds for each $j \in \aleph$, we have $p \leqslant P_A$. □

490 *7.3. The structure of the solution error*

Now substitute the representation of $A(\phi)$ given by Theorem 7.1 into (6.16). We have

$$\hat{\varepsilon}(\phi, \nu, \Pi_h) = S(\phi) \left( e^{\nu M(\phi)} - I \right) S^{-1}(\phi) v(\phi, \Pi_h) = S(\phi) \begin{pmatrix} \vdots \\ E_j(\phi, \nu, \Pi_h) \\ \vdots \end{pmatrix},$$

where

$$E_j(\phi, \nu, \Pi_h) = \left( e^{\nu \phi^j M_j(\phi)} - I \right) v_j(\phi, \Pi_h) \tag{7.7}$$

and $v_j(\phi, \Pi_h)$ is defined in Section 7.1. For $j \in \bar{\aleph} \setminus \aleph$ in a neighborhood of $\phi = 0$ we have $E_j(\phi, \nu, \Pi_h) \equiv 0$. For $j \in \aleph$ put $Y = \nu \phi^j M_j(\phi)$. By the stability condition (6.21) we have $\|e^Y\| \leqslant K\varkappa(\phi)$, where $\varkappa(\phi)$ is the condition number of $S(\phi)$. By Lemma 5.11 from (7.3) we get

$$\|E_j(\phi, \nu, \Pi_h)\| \leqslant (K\varkappa(\phi) + e) \min\left\{1, \nu|\phi|^j \, \|M_j(\phi)\|\right\} 2c_j|\phi|^{p_j}. \tag{7.8}$$

**Lemma 7.5.** *Let the scheme possess the formal order of accuracy $P$ and the long-time simulation order $Q \geqslant P$. Then for each $j \in \aleph$ there holds*

$$p_j \geqslant \min\{P, Q + 1 - j\}. \tag{7.9}$$

*Proof.* By Lemma 6.11 in a neighborhood of $\phi = 0$ there holds

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| \leqslant \tilde{C}_1|\phi|^P + C_2\nu|\phi|^{Q+1}.$$

Since $S(0)$ is non-degenerate, we get the estimate for each error component:

$$\|E_j(\phi, \nu, \Pi_h)\| \leqslant C_1'|\phi|^P + C_2'\nu|\phi|^{Q+1} \tag{7.10}$$

in a neighborhood of $\phi = 0$. Particularly, for $\nu = |\phi|^{-j}\|M_j(0)\|^{-1}$ we get

$$\|E_j(\phi, |\phi|^{-j}\|M_j(0)\|^{-1}, \Pi_h)\| \leqslant C'|\phi|^{\min\{P, Q+1-j\}}.$$

From (7.7) we have

$$E_j(\phi, |\phi|^{-j}\|M_j(0)\|^{-1}, \Pi_h) = \left( \exp\left( \frac{M_j(\phi)}{\|M_j(0)\|} \right) - I \right) v_j(\phi, \Pi_h).$$

By Lemma 5.13 the matrix $\exp(M_j(\phi)/\|M_j(0)\|) - I$ is non-degenerate for $\phi = 0$. Then it is non-degenerate in a neighborhood of $\phi = 0$ and

$$\|v_j(\phi, \Pi_h)\| \leqslant \left\| \left(\exp(M_j(\phi)/\|M_j(0)\|) - I\right)^{-1} \right\| \, \|E_j(\phi, |\phi|^{-j}\|M_j(0)\|^{-1}, \Pi_h)\| \leqslant C''|\phi|^{\min\{P, Q+1-j\}}.$$

Now (7.9) is by the definition of $p_j$. □

**Lemma 7.6.** *If $p_0 = 0$ or $p_1 = 0$, then the scheme does not possess any order of accuracy. Otherwise the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ where*

$$P = \min_{j\in\aleph} (\max\{p_j + j - 1, p_j\}) = \min\{p_0, \min_{j\geqslant 1}(p_j + j - 1)\}, \tag{7.11}$$

$$Q = \min_{j\in\aleph:\ p_j<P}(p_j + j - 1). \tag{7.12}$$

*Minimum over the empty set is assumed to be $+\infty$. The values $P$ and $Q$ given by* (7.11) *and* (7.12) *are optimal in the sense of Definition 5.*

*Proof.* First consider the case when $p_0 = 0$ or $p_1 = 0$. Assume that the scheme possess the formal order of accuracy $P$ and the long-time simulation order $Q \geqslant P$. Then by Lemma 7.5 there holds $p_0 \geqslant \min\{P, Q + 1\}$, $p_1 \geqslant \min\{P, Q\}$. Since $Q \geqslant P$, each of the assumptions $p_0 = 0$ or $p_1 = 0$ immediately gives $P \leqslant 0$.

Now assume $p_0 \neq 0$, $p_1 \neq 0$. This yields $P \neq 0$. Let us show that $Q \geqslant P$. By (7.11) there holds $p_0 \geqslant P$. Thus

$$Q = \min_{j:p_j<P}(p_j + j - 1) = \min_{j\geqslant 1:p_j<P}(p_j + j - 1) \geqslant \min_{j\geqslant 1}(p_j + j - 1) \geqslant P.$$

Now we prove that the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$. Indeed, by (7.8) in a neighborhood of $\phi = 0$ there holds

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| \leqslant \hat{C} \sum_{j\in\aleph} \|E_j(\phi, \nu, \Pi_h)\| \leqslant C \sum_{j\in\aleph} \min\{|\phi|^{p_j}, \nu|\phi|^{p_j+j}\}. \tag{7.13}$$

Therefore,

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| \leqslant C \sum_{j:p_j<P} \nu|\phi|^{p_j+j} + C \sum_{j:p_j\geqslant P} |\phi|^{p_j} \leqslant C'\nu|\phi|^{Q+1} + C'|\phi|^P. \tag{7.14}$$

By Theorem 6.17 we get the desired estimate (2.17).

Now we show that $P$ and $Q$ are optimal. Let estimate (2.17) be valid under the substitutions $Q'$ for $Q$ and $P'$ for $P$ where $Q' \geqslant P' > 0$. Then for each error component given by (7.7) there holds (7.10) with the same substitution. By Lemma 7.5 for each $j$ there holds one of the inequalities

$$\left[\begin{array}{l} p_j \geqslant P', \\ p_j + j - 1 \geqslant Q'. \end{array}\right. \tag{7.15}$$

Since $Q' \geqslant P'$, (7.15) for each $j$ yields $P' \leqslant \max\{p_j, p_j + j - 1\}$. Taking minimum over $j$, using (7.11) we get $P' \leqslant P$.

Now assume that $P' = P$. Then (7.15) yields

$$Q' \leqslant \min_{p_j<P'}(p_j + j - 1) = \min_{p_j<P}(p_j + j - 1) = Q.$$

Thus estimate (2.17) under the substitutions $Q'$ for $Q$ and $P'$ for $P$ yields the alternative

$$\left[\begin{array}{ll} P' < P, & \\ P' = P, & Q' \leqslant Q. \end{array}\right.$$

This means that $P$ and $Q$ are optimal in the sense of Definition 5. □

### *7.4. The main result*

**Theorem 7.7.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with the kernels $\mu_\xi, \hat{\mu}_\xi$, correspondingly, where $\langle\hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$. Suppose the scheme possesses the formal order of accuracy $P \in \mathbb{N}$ and long-time simulation order $Q \in \mathbb{N}$ in the sense of $\Pi_h$. Then there exist diagonal matrices $\mathfrak{C}^{(m)} \in \mathbb{R}^{M^0}$, $m = P, \dots, Q$, such that the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_h^{(P,Q)}$ given by* (2.12).

*Proof.* By Lemma 6.11 there exist $C_1, C_2 \geqslant 0$ such that in a neighborhood of $\phi = 0$ there holds

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| = \left\| \left( e^{\nu A(\phi)} - I \right) v(\phi, \Pi_h) \right\| \leqslant C_1 |\phi|^P + C_2 \nu |\phi|^{Q+1}.$$

Using representation (7.1), (7.2) provided by Theorem 7.1, in a neighborhood of $\phi = 0$ define the holomorphic function

$$w(\phi) = S(\phi) \begin{pmatrix} \delta_0 I & 0 & \dots & 0 & 0 \\ 0 & \delta_1 I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_m I & 0 \\ 0 & 0 & \dots & 0 & \delta_\infty I \end{pmatrix} S^{-1}(\phi) v(\phi, \Pi_h),$$

where the blocks correspond to the blocks of $M(\phi)$, $\delta_j = 0$ if $p_j \geqslant P$ and $\delta_j = 1$ if $p_j < P$. Since $S(\phi)$, $S^{-1}(\phi)$, and $v(\phi, \Pi_h)$ are real-valued for $i\phi \in \mathbb{R}$, so is $w(\phi)$. Then $\overline{w(-\phi)} = w(\overline{\phi})$. By construction $w(\phi)$ satisfies

$$\|w(\phi) - v(\phi, \Pi_h)\| = O(|\phi|^P).$$

Since

$$A(\phi) w(\phi) = S(\phi) \begin{pmatrix} \delta_0 M_0(\phi) & 0 & \dots & 0 & 0 \\ 0 & \delta_1 \phi M_1(\phi) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_m \phi^m M_m(\phi) & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} S^{-1}(\phi) v(\phi, \Pi_h),$$

by Lemma 7.5 in a neighborhood of $\phi = 0$ there holds

$$\|A(\phi) w(\phi)\| \leqslant \tilde{C} \max_{j : p_j < P} |\phi|^j \|M_j(\phi)\| \, \|v_j(\phi, \Pi_h)\| = O(|\phi|^{Q+1}). \tag{7.16}$$

By Lemma 6.7 there exist real-valued diagonal matrices $\mathfrak{C}^{(m)}$, $m = P, \dots, Q$, such that $v(\phi, \Pi_h^{(P,Q)}) = w(\phi) + O(|\phi|^{Q+1})$. Now (7.16) yields

$$\hat{\epsilon}(\phi, \Pi_h^{(P,Q)}) = A(\phi) v(\phi, \Pi_h^{(P,Q)}) = O(|\phi|^{Q+1}).$$

By Theorem 6.15 the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_h^{(P,Q)}$. □

*7.5. Construction of a scheme with specified properties*

**Proposition 7.8.** *Consider the transport equation $\partial v/\partial t + \omega \partial v/\partial x = 0$, $\omega \neq 0$. Let $\bar{\aleph}$ be a finite subset of $\mathbb{N} \cup \{0\}$. Then for any set of $p_j \in \mathbb{N} \cup \{0, \infty\}$, $j \in \bar{\aleph}$, containing at least one zero there exists a scheme of the form* (2.8) *and a local mapping such that the values $p_j$ coincide with those defined by* (7.3).

*Proof.* Let $c_{m,\eta}$, $m \in \mathbb{N}$, be coefficients of a finite-difference approximation of the first derivative on the uniform mesh with unit step possessing exactly the order $m - 1$, i. e.

$$i\phi - \sum_{\eta \in \mathcal{S}_m} c_{m,\eta} e^{i\phi\eta} = i\phi - \sum_{r=0}^{\infty} \frac{(i\phi)^r}{r!} \sum_{\eta \in \mathcal{S}_m} c_{m,\eta} \eta^r = \gamma_m \phi^m + O(\phi^{m+1}) \tag{7.17}$$

where $\gamma_m \neq 0$, and $\mathcal{S}_m \subset \mathbb{Z}$ is a finite set. For $m = 0$ put $\mathcal{S}_0 = \{0\}$, $c_{0,0} = \operatorname{sign} \omega$, so (7.17) also holds. For $\eta \notin \mathcal{S}_m$ put $c_{m,\eta} = 0$.

Now we define a scheme of the form (2.8) and a local mapping. Put $M^0 = \bar{\aleph}$,

$$\mathcal{S} = \bigcup_{m \in \bar{\aleph}} \mathcal{S}_m; \quad Z_0 = I; \quad Z_\eta = 0, \eta \neq 0; \quad L_\eta = \operatorname{diag} \{\omega c_{\xi,\eta}, \xi \in M^0\};$$

$$(\Pi_h f)_{\eta,\xi} = h^{p_\xi}\frac{d^{p_\xi} f}{dx^{p_\xi}}(\eta h), \quad p_\xi < \infty; \qquad (\Pi_h f)_{\eta,\xi} = 0, \quad p_\xi = \infty.$$

Then $(v(\phi,\Pi_h))_\xi = (i\phi)^{p_\xi}$ and

$$A(\phi) = i\omega\phi - \sum_{\eta\in\mathcal{S}} L_\eta e^{i\phi\eta} = \operatorname{diag}\left\{ i\omega\phi - \omega\sum_{\eta\in\mathcal{S}} c_{\xi,\eta}e^{i\phi\eta}, \xi\in M^0\right\} = \operatorname{diag}\left\{\omega\gamma_\xi\phi^\xi + O(\phi^{\xi+1}), \xi\in M^0\right\}.$$

Thus we have a representation of the form (7.1)–(7.2), with the blocks $j\in\bar{\aleph}$ of the size 1 and no other blocks, and the values $p_j$ coincide with those defined by (7.3). If the finite-difference schemes with the coefficients $c_{m,\eta}$ are stable then the scheme we constructed is also stable. □

*7.6. One-dimensional case: summary*

Consider a stable scheme of the form (2.8) and a local mapping $\Pi_h$. In the sense of $\Pi_h$, the truncation error and the solution error have spectral representations (7.6) and (7.13), correspondingly. Then by Lemma 6.10 for $v_0 = \exp(i\alpha x)$, $\alpha/\pi\in\mathbb{Q}$, there holds

$$\|\epsilon_h(v_0,\Pi_h)\| \leqslant C\sum_{j\in\aleph}\|\nabla^{p_j+j}v_0\|h^{p_j+j-1} \leqslant C'\|\nabla^{P_A+1}v_0\|h^{P_A},$$

$$\|\varepsilon_h(t,v_0,\Pi_h)\| \leqslant C\sum_{j\in\aleph}\min\left\{\|\nabla^{p_j}v_0\|h^{p_j}, \|\nabla^{p_j+j}v_0\|h^{p_j+j-1}t\right\} \leqslant C'\|\nabla^P v_0\|h^P + C'\|\nabla^{Q+1}v_0\|h^Q t, \quad (7.18)$$

where $\aleph$ and $p_j$ are defined in Section 7.1 and $P_A$, $P$, and $Q$ are given by (7.5), (7.11), and (7.12):

$$P_A = \min_{j\in\aleph}\{p_j + j - 1\}, \quad P = \min\{p_0, \min_{j\geqslant 1}(p_j + j - 1)\}, \quad Q = \min_{j\in\aleph:\ p_j<P}(p_j + j - 1).$$

Theorem 7.7 states that there exists $\Pi_h^{(P,Q)}$ in the sense of which the truncation error satisfies

$$\|\epsilon_h(v_0,\Pi_h^{(P,Q)})\| \leqslant \tilde{C}\|\nabla^{Q+1}v_0\|h^Q.$$

By the Lax – Ryabenkii theorem (Theorem 4.1) the solution error satisfies

$$\|\varepsilon_h(t,v_0,\Pi_h^{(P,Q)})\| \leqslant \tilde{C}K\|Z^{-1}\|\,\|\nabla^{Q+1}v_0\|h^Q t.$$

Theorems 6.15 and 6.17 allow to extend these estimates to $v_0$ smooth enough.

*Example.* Put $\bar{\aleph} = \{3,5\}$, $p_3 = 1$, and $p_5 = 0$. A scheme with these parameters exists by Proposition 7.8. By Lemma 7.4 this scheme possesses the truncation error of order $P_A = 3$. For $v_0 = \exp(i\alpha x)$, $\alpha/\pi\in\mathbb{Q}$, the error estimate (7.18) reads as

$$\|\varepsilon_h(t,v_0,\Pi_h)\| \leqslant C_1\min\{h\|\nabla v_0\|, h^3 t\|\nabla^4 v_0\|\} + C_2\min\{\|v_0\|, h^4 t\|\nabla^5 v_0\|\}.$$

By Lemma 7.6 the scheme possesses the formal order of accuracy $P = 3$ and the long-time simulation order $Q = 3$, these values being optimal in the sense of Definition 5. However we can write the estimate

$$\|\varepsilon_h(t,v_0,\Pi_h)\| \leqslant C_1 h\|\nabla v_0\| + C_2 h^4 t\|\nabla^5 v_0\|,$$

so the scheme possesses the 1st formal order of accuracy and the 4th order in the long-time simulation. The values $P = 1$ and $Q = 4$ are not optimal in the sense of our definition. Note that due to Theorem 7.7 there exists a local mapping that differs from $\Pi_h$ by $O(h)$ and gives the 4th order of the truncation error, the 4th formal order of accuracy and the 4th long-time simulation order.

## 8. The general case

In this section we prove Theorem 2 and state some of its corollaries.

*8.1. The method of auxilliary mapping*

The following lemma describes the method of auxiliary mapping (see Proposition 4.2) using the spectral representation of a scheme.

**Lemma 8.1.** *Consider a scheme of the form* (2.8) *stable with a constant $K$ and a bounded homogeneous mapping $\Pi_h$ of $H^q_{per}(\mathbb{R}^d)$ to $V_{per}$ or a local mapping $\Pi_h$ with $\mu_\xi \in (C^{\lceil q\rceil}(G))^*$. Suppose there exists a function $w(\boldsymbol{\phi})$ and $c_1, c_2 > 0$ such that in a neighborhood of $\boldsymbol{\phi} = 0$ there holds*

$$\|v(\boldsymbol{\phi}, \Pi_h) - w(\boldsymbol{\phi})\| \leqslant c_1|\boldsymbol{\phi}|^P, \quad \|A(\boldsymbol{\phi})w(\boldsymbol{\phi})\| \leqslant c_2|\boldsymbol{\phi}|^{Q+1}, \tag{8.1}$$

*where $0 < P \leqslant Q \leqslant \infty$ (we assume $|\boldsymbol{\phi}|^\infty \equiv 0$). Let $r \geqslant \max\{Q+1, q\}$ (if $P \leqslant Q < \infty$) or $r \geqslant \max\{P, q\}$, (if $P < Q = \infty$) or $r \geqslant q$ (if $P = Q = \infty$). Then for $\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)$ defined by* (6.16) *there holds*

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant (K+1)c_1|\boldsymbol{\phi}|^P + \nu K c_2|\boldsymbol{\phi}|^{Q+1} \tag{8.2}$$

*and the scheme possesses the error estimate* (2.17) *on $H^r_{per}(\mathbb{R}^d)$ (or $C^{\lceil r\rceil}_{per}(\mathbb{R}^d)$) with the same $P$ and $Q$, $C_1 = (K+1)c_1$, $C_2 = Kc_2$, and $C_3 > 0$ depending only on $\Pi_h, r, T, K$, the neighborhood of $\boldsymbol{\phi} = 0$ in the conditions of the lemma, and the norm on $\mathbb{C}^{M^0}$.*

*Proof.* We have

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| = \left\|\left(e^{\nu A(\boldsymbol{\phi})} - I\right) v(\boldsymbol{\phi}, \Pi_h)\right\| \leqslant \left\|\left(e^{\nu A(\boldsymbol{\phi})} - I\right)(w(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h))\right\| + \left\|\left(e^{\nu A(\boldsymbol{\phi})} - I\right) w(\boldsymbol{\phi})\right\| \leqslant$$

$$\leqslant \|\exp(\nu A(\boldsymbol{\phi})) - I\| \, \|w(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h)\| + \left\|\int_0^\nu e^{\tau A(\boldsymbol{\phi})} d\tau \; A(\boldsymbol{\phi})w(\boldsymbol{\phi})\right\| \leqslant$$

$$\leqslant (\|\exp(\nu A(\boldsymbol{\phi}))\| + 1) \, \|w(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h)\| + \int_0^\nu \left\|e^{\tau A(\boldsymbol{\phi})}\right\| d\tau \; \|A(\boldsymbol{\phi})w(\boldsymbol{\phi})\| \leqslant (K+1)c_1|\boldsymbol{\phi}|^P + \nu K c_2|\boldsymbol{\phi}|^{Q+1}.$$

Thus we get (8.2). It remains to use Theorem 6.17. □

*8.2. The existence of an auxiliary mapping*

Let $\mathbb{C}^{n\times n}$ be the space of complex matrices of size $n$. We need the following result.

**Theorem 8.2** ([28])**.** *Let $A \in \mathbb{C}^{n\times n}$ and $v \in \mathbb{C}^n$. Suppose for each $\nu \geqslant 0$ there hold $\|e^{\nu A}\| \leqslant K$ and*

$$\|(e^{\nu A} - I)v\| \leqslant (\tilde{C}_1 + \tilde{C}_2\nu)\|v\|. \tag{8.3}$$

*Then for $w = (A^*A + \varepsilon^2)^{-1}\varepsilon^2 v$ where $\varepsilon = \tilde{C}_2/\tilde{C}_1$ (if $\tilde{C}_1 = 0$ or $\tilde{C}_2 = 0$ then the vector $w$ is the corresponding limit) there hold*

$$\|v - w\| \leqslant \delta\tilde{C}_1\|v\|, \quad \|Aw\| \leqslant \delta\tilde{C}_2\|v\|, \tag{8.4}$$

*where $\delta$ depends on $n$ and $K$ only.*

Note that $w = \operatorname{argmin}(\tilde{C}_2^2\|v - w\|^2 + \tilde{C}_1^2\|Aw\|^2)$.

Denote

$$\mathcal{W}^{P,Q}(\boldsymbol{\phi}) = \left(A^*(\boldsymbol{\phi})A(\boldsymbol{\phi}) + |\boldsymbol{\phi}|^{2(Q+1-P)}\right)^{-1} |\boldsymbol{\phi}|^{2(Q+1-P)} v(\boldsymbol{\phi}, \Pi_h), \tag{8.5}$$

where $v(\boldsymbol{\phi}, \Pi_h)$ and $A(\boldsymbol{\phi})$ are defined by (6.7) and (6.14).

**Corollary 8.3.** *Let the scheme* (2.8) *be stable, $\Pi_h$ be a local mapping, and* (2.17) *hold. In a neighborhood of $\boldsymbol{\phi} = 0$ there holds*

$$\|v(\boldsymbol{\phi}, \Pi_h) - \mathcal{W}^{P,Q}(\boldsymbol{\phi})\| \leqslant \delta C_1|\boldsymbol{\phi}|^P, \quad \|A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi})\| \leqslant \delta C_2|\boldsymbol{\phi}|^{Q+1}, \tag{8.6}$$

*where $\delta$ depends only on $M^0$, $\|\Pi_h 1\|$, and the stability constant.*

*Proof.* We have

$$\|v(\boldsymbol{\phi},\Pi_h)\| = \|(\Pi_1 \exp(i\boldsymbol{\phi}\cdot\boldsymbol{r}))_0\| \to \|\Pi_h 1\| \quad \text{as} \quad \boldsymbol{\phi}\to 0.$$

From (2.17) by Lemma 6.11 in a neighborhood of $\boldsymbol{\phi}=0$ there holds

$$\|\hat{\varepsilon}(\boldsymbol{\phi},\nu,\Pi_h)\| \leqslant \frac{3}{2}C_1|\boldsymbol{\phi}|^P + \nu C_2|\boldsymbol{\phi}|^{Q+1}.$$

By (6.16) in a neighborhood of $\boldsymbol{\phi}=0$ there holds

$$\|(e^{\nu A(\boldsymbol{\phi})}-I)v(\boldsymbol{\phi},\Pi_h)\| \leqslant (2C_1|\boldsymbol{\phi}|^P\|\Pi_h 1\|^{-1} + 2C_2|\boldsymbol{\phi}|^{Q+1}\|\Pi_h 1\|^{-1}\nu)\|v(\boldsymbol{\phi},\Pi_h)\|.$$

It remains to use Theorem 8.2. □

Now we are ready to prove Theorem 2, which we repeat here.

**Theorem 2.** *Let $\Pi_h$ be a local mapping with $\mu_\xi \in (W_2^q(G))^*$ or $(C^q(G))^*$. Let $P,Q>0$, $r\geqslant\max\{P,q\}$, and $R\geqslant\max\{Q+1,r\}$. Let the scheme* (2.8) *be stable and possess the error estimate* (2.17) *on $H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$). Then there exists a homogeneous mapping $\tilde{\Pi}_h : L_{2,per}(\mathbb{R}^d)\to V_{per}$ such that*

$$\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant C(h^P\|\nabla^P f\| + h^r\|\nabla^r f\|_*), \quad \|\epsilon_h(f,\tilde{\Pi}_h)\| \leqslant Ch^Q\|\nabla^{Q+1}f\| \tag{8.7}$$

*for each $h$ and $f\in H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$), where $C$ does not depend on $h$ and $f$, and $\|\cdot\|_*$ means either $\|\cdot\|$ or $\|\cdot\|_\infty$ depending on the case.*

*Proof.* First we consider the case $\mu_\xi\in(W_2^q(G))^*$. Recall the notation $\mathcal{A}_\beta=\{\boldsymbol{\alpha}\in\mathcal{A} : T^*\boldsymbol{\alpha}\in[-\beta,\beta)^d\}$ for $\beta>0$. By Corollary 8.3 there exists $\beta>0$ such that (8.6) holds for $\boldsymbol{\phi}\in\mathcal{A}_\beta$. Since $v(-\boldsymbol{\phi},\Pi_h)=\overline{v(\boldsymbol{\phi},\Pi_h)}$ and $A(-\boldsymbol{\phi})=\overline{A(\boldsymbol{\phi})}$, by construction we have $\mathcal{W}^{P,Q}(-\boldsymbol{\phi})=\overline{\mathcal{W}^{P,Q}(\boldsymbol{\phi})}$. Using the continuity of $v(\boldsymbol{\phi},\Pi_h)$ and the first inequality in (8.6) we get

$$\sup_{\boldsymbol{\phi}\in\mathcal{A}_\beta}\|\mathcal{W}^{P,Q}(\boldsymbol{\phi})\| \leqslant \sup_{\boldsymbol{\phi}\in\mathcal{A}_\beta}\|v(\boldsymbol{\phi},\Pi_h)\|$$

By Lemma 6.6 with $G=\mathcal{A}_\beta$ there exists a bounded homogeneous mapping $\tilde{\Pi}_h$ such that $\mathcal{W}^{P,Q}(\boldsymbol{\phi})=v(\boldsymbol{\phi},\tilde{\Pi}_h)$. Let $F_h=\tilde{\Pi}_h-\Pi_h$. For $\boldsymbol{\alpha}\in\mathcal{A}_{\beta/h}$ using the first inequality in (8.6) we obtain

$$\|F_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}\| = \|(F_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}})_0\| = \|(\tilde{\Pi}_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}})_0 - (\Pi_h e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}})_0\| = \|\mathcal{W}^{P,Q}(\boldsymbol{\alpha}h) - v(\boldsymbol{\alpha}h,\Pi_h)\| \leqslant \delta C_1 h^P|\boldsymbol{\alpha}|^P. \tag{8.8}$$

Applying Lemma 6.14 we obtain the first inequality in (8.7). The second inequality in (8.6) yields the second inequality in (8.7) by Theorem 6.15.

Now consider the case $\mu_\xi\in(C^q(G))^*$. By Lemma 5.9 there exists a local mapping $\hat{\Pi}_h$ with the kernel $\hat{\mu}_\xi\in L_2(G+B_1(0))$ such that

$$\|\Pi_h f - \hat{\Pi}_h f\| \leqslant Ch^r\|\nabla^r f\|_\infty.$$

By the above proof we have (8.7) with the substitution of $\hat{\Pi}_h$ for $\Pi_h$, and it remains to use the triangle inequality. □

### *8.3. An analytical criterion*

For a given scheme (2.8), a mapping $\Pi_h$, and two numbers $0<P\leqslant Q<\infty$ denote

$$\mathcal{F}^{P,Q}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2P}v^*(\boldsymbol{\phi},\Pi_h)\left(A^*(\boldsymbol{\phi})A(\boldsymbol{\phi}) + |\boldsymbol{\phi}|^{2(Q+1-P)}\right)^{-1}A^*(\boldsymbol{\phi})A(\boldsymbol{\phi})v(\boldsymbol{\phi},\Pi_h). \tag{8.9}$$

**Lemma 8.4.** *For $\mathcal{W}^{P,Q}$ given by* (8.5) *there holds*

$$\mathcal{F}^{P,Q}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2P}\|v(\boldsymbol{\phi},\Pi_h)-\mathcal{W}^{P,Q}(\boldsymbol{\phi})\|^2 + |\boldsymbol{\phi}|^{-2(Q+1)}\|A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi})\|^2. \tag{8.10}$$

*Proof.* We have

$$|\boldsymbol{\phi}|^{-2(Q+1-P)}A^*(\boldsymbol{\phi})A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi}) = v(\boldsymbol{\phi},\Pi_h) - \mathcal{W}^{P,Q}(\boldsymbol{\phi}).$$

Thus

$$|\boldsymbol{\phi}|^{-2(Q+1-P)}\|A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi})\|^2 =$$
$$= |\boldsymbol{\phi}|^{-2(Q+1-P)}(A^*(\boldsymbol{\phi})A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi}), \mathcal{W}^{P,Q}(\boldsymbol{\phi})) = (v(\boldsymbol{\phi},\Pi_h) - \mathcal{W}^{P,Q}(\boldsymbol{\phi}), \mathcal{W}^{P,Q}(\boldsymbol{\phi})).$$

On the other hand,

$$|\boldsymbol{\phi}|^{2P}\mathcal{F}^{P,Q}(\boldsymbol{\phi}) = (|\boldsymbol{\phi}|^{-2(Q+1-P)}A^*(\boldsymbol{\phi})A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi}), v(\boldsymbol{\phi},\Pi_h)) = (v(\boldsymbol{\phi},\Pi_h) - \mathcal{W}^{P,Q}(\boldsymbol{\phi}), v(\boldsymbol{\phi},\Pi_h)).$$

Now (8.10) is obvious. □

The following theorem establishes an analytical criterion which allows us (at least in theory) to find the optimal values of the formal order of accuracy and the long-time simulation order.

**Theorem 8.5.** *Let the scheme* (2.8) *be stable,* $\Pi_h$ *be a local mapping with kernel* $\mu_\xi \in (W_2^q(G))^*$ *(or* $(C^q(G))^*$*),* $0 < P \leqslant Q < \infty$*,* $R \geqslant \max\{Q, q+1\}$*. The scheme possesses the formal order of accuracy* $P$ *and the long-time simulation order* $Q$ *on* $H^R_{per}(\mathbb{R}^d)$ *(or* $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$*) if and only if* $\mathcal{F}^{P,Q}(\boldsymbol{\phi})$ *is bounded in a neighborhood of* $\boldsymbol{\phi} = 0$*.*

*Proof.* Assume the error estimate (2.17). Then in an neighborhood of $\boldsymbol{\phi} = 0$ we have (8.6) and thus by (8.10) there holds

$$\mathcal{F}^{P,Q}(\boldsymbol{\phi}) \leqslant |\boldsymbol{\phi}|^{-2P}\left(\delta^2 C_1^2|\boldsymbol{\phi}|^{2P}\right) + |\boldsymbol{\phi}|^{-2(Q+1)}\left(\delta^2 C_2^2|\boldsymbol{\phi}|^{2(Q+1)}\right) = \delta^2(C_1^2 + C_2^2).$$

On the other hand, assuming that $\mathcal{F}(\boldsymbol{\phi}) \leqslant C^2$ from (8.10) we get

$$\|v(\boldsymbol{\phi},\Pi_h) - \mathcal{W}^{P,Q}(\boldsymbol{\phi})\| \leqslant C|\boldsymbol{\phi}|^P, \quad \|A(\boldsymbol{\phi})\mathcal{W}^{P,Q}(\boldsymbol{\phi})\| \leqslant C|\boldsymbol{\phi}|^{Q+1}.$$

By Lemma 8.1 the scheme possesses error estimate (2.17) on $H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$). □

**Lemma 8.6.** *Let the scheme* (2.8) *be stable,* $\Pi_h$ *be a local mapping with kernel* $\mu_\xi \in (W_2^q(G))^*$ *(or* $(C^q(G))^*$*),* $0 < P \leqslant Q < \infty$*,* $R \geqslant \max\{Q, q+1\}$*. Let* $\tilde{\boldsymbol{e}} \in \mathring{\Omega}$*. The scheme possesses the formal order of accuracy* $P$ *and the long-time simulation order* $Q$ *on* $H^R_{per,\tilde{\boldsymbol{e}}}(\mathbb{R}^d)$ *(or* $C^R_{per,\tilde{\boldsymbol{e}}}(\mathbb{R}^d)$*) if and only if* $\mathcal{F}^{P,Q}(\tilde{\boldsymbol{e}}\psi)$ *is bounded as* $\psi$ *tends to zero.*

The proof is similar to Theorem 8.5.

**Proposition 8.7.** *Let* $\Pi_h$ *and* $\mathcal{P}_h$ *be local mappings with kernels* $\mu_\xi, \hat{\mu}_\xi \in (W_2^q(G))^*$ *(or* $(C^q(G))^*$*), and* $\langle\hat{\mu}_\xi, 1\rangle \neq 0$ *for each* $\xi \in M^0$*. Let* $\boldsymbol{e} \in \mathring{\Omega}$*. Suppose a stable scheme of the form* (2.8) *possesses the formal order of accuracy* $P$ *and the long-time simulation order* $Q < \infty$ *in the sense of* $\Pi_h$ *on* $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ *(or* $C^{\lceil R\rceil}_{per,\boldsymbol{e}}(\mathbb{R}^d)$*), where* $R = \max\{Q+1, q\}$*. Then there exist real-valued diagonal matrices* $\mathfrak{C}^{(m)}_{\boldsymbol{e}}$*,* $m = \lceil P\rceil, \ldots, \lceil Q\rceil$*, such that the scheme possesses the truncation error of order* $\lceil Q\rceil$ *in the sense of* $\Pi^{(\lceil P\rceil,\lceil Q\rceil)}_{h,\boldsymbol{e}}$ *given by* (2.12) *on the direction* $\boldsymbol{e}$*. As a corollary, the optimal values of the order of accuracy and of the long-time simulation order on* $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ *(or* $C^{\lceil R\rceil}_{per,\boldsymbol{e}}(\mathbb{R}^d)$*) are integers.*

*Proof.* Let $\mathcal{W}^{P,Q}(\boldsymbol{\phi})$ be given by (8.5) and $W(\psi) = \mathcal{W}^{P,Q}(\boldsymbol{e}\psi)$. By Lemma 8.6 the function $\mathcal{F}^{P,Q}(\boldsymbol{e}\psi)$, where $\mathcal{F}^{P,Q}$ is defined by (8.9), is bounded as $\psi \to 0$. Using (8.10) for $\boldsymbol{\phi} = \boldsymbol{e}\psi$ we have

$$\|v(\psi\boldsymbol{e},\Pi_h) - W(\psi)\| = O(|\psi|^P), \quad \|A(\psi\boldsymbol{e})W(\psi)\| = O(|\psi|^{Q+1}). \tag{8.11}$$

By construction each component of $W(\psi)$ is a ratio of two analytical functions of the real argument $\psi$; by (8.11) it is bounded as $\psi \to 0$, thus $W(\psi)$ is an analytical function. Thus both expressions under norm signs in (8.11) are analytical at $\psi = 0$ and so $P$ and $Q$ can be replaced by $\lceil P\rceil$ and $\lceil Q\rceil$, correspondingly. By Lemma 6.8 there exist

real-valued diagonal matrices $\mathfrak{C}_{\boldsymbol{e}}^{(m)}$, $m = \lceil P \rceil, \dots, \lceil Q \rceil$, such that for the mapping $\Pi_{h,\boldsymbol{e}}^{(\lceil P \rceil, \lceil Q \rceil)}$ given by (2.13) there holds

$$v(\psi \boldsymbol{e}, \Pi_{h,\boldsymbol{e}}^{(\lceil P \rceil, \lceil Q \rceil)}) = W(\psi) + O(|\psi|^{\lceil Q \rceil + 1}),$$

thus by the triangle inequality we have

$$\|\hat{\epsilon}(\psi \boldsymbol{e}, \Pi_{h,\boldsymbol{e}}^{(\lceil P \rceil, \lceil Q \rceil)})\| = \|A(\psi \boldsymbol{e}) v(\psi \boldsymbol{e}, \Pi_{h,\boldsymbol{e}}^{(\lceil P \rceil, \lceil Q \rceil)})\| = O(|\psi|^{\lceil Q \rceil + 1}).$$

It remains to use Theorem 6.15. □

Given a scheme with enhanced accuracy in the long-time simulation, Theorem 2 states that there exists an auxiliary mapping $\tilde{\Pi}_h$ explaining this fact. This mapping is generally not of the form (2.11). For the 1D case, Proposition 8.7 states the existence of an auxiliary mapping of the form (2.11). Hovewer, in the multidimensional case, an auxiliary mapping of the form (2.11) generally does not exist, see a counter-example in Section 12.3.

## 9. The good

Throughout this section we consider a stable scheme of the form (2.8) with stability constant $K$ and a local mapping $\Pi_h$. In this section we will prove Theorem 1 and Theorem 3, which is splitted into Theorems 3A and 3B.

### *9.1. The order of accuracy*

In this subsection we establish the optimal value of the formal order of accuracy.

**Lemma 9.1.** *Let the matrix $A(\boldsymbol{\phi})$ be given by* (6.14)*. Then the eigenvalue $\lambda = 0$ of $A(0)$ is semisimple, i. e. there are no Jordan cells of size greater than one corresponding to $\lambda = 0$.*

*Proof.* Assume the converse. Represent the matrix $A(0)$ in the form $A(0) = SJS^{-1}$ where $J$ is its Jordan normal form. Then $e^{\nu A(0)} = Se^{\nu J}S^{-1}$. The explicit expression for $e^{\nu J}$ shows that $\|e^{\nu J}\|$ grows unlimitedly as $\nu \to \infty$ and so does $\|e^{\nu A(0)}\|$. This contradicts the stability condition (6.21). □

Consider $\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)$ given by (6.16). By Theorem 5.19 in a neighborhood of $\boldsymbol{\phi} = 0$ there holds $A(\boldsymbol{\phi}) = S(\boldsymbol{\phi})M(\boldsymbol{\phi})S^{-1}(\boldsymbol{\phi})$ where $M(\boldsymbol{\phi})$ has the form (5.10). Let $v_j(\boldsymbol{\phi}, \Pi_h)$ be the components of $S^{-1}(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h)$ corresponding to the blocks $M^{(j)}(\boldsymbol{\phi})$. Denote

$$E_j(\boldsymbol{\phi}, \nu, \Pi_h) = \left(e^{\nu M^{(j)}(\boldsymbol{\phi})} - I\right) v_j(\boldsymbol{\phi}, \Pi_h), \tag{9.1}$$

then

$$\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h) = S(\boldsymbol{\phi})(e^{M(\boldsymbol{\phi})\nu} - I)S^{-1}(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h) = S(\boldsymbol{\phi}) \begin{pmatrix} \vdots \\ E_j(\boldsymbol{\phi}, \nu, \Pi_h) \\ \vdots \end{pmatrix}. \tag{9.2}$$

Note that the notation is similar to the 1D case but the decomposition is different. We will also use the notation $M^{(*)}(\boldsymbol{\phi})$ for the submatrix of $M(\boldsymbol{\phi})$ containing the blocks $M^{(j)}(\boldsymbol{\phi})$, $j \neq 0$, and the notation $v_*(\boldsymbol{\phi}, \Pi_h)$, $E_*(\boldsymbol{\phi}, \nu, \Pi_h)$ for the unions of all blocks excluding $j = 0$ of the corresponding vectors. Below we assume that the norms of subvectors are inherited from $\mathbb{C}^{M^0}$, i. e., for example, $\|v_j\| = \|(0, \dots, 0, v_j, 0, \dots, 0)^T\|$.

Since $M^{(0)}(0) = 0$, there holds $M^{(0)}(\boldsymbol{\phi}) = \sum_{k=1}^{d} M_k^{(0)}(\boldsymbol{\phi})\phi_k$ where $M_k^{(0)}(\boldsymbol{\phi})$ are holomorphic. Thus

$$M^{(0)}(\boldsymbol{\alpha} h)\frac{t}{h} = \sum_{k=1}^{d} M_k^{(0)}(\boldsymbol{\alpha} h)\alpha_k t$$

is a holomorphic function of $t$, $h$, and $\boldsymbol{\alpha}$. Particularly, if $A(0) = 0$ then $\hat{\varepsilon}(\boldsymbol{\alpha} h, t/h, \Pi_h)$ is a holomorphic function of $t$, $h$, and $\boldsymbol{\alpha}$ for each $t$ in a neighborhood of $h = 0$ and $\boldsymbol{\alpha} = 0$.

**Lemma 9.2.** *Suppose the scheme* (2.8) *possesses the estimate* (2.17) *in the sense of a local mapping* $\Pi_h$. *Then in a neighborhood of* $\boldsymbol{\phi} = 0$ *for each* $\nu \geqslant 0$ *there holds*

$$\|v_*(\boldsymbol{\phi}, \Pi_h)\| \leqslant \tilde{\delta} C_1 |\boldsymbol{\phi}|^{\lceil P \rceil}, \quad \|E_*(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant (K+1)\tilde{\delta} C_1 |\boldsymbol{\phi}|^{\lceil P \rceil}. \tag{9.3}$$

*Moreover, if* $P$ *is integer, then* $\tilde{\delta}$ *depends only on* $|M^0|$, $K$, *and the norm on* $\mathbb{C}^{M^0}$.

*Proof.* By Lemma 6.11 in a neighborhood of $\boldsymbol{\phi} = 0$ there holds

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant 2C_1 |\boldsymbol{\phi}|^P + C_2 \nu |\boldsymbol{\phi}|^{Q+1}.$$

Thus for each $j$ we have

$$\|E_j(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant c_n \|S^{-1}(\boldsymbol{\phi})\| \left(2C_1 |\boldsymbol{\phi}|^P + C_2 \nu |\boldsymbol{\phi}|^{Q+1}\right),$$

where $c_n$ depends on the norms in use. Directly from (9.1) for each $j \neq 0$ and each $\nu \geqslant 0$ we get

$$\|v_j(\boldsymbol{\phi}, \Pi_h)\| \leqslant \left\| \left[ \exp\left(M^{(j)}(\boldsymbol{\phi})\nu\right) - I \right]^{-1} \right\| \|E_j(\boldsymbol{\phi}, \nu, \Pi_h)\|.$$

Put $\nu = 1/(2\|M^{(j)}(0)\|)$. Then by Lemma 5.13 in a neighborhood of $\boldsymbol{\phi} = 0$ we have

$$\begin{aligned} \|v_j(\boldsymbol{\phi}, \Pi_h)\| &\leqslant 4 \left\| \left( \frac{M^{(j)}(\boldsymbol{\phi})}{2\|M^{(j)}(0)\|} \right)^{-1} \right\| \|E_j(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant 10 \varkappa(M^{(j)}(0)) \left\| E_j\left(\boldsymbol{\phi}, \frac{1}{2\|M^{(j)}(0)\|}, \Pi_h\right) \right\| \leqslant \\ &\leqslant 20 c_n \delta^2 \left( 2C_1 \, |\boldsymbol{\phi}|^P + C_2 \, \frac{1}{2\|M^{(j)}(0)\|} |\boldsymbol{\phi}|^{Q+1} \right) \leqslant 100 \delta^2 c_n C_1 \, |\boldsymbol{\phi}|^P. \end{aligned}$$

Since $v_j(\boldsymbol{\phi}, \Pi_h)$ is holomorphic then the degree $P$ in the right-hand side can be improved to $\lceil P \rceil$. This leads to the first inequality in (9.3). The second inequality follows from the first by stability. □

**Lemma 9.3.** *Let the scheme* (2.8) *possess the estimate* (2.17) *in the sense of a local mapping* $\Pi_h$. *Then in a neighborhood of* $\boldsymbol{\phi} = 0$ *there holds*

$$\|M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h)\| \leqslant C |\boldsymbol{\phi}|^{\lceil P \rceil + 1}. \tag{9.4}$$

*Proof.* Similar to the previous lemma, we have

$$\|E_0(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant c_n \|S^{-1}(\boldsymbol{\phi})\| \left(2C_1 |\boldsymbol{\phi}|^P + C_2 \nu |\boldsymbol{\phi}|^{Q+1}\right). \tag{9.5}$$

Using the function $f$ given by (5.7) ($f(x) = (e^x - 1)/x$), formula (9.1) is equivalent to

$$E_0(\boldsymbol{\phi}, \nu, \Pi_h) = f\left(M^{(0)}(\boldsymbol{\phi})\nu\right) M^{(0)}(\boldsymbol{\phi}) \, \nu \, v_0(\boldsymbol{\phi}, \Pi_h).$$

Assuming $M^{(0)}(\boldsymbol{\phi}) \neq 0$ put $\nu = 1/\|M^{(0)}(\boldsymbol{\phi})\|$. By Lemma 5.12 we get $\|(f(\nu M^{(0)}(\boldsymbol{\phi})))^{-1}\| \leqslant 4$, thus

$$\left\| M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h) \right\| \leqslant \frac{4}{\nu} \|E_0(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant C_1' |\boldsymbol{\phi}|^P \|M^{(0)}(\boldsymbol{\phi})\| + C_2' |\boldsymbol{\phi}|^{Q+1}.$$

If $M^{(0)}(\boldsymbol{\phi}) = 0$ then $M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h) = 0$. Since $M^{(0)}(\boldsymbol{\phi}) = O(|\boldsymbol{\phi}|)$ as $\boldsymbol{\phi} \to 0$ and $Q \geqslant P$ we get $\|M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h)\| = O(|\boldsymbol{\phi}|^{P+1})$ as $\boldsymbol{\phi} \to 0$. Taking into account that $M^{(0)}(\boldsymbol{\phi})$ and $v_0(\boldsymbol{\phi}, \Pi_h)$ are holomorphic, we arrive at (9.4). □

Now we are ready to prove Theorem 1, which we repeat here. Recall that for $n \in \mathbb{N}$ by $\mathring{\Omega}_n$ we denote a set of vectors $\{\boldsymbol{e}_k \in \mathring{\Omega}, k = 1, \ldots, C_{n+d-1}^{d-1}\}$ such that $\{(\boldsymbol{e}_k \cdot \boldsymbol{r})^n\}$ form a basis in the set of homogeneous polynomials of order $n$ (see Lemma A.6).

**Theorem 1.** *Consider a scheme of the form* (2.8)*, stable with a constant $K$, and local mappings $\Pi_h$, $\mathcal{P}_h$ with kernels $\mu_\xi, \hat{\mu}_\xi \in (W_2^q(G))^*$ (or $(C^q(G))^*$). Let $P_A$ and $P$ be the optimal orders of the truncation error and accuracy, correspondingly, in the sense of $\Pi_h$. Let $\hat{\varkappa} = \min_{\xi \in M^0} |\langle \hat{\mu}_\xi, 1\rangle| > 0$ and $R = \max\{q, P\} + 1$. Then the following holds.*

1. *$P_A$ and $P$ are integers.*
2. *Either $P = P_A$ or $P = P_A + 1$.*
3. *If $P = P_A + 1$ then there exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}| = P$, such that the scheme possesses the truncation error of order $P$ in the sense of $\Pi_h^{(P,P)}$ given by* (2.12)*. Moreover, $\|\mathfrak{C}^{(\boldsymbol{m})}\| \leqslant \tilde{\delta} C_1$, where $C_1$ is the constant in estimate* (2.17) *and $\tilde{\delta}$ depends only on $K$, $\hat{\varkappa}$, $P$, $|M^0|$, and the norm on $\mathbb{C}^{M^0}$.*
4. *If $P = P_A$ then there exists no set of matrices $\{\mathfrak{C}^{(\boldsymbol{m})}, |\boldsymbol{m}| = P_A + 1\}$, such that the scheme possesses the truncation error of order $P_A + 1$ in the sense of $\Pi_h^{(P_A+1,P_A+1)}$ given by* (2.12)*.*
5. *If $L(0) = 0$ then $P = P_A$.*
6. *If the scheme possesses the formal order of accuracy $P_A+1$ on $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ (or $C^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$) for each $\boldsymbol{e} \in \mathring{\Omega}_{P_A+1}$, then it possesses the formal order of accuracy $P = P_A + 1$ on $H^R_{per}(\mathbb{R}^d)$ (or $C^R_{per}(\mathbb{R}^d)$).*
7. *$P$ coincides with the optimal order of accuracy in the weak sense.*

*Proof.* 1. Let the scheme possess the order of the truncation error $P_A$. Then by Corollary 6.21 there holds (6.31). Taking $v_0 = \exp(i\boldsymbol{\alpha} \cdot \boldsymbol{r})$, by Lemma 6.10 in a neighborhood of $\boldsymbol{\phi} = 0$ we get $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant c|\boldsymbol{\phi}|^{P_A+1}$. Since $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)$ is a holomorphic function, then we have $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant \tilde{c}|\boldsymbol{\phi}|^{\lceil P_A \rceil + 1}$. By Theorem 6.15 the scheme possesses the truncation error of order $\lceil P_A \rceil$. Thus the optimal value of the order of the truncation error is an integer.

Let the scheme possess the order of accuracy $p$ (possibly not optimal). Put by definition

$$w(\boldsymbol{\phi}) = S(\boldsymbol{\phi}) \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} S^{-1}(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \Pi_h), \tag{9.6}$$

where the identity matrix corresponds to the block $M^{(0)}(\boldsymbol{\phi})$. By Lemma 9.2 we have

$$\|w(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h)\| = \left\| S(\boldsymbol{\phi}) \begin{pmatrix} v_*(\boldsymbol{\phi}, \Pi_h) \\ 0 \end{pmatrix} \right\| \leqslant c_n \|S(\boldsymbol{\phi})\| \|v_*(\boldsymbol{\phi}, \Pi_h)\| \leqslant \delta C_1 |\boldsymbol{\phi}|^{\lceil p \rceil}. \tag{9.7}$$

If $p$ is integer, $\delta$ in (9.7) depends only on $K$, $|M^0|$ and the choice of norms. By Lemma 9.3 there holds

$$A(\boldsymbol{\phi}) w(\boldsymbol{\phi}) = S(\boldsymbol{\phi}) \begin{pmatrix} 0 \\ M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h) \end{pmatrix} = O(|\boldsymbol{\phi}|^{\lceil p \rceil + 1}). \tag{9.8}$$

By Lemma 8.1 the scheme possesses the formal order of accuracy $\lceil p \rceil$. This proves that the optimal order of accuracy is an integer.

2. By Lemma 9.2 and Lemma 9.3 there holds

$$\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| = \left\| S(\boldsymbol{\phi}) \begin{pmatrix} M^{(*)}(\boldsymbol{\phi}) v_*(\boldsymbol{\phi}, \Pi_h) \\ M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h) \end{pmatrix} \right\| = O(|\boldsymbol{\phi}|^P).$$

By Theorem 6.15 the scheme possesses the order of the truncation error $P - 1$. Thus $P \leqslant P_A + 1$. On the other hand, by Theorem 4.1 we have $P \geqslant P_A$. Since both $P$ and $P_A$ are integers there holds either $P = P_A$ or $P = P_A + 1$.

3. Now assume $P = P_A + 1$. Let $w$ be given by (9.6). Since $v(\boldsymbol{\phi}, \Pi_h) = \overline{v(-\boldsymbol{\phi}, \Pi_h)}$ and $A(\boldsymbol{\phi}) = \overline{A(-\boldsymbol{\phi})}$, for $W(\boldsymbol{\phi}) = (w(\boldsymbol{\phi}) + \overline{w(-\boldsymbol{\phi})})/2$ by (9.7) and (9.8) with $p = P$ we have

$$\|W(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h)\| \leqslant \delta c_n C_1 |\boldsymbol{\phi}|^P, \quad A(\boldsymbol{\phi}) W(\boldsymbol{\phi}) = O(|\boldsymbol{\phi}|^{P+1}). \tag{9.9}$$

By Lemma 6.7 there exist $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}| = P$, such that for $\Pi_h^{(P,P)}$ there holds

$$v(\boldsymbol{\phi}, \Pi_h^{(P,P)}) = W(\boldsymbol{\phi}) + O(|\boldsymbol{\phi}|^{P+1}).$$

By the triangle inequality we get

$$\hat\epsilon(\boldsymbol\phi,\Pi_h^{(P,P)}) = A(\boldsymbol\phi)v(\boldsymbol\phi,\Pi_h^{(P,P)}) = O(|\boldsymbol\phi|^{P+1}),$$

so the coefficients $\mathfrak{C}^{(\boldsymbol m)}$ satisfy the statement of the theorem. By construction (see (6.12) and (6.13)),

$$(\mathfrak{C}^{(\boldsymbol m)})_{\xi,\xi} = \frac{1}{i^{|\boldsymbol m|}\boldsymbol m!}\left(v_\xi(0,\mathcal{P}_h)\right)^{-1}\left(D^{\boldsymbol m}\left(W(\boldsymbol\phi)-v(\boldsymbol\phi,\Pi_h)\right)|_{\boldsymbol\phi=0}\right)_\xi.$$

Using (9.9) and Lemma 5.2 we get $|(\mathfrak{C}^{(\boldsymbol m)})_{\xi,\xi}| \leqslant \hat\varkappa^{-1}\delta c_n c_P C_1$, where $c_P$ depends on $P$ only. From here, the estimate for $\|\mathfrak{C}^{(\boldsymbol m)}\|$ is obvious.

4. Suppose there exists a set of matrices $\mathfrak{C}^{(\boldsymbol m)}$ such that the scheme possesses the truncation error of order $P_A+1$ in the sense of $\Pi_h^{(P_A+1,P_A+1)}$. Then for $w(\boldsymbol\phi) = v(\boldsymbol\phi,\Pi_h^{(P_A+1,P_A+1)})$ we have (8.1) with $P=Q=P_A+1$ and thus by Lemma 8.1 the scheme possesses the formal order of accuracy $P_A+1$ and the long-time simulation order $P_A+1$.

5. If $L(0)=0$, the block $M^{(*)}(\boldsymbol\phi)$ does not exist, and by Lemma 9.3 we have $\|\hat\epsilon(\boldsymbol\phi,\Pi_h)\| = O(|\boldsymbol\phi|^{P+1})$. By Theorem 6.15 the scheme possesses the order of the truncation error $P$, i. e. there holds $P=P_A$.

6. Now let the scheme possess the order of accuracy $p=P_A+1$ on $H^R_{per,\boldsymbol e}(\mathbb{R}^d)$ for each $\boldsymbol e\in\mathring\Omega_p$. Arguing as in the proof of Lemma 6.11 for $\boldsymbol\phi\in\operatorname{span}\{\boldsymbol e\}$ in a neighborhood of $\boldsymbol\phi=0$ we get

$$\|\hat\varepsilon(\boldsymbol\phi,\nu,\Pi_h)\| \leqslant c_1(\boldsymbol e)|\boldsymbol\phi|^p + c_2(\boldsymbol e)\nu|\boldsymbol\phi|^{p+1}.$$

Following proofs of Lemmas 9.2 and 9.3 for $\boldsymbol\phi\in\operatorname{span}\{\boldsymbol e\}$ in a neighborhood of $\boldsymbol\phi=0$ there hold (9.3) and (9.4). By Lemma 5.3 the inequalities (9.3) and (9.4) hold for each $\boldsymbol\phi$ in a neighborhood of $\boldsymbol\phi=0$ with another multiplicative constants. Substituting this into (9.1) by stability we get $E_*(\boldsymbol\phi,\nu,\Pi_h) = O(|\boldsymbol\phi|^{\lceil p\rceil})$ and $E_0(\boldsymbol\phi,\nu,\Pi_h) = O(\nu|\boldsymbol\phi|^{\lceil p\rceil+1})$, thus $\hat\varepsilon(\boldsymbol\phi,\nu,\Pi_h) = O(|\boldsymbol\phi|^{\lceil p\rceil}+\nu|\boldsymbol\phi|^{\lceil p\rceil+1})$. It remains to apply Theorem 6.17.

7. If the scheme possesses the order of accuracy $P$ in the weak sense on $H^R_{per}(\mathbb{R}^d)$, then by Proposition 6.22 for each $\boldsymbol e\in\mathring\Omega$ it possesses the order of accuracy $P$ in the strong sense on $H^R_{per,\boldsymbol e}(\mathbb{R}^d)$, and it remains to use the previous statement. □

An auxiliary mapping that provides (8.7) can't be generally found in the form (2.12), see Section 12.2 for the example. We just proved that this is possible if $Q=P$. Now we consider another two cases when this is possible, namely, the simple case and the quasi-1D case.

*9.2. The simple case*

Below we assume that the scheme is *0-exact*, i. e. the scheme preserves constant solutions, or, that is the same, $\sum L_{\boldsymbol\eta}(\Pi_1 1)_0 = 0$. Here the subscript 1 means the substitution 1 for $h$ and the subscript 0 means the substitution 0 for $\boldsymbol\eta$. Otherwise the order of accuracy is $P_A=-1$ and by Theorem 1 the numerical solution does not converge to the exact one.

Since the scheme is 0-exact, then the operator $A(0)$ has nontrivial kernel: $A(0)\vec{\mathfrak{e}}=0$, where $\vec{\mathfrak{e}} = v(0,\Pi_h)$. We shall say that the scheme is *simple* if the zero eigenvalue of the operator $A(0)$ is simple or, this is the same, $\dim\operatorname{Ker}L(0)=1$. The scheme is simple iff the vector $v_0(\boldsymbol\phi,\Pi_h)$ defined in Section 9.1 is of size 1.

**Proposition 9.4.** *Suppose a scheme has no steady solutions but the constant. Then it is simple.*

*Proof.* Let $\mathfrak{E}\in V^1_{per}$ be the constant sequence formed by the element $\vec{\mathfrak{e}}$. By assumption the scheme has no steady solutions but $c\mathfrak{E}$, $c\in\mathbb{R}$. So $\operatorname{Ker}L\cap V_{per} = \operatorname{span}\{\mathfrak{E}\}$. By (6.14) we have $A(0) = -Z^{-1}(0)L(0)$. Consider a vector $w_0\in\mathbb{C}^{M^0}$ such that $w_0\notin\operatorname{span}\{\vec{\mathfrak{e}}\}$. Let $w\in V^1_{per}$ be the constant sequence all elements of which are equal to $w_0$. Then $Lw\neq 0$. By (6.4) all the block components of $Lw$ are equal to $L(0)w_0$, thus we have $L(0)w_0\neq 0$. Then $A(0)w_0 = -Z^{-1}(0)L(0)w_0\neq 0$. Hence, $w_0=\vec{\mathfrak{e}}$ is the only (up to a factor) solution of $A(0)w_0=0$. Therefore by Lemma 9.1 we obtain that $\lambda=0$ is a simple eigenvalue of $A(0)$. □

The discontinuous Galerkin method with the upwind flux in 1D is a simple scheme unless the transport velocity is zero. On a multidimensional simplicial translationally-invariant (i. e. invariant with respect to the translation by the vector of any mesh edge) mesh the DG method with the upwind flux is a simple scheme unless the transport velocity $\boldsymbol\omega$ is parallel to one of the mesh faces.

**Lemma 9.5.** *Suppose a simple scheme* (2.8) *possesses the estimate* (2.17) *in the sense of a local mapping* $\Pi_h$. *Then in a neighborhood of* $\boldsymbol{\phi} = 0$ *there holds*

$$\|M^{(0)}(\boldsymbol{\phi})\| \leqslant C|\boldsymbol{\phi}|^{\lceil Q\rceil+1}. \tag{9.10}$$

*Proof.* Denote by $\lambda(\boldsymbol{\phi})$ the only element of the matrix $M^{(0)}(\boldsymbol{\phi})$. It is a holomorphic function. By Lemma 9.2 in a neighborhood of $\boldsymbol{\phi} = 0$ we have $\|v_*(\boldsymbol{\phi}, \Pi_h)\| \leqslant C|\boldsymbol{\phi}|^P$. Then $v_0(0, \Pi_h) \neq 0$; assuming the converse, we immediately get $(\Pi_1 1)_0 = v(0, \Pi_h) = 0$, which contradicts the definition of the local mapping. Recall that $v_0(\boldsymbol{\phi}, \Pi_h)$ is a 1-component quantity. Hence for small $|\boldsymbol{\phi}|$ by (9.5) there holds

$$\left|e^{\nu\lambda(\boldsymbol{\phi})} - 1\right| \leqslant \frac{2c_n}{|v_0(0, \Pi_h)|}(C_1|\boldsymbol{\phi}|^P + C_2\nu|\boldsymbol{\phi}|^{Q+1}).$$

Put $\nu = |\boldsymbol{\phi}|^{-\lceil Q\rceil}$; then the right-hand side tends to zero as $\boldsymbol{\phi} \to 0$. Hence, $\lambda(\boldsymbol{\phi})|\boldsymbol{\phi}|^{-\lceil Q\rceil} \to 0$. Since $\lambda(\boldsymbol{\phi})$ is a holomorphic function of $\boldsymbol{\phi}$ at $\boldsymbol{\phi} = 0$, then $\lambda(\boldsymbol{\phi}) = O(|\boldsymbol{\phi}|^{\lceil Q\rceil+1})$. □

**Lemma 9.6.** *Let* $\Pi_h$ *be a local mapping. Let a simple scheme possess the formal order of accuracy* $P$ *and the long-time simulation order* $Q$ *and, at the same time, the formal order of accuracy* $P'$ *and the long-time simulation order* $Q'$. *Then it possesses the formal order of accuracy* $\lceil\max\{P, P'\}\rceil$ *and the long-time simulation order* $\lceil\max\{Q, Q'\}\rceil$.

*Proof.* By Lemma 9.2 and Lemma 9.5 we have $|\lambda(\boldsymbol{\phi})| \leqslant C|\boldsymbol{\phi}|^{\lceil Q\rceil+1}$ and $\|v^*(\boldsymbol{\phi}, \Pi_h)\| \leqslant C|\boldsymbol{\phi}|^{\lceil P\rceil}$. At the same time, we have similar inequalities with the substitution $P'$ for $P$ and $Q'$ for $Q$. Thus we have the same inequalities with the substitution $\max\{P, P'\}$ for $P$ and $\max\{Q, Q'\}$ for $Q$. By direct substitution into (9.2) and (9.1) we get

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant C|\boldsymbol{\phi}|^{\lceil\max\{P,P'\}\rceil} + C\nu|\boldsymbol{\phi}|^{\lceil\max\{Q,Q'\}\rceil+1}.$$

It remains to apply Theorem 6.17. □

Lemma 9.6 shows that the situation mentioned in the comments to Definition 5 and showed in Section 7.6 is not possible in the simple case.

**Lemma 9.7.** *Consider a simple scheme and a local mapping* $\Pi_h$. *Then the optimal value of the long-time simulation order is integer.*

*Proof.* This follows from the previous lemma with $P = P'$ and $Q = Q'$. □

**Lemma 9.8.** *Consider a simple scheme and a local mapping* $\Pi_h$. *Let* $p \in \mathbb{N}$. *If for each* $q \in \mathbb{N}$ *the scheme possesses the formal order of accuracy* $p$ *and the long-time simulation order* $q$, *then it possesses the formal order of accuracy* $p$ *and the long-time simulation order* $\infty$.

*Proof.* By Lemma 9.5 for each $s \in \mathbb{N}$ we have $\|M^{(0)}(\boldsymbol{\phi})\| \leqslant C_s|\boldsymbol{\phi}|^{s+1}$; since $M^{(0)}(\boldsymbol{\phi})$ is holomorphic at $\boldsymbol{\phi} = 0$, we have $M^{(0)}(\boldsymbol{\phi}) = 0$ in a neighborhood of zero. Using Lemma 9.2 and representation (9.2), (9.1), we have $\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant c|\boldsymbol{\phi}|^p$, where $c$ does not depend on $\nu$. By Theorem 6.17 the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q = \infty$. □

**Proposition 9.9.** *For a stable scheme* (2.8) *with* $|M^0| = 1$ *and a local mapping* $\Pi_h$, *the optimal values of the order of truncation error, the order of accuracy, and the long-time simulation order coincide.*

*Proof.* Let $P$ and $Q \geqslant P$ be the optimal values of the order of accuracy and of the long-time simulation order. Then by Lemma 9.5 we have $\|M^{(0)}(\boldsymbol{\phi})\| \leqslant C|\boldsymbol{\phi}|^{Q+1}$. Since $M^{(0)}$ is a $1 \times 1$-matrix and there is no block $\|M^{(*)}(\boldsymbol{\phi})\|$, we have $\|A(\boldsymbol{\phi})\| = \|M(\boldsymbol{\phi})\| = \|M^{(0)}(\boldsymbol{\phi})\| \leqslant C|\boldsymbol{\phi}|^{Q+1}$. By (6.15) this yields $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \leqslant C|\boldsymbol{\phi}|^{Q+1}\|v(0, \Pi_h)\|$. Then Theorem 6.15 states that the scheme possesses the order $Q$ of the truncation error. By stability $P \geqslant Q$, thus $P = Q$. Clearly, $Q$ is the optimal value of the order of the truncation error, otherwise by stability the scheme possesses the formal order of accuracy and the order of the long-time simulation greater than $Q$, which contradicts the assumption. □

**Lemma 9.10.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with kernels $\mu_\xi, \hat{\mu}_\xi$ such that $\langle \hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$. Let a simple stable scheme of the form* (2.8) *possess the formal order of accuracy $P \in \mathbb{N}$ and the long-time simulation order $Q$ in the sense of $\Pi_h$. Then there exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$ such that the scheme possesses the truncation error of order $\lceil Q \rceil$ in the sense of $\Pi_h^{(P,\lceil Q\rceil)}$ given by* (2.12)*.*

*Proof.* We repeat a fragment of the proof of Theorem 1, using Lemma 9.5 instead of Lemma 9.3. By Lemma 9.2 and Lemma 9.5 for

$$w(\boldsymbol{\phi}) = S(\boldsymbol{\phi}) \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} S^{-1}(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \Pi_h).$$

we have

$$\|w(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h)\| = \left\| S(\boldsymbol{\phi}) \begin{pmatrix} v_*(\boldsymbol{\phi}, \Pi_h) \\ 0 \end{pmatrix} \right\| \leqslant c_n \|S(\boldsymbol{\phi})\| \|v(\boldsymbol{\phi}, \Pi_h)\| = O(|\boldsymbol{\phi}|^P),$$

$$A(\boldsymbol{\phi}) w(\boldsymbol{\phi}) = S(\boldsymbol{\phi}) \begin{pmatrix} 0 \\ M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h) \end{pmatrix} = O(|\boldsymbol{\phi}|^{\lceil Q\rceil + 1}).$$

By Lemma 6.7 there exist $\mathfrak{C}^{(\boldsymbol{m})}$, $P \leqslant |\boldsymbol{m}| \leqslant \lceil Q \rceil$, such that for $\Pi_h^{(P,\lceil Q\rceil)}$ there holds

$$v(\boldsymbol{\phi}, \Pi_h^{(P,\lceil Q\rceil)}) = w(\boldsymbol{\phi}) + O(|\boldsymbol{\phi}|^{\lceil Q\rceil + 1}).$$

By the triangle inequality we get

$$\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h^{(P,Q)}) = A(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \Pi_h^{(P,Q)}) = O(|\boldsymbol{\phi}|^{\lceil Q\rceil + 1}).$$

By Theorem 6.15 the scheme possesses the truncation error of order $\lceil Q \rceil$ in the sense of $\Pi_h^{(P,Q)}$, so the coefficients $\mathfrak{C}^{(\boldsymbol{m})}$ satisfy the statement of the lemma. □

**Lemma 9.11.** *Let the scheme* (2.8) *be simple, $\Pi_h$ be a local mapping with a kernel $\mu_\xi \in (W_2^r(G))^*$ (or $(C^r(G))^*$), and let $p, q > 0$. Suppose for each $\boldsymbol{e} \in \mathring{\Omega}_{\lceil q\rceil}$ the scheme possesses the formal order of accuracy $p$ and the long-time simulation order $q$ on $H^R_{per,\boldsymbol{e}}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per,\boldsymbol{e}}(\mathbb{R}^d)$), Then the scheme possesses the formal order of accuracy $p$ and the long-time simulation order $q$ on $H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R\rceil}_{per}(\mathbb{R}^d)$), $R = \max\{r, q+1\}$.*

*Proof.* By assumption for each $\boldsymbol{e} \in \mathring{\Omega}_p$ and each $\boldsymbol{\phi}$ aligned with $\boldsymbol{e}$ we have

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\| \leqslant c_1(\boldsymbol{e}) |\boldsymbol{\phi}|^p + c_2(\boldsymbol{e}) \nu |\boldsymbol{\phi}|^{q+1}.$$

Then for $\boldsymbol{\phi}$ aligned with $\boldsymbol{e}$ there hold (9.3) and (9.10). By Lemma 5.3 the inequalities (9.3) and (9.10) hold for each $\boldsymbol{\phi}$ with some other multiplicative constants. Substituting this into (9.1) by stability we get $E_*(\boldsymbol{\phi}, \nu, \Pi_h) = O(|\boldsymbol{\phi}|^{\lceil p\rceil})$ and $E_0(\boldsymbol{\phi}, \nu, \Pi_h) = O(\nu |\boldsymbol{\phi}|^{\lceil q\rceil + 1})$, thus $\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h) = O(|\boldsymbol{\phi}|^{\lceil p\rceil} + \nu |\boldsymbol{\phi}|^{\lceil q\rceil + 1})$. It remains to apply Theorem 6.17. □

**Lemma 9.12.** *Let the scheme be simple. Then the optimal values of the formal order of accuracy and of the long-time simulation order coincide with the ones in the weak sense.*

*Proof.* This directly follows from Proposition 6.22 and Lemma 9.11. □

**Lemma 9.13.** *Suppose the scheme is simple, and $\Pi_h$ is a local mapping. Let $P_A$, $P > 0$, and $Q$ be the optimal values of the truncation error, of the formal order of accuracy and of the long-time simulation order. Then either $P_A = P = Q$ or $Q \geqslant P = P_A + 1$.*

*Proof.* If $|M^0| = 1$, then the statement of the lemma follows from Proposition 9.9. So without loss we assume $|M^0| > 1$. By Theorem 1 the values $P$ and $Q$ are integers and there holds either $P = P_A$ or $P = P_A + 1$. So we need only to prove that $Q \geqslant P_A + 1$ implies $P = P_A + 1$.

Since

$$\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| = \left\| S(\boldsymbol{\phi}) \begin{pmatrix} M^{(*)}(\boldsymbol{\phi}) v_*(\boldsymbol{\phi}, \Pi_h) \\ M^{(0)}(\boldsymbol{\phi}) v_0(\boldsymbol{\phi}, \Pi_h) \end{pmatrix} \right\| = O(|\boldsymbol{\phi}|^{P_A + 1}),$$

we have $\|M^{(*)}(\boldsymbol{\phi})v_*(\boldsymbol{\phi},\Pi_h)\| = O(|\boldsymbol{\phi}|^{P_A+1})$. Then $\|v_*(\boldsymbol{\phi},\Pi_h)\| = O(|\boldsymbol{\phi}|^{P_A+1})$ and by stability

$$\|E_*(\boldsymbol{\phi},\nu,\Pi_h)\| \leqslant C|\boldsymbol{\phi}|^{P_A+1}.$$

Lemma 9.5 yields $\|M^{(0)}(\boldsymbol{\phi})\| = O(|\boldsymbol{\phi}|^{Q+1})$ and hence

$$\|E_0(\boldsymbol{\phi},\nu,\Pi_h)\| \leqslant \tilde{C}|\boldsymbol{\phi}|^{Q+1}\nu.$$

Combining these estimates we get

$$\|\hat{\varepsilon}(\boldsymbol{\phi},\nu,\Pi_h)\| \leqslant C'|\boldsymbol{\phi}|^{P_A+1} + C'\nu|\boldsymbol{\phi}|^{Q+1}.$$

If $Q \geqslant P_A+1$ then by Theorem 6.17 the scheme possesses the formal order of accuracy $P_A+1$, thus $P = P_A+1$. □

**Theorem 3A.** *For a simple scheme, Theorem 3 holds.*

*Proof.* For a simple scheme, the five statements of Theorem 3 follow from Lemmas 9.7, 9.10, 9.11, 9.12, 9.13, in respective order. □

The following proposition may be useful in a computer-based analysis of a scheme.

**Proposition 9.14.** *Let the scheme* (2.8) *be simple and $\Pi_h$ be a local mapping. Let $Q \in \mathbb{N}$ be such that $|\det A(\boldsymbol{\phi})| \leqslant c|\boldsymbol{\phi}|^{q+1}$ holds for $q = Q$ and does not hold for $q = Q+1$. Let $P_A$ be the optimal value of the order of the truncation error. Let $P = P_A + 1$ if $Q > P_A$ and $P = P_A$ otherwise. Then $P$ and $Q$ the optimal values of the formal order of accuracy and of the long-time simulation order.*

*Proof.* Denote by $\lambda(\boldsymbol{\phi})$ the only element of the matrix $M^{(0)}(\boldsymbol{\phi})$. Since the eigenvalue $\lambda = 0$ of $A(0)$ is simple, in a neighborhood of $\boldsymbol{\phi} = 0$ there holds $|\lambda(\boldsymbol{\phi})| \leqslant 2cc_\Pi^{-1}|\boldsymbol{\phi}|^{Q+1}$, where $c_\Pi$ is the product of all nonzero eigenvalues of $A(0)$ taking into account their algebraic multiplicity. Substituting this estimate for $M^{(0)}(\boldsymbol{\phi}) = \lambda(\boldsymbol{\phi})$ and (9.3) into (9.2) and (9.1) we get

$$\|\hat{\varepsilon}(\boldsymbol{\phi},\nu,\Pi_h)\| \leqslant C|\boldsymbol{\phi}|^p + C\nu|\boldsymbol{\phi}|^{Q+1},$$

where $p$ is the optimal value of the formal order of accuracy. Thus by Theorem 6.17 the scheme possesses the formal order of accuracy $p$ and the long-time simulation order $Q$.

Now we claim that the values $p$ and $Q$ are optimal. Assume the converse. Let $Q' > Q$ be the optimal value of the long-time simulation order. Then by Lemma 9.5 there holds $|\lambda(\boldsymbol{\phi})| \leqslant C|\boldsymbol{\phi}|^{Q'+1}$, that contradicts the assumption.

The fact that $p = P$ follows from Lemma 9.13. □

### *9.3. The quasi-one-dimensional case*

Recall that a scheme of the form (2.8) is *quasi-1D* if the stencil $\mathcal{S}$ of the scheme belongs to a 1D subset of $\mathbb{Z}^d$, i. e. there exists $\boldsymbol{\eta} \in \mathbb{Z}^d$ such that $\mathcal{S} \subset \{m\boldsymbol{\eta}, m = -M, \dots, M\}$. The scheme is quasi-1D iff the matrix $A(\boldsymbol{\phi})$ depends only on $\boldsymbol{\phi}\cdot\boldsymbol{e}$ for some $\boldsymbol{e} \in \mathbb{R}^d$. In this section we will use the notation $\boldsymbol{e}$ for this vector.

In 1D case (when $d = 1$) every scheme of the form (2.8) is quasi-1D. In the multidimensional case an example of a quasi-1D scheme is the DG method on simplicial translationally-invariant meshes when the vector of the transport velocity is collinear to one of the mesh edges.

**Theorem 9.15** ([30], §3.5, Corollary 3; [26], §2.6.2)**.** *Let $A(\psi)$ be a holomorphic and selfadjoint $N \times N$-matrix defined on the interval $[\psi_1,\psi_2]$. Then there exists an unitary matrix $\mathcal{U}(\psi)$, holomorphic on $[\psi_1,\psi_2]$, such that $\mathcal{U}^{-1}(\psi)A(\psi)\mathcal{U}(\psi) = \operatorname{diag}\{\lambda_1(\psi),\dots,\lambda_N(\psi)\}$.*

**Lemma 9.16.** *Let $f(\psi)$ and $g(\boldsymbol{\phi})$ be holomorphic functions at $\psi = 0$ and $\boldsymbol{\phi} = 0$, correspondingly. Let $\boldsymbol{e} \in \Omega$. Suppose there hold $f(\psi) \sim \psi^a$ as $\psi \to 0$ and $|f(\boldsymbol{\phi}\cdot\boldsymbol{e})g(\boldsymbol{\phi})| \leqslant c|\boldsymbol{\phi}|^{a+b}$, where $a, b \in \mathbb{N}\cup\{0\}$. Then in a neighborhood of $\boldsymbol{\phi} = 0$ there holds*

$$|f(\boldsymbol{\phi}\cdot\boldsymbol{e})g(\boldsymbol{\phi})| \leqslant \delta c|\boldsymbol{\phi}\cdot\boldsymbol{e}|^a|\boldsymbol{\phi}|^b, \tag{9.11}$$

*where $\delta$ depends on $a+b$ and the space dimension $d$ only.*

*Proof.* Let $\breve{C} = \{\boldsymbol{x} \in \mathbb{R}^d \,:\, \boldsymbol{e}\cdot\boldsymbol{x} > |\boldsymbol{x}|/2\}$. By Lemma A.6 there exist $\boldsymbol{e}_k \in \breve{C}\cap\Omega$, $k = 1,\dots,C^{d-1}_{b+d-1}$, such that $(\boldsymbol{e}_k\cdot\boldsymbol{r})^b$ form a basis in the space of homogeneous polynomials of order $b$. For different $\boldsymbol{e}$ the set of vectors $\boldsymbol{e}_k$ can be chosen the same modulo rotation. By assumption for each $\boldsymbol{e}'\in\breve{C}\cap\Omega$ in a neighborhood of $\psi = 0$ there holds

$$|g(\boldsymbol{e}'\psi)| \leqslant \frac{c|\psi|^{a+b}}{f((\boldsymbol{e}\cdot\boldsymbol{e}')\psi)} \leqslant 2\frac{c|\psi|^b}{(\boldsymbol{e}\cdot\boldsymbol{e}')^a} \leqslant 2^{1+a}c|\psi|^b.$$

Thus $|\partial^n g(0)/\partial\boldsymbol{e}'^n| = 0$ for $n < b$ and $|\partial^b g(0)/\partial\boldsymbol{e}'^b| \leqslant 2^{1+a}b!c$. By Lemma 5.1 we have $D^{\boldsymbol{m}}g(0) = 0$ for $|\boldsymbol{m}| < b$ and $|D^{\boldsymbol{m}}g(0)| \leqslant \delta c$ for $|\boldsymbol{m}| = b$, where $\delta = 2^{1+a}b!\sum_k|\gamma_k^{(\boldsymbol{m})}|$ does not depend on $\boldsymbol{e}$. □

**Lemma 9.17.** *Let the scheme be quasi-1D, $\Pi_h$ be a local mapping, $0 < P \leqslant Q < \infty$. Then the function $\mathcal{F}^{P,Q}(\boldsymbol{\phi})$ defined by* (8.9) *is bounded as $\boldsymbol{\phi}\to 0$ iff the function*

$$\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2\lceil P\rceil}v^*(\boldsymbol{\phi},\Pi_h)\left(A^*(\boldsymbol{\phi})A(\boldsymbol{\phi}) + |\boldsymbol{\phi}\cdot\boldsymbol{e}|^{2(\lceil Q\rceil+1-\lceil P\rceil)}\right)^{-1}A^*(\boldsymbol{\phi})A(\boldsymbol{\phi})v(\boldsymbol{\phi},\Pi_h).$$

*is bounded as $\boldsymbol{\phi}\to 0$.*

*Proof.* Throughout this proof we will use the notation $\psi \equiv \boldsymbol{\phi}\cdot\boldsymbol{e}$. By Theorem 9.15 there exist holomorphic matrix function $\mathcal{U}$ and holomorphic functions $\lambda_j$ such that $\mathcal{U}^{-1}(\psi)A^*(\boldsymbol{\phi})A(\boldsymbol{\phi})\mathcal{U}(\psi) = \operatorname{diag}\{\lambda_j(\psi)\}$. Therefore,

$$\mathcal{F}^{P,Q}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2P}\sum_{j\in M^0}|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2\frac{\lambda_j(\psi)}{\lambda_j(\psi)+|\boldsymbol{\phi}|^{2(Q+1-P)}},$$

$$\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2\lceil P\rceil}\sum_{j\in M^0}|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2\frac{\lambda_j(\psi)}{\lambda_j(\psi)+|\psi|^{2(\lceil Q\rceil+1-\lceil P\rceil)}}.$$

First assume that $\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi})$ is bounded. Then we have

$$\mathcal{F}^{\lceil P\rceil,\lceil Q\rceil}(\boldsymbol{\phi}) \leqslant \tilde{\mathcal{F}}^{\lceil P\rceil,\lceil Q\rceil}(\boldsymbol{\phi}) = \tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi}) < \infty.$$

Thus by Theorem 8.5 the scheme possesses the formal order of accuracy $\lceil P\rceil$ and the long-time simulation order $\lceil Q\rceil$. By Corollary 6.18 it possesses the formal order of accuracy $P$ and the long-time simulation order $Q$. Using again Theorem 8.5, we obtain that $\mathcal{F}^{P,Q}(\boldsymbol{\phi})$ then is bounded as $\boldsymbol{\phi}$ tends to zero.

Now assume that $\mathcal{F}^{P,Q}(\boldsymbol{\phi})$ is bounded in a neighborhood of $\boldsymbol{\phi} = 0$. Then for each $j\in M^0$ there holds

$$\lambda_j(\psi)|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2 \leqslant C|\boldsymbol{\phi}|^{2P}(\lambda_j(\psi)+|\boldsymbol{\phi}|^{2(Q+1-P)}) \tag{9.12}$$

in a neighborhood of $\boldsymbol{\phi} = 0$. The function $\lambda_j(\psi)$ is holomorphic at $\psi = 0$ and non-negative (as the eigenvalue of matrix $A^*(\psi)A(\psi)$). Thus either $\lambda_j(\psi) = c_\lambda\psi^{2q_j}(1+O(\psi))$ as $\psi\to 0$ with some $c_\lambda > 0$ and $q_j\in\mathbb{N}\cup\{0\}$ or $\lambda_j(\psi)\equiv 0$ (in this case we put $q_j = +\infty$).

The function $(\mathcal{U}^{-1}(\boldsymbol{\phi}\cdot\boldsymbol{e})v(\boldsymbol{\phi},\Pi_h))_j$ is holomorphic at $\boldsymbol{\phi} = 0$, let $r_j$ be the lowest order of the terms in its Taylor series. From (9.12) we have

$$2q_j + 2r_j \geqslant 2P + \min\{2q_j, 2(Q+1-P)\}$$

or, equivalently,

$$q_j + r_j \geqslant \min\{q_j + P, Q+1\}.$$

Since $r_j, q_j\in\mathbb{N}$, we have

$$q_j + r_j \geqslant \min\{q_j + \lceil P\rceil, \lceil Q\rceil + 1\}.$$

If $q_j + \lceil P\rceil < \lceil Q\rceil + 1$, then $r_j \geqslant \lceil P\rceil$, so $|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2 \leqslant \tilde{c}|\boldsymbol{\phi}|^{2\lceil P\rceil}$ and thus

$$\lambda_j(\psi)|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2 \leqslant \tilde{c}|\boldsymbol{\phi}|^{2\lceil P\rceil}\lambda_j(\psi).$$

If $q_j + \lceil P \rceil \geqslant \lceil Q \rceil + 1$, then $r_j \geqslant \lceil Q \rceil + 1 - q_j$ and

$$\lambda_j(\psi)|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2 \leqslant \tilde{c}\psi^{2q_j}|\boldsymbol{\phi}|^{2(\lceil Q\rceil+1-q_j)} =$$
$$= \tilde{c}\psi^{2(\lceil Q\rceil+1-\lceil P\rceil)}\psi^{2(q_j-(\lceil Q\rceil+1-\lceil P\rceil))}|\boldsymbol{\phi}|^{2(\lceil Q\rceil+1-q_j)} \leqslant \tilde{c}\psi^{2(\lceil Q\rceil+1-\lceil P\rceil)}|\boldsymbol{\phi}|^{2\lceil P\rceil}.$$

Combining the results for both cases, in a neighborhood of $\boldsymbol{\phi} = 0$ we get

$$\lambda_j(\psi)|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h))_j|^2 \leqslant \tilde{c}|\boldsymbol{\phi}|^{2\lceil P\rceil}(\lambda_j(\psi) + |\psi|^{2(\lceil Q\rceil+1-\lceil P\rceil)}). \tag{9.13}$$

From here, the statement of the lemma is obvious. □

**Lemma 9.18.** *Suppose the scheme is quasi-1D and $\Pi_h$ is a local mapping. Then the optimal values of the formal order of accuracy and of the long-time simulation order are integers.*

*Proof.* The optimal order of accuracy $P$ is integer by Theorem 1. Let the scheme possess the formal order of accuracy $P$ and a long-time simulation order $Q$. By Theorem 8.5, $\mathcal{F}^{P,Q}(\boldsymbol{\phi})$ is bounded as $\boldsymbol{\phi} \to 0$. By Lemma 9.17 so is $\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi})$ and, applying this Lemma again, so is $\mathcal{F}^{P,\lceil Q\rceil}(\boldsymbol{\phi})$. By Theorem 8.5 the scheme possesses the formal order $P$ and the long-time simulation order $\lceil Q \rceil$. □

**Lemma 9.19.** *Let the scheme be quasi-1D, $\Pi_h$ be a local mapping, $P \in \mathbb{N}$. Let the scheme possess the formal order of accuracy $P$ and the long-time simulation order $s$ for each $s \geqslant P$. Then the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $\infty$.*

*Proof.* By Theorem 8.5 the functional $\mathcal{F}^{P,s}(\boldsymbol{\phi})$ is bounded in a neighborhood of $\boldsymbol{\phi} = 0$. Thus using the notation of Lemma 9.17 for each $j \in M^0$ and $s \geqslant P$ there holds

$$q_j + r_j \geqslant \min\{q_j + P, s + 1\}.$$

Thus for each $j$ there holds either $q_j = +\infty$ or $r_j \geqslant P$. Put $w(\boldsymbol{\phi}) = \mathcal{U}(\psi)\mathrm{diag}\{\chi_j\}\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi},\Pi_h)$ where $\chi_j = 0$ if $r_j \geqslant P$ and $\chi_j = 1$ otherwise. Then

$$\|w(\boldsymbol{\phi}) - v(\boldsymbol{\phi},\Pi_h)\| \leqslant c|\boldsymbol{\phi}|^P, \quad A(\boldsymbol{\phi})w(\boldsymbol{\phi}) = 0,$$

and it remains to apply Lemma 8.1. □

**Lemma 9.20.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with kernels $\mu_\xi, \hat{\mu}_\xi \in (W_2^q(G))^*$ (or $(C^q(G))^*$), such that $\langle\hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$. Suppose the scheme* (2.8) *is quasi-1D and possesses the formal order of accuracy $P \in \mathbb{N}$ and the long-time simulation order $Q \in \mathbb{N}$ in the sense of $\Pi_h$ on $H^R_{per}(\mathbb{R}^d)$ (or $C^R_{per}(\mathbb{R}^d)$), where $R = \max\{Q+1, q\}$. Then there exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$ such that the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_h^{(P,Q)}$ given by* (2.12).

*Proof.* By Theorem 8.5, $\mathcal{F}^{P,Q}(\boldsymbol{\phi})$ is bounded as $\boldsymbol{\phi} \to 0$. By Lemma 9.17 so does $\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi})$. Put

$$w(\boldsymbol{\phi}) = \left(A^*(\boldsymbol{\phi})A(\boldsymbol{\phi}) + |\boldsymbol{\phi}\cdot\boldsymbol{e}|^{2(Q+1-P)}\right)^{-1}|\boldsymbol{\phi}\cdot\boldsymbol{e}|^{2(Q+1-P)}v(\boldsymbol{\phi},\Pi_h) \tag{9.14}$$

and note that

$$\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2P}\|v(\boldsymbol{\phi},\Pi_h) - w(\boldsymbol{\phi})\|^2 + |\boldsymbol{\phi}|^{-2P}|\boldsymbol{\phi}\cdot\boldsymbol{e}|^{-2(Q-P+1)}\|A(\boldsymbol{\phi})w(\boldsymbol{\phi})\|^2. \tag{9.15}$$

(the proof of this fact repeats the one of Lemma 8.4). By construction, $w(\boldsymbol{\phi}) = W(\boldsymbol{\phi})/\mathfrak{d}(\psi)$, there $W(\boldsymbol{\phi})$ is holomorphic at $\boldsymbol{\phi} = 0$ and $\mathfrak{d}(\psi) = \det(A^*(\psi\boldsymbol{e})A(\psi\boldsymbol{e}) + \psi^{2(Q+1-P)})$. Let $\mathfrak{d}(\psi) \sim c\psi^q$ as $\psi \to 0$, where $c \neq 0$. All the terms in the Taylor expansion of the numerator contain the multiplier $\psi^q$, otherwise $w(\boldsymbol{\phi})$ is unbounded and so does $\tilde{\mathcal{F}}^{P,Q}(\boldsymbol{\phi})$, which contradicts the assumption. This proves that $w(\boldsymbol{\phi})$ is analytical at $\boldsymbol{\phi} = 0$.

From (9.15) and the boundedness of $\tilde{\mathcal{F}}^{P,Q}$ it follows that $\|v(\boldsymbol{\phi},\Pi_h) - w(\boldsymbol{\phi})\| \leqslant \tilde{C}|\boldsymbol{\phi}|^P$ and

$$\|A(\boldsymbol{\phi})w(\boldsymbol{\phi})\||\boldsymbol{\phi}\cdot\boldsymbol{e}|^P \leqslant \tilde{C}|\boldsymbol{\phi}\cdot\boldsymbol{e}|^{Q+1}|\boldsymbol{\phi}|^P \leqslant \tilde{C}|\boldsymbol{\phi}|^{Q+1+P}.$$

Since $A(\boldsymbol{\phi})w(\boldsymbol{\phi})$ is holomorphic, this implies $A(\boldsymbol{\phi})w(\boldsymbol{\phi}) = O(|\boldsymbol{\phi}|^{Q+1})$.

By Lemma 6.7 there exist $\mathfrak{C}^{(\boldsymbol{m})}$, $P \leqslant |\boldsymbol{m}| \leqslant Q$, such that for $\Pi_h^{(P,Q)}$ there holds

$$v(\boldsymbol{\phi}, \Pi_h^{(P,Q)}) = w(\boldsymbol{\phi}) + O(|\boldsymbol{\phi}|^{Q+1}).$$

By the triangle inequality we get

$$\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h^{(P,Q)}) = A(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h^{(P,Q)}) = O(|\boldsymbol{\phi}|^{Q+1}).$$

By Theorem 6.15 the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_h^{(P,Q)}$, so the coefficients $\mathfrak{C}^{(\boldsymbol{m})}$ satisfy the statement of the lemma. □

**Lemma 9.21.** *Let the scheme* (2.8) *be quasi-1D, $\Pi_h$ be a local mapping with a kernel $\mu_\xi \in (W_2^r(G))^*$ (or $(C^r(G))^*$), and let $0 < P \leqslant Q < \infty$. Suppose for each $\tilde{\boldsymbol{e}} \in \mathring{\Omega}_{\lceil Q \rceil}$ the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ on $H^R_{per,\tilde{\boldsymbol{e}}}(\mathbb{R}^d)$ (or $C^{\lceil R \rceil}_{per,\tilde{\boldsymbol{e}}}(\mathbb{R}^d)$). Then the scheme possesses the formal order of accuracy $\lceil P \rceil$ and the long-time simulation order $\lceil Q \rceil$ on $H^R_{per}(\mathbb{R}^d)$ (or $C^{\lceil R \rceil}_{per}(\mathbb{R}^d)$).*

*Proof.* We will use the notation of Lemma 9.17. By assumption, for each direction $\tilde{\boldsymbol{e}} \in \mathring{\Omega}_{\lceil Q \rceil}$ the scheme possesses the order of accuracy $P$ and the long-time simulation order $Q$ on $H^s_{per,\tilde{\boldsymbol{e}}}(\mathbb{R}^d)$ for $s$ large enough. By Lemma 8.6, the functional $\mathcal{F}^{P,Q}(\psi\tilde{\boldsymbol{e}})$ is bounded as $\psi$ tends to zero. Thus for each $j \in M^0$ and each $\tilde{\boldsymbol{e}} \in \mathring{\Omega}_Q$ there holds

$$\sqrt{\lambda_j(\psi)}|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi}, \Pi_h))_j| \leqslant C_{\tilde{\boldsymbol{e}}}|\boldsymbol{\phi}|^P(\sqrt{\lambda_j(\psi)} + |\boldsymbol{\phi}|^{Q+1-P}) \leqslant 2C_{\tilde{\boldsymbol{e}}}|\boldsymbol{\phi}|^{\min\{Q+1, P+q_j\}},$$

where $\boldsymbol{\phi}$ belongs to a neighborhood of zero in $\operatorname{span}\{\tilde{\boldsymbol{e}}\}$. Since the function on the left-hand side is the absolute value of a function analytical at $\boldsymbol{\phi} = 0$, by Lemma 5.3 in a neighborhood of zero we have

$$\sqrt{\lambda_j(\psi)}|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi}, \Pi_h))_j| \leqslant C|\boldsymbol{\phi}|^{\min\{\lceil Q \rceil + 1, \lceil P \rceil + q_j\}},$$

where $C$ does not depend on $\boldsymbol{e}$.

If for a particular $j$ there holds $\lceil P \rceil \leqslant \lceil Q \rceil + 1 - q_j$, then we have

$$|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi}, \Pi_h))_j| \leqslant \tilde{c}|\boldsymbol{\phi}|^{\lceil P \rceil}$$

and hence (9.13). If $q_j > \lceil Q \rceil + 1 - \lceil P \rceil$, then by Lemma 9.16 we have

$$\lambda_j(\psi)|(\mathcal{U}^{-1}(\psi)v(\boldsymbol{\phi}, \Pi_h))_j|^2 \leqslant c'\psi^{2q_j}|\boldsymbol{\phi}|^{2(\lceil Q \rceil + 1 - q_j)} \leqslant c'\psi^{2(\lceil Q \rceil + 1 - \lceil P \rceil)}|\boldsymbol{\phi}|^{2\lceil P \rceil},$$

and we again have (9.13). Thus $\tilde{\mathcal{F}}^{\lceil P \rceil, \lceil Q \rceil}(\boldsymbol{\phi})$ is bounded as $\boldsymbol{\phi} \to 0$. By Lemma 9.17 so does $\mathcal{F}^{\lceil P \rceil, \lceil Q \rceil}(\boldsymbol{\phi})$ and by Theorem 8.5 the scheme possesses the formal order of accuracy $\lceil P \rceil$ and the long-time simulation order $\lceil Q \rceil$. □

**Lemma 9.22.** *Suppose the scheme is quasi-1D and $\Pi_h$ is a local mapping. Then the optimal values of the formal order of accuracy and of the long-time simulation order coincide with the ones in the weak sense.*

*Proof.* Let the scheme possess the formal order of accuracy $P$ and the long-time simulation order $Q$ in the weak sense. By Proposition 6.22 for each $\tilde{\boldsymbol{e}} \in \mathring{\Omega}$ the scheme possesses the formal order of accuracy $P$ and the long-time simulation order in the strong sense on $H^R_{per,\tilde{\boldsymbol{e}}}(\mathbb{R}^d)$. It remains to use Lemma 9.21. □

**Theorem 3B.** *For a quasi-1D scheme, Theorem 3 holds.*

*Proof.* For a quasi-1D scheme, the first four statements of Theorem 3 follow from Lemmas 9.18, 9.20, 9.21, 9.22, in respective order. The fifth statement of Theorem 3, which only concerns only simple schemes, has been already proved in Theorem 3A. □

## 10. Algorithms for scheme analysis

Throughout this section we consider a scheme of the form (2.8), a local mapping $\Pi_h$, and assume that the scheme is stable and 0-exact. For the cases where the existence of a local mapping of the form (2.12) is proved it is possible to construct algorithms for the analysis of the scheme.

### *10.1. The order of the truncation error*

**Lemma 10.1.** *Let $P_A \in \mathbb{N} \cup \{0\}$. Then the following statements are equivalent.*

1. *The scheme possesses the truncation error of order $P_A$.*
2. *There holds $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| = O(|\boldsymbol{\phi}|^{P_A+1})$ as $\boldsymbol{\phi} \to 0$.*
3. *For each multiindex $\boldsymbol{m}$ such that $|\boldsymbol{m}| \leqslant P_A$ there holds $(\epsilon_1(\boldsymbol{r}^{\boldsymbol{m}}/\boldsymbol{m}!, \Pi_1))_0 = 0$.*

*Proof.* The equivalence of statements 1 and 2 is by Lemma 6.10 and boundedness of $Z(\boldsymbol{\phi})$ and $Z^{-1}(\boldsymbol{\phi})$. Now we prove the equivalence of statements 2 and 3.

Denote by $\tilde{\Pi}_h$ the operator taking $f$ to $\tilde{\Pi}_h f = h\epsilon_h(f, \Pi_h)$. By (6.19) we have

$$Z(\boldsymbol{\phi})\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = -F[h\epsilon_h(e^{i\boldsymbol{\phi}\cdot\boldsymbol{r}/h}, \Pi_h)](\boldsymbol{\phi}) = -F[\tilde{\Pi}_h e^{i\boldsymbol{\phi}\cdot\boldsymbol{r}/h}](\boldsymbol{\phi}).$$

The mapping $\tilde{\Pi}_h$ is of the form (2.11) and enjoys all the properties of a local mapping with the only exception that $\tilde{\Pi}_h 1 = 0$ (recall that $P_A \geqslant 0$). Particularly, Lemma 6.5 remains valid for $\tilde{\Pi}_h$. Then by Lemma 6.5 we obtain

$$Z(\boldsymbol{\phi})\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = -v(\boldsymbol{\phi}, \tilde{\Pi}_h) = -(\tilde{\Pi}_1 e^{i\boldsymbol{\phi}\cdot\boldsymbol{r}})_0 =$$
$$= -\sum_{0\leqslant|\boldsymbol{m}|<\infty} \boldsymbol{\phi}^{\boldsymbol{m}} \left(\tilde{\Pi}_1 \frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_0 = -\sum_{0\leqslant|\boldsymbol{m}|<\infty} \boldsymbol{\phi}^{\boldsymbol{m}} \left(\epsilon_1\left(\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}, \Pi_1\right)\right)_0.$$

Since $Z(\boldsymbol{\phi})$ and $Z^{-1}(\boldsymbol{\phi})$ are uniformly bounded, the equivalence of statements 2 and 3 is obvious. □

**Lemma 10.2.** *Let $P_A \in \mathbb{N} \cup \{0\}$ and $\boldsymbol{e} \in \Omega$. Then $\|\hat{\epsilon}(\boldsymbol{e}\psi, \Pi_h)\| = O(|\psi|^{P_A+1})$ as $\psi \to 0$ iff for each $m = 0, \dots, P_A$ there holds $(\epsilon_1((\boldsymbol{r}\cdot\boldsymbol{e})^m/m!, \Pi_1))_0 = 0$.*

*Proof.* The proof repeats the one of the previous lemma. □

Recall the following standard algorithm.

**Algorithm 1** (to detect the optimal order of the truncation error)**.**

1. *Put $P_A = 0$.*
2. *Compute $f^{\boldsymbol{m}} = -(\epsilon_1(\boldsymbol{r}^{\boldsymbol{m}}/\boldsymbol{m}!, \Pi_1))_0$ for each $\boldsymbol{m}$ such that $|\boldsymbol{m}| = P_A + 1$.*
3. *If for each $\boldsymbol{m}$ with $|\boldsymbol{m}| = P_A + 1$ there holds $f^{\boldsymbol{m}} = 0$ then put $P_A = P_A + 1$ and return to the step 2.*

**Proposition 10.3.** *If Algorithm 1 returns the value $P_A$, then $P_A$ is the optimal order of the truncation error. If Algorithm 1 loops endlessly, then the scheme is exact, i. e. $\epsilon_h(f, \Pi_h) = 0$ for each $f$.*

*Proof.* By Theorem 1 the optimal order of the truncation error is integer, so the first statement follows from Lemma 10.1. If the algorithm loops endlessly, then by Lemma 10.1 there holds $\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| = O(|\boldsymbol{\phi}|^s)$ for each $s \in \mathbb{N}$. Since $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)$ is holomorphic at $\boldsymbol{\phi} = 0$, we have $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = 0$ in a neighborhood of $\boldsymbol{\phi} = 0$, and it remains to apply Lemma 6.16. □

*10.2. The order of the truncation error in the sense of auxiliary mappings*

Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings, and the kernel of $\mathcal{P}_h$ satisfy $\langle \hat{\mu}_\xi, 1\rangle \neq 0$ for each $\xi \in M^0$. Let $\mathfrak{C}^{(\boldsymbol{m})}$ be real-valued diagonal matrices. Now we discuss the truncation error in the sense of the mappings $\Pi_h^{(p,q)}$ given by (2.12).

**Lemma 10.4.** *Let $\Pi_h^{(p,q)}$ be an operator of the form* (2.12) *with coefficients $\mathfrak{C}^{(\boldsymbol{m})}$. The scheme possesses the truncation error of order $P_A \in \mathbb{N}$ in the sense of $\Pi_h^{(p,q)}$ if and only if for each $|\boldsymbol{m}| \leqslant P_A$ there holds*

$$L(0)\mathfrak{C}^{(\boldsymbol{m})}(\mathcal{P}_1 1)_0 = -\left(\epsilon_1\left(\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}, \Pi_1^{(p,\min\{|\boldsymbol{m}|-1,q\})}\right)\right)_0, \tag{10.1}$$

*where the mappings $\Pi^{(p,\min\{|\boldsymbol{m}|-1,q\})}$ are based on the same coefficients $\mathfrak{C}^{(\boldsymbol{m})}$.*

*Proof.* If $q < P_A$, put $\mathfrak{C}^{\boldsymbol{m}} = 0$ for each $q < |\boldsymbol{m}| \leqslant P_A$, so without loss we can assume that $q \geqslant P_A$.

For each $|\boldsymbol{m}| \leqslant P_A$ we have

$$\left(\epsilon_1\left(\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}, \Pi_1^{(p,P_A)}\right)\right)_0 = \left(-Z\Pi_1^{(p,P_A)}(\boldsymbol{\omega}\cdot\nabla)\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!} + L\Pi_1^{(p,P_A)}\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_0 =$$

$$= -\sum_{\boldsymbol{\eta}\in\mathcal{S}} Z_{\boldsymbol{\eta}}\left(\Pi_1^{(p,P_A)}(\boldsymbol{\omega}\cdot\nabla)\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_{\boldsymbol{\eta}} + \sum_{\boldsymbol{\eta}\in\mathcal{S}} L_{\boldsymbol{\eta}}\left(\Pi_1^{(p,P_A)}\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_{\boldsymbol{\eta}}.$$

The function $(\boldsymbol{\omega}\cdot\nabla)\boldsymbol{r}^{\boldsymbol{m}}$ is a polynomial of order not higher than $|\boldsymbol{m}|-1$; thus $\Pi_1^{(p,P_A)}$ acts on it the same way as $\Pi_1^{(p,|\boldsymbol{m}|-1)}$. Besides,

$$\left(\Pi_1^{(p,P_A)}\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_{\boldsymbol{\eta}} = \left(\Pi_1^{(p,|\boldsymbol{m}|-1)}\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_{\boldsymbol{\eta}} + \mathfrak{C}^{(\boldsymbol{m})}(\mathcal{P}_1 1)_{\boldsymbol{\eta}} = \left(\Pi_1^{(p,|\boldsymbol{m}|-1)}\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)_{\boldsymbol{\eta}} + \mathfrak{C}^{(\boldsymbol{m})}(\mathcal{P}_1 1)_0.$$

Therefore, using $L(0) = \sum_{\boldsymbol{\eta}} L_{\boldsymbol{\eta}}$ we have

$$\left(\epsilon_1\left(\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}, \Pi_1^{(p,P_A)}\right)\right)_0 = \left(\epsilon_1\left(\frac{\boldsymbol{r}^{\boldsymbol{m}}}{\boldsymbol{m}!}, \Pi_1^{(p,|\boldsymbol{m}|-1)}\right)\right)_0 + L(0)\mathfrak{C}^{(\boldsymbol{m})}(\mathcal{P}_1 1)_0.$$

By Lemma 10.1 the scheme possesses the truncation error of order $P_A$ iff the left-hand side of this identity is zero for each $|\boldsymbol{m}| \leqslant P_A$. Thus we get the statement of the lemma. □

In 1D case equation (10.1) takes a form

$$L(0)\mathfrak{C}^{(m)}(\mathcal{P}_1 1)_0 = \omega\sum_{\eta\in\mathcal{S}} Z_\eta\left[\left(\Pi_1\frac{x^{m-1}}{(m-1)!}\right)_\eta + \sum_{n=p}^{\min\{m-1,q\}}\mathfrak{C}^{(n)}\left(\mathcal{P}_1\frac{x^{m-n-1}}{(m-n-1)!}\right)_\eta\right] -$$

$$-\sum_{\eta\in\mathcal{S}} L_\eta\left[\left(\Pi_1\frac{x^m}{m!}\right)_\eta + \sum_{n=p}^{\min\{m-1,q\}}\mathfrak{C}^{(n)}\left(\mathcal{P}_1\frac{x^{m-n}}{(m-n)!}\right)_\eta\right].$$

*10.3. The order of accuracy*

**Algorithm 2** (to detect the order of accuracy)**.**

1. *Detect the order of the truncation error $P_A$ using Algorithm 1.*
2. *Compute $f^{\boldsymbol{m}} = -(\epsilon_1(\boldsymbol{r}^{\boldsymbol{m}}/\boldsymbol{m}!, \Pi_1))_0$ for each $\boldsymbol{m}$ such that $|\boldsymbol{m}| = P_A + 1$.*
3. *If for each $\boldsymbol{m}$ with $|\boldsymbol{m}| = P_A + 1$ there holds $f^{\boldsymbol{m}} \in \mathrm{Im}L(0)$ then put $P = P_A + 1$, otherwise put $P = P_A$.*

**Theorem 10.5.** *The value $P$ given by Algorithm 2 coincides with the optimal value of the order of accuracy.*

*Proof.* Let $\mathcal{P}_h$ be any local mapping such that $(\mathcal{P}_1 1)_{0,\xi} \neq 0$ for each $\xi \in M^0$. Obviously, the algorithm returns the value $P = P_A + 1$ iff for each multiindex $\boldsymbol{m}$ such that $|\boldsymbol{m}| = P_A + 1$ the system $L(0)\mathfrak{C}^{(\boldsymbol{m})}(\mathcal{P}_1 1)_0 = f^{(\boldsymbol{m})}$ (as a system for the coefficients of $\mathfrak{C}^{(\boldsymbol{m})}$) is consistent. By Lemma 10.4 this is equivalent to the existence of $\mathfrak{C}^{(\boldsymbol{m})}$ such that the scheme possesses the order $P_A + 1$ of the truncation error in the sense of $\Pi_h^{(P_A+1,P_A+1)}$ given by (2.12) with the chosen mapping $\mathcal{P}_h$. By Theorem 1 this holds iff the scheme possesses the order of accuracy $P_A + 1$. □

*10.4. The quasi-one-dimensional case*

**Algorithm 3** (to detect the formal and the long-time simulation orders in the quasi-1D case)**.**

1. *Detect the optimal value of the order of accuracy $P$ using Algorithm 2.*
2. *Choose any local mapping $\mathcal{P}_h$ such that $(\mathcal{P}_1 1)_{0,\xi} \neq 0$ for each $\xi \in M^0$.*
3. *Put $Q' = P$.*
4. *While the set of equations $(\epsilon_1(\boldsymbol{r}^{\boldsymbol{m}}/\boldsymbol{m}!, \Pi_1^{(P,Q'+1)}))_0 = 0$, where $0 \leqslant |\boldsymbol{m}| \leqslant Q'+1$, is consistent as a system for diagonal matrices $\{\mathfrak{C}^{(\boldsymbol{m})}, P \leqslant |\boldsymbol{m}| \leqslant Q'+1\}$, increment $Q'$.*

**Theorem 10.6.** *Let the scheme* (2.8) *be stable and quasi-1D. Let $P$ and $Q$ be its optimal values of the formal order of accuracy and the long-time simulation order (see Definition 5). If $Q = \infty$, then the algorithm loops endlessly. Otherwise the values $P$ and $Q'$ given by Algorithm 3 coincide with $P$ and $Q$.*

*Proof.* First assume $Q < \infty$. By Theorem 3 there exist $\mathfrak{C}^{(\boldsymbol{m})}$, $P \leqslant |\boldsymbol{m}| \leqslant Q$, such that the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_h^{(P,Q)}$. Then $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h^{(P,Q)}) = O(|\boldsymbol{\phi}|^{Q+1})$ and by Lemma 10.1 the matrices $\mathfrak{C}^{(\boldsymbol{m})}$ satisfy the system checked at step 4 for $Q' = Q-1$. Thus the algorithm either returns $Q' \geqslant Q$ or loops endlessly.

If $Q = \infty$, then the for each $q \in \mathbb{N}$ the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $q$. By the argument above, the algorithm will find coefficients $\mathfrak{C}^{(\boldsymbol{m})}$ for each $P \leqslant |\boldsymbol{m}| \leqslant q$. Since $q$ may be chosen arbitrary high, the algorithm will loop endlessly.

Now suppose that the algorithm returns values $P$ and $Q' < \infty$. Let $\{\mathfrak{C}^{(\boldsymbol{m})}, P \leqslant |\boldsymbol{m}| \leqslant Q'\}$ be a solution of the last consistent system and let $\Pi_h^{(P,Q')}$ be given by (2.12) with these coefficients (if no systems checked at step 4 were consistent, put $\Pi_h^{(P,Q')} = \Pi_h$). By Lemma 10.1 there holds $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h^{(P,Q')}) = A(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \Pi_h^{(P,Q')}) = O(|\boldsymbol{\phi}|^{Q'+1})$. Obviously, $v(\boldsymbol{\phi}, \Pi_h^{(P,Q')}) - v(\boldsymbol{\phi}, \Pi_h) = O(|\boldsymbol{\phi}|^P)$. Thus by Lemma 8.1 the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q'$ in the sense of $\Pi_h$, i. e. there holds $Q \geqslant Q'$.

If the algorithm loops endlessly, then arguing as above we obtain that the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $s$ for each $s \in \mathbb{N}$, $s \geqslant P$. By Lemma 9.19 it possesses the formal order of accuracy $P$ and the long-time simulation order $Q = \infty$. □

The following proposition simplifies the analysis in practice allowing to reduce the number of unknowns while processing Algorithm 3. Note that it assumes the additional condition $\langle \mu_\xi, 1 \rangle \neq 0$ for each $\xi \in M^0$.

**Proposition 10.7.** *Let $\Pi_h$ and $\mathcal{P}_h$ be local mappings with kernels $\mu_\xi$, $\hat{\mu}_\xi$ such that $\langle \mu_\xi, 1 \rangle \neq 0$ and $\langle \hat{\mu}_\xi, 1 \rangle \neq 0$ for each $\xi \in M^0$. Let $p, q \in \mathbb{N}$. Suppose there exist real-valued diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$, $p \leqslant |\boldsymbol{m}| \leqslant q$, such that the scheme possesses the truncation error of order $q$ in the sense of $\Pi_h^{(p,q)}$ given by* (2.12)*. For each $\boldsymbol{m}$, $p \leqslant |\boldsymbol{m}| \leqslant q$, take any $\xi_{\boldsymbol{m}} \in M^0$ and $c_{\boldsymbol{m}} \in \mathbb{R}$. Then there exist real-valued diagonal matrices $\tilde{\mathfrak{C}}^{(\boldsymbol{m})}$, $p \leqslant |\boldsymbol{m}| \leqslant q$, such that $\tilde{\mathfrak{C}}^{(\boldsymbol{m})}_{\xi_m, \xi_m} = c_{\boldsymbol{m}}$ and the scheme possesses the truncation error of order $q$ in the sense of $\tilde{\Pi}_h^{(p,q)}$ given by* (2.12) *with the substitution $\tilde{\mathfrak{C}}^{(\boldsymbol{m})}$ for $\mathfrak{C}^{(\boldsymbol{m})}$.*

*Proof.* Denote $f_{p-1}(\boldsymbol{\phi}) = v(\boldsymbol{\phi}, \Pi_h^{(p,q)})$. The function $f_{p-1}(\boldsymbol{\phi})$ is holomorphic by construction; by assumption $(f_{p-1}(0))_\xi \neq 0$ for each $\xi \in M^0$. Denote

$$Y_{\boldsymbol{m}}[f] = \frac{1}{\boldsymbol{m}!} \frac{1}{i^{|\boldsymbol{m}|}} D^{\boldsymbol{m}} \left( \frac{(f(\boldsymbol{\phi}))_{\xi_m}}{(v(\boldsymbol{\phi}, \mathcal{P}_h))_{\xi_m}} \right) \Bigg|_{\boldsymbol{\phi}=0}, \quad g(\boldsymbol{\phi}) = v(\boldsymbol{\phi}, \Pi_h).$$

For $k = p, \ldots, q$ define inductively

$$f_k(\boldsymbol{\phi}) = f_{k-1}(\boldsymbol{\phi}) \Big( 1 + \sum_{|\boldsymbol{m}|=k} a_{\boldsymbol{m}} (i\boldsymbol{\phi})^{\boldsymbol{m}} \Big),$$

choosing constants $a_{\boldsymbol{m}}$, $|\boldsymbol{m}| = k$, to satisfy

$$Y_{\boldsymbol{m}}[f_k - g] = c_{\boldsymbol{m}}. \tag{10.2}$$

Each of these equations easily reduces to the linear equation

$$\frac{(v(0,\Pi_h))_{\xi_{\boldsymbol{m}}}}{(v(0,\mathcal{P}_h))_{\xi_{\boldsymbol{m}}}}a_{\boldsymbol{m}} + Y_{\boldsymbol{m}}[f_{k-1} - g] = c_{\boldsymbol{m}},$$

which is uniquely solvable since by assumption $(v(0,\Pi_h))_{\xi_{\boldsymbol{m}}} \neq 0$. Moreover, $f_k$ satisfies (10.2) for each $|\boldsymbol{m}| < k$. Indeed, for $k = p$ this is by construction. Let it be true for $f_{k-1}$. If $|\boldsymbol{m}| \leq k-1$ then

$$Y_{\boldsymbol{m}}[f_k - g] = Y_{\boldsymbol{m}}[f_{k-1} - g] + Y_{\boldsymbol{m}}\left[f_{k-1}(\boldsymbol{\phi}) \sum_{|\boldsymbol{n}|=k} a_{\boldsymbol{n}}(i\boldsymbol{\phi})^{\boldsymbol{n}}\right] = c_{\boldsymbol{m}} + \sum_{|\boldsymbol{n}|=k} a_{\boldsymbol{n}} Y_{\boldsymbol{m}}[f_{k-1}(\boldsymbol{\phi})(i\boldsymbol{\phi})^{\boldsymbol{n}}].$$

It remains to note that $Y_{\boldsymbol{m}}[f(\boldsymbol{\phi})(i\boldsymbol{\phi})^{\boldsymbol{n}}] = 0$ if $|\boldsymbol{n}| > |\boldsymbol{m}|$.

Denote $W(\boldsymbol{\phi}) = f_q(\boldsymbol{\phi})$. By construction, there holds $W(\boldsymbol{\phi}) = v(\boldsymbol{\phi},\Pi_h)(1 + O(|\boldsymbol{\phi}|^p))$ and $W(\boldsymbol{\phi}) = v(\boldsymbol{\phi},\Pi_h^{(p,q)})(1 + O(|\boldsymbol{\phi}|^p))$. By Lemma 6.7 there exist real-valued diagonal matrices $\tilde{\mathfrak{C}}^{(\boldsymbol{m})}$, $p \leqslant |\boldsymbol{m}| \leqslant q$, such that $W(\boldsymbol{\phi}) = v(\boldsymbol{\phi},\tilde{\Pi}_h^{(p,q)}) + O(|\boldsymbol{\phi}|^{q+1})$; by construction from (10.2) (see (6.13)) we have $\tilde{\mathfrak{C}}^{(\boldsymbol{m})}_{\xi_{\boldsymbol{m}},\xi_{\boldsymbol{m}}} = c_{\boldsymbol{m}}$. Then

$$\hat{\epsilon}(\boldsymbol{\phi},\tilde{\Pi}_h^{(p,q)}) = A(\boldsymbol{\phi})v(\boldsymbol{\phi},\tilde{\Pi}_h^{(p,q)}) = A(\boldsymbol{\phi})W(\boldsymbol{\phi}) + O(|\boldsymbol{\phi}|^{q+1}) =$$
$$= A(\boldsymbol{\phi})v(\boldsymbol{\phi},\Pi_h^{(p,q)})(1 + O(|\boldsymbol{\phi}|^p)) + O(|\boldsymbol{\phi}|^{q+1}) = O(|\boldsymbol{\phi}|^{q+1}).$$

By Theorem 6.15 the scheme possesses the truncation error of order $q$ in the sense of $\tilde{\Pi}_h^{(p,q)}$. □

*10.5. The simple case*

For a simple scheme the optimal values of the formal order of accuracy and the long-time simulation order can be found by the following algorithm.

**Algorithm 4** (to detect the formal and long-time simulation orders in the simple case)**.**

1. *Detect the order of accuracy $P_A$ using Algorithm 1.*
2. *Choose any local mapping $\mathcal{P}_h$ such that $(\mathcal{P}_1 1)_{0,\xi} \neq 0$ for each $\xi \in M^0$.*
3. *Put $m = P_A + 1$.*
4. *Compute $f^{\boldsymbol{m}} = -(\epsilon_1(\boldsymbol{r}^{\boldsymbol{m}}/\boldsymbol{m}!, \Pi_1^{(P_A+1,m-1)}))_0$ for each $\boldsymbol{m}$ such that $|\boldsymbol{m}| = m$, substituting into $\Pi_h^{(P_A+1,m-1)}$ the previously found coefficients $\mathfrak{C}^{(\boldsymbol{n})}$, $P_A + 1 \leqslant |\boldsymbol{n}| \leqslant m-1$.*
5. *If for each $\boldsymbol{m}$ with $|\boldsymbol{m}| = m$ there holds $f^{\boldsymbol{m}} \in \mathrm{Im} L(0)$, then:*
   - *find any diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$ satisfying $L(0)\mathfrak{C}^{(\boldsymbol{m})}(\mathcal{P}_1 1)_0 = f^{\boldsymbol{m}}$;*
   - *increase $m$ by one;*
   - *return to the step 4.*
6. *Put $Q' = m - 1$.*
7. *Put $P' = P_A$ if $Q' = P_A$ and $P' = P_A + 1$ otherwise.*

Note that at step 5, since $\mathrm{Rank}\, L(0) = |M^0| - 1$, the equation specifies a one-parametric family of diagonal matrices $\mathfrak{C}^{\boldsymbol{m}}$. Any matrix of this family may be chosen.

**Theorem 10.8.** *Consider a simple scheme of the form* (2.8) *and a local mapping $\Pi_h$. Let $P$ and $Q$ be its optimal values of the formal order of accuracy and the long-time simulation order (see Definition 5). If $Q = \infty$, then the algorithm loops endlessly. Otherwise Algorithm 4 returns values $P' = P$ and $Q' = Q$.*

*Proof.* Let $P_A$ be the optimal order of the truncation error. Let $P'$ and $Q'$ be the values returned by the algorithm; if it loops endlessly, formally put $P' = P_A + 1$, $Q' = \infty$. We need to prove that $Q' = Q$, then the equality $P' = P$ will follow from statement 5 of Theorem 3.

First we prove that $Q \geqslant Q'$. If $Q' = P_A$, this is obvious by stability. If $P_A < Q' < \infty$, arguing as in the proof of Theorem 10.6 we get that the scheme possesses the orders $P'$ and $Q'$. By Lemma 9.6 it possesses the orders $P$ and

$Q'$, hence $Q \geqslant Q'$. If $Q' = \infty$, using the same argument, for each $s \geqslant P'$ the scheme possesses the formal order of accuracy $P'$ and the long-time simulation order $s$. By Lemma 9.8 it possesses the formal order of accuracy $P'$ and the long-time simulation order $Q = \infty$.

Now we prove that $Q' \geqslant Q$. Let $\Pi_h^{(P_A+1,Q')}$ be given by (2.12) with the coefficients $\mathfrak{C}^{(\boldsymbol{m})}$ found by the algorithm. By construction and Lemma 10.4,

$$\left(\epsilon_1\left(\frac{\boldsymbol{r}^m}{\boldsymbol{m}!}, \Pi_1^{(P_A+1,Q')}\right)\right)_0 = 0$$

for each $|\boldsymbol{m}| \leqslant Q'$, i. e. the scheme possesses the truncation error of order $Q'$ in the sense of $\Pi_h^{(P_A+1,Q')}$. Assume that $Q > Q'$. If $Q'$ is an optimal value of the truncation error in the sense of $\Pi_h^{(P_A+1,Q')}$, then by statement 5 of Theorem 3 the scheme possesses the formal order of accuracy $Q'+1$ in the sense of $\Pi_h^{(P_A+1,Q')}$. By statement 3 of Theorem 1 there exist $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}| = Q'+1$, such that the scheme possesses the truncation error of order $Q'+1$ in the sense of $\Pi_h^{(P_A+1,Q'+1)}$. If $Q'$ is not an optimal value, then the scheme possesses the truncation error of order $Q'+1$ in the sense of $\Pi_h^{(P_A+1,Q'+1)}$ with $\mathfrak{C}^{(\boldsymbol{m})} = 0$, $|\boldsymbol{m}| = Q'+1$. These coefficients $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}| = Q'+1$, give a solution of the equation to solve at step 5, thus the algorithm makes one more step. This contradiction proves that $Q' \geqslant Q$ and thus the whole Theorem. □

Note that in the quasi-1D case Algorithm 4 may give a wrong result (see Section 11.3).

## 11. 1D examples

Throughout this section we consider 1D transport equation (2.6) with transport velocity $\omega = 1$, i. e. $\partial v/\partial t + \partial v/\partial x = 0$, and put $a_1 = 1$, i. e. the parameter $h$ coincides with the mesh step.

### *11.1. P1 discontinuous Galerkin method*

Consider the discontinuous Galerkin method (its definition may be found, for example, in [31]) based on the piecewise-linear polynomials on the uniform mesh with the nodes $x_j = jh$. On each cell $(x_j, x_{j+1})$ define two basis functions:

$$\phi_j^L(x) = \frac{x_{j+1} - x}{h}, \quad \phi_j^R(x) = \frac{x - x_j}{h},$$

and extend these functions by zero outside $(x_j, x_{j+1})$. For the numerical solution

$$u(t,x) = \sum_{j=0}^{N-1}\Big(u_j^L(t)\phi_j^L(x) + u_j^R(t)\phi_j^R(x)\Big),$$

the DG scheme gives

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix} \frac{du_j}{dt} + \frac{1}{h}\left[\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} u_j + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} u_{j-1}\right] = 0. \tag{11.1}$$

It has the form (2.8) with $\mathcal{S} = \{0, -1\}$,

$$Z_{-1} = 0, \quad Z_0 = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}, \quad L_{-1} = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \quad L_0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

By definition (6.14) we have

$$A(\phi) = i\phi I - \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} - e^{-i\phi} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} i\phi - 3 & -1 + 4e^{-i\phi} \\ 3 & i\phi - 1 - 2e^{-i\phi} \end{pmatrix}.$$

We see that $A(0) \neq 0$ and

$$\det A(\phi) = (-2i\phi - 6)e^{-i\phi} + (-\phi^2 - 4i\phi + 6) = \frac{\phi^4}{12} + O(\phi^5).$$

The eigenvalues of the matrix $A(\phi)$ are $\lambda_\pm(\phi) = i\phi - 2 - e^{-i\phi} \pm \varkappa(\phi)$, where $\varkappa(\phi) = (e^{-2i\phi} + 10e^{-i\phi} - 2)^{1/2}$. The branch of the root is such that $\varkappa(0) = 3$. The right eigenvectors matrix is

$$S(\phi) = \frac{1}{4e^{-i\phi} - 1} \begin{pmatrix} 4e^{-i\phi} - 1 & -4e^{-i\phi} + 1 \\ 1 - e^{-i\phi} + \varkappa(\phi) & -1 + e^{-i\phi} + \varkappa(\phi) \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_+(\phi) & \boldsymbol{r}_-(\phi) \end{pmatrix},$$

the left eigenvectors matrix is

$$S^{-1}(\phi) = \frac{1}{2\varkappa(\phi)} \begin{pmatrix} -1 + e^{-i\phi} + \varkappa(\phi) & 4e^{-i\phi} - 1 \\ -1 + e^{-i\phi} - \varkappa(\phi) & 4e^{-i\phi} - 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{l}_+(\phi) \\ \boldsymbol{l}_-(\phi) \end{pmatrix}.$$

The first column in $S(\phi)$ and, correspondingly, the first row in $S^{-1}(\phi)$ correspond to $\lambda_+(\phi)$, the second ones to $\lambda_-(\phi)$. The eigenvalues can be approximated as $\lambda_+(\phi) \approx -\phi^4/72 + O(\phi^5)$, $\lambda_-(\phi) = -6 + O(\phi)$. Clearly, $\lambda_+(\phi)$ corresponds to the physical mode

$$\begin{pmatrix} u_j^L(t) \\ u_j^R(t) \end{pmatrix} = \exp\left(i\frac{\phi}{h}(jh - t)\right) \exp\left(\lambda_+(\phi)\frac{t}{h}\right) \boldsymbol{r}_+(\phi)$$

and $\lambda_-(\phi)$ to the spurious one (the same expression with $\lambda_+(\phi)$, $\boldsymbol{r}_+(\phi)$ replaced by $\lambda_-(\phi)$, $\boldsymbol{r}_-(\phi)$).

In order to speak about the accuracy we need to specify a map $\Pi_h$. For simplicity we use the pointwise map $(\Pi_h f)_j^L = f(x_j)$, $(\Pi_h f)_j^R = f(x_{j+1})$. Then $v(\phi, \Pi_h) = (1, e^{i\phi})^T$ and

$$S^{-1}(\phi)v(\phi, \Pi_h) = \frac{1}{2\varkappa(\phi)} \begin{pmatrix} 3 + e^{-i\phi} - e^{i\phi} + \varkappa(\phi) \\ 3 + e^{-i\phi} - e^{i\phi} - \varkappa(\phi) \end{pmatrix} = \begin{pmatrix} 1 + O(\phi) \\ \frac{1}{12}\phi^2 + O(\phi^3) \end{pmatrix}.$$

Using the notation used in Section 7 we have $\bar{\aleph} = \{0, 4\}$, $p_0 = 2$, $p_4 = 0$. Thus by Lemma 7.6 the scheme possesses the 1st order of the truncation error and the optimal error estimate has the form $O(h^2 + h^3 t)$.

This result can be also obtained with the help of Proposition 9.14 without the calculation of eigenvalues. Since $\dim \operatorname{Ker} A(0) = 1$, the scheme is simple. By construction, the scheme is exact on linear functions and thus possesses the truncation error of the first order. Using the fact that $\det A(\phi) = c\phi^4 + O(|\phi|^5)$, $c \neq 0$, by Proposition 9.14 the scheme possesses the second order of accuracy and the third order in the long-time simulation. If we additionally verify that the scheme is not exact on quadratic polynomials, then by Proposition 9.14 we get that these values are optimal.

Replacing $\Pi_h$ by $\tilde{\Pi}_h$ it is possible to improve $p_0$ and hence the order of the truncation error and the formal order of accuracy. For this purpose we need to reduce the second component of $S^{-1}(\phi)v(\phi, \tilde{\Pi}_h)$. We can nullify it if we put $w(\phi) = \gamma(\phi)\boldsymbol{r}_+(\phi)$ for some $\gamma(\phi) \in \mathbb{C} \setminus \{0\}$ such that $\gamma(-\phi) = \overline{\gamma(\phi)}$ (for example, $\gamma(\phi) \equiv 1$) and get $\tilde{\Pi}_h$ by Lemma 6.6 such that $v(\phi, \tilde{\Pi}_h) = w(\phi)$ in a neighborhood of $\phi = 0$. This map is nonlocal; to specify a local map we need to approximate $w(\phi)$. For example, if we put $\gamma(\phi) \equiv 1$, define $\mathcal{P}_h$ as $(\mathcal{P}_h f)_{\eta,L} = (\mathcal{P}_h f)_{\eta,R} = f(h\eta)$, and use Lemma 6.7 then we get the mapping $\Pi_h^{(2,3)}$ given by

$$(\Pi_h^{(2,3)} f)_{\eta,L} = f(\eta h), \quad (\Pi_h^{(2,3)} f)_{\eta,R} = f((\eta+1)h) + \frac{1}{6}h^2 f''(\eta h) + \frac{5}{18}h^3 f'''(\eta h). \tag{11.2}$$

In this case $v(\phi, \Pi_h^{(2,3)}) = (1, e^{i\phi} - \phi^2/6 - 5i\phi^3/18)^T$. The scheme (11.1) possesses the truncation error of order 3 in the sense of $\Pi_h^{(2,3)}$. Doing the same with $\mathcal{P}_h = \Pi_h$ we get the mapping $\tilde{\Pi}_h^{(2,3)}$ given by

$$(\tilde{\Pi}_h^{(2,3)} f)_{\eta,L} = f(\eta h), \quad (\tilde{\Pi}_h^{(2,3)} f)_{\eta,R} = f((\eta+1)h) + \frac{1}{6}h^2 f''((\eta+1)h) + \frac{1}{9}h^3 f'''((\eta+1)h),$$

and thus $v(\phi, \tilde{\Pi}_h^{(2,3)}) = (1, e^{i\phi}(1 - \phi^2/6 - i\phi^3/9))^T$. Approximations of derivatives in (11.2) with enough order yield other suitable mappings.

Now we demonstrate the method of auxiliary mapping with no use of the eigenvectors. We have

$$L(0) = L_0 + L_{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Since $\dim \operatorname{Ker} L(0) = 1$, we will use Algorithm 4. We will use $\mathcal{P}_h$ defined as $(\mathcal{P}_h f)_{\eta,L} = (\mathcal{P}_h f)_{\eta,R} = f(h\eta)$, then $\vec{\mathfrak{e}} = v(0, \mathcal{P}_h) = (1,1)^T$. Substituting a linear function into the scheme we see that the scheme possesses the first order of the truncation error. Doing the same with a quadratic polynomial we see the optimal value of the order of the truncation error is equal to 1.

We need to compute $f^2 = -(\epsilon_1(x^2/2, \Pi_1))_0$. We have

$$f^2 = \big(Z\Pi_1 x - L\Pi_1(x^2/2)\big)_0 = Z_0(\Pi_1 x)_0 - L_0(\Pi_1(x^2/2))_0 - L_{-1}(\Pi_1(x^2/2))_{-1}$$

and

$$f^2 = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{bmatrix} 0 \\ 1/2 \end{bmatrix} - \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} = \begin{bmatrix} -1/12 \\ 1/12 \end{bmatrix}.$$

The system

$$\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \mathfrak{C}^{(2)} \vec{\mathfrak{e}} = \begin{pmatrix} -1/12 \\ 1/12 \end{pmatrix}$$

has the solution

$$\mathfrak{C}^{(2)} = \begin{pmatrix} 0 & 0 \\ 0 & 1/6 \end{pmatrix} + c_2 I.$$

According to the algorithm the value of $c_2$ can be set arbitrarily; put $c_2 = 0$. Now we compute $f^3 = -(\epsilon_1(x^3/6, \Pi_1^{(2,2)}))_0$ where $\Pi_h^{(2,2)}$ is prescribed by (2.12) with the coefficients $\mathfrak{C}^{(2)}$ we just found. We have

$$f^3 = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix} \left[ \begin{pmatrix} 0 \\ 1/2 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] - \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left[ \begin{pmatrix} 0 \\ 1/6 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right] -$$
$$- \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \left[ \begin{pmatrix} -1/6 \\ 0 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right] = \begin{pmatrix} -5/36 \\ 5/36 \end{pmatrix}.$$

The system

$$\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \mathfrak{C}^{(3)} \vec{\mathfrak{e}} = \begin{pmatrix} -5/36 \\ 5/36 \end{pmatrix}$$

has the solution

$$\mathfrak{C}^{(3)} = \begin{pmatrix} 0 & 0 \\ 0 & 5/18 \end{pmatrix} + c_3 I.$$

The map $\Pi_h^{(2,3)}$ generated by chosen $\Pi_h$, $\mathcal{P}_h$ and matrices $\mathfrak{C}^{(2)}$ and $\mathfrak{C}^{(3)}$ with $c_2 = c_3 = 0$ coincides with (11.2).

Repeat the operation for $f^4 = -(\epsilon_1(x^4/24, \Pi_1^{(2,3)}))_0$.

$$f^4 = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix} \left[ \begin{pmatrix} 0 \\ 1/6 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} 1 \\ 1/2 \end{pmatrix} + \mathfrak{C}^{(3)} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] -$$
$$- \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left[ \begin{pmatrix} 0 \\ 1/24 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \mathfrak{C}^{(3)} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right] -$$
$$- \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \left[ \begin{pmatrix} -1/24 \\ 0 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} -1/2 \\ -1/2 \end{pmatrix} + \mathfrak{C}^{(3)} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right] = \begin{pmatrix} -127/432 \\ 67/432 \end{pmatrix}.$$

The system $L(0)\mathfrak{C}^{(4)}\vec{\mathfrak{e}} = f^4$ is inconsistent. Thus $P = 2$ and $Q = 3$ are the optimal values of the formal order of accuracy and of the long-time simulation order.

### *11.2. Arbitrary order DG method*

In this subsection we demonstrate Algorithm 4 on the discontinuous Galerkin method based on the $p$-th order polynomials, where $p \in \mathbb{N}$. The discontinuous Galerkin method gives a scheme of the form

$$\sum_{\xi=0}^{p} \frac{du_{\eta,\xi}}{dt} \int_0^1 \phi_\xi(x)\psi_j(x)dx - \sum_{\xi=0}^{p} u_{\eta,\xi} \int_0^1 \phi_\xi(x)\psi_j'(x)dx + \sum_{\xi=0}^{p} u_{\eta,\xi}\phi_\xi(1)\psi_j(1) - \sum_{\xi=0}^{p} u_{\eta-1,\xi}\phi_\xi(1)\psi_j(0) = 0,$$

where $\phi_\xi$, $\xi = 0, \dots, p$ and $\psi_j$, $j = 0, \dots, p$ are two bases in the space of polynomials of order $p$.

In this scheme we can choose the set of basis functions in the polynomial space and the initial local mapping $\Pi_h$ at our convenience. Let $x_k$ and $w_k$ be the nodes and weights of the Gauss – Jacobi quadrature rule with $p+1$ nodes, i. e. the quadrature rule on $[0,1]$ exact on the polynomials of order $2p$ such that $x_p = 1$. Let $\{\phi_\xi(x), x \in [0,1], \xi = 0, \dots, p\}$ be the collocation basis with the nodes $x_\xi$, i. e. $\phi_\xi(x_\zeta) = 1$ for $\xi = \zeta$ and 0 otherwise. Let the test functions be $\psi_j(x) = x^j/j!$, $j = 0, \dots, p$. Then we have

$$\int_0^1 \phi_\xi(x)\psi_j(x)dx = \sum_{k=0}^{p} w_k \phi_\xi(x_k)\psi_j(x_k) = w_\xi \frac{x_\xi^j}{j!}, \quad \int_0^1 \phi_\xi(x)\psi_j'(x)dx = \sum_{k=0}^{p} w_k \phi_\xi(x_k)\psi_j'(x_k) = w_\xi \frac{x_\xi^{j-1}}{(j-1)!},$$

and $\phi_\xi(1) = 1$ for $\xi = p$ and 0 otherwise. So the scheme is of the form (2.8), namely,

$$Z_0 \frac{du_\eta}{dt} + L_0 u_\eta + L_{-1} u_{\eta-1} = 0,$$

$$Z_0 = \begin{pmatrix} w_0 & \dots & w_{p-1} & w_p \\ w_0 x_0 & \dots & w_{p-1}x_{p-1} & w_p x_p \\ w_0 x_0^2/2 & \dots & w_{p-1}x_{p-1}^2/2 & w_p x_p^2/2 \\ \vdots & \ddots & \vdots & \vdots \\ w_0 x_0^p/p! & \dots & w_{p-1}x_{p-1}^p/p! & w_p x_p^p/p! \end{pmatrix}, \quad L_{-1} = \begin{pmatrix} 0 & \dots & 0 & -1 \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 0 \end{pmatrix},$$

$$L_0 = \begin{pmatrix} 0 & \dots & 0 & 1 \\ -w_0 & \dots & -w_{p-1} & 1 - w_p \\ -w_0 x_0 & \dots & -w_{p-1}x_{p-1} & 1/2 - w_p x_p \\ \vdots & \ddots & \vdots & \vdots \\ -w_0 x_0^{p-1}/(p-1)! & \dots & -w_{p-1}x_{p-1}^{p-1}/(p-1)! & 1/p! - w_p x_p^{p-1}/(p-1)! \end{pmatrix}. \tag{11.3}$$

We equip the scheme with the mapping $\Pi_h$ defined by $(\Pi_h f)_{\eta,\xi} = f(h(\eta + x_\xi))$.

The bottom-left $(p \times p)$-submatrix of $L(0) = L_{-1} + L_0$ is the Vandermonde matrix with scaled columns, so it is nonsingular. Thus $\dim \operatorname{Ker} L(0) = 1$ and $\operatorname{Im} L(0)$ consists of the vectors with the first component equal to zero. Denote $\vec{\mathfrak{e}} = (1, \dots, 1)^T$. We claim that $L(0)\vec{\mathfrak{e}} = 0$ and thus $\operatorname{Ker} L(0) = \operatorname{span}\{\vec{\mathfrak{e}}\}$. Indeed, for $j = 0$ the equality $(L(0)\vec{\mathfrak{e}})_j = 0$ is obvious; for $j > 0$ we have

$$(L(0)\vec{\mathfrak{e}})_j = \frac{1}{j!} - \sum_{\xi=0}^{p} w_\xi \frac{x_\xi^{j-1}}{(j-1)!} = 0.$$

The scheme is simple ($\dim \operatorname{Ker} L(0) = 1$) and we can use Algorithm 4. We choose the mapping $\mathcal{P}_h$ defined by $(\mathcal{P}_h f)_{\eta,\xi} = f(h\eta)$, $\xi = 0, \dots, p$. Step 5 of the algorithm defines the diagonal matrix $\mathfrak{C}^{(m)}$ as a solution of $L(0)\mathfrak{C}^{(m)}\vec{\mathfrak{e}} = f^m$. Since $L(0)\vec{\mathfrak{e}} = 0$, the general solution of this equation has the form $\mathfrak{C}^{(m)} + \alpha I$, $\alpha \in \mathbb{R}$, so we will specify $\mathfrak{C}^{(m)}$ by $\mathfrak{C}^{(n)}_{p,p} = 0$.

Step 1. We need to get the order of the truncation error. We have

$$-\left(\epsilon_1\left(\frac{x^m}{m!}, \Pi_1\right)\right)_0 = Z_0\left(\Pi_1 \frac{x^{m-1}}{(m-1)!}\right)_0 - L_0\left(\Pi_1 \frac{x^m}{m!}\right)_0 - L_{-1}\left(\Pi_1 \frac{x^m}{m!}\right)_{-1}.$$

Componentwise,

$$-\left(\epsilon_1\left(\frac{x^m}{m!}, \Pi_1\right)\right)_{0,j} = \sum_\xi (Z_0)_{j,\xi}\left(\Pi_1 \frac{x^{m-1}}{(m-1)!}\right)_{0,\xi} - \sum_\xi (L_0)_{j,\xi}\left(\Pi_1 \frac{x^m}{m!}\right)_{0,\xi} - \sum_\xi (L_{-1})_{j,\xi}\left(\Pi_1 \frac{x^m}{m!}\right)_{-1,\xi}.$$

For $m = 0$ this is zero, so assume $m > 0$. For $j = 0$ we have

$$-\left(\epsilon_1\left(\frac{x^m}{m!}, \Pi_1\right)\right)_{0,0} = \sum_\xi w_\xi \frac{x_\xi^{m-1}}{(m-1)!} - \frac{1}{m!}$$

and for $j = 1, \dots, p$

$$-\left(\epsilon_1\left(\frac{x^m}{m!}, \Pi_1\right)\right)_{0,j} = \sum_\xi w_\xi \frac{x_\xi^j}{j!}\frac{x_\xi^{m-1}}{(m-1)!} + \sum_\xi w_\xi \frac{x_\xi^{j-1}}{(j-1)!}\frac{x_\xi^m}{m!} - \frac{1}{j!\,m!} = \frac{m+j}{j!\,m!}\left(\sum_\xi w_\xi x_\xi^{j+m-1} - \frac{1}{m+j}\right).$$

Since $x_\xi$ and $w_\xi$ are nodes and weights of the quadrature rule of order $2p$, then

$$\left(\epsilon_1\left(\frac{x^m}{m!}, \Pi_1\right)\right)_{0,j} = 0, \qquad j + m - 1 \leqslant 2p. \tag{11.4}$$

Since the maximal value for $j$ is $p$, then the truncation error is zero for $m \leqslant p+1$. Thus the scheme possesses the truncation error of order $P_A = p+1$. The value $P_A = p+1$ is optimal, otherwise the quadrature formula would be exact on all polynomials of order $2p+1$.

Step 2. Put $m = P_A + 1 = p + 2$.

Steps 3, 4. Let $\mathfrak{C}^{(n)}$, $n = p+2, \dots, m-1$ be the coefficients already found by the algorithm by solving

$$L(0)\mathfrak{C}^{(n)}\vec{\mathfrak{e}} = f^n, \quad f^n = -\left(\epsilon_1\left(\frac{x^n}{n!}, \Pi_1^{(p+2,n-1)}\right)\right)_0 \tag{11.5}$$

for $\mathfrak{C}^{(n)}$. We claim that the first $2p+2-n$ components of the right-hand side of (11.5) are equal to zero. Indeed, for $n = p+2$ this follows from (11.4). Assume this holds for all $n' = p+2, \dots, n-1$. Then

$$f^n = -\left(\epsilon_1\left(\frac{x^n}{n!}, \Pi_1^{(p+2,n-1)}\right)\right)_0 = -\left(\epsilon_1\left(\frac{x^n}{n!}, \Pi_1\right)\right)_0 + Z_0 \sum_{n'=p+2}^{n-1} \mathfrak{C}^{(n')}\left(\mathcal{P}_1\frac{x^{n-1-n'}}{(n-1-n')!}\right)_0 -$$
$$-L_0 \sum_{n'=p+2}^{n-1} \mathfrak{C}^{(n')}\left(\mathcal{P}_1\frac{x^{n-n'}}{(n-n')!}\right)_0 - L_{-1} \sum_{n'=p+2}^{n-1} \mathfrak{C}^{(n')}\left(\mathcal{P}_1\frac{x^{n-n'}}{(n-n')!}\right)_{-1}.$$

Recall that $\mathcal{P}_1$ takes the point value at $x = 0$ and $\mathfrak{C}^{(n)}_{p,p} = 0$ for each $n$. Thus

$$f^n = -\left(\epsilon_1\left(\frac{x^n}{n!}, \Pi_1\right)\right)_0 + Z_0\mathfrak{C}^{(n-1)}\vec{\mathfrak{e}}.$$

For the first term on the right-hand side the first $2p+2-n$ components are zero (see (11.4)). For the second term on the right-hand side by (11.3) and $\mathfrak{C}^{(n')}_{p,p} = 0$ for each $n'$ we have

$$(Z_0\mathfrak{C}^{(n-1)}\vec{\mathfrak{e}})_j = (-L_0\mathfrak{C}^{(n-1)}\vec{\mathfrak{e}})_{j+1}, \quad j = 0, \dots, p-1.$$

So by induction assumption the first $2p+2-n$ components of $Z_0\mathfrak{C}^{(n-1)}\vec{\mathfrak{e}}$ are also zero. Thus the first $2p+2-n$ components of $f^n$ are equal to zero. In particular, for each $m \leqslant 2p+1$ there holds $f^m_0 = 0$, i. e. $f^m \in \mathrm{Im}L(0)$. For $m = 2p+2$ this possibly will not hold (in fact – not possibly but definitely, but we have no proof for this).

Step 5. Put $Q' = m - 1$. We know that at the last stage $m$ was greater than or equal to $2p+2$, thus $Q' \geqslant 2p+1$.

Step 6. Since $Q' \geqslant 2p+1 > P_A = p+1$, we have $P' = P_A + 1 = p+2$.

Since the algorithm returns $Q' \geqslant 2p+1$, by Theorem 10.8 the DG scheme based on $p$-th order polynomials possesses the long-time simulation order $2p+1$. This proof can be considered as a variation of the proof in [15].

### *11.3. Alternating central difference scheme*

In this section we demonstrate Algorithm 3.

Let the set of the DOFs be $M^0 = \{$"L", "R"$\}$, $|M^0| = 2$. We will treat the first DOF as a value in the node and the second one as a value in the cell center. Correspondingly we define the pointwise map $\overline{\Pi}_h$ as $(\overline{\Pi}_h f)_{j,L} = f(jh)$, $(\overline{\Pi}_h f)_{j,R} = f(jh + h/2)$, $j \in \mathbb{Z}$.

Derivatives at nodes are approximated by the 2-nd order central difference and derivatives at cell centers are approximated by the 4-th order central difference. For brevity denote $(u_j)_L = v_j$, $(u_j)_R = v_{j+1/2}$, then

$$\begin{gathered} \frac{dv_j}{dt} + \frac{v_{j+1/2} - v_{j-1/2}}{h} = 0, \quad j \in \mathbb{Z}; \\ \frac{dv_{j+1/2}}{dt} + \frac{4}{3}\frac{v_{j+1} - v_j}{h} - \frac{1}{3}\frac{v_{j+3/2} - v_{j-1/2}}{2h} = 0, \quad j \in \mathbb{Z}. \end{gathered} \tag{11.6}$$

In the sense of $\overline{\Pi}_h$ the scheme (11.6) possesses the second order of truncation error.

The scheme can be rewritten in the block form

$$\frac{du_j}{dt} + L_{-1}u_{j-1} + L_0 u_j + L_1 u_{j+1} = 0,$$

$$L_{-1} = \begin{pmatrix} 0 & -1 \\ 0 & 1/6 \end{pmatrix}, \quad L_0 = \begin{pmatrix} 0 & 1 \\ -4/3 & 0 \end{pmatrix}, \quad L_1 = \begin{pmatrix} 0 & 0 \\ 4/3 & -1/6 \end{pmatrix}.$$

The matrix $L(\phi) = e^{-i\phi}L_{-1} + L_0 + e^{i\phi}L_1$ is

$$L(\phi) = \begin{pmatrix} 0 & 1 - e^{-i\phi} \\ \frac{4}{3}(-1 + e^{i\phi}) & -\frac{1}{6}(e^{i\phi} - e^{-i\phi}) \end{pmatrix} = \begin{pmatrix} 0 & 2ie^{-i\phi/2}\sin(\phi/2) \\ \frac{8}{3}ie^{i\phi/2}\sin(\phi/2) & -\frac{1}{3}i\sin(\phi) \end{pmatrix}.$$

The matrix $A(\phi)$ defined by (6.14) is

$$A(\phi) = \begin{pmatrix} i\phi & -2ie^{-i\phi/2}\sin(\phi/2) \\ -\frac{8}{3}ie^{i\phi/2}\sin(\phi/2) & i\phi + \frac{1}{3}i\sin(\phi) \end{pmatrix}.$$

The eigenvalues of $A(\phi)$ are equal to

$$\lambda_\pm(\phi) = i\phi + \frac{1}{6}i\sin(\phi) \pm \frac{1}{6}i\sin(\phi/2)\varkappa(\phi), \quad \varkappa(\phi) = \sqrt{194 + 2\cos\phi} > 0.$$

The eigenvaules $\lambda_\pm(\phi)$ are pure imaginary. For $\phi = 0$ there holds $A(\phi) = 0$. For $\phi \neq 0$ there holds $\lambda_+(\phi) \neq \lambda_-(\phi)$. Thus for each $\phi$ the matrix $A(\phi)$ has two linearly independent eigenvectors. The matrices of the right and left eigenvectors are

$$S(\phi) = \frac{1}{4\varkappa(\phi)}\begin{pmatrix} 2\exp(-i\phi/2) & 2\exp(-i\phi/2) \\ i\lambda_+/\sin(\phi/2) & i\lambda_-/\sin(\phi/2) \end{pmatrix}, \quad S^{-1}(\phi) = \begin{pmatrix} \exp(i\phi/2)(\varkappa(\phi) - 2\cos(\phi/2)) & -12 \\ \exp(i\phi/2)(\varkappa(\phi) + 2\cos(\phi/2)) & 12 \end{pmatrix}.$$

For $\phi = 0$ we take the corresponding limits. The first column of $S(\phi)$ and the first row of $S^{-1}(\phi)$ correspond to $\lambda_+(\phi)$, the second ones to $\lambda_-(\phi)$.

Since matrices $S(\phi)$ and $S^{-1}(\phi)$ are bounded uniformly in $\phi \in \mathbb{R}$, this yields $\|\exp(A(\phi)\nu)\| \leqslant K$ for $\nu \geqslant 0$. Thus the scheme (11.6) is stable.

The Taylor expansions of the eigenvalues are

$$\lambda_+(\phi) = \frac{7}{3}i\phi + O(\phi^3), \quad \lambda_-(\phi) = \frac{1}{42}i\phi^3 + O(\phi^5).$$

The Taylor expansions of the matrix $S^{-1}$ is

$$S^{-1}(\phi) = \begin{pmatrix} 12 + 6i\phi - \frac{9}{7}\phi^2 + \ldots & -12 \\ 16 + 8i\phi - \frac{16}{7}\phi^2 + \ldots & 12 \end{pmatrix}.$$

Now consider the "block" map $\Pi_h$ defined as

$$(\Pi_h f)_{j,L} = (\Pi_h f)_{j,R} = f(jh).$$

Then we have $v(\phi,\Pi_h)=(1,1)^T$ and

$$S^{-1}(\phi)v(\phi,\Pi_h)=\begin{pmatrix} 6i\phi+O(\phi^2)\\ 28+O(\phi)\end{pmatrix}.$$

945 Thus we get $\bar{\aleph}=\{1,3\}, p_1=1, p_3=0$. By Lemma 7.6 this yields $\|\varepsilon_h(t,v_0,\Pi_h)\|\leqslant C_1h\|\nabla v_0\|_\infty+C_2th^2\|\nabla^3v_0\|_\infty$ for each $v_0\in C^3_{per}(\mathbb{R})$.

Now we demonstrate Algorithm 3. Using Algorithm 1 it is easy to check in the sense of $\Pi_h$ the scheme possesses the truncation error of order $P_A=1$ and not of order 2. Since $A(0)=0$, by statement 5 of Theorem 1 there holds $P=P_A=1$, and we can skip step 1 of the algorithm. Now put $\mathcal{P}_h=\Pi_h$ and denote $\vec{\mathfrak{e}}=(1,1)^T$. Consider the system

$$\begin{cases}\left(\epsilon_1(x,\Pi_1^{(1,2)})\right)_0=0;\\ \left(\epsilon_1(x^2/2,\Pi_1^{(1,2)})\right)_0=0.\end{cases}$$

Since $(\epsilon_1(x,\Pi_1))_0=0$ and $L(0)=0$, we have $(\epsilon_1(x,\Pi_1^{(1,2)}))_0=(\epsilon_1(x,\Pi_1))_0+L(0)\mathfrak{C}^{(1)}\vec{\mathfrak{e}}=0$, so the first equation holds for each $\mathfrak{C}^{(1)}$ and $\mathfrak{C}^{(2)}$. The second equation has the form

$$-\left(\Pi_1^{(1,2)}(x)\right)_0+L_{-1}\left(\Pi_1^{(1,2)}(x)\right)_{-1}+L_0\left(\Pi_1^{(1,2)}(x)\right)_0+L_1\left(\Pi_1^{(1,2)}(x)\right)_1=0,$$

which expands to

$$\begin{aligned}&-\mathfrak{C}^{(1)}\begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}0&-1\\0&1/6\end{pmatrix}\left[\begin{pmatrix}1/2\\1/2\end{pmatrix}+\mathfrak{C}^{(1)}\begin{pmatrix}-1\\-1\end{pmatrix}+\mathfrak{C}^{(2)}\vec{\mathfrak{e}}\right]+\\ &+\begin{pmatrix}0&1\\-4/3&0\end{pmatrix}\mathfrak{C}^{(2)}\vec{\mathfrak{e}}+\begin{pmatrix}0&0\\4/3&-1/6\end{pmatrix}\left[\begin{pmatrix}1/2\\1/2\end{pmatrix}+\mathfrak{C}^{(1)}\begin{pmatrix}1\\1\end{pmatrix}+\mathfrak{C}^{(2)}\vec{\mathfrak{e}}\right]=0.\end{aligned}$$

Since $L(0)=L_{-1}+L_0+L_1=0$, the terms with $\mathfrak{C}^{(2)}$ negate each other. Simplifying, we obtain the system for $\mathfrak{C}^{(1)}=\operatorname{diag}\{\mathfrak{C}_L^{(1)},\mathfrak{C}_R^{(1)}\}$:

$$\begin{cases}-\mathfrak{C}_L^{(1)}+\mathfrak{C}_R^{(1)}-\frac12=0;\\ \frac43\mathfrak{C}_L^{(1)}-\frac43\mathfrak{C}_R^{(1)}+\frac23=0.\end{cases}$$

The solution of this system is

$$\mathfrak{C}^{(1)}=\begin{pmatrix}0&0\\0&1/2\end{pmatrix}+c_1I, \tag{11.7}$$

and the matrix $\mathfrak{C}^{(2)}$ is arbitrary. For $c_1=0$ and $\mathfrak{C}^{(2)}=0$ we obtain the map $\Pi_h^{(1,2)}$ prescribed by $(\Pi_hf)^{(1,2)}_{j,L}=f(jh)$, $(\Pi_h^{(1,2)}f)_{j,R}=f(jh)+(h/2)f'(jh)$. Note that $\Pi_h^{(1,2)}f=\overline{\Pi}_hf+O(h^2)$ for $f$ smooth enough.

Following the algorithm write the system

$$\begin{cases}\left(\epsilon_1(x,\Pi_1^{(1,3)})\right)_0=0;\\ \left(\epsilon_1(x^2/2,\Pi_1^{(1,3)})\right)_0=0;\\ \left(\epsilon_1(x^3/6,\Pi_1^{(1,3)})\right)_0=0.\end{cases}$$

for $\mathfrak{C}^{(1)},\mathfrak{C}^{(2)},\mathfrak{C}^{(3)}$. The first equation is valid for each $\mathfrak{C}^{(m)}$. The second one is equivalent to (11.7). By Lemma 10.7 we can drop the identity term in the expression (11.7) for $\mathfrak{C}^{(1)}$ by putting $c_1=0$. The third one is

$$\begin{aligned}&-\mathfrak{C}^{(2)}\begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}0&-1\\0&1/6\end{pmatrix}\left[\begin{pmatrix}-1/6\\-1/6\end{pmatrix}+\mathfrak{C}^{(1)}\begin{pmatrix}1/2\\1/2\end{pmatrix}+\mathfrak{C}^{(2)}\begin{pmatrix}-1\\-1\end{pmatrix}+\mathfrak{C}^{(3)}\vec{\mathfrak{e}}\right]+\\ &+\begin{pmatrix}0&1\\-4/3&0\end{pmatrix}\mathfrak{C}^{(3)}\vec{\mathfrak{e}}+\begin{pmatrix}0&0\\4/3&-1/6\end{pmatrix}\left[\begin{pmatrix}1/6\\1/6\end{pmatrix}+\mathfrak{C}^{(1)}\begin{pmatrix}1/2\\1/2\end{pmatrix}+\mathfrak{C}^{(2)}\begin{pmatrix}1\\1\end{pmatrix}+\mathfrak{C}^{(3)}\vec{\mathfrak{e}}\right]=0.\end{aligned}$$

Since $L(0)=0$, the terms with $\mathfrak{C}^{(3)}$ negate each other. Substituting $\mathfrak{C}_L^{(1)}=0$, $\mathfrak{C}_R^{(1)}=1/2$ and simplifying,

$$\begin{cases} -\mathfrak{C}_L^{(2)}+\mathfrak{C}_R^{(2)}-\frac{1}{12}=0; \\ \frac{4}{3}\mathfrak{C}_L^{(2)}-\frac{4}{3}\mathfrak{C}_R^{(2)}+\frac{1}{6}=0. \end{cases}$$

This system is inconsistent. This proves that the values $P=1$ and $Q=2$ are optimal.

Note that the use of Algorithm 4 would give the wrong result $P'=Q'=1$. Indeed, following Algorithm 4 we write

$$f^2=-\left(\epsilon_1\left(x^2/2,\Pi_1\right)\right)_0\neq 0$$

and $f^2\notin \mathrm{Im}L(0)$ since $\mathrm{Im}L(0)=\{0\}$.

*11.4. Family of schemes with unlimited growth of* $\mathfrak{C}$

In Theorem 1 for a scheme with the truncation error of order $P_A$ and the formal order of accuracy $P=P_A+1$ we proved the existence of $\Pi_h^{(P,P)}$ that gives the truncation error of order $P$ with an estimate for $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}|=P$. Theorem 3 does not give an analogous estimate on $\mathfrak{C}^{(\boldsymbol{m})}$. In this section we show that this estimate is not possible without additional assumptions.

Consider the family of schemes parametrized by $\gamma>0$ with 3 DOFs per cell:

$$h\frac{du_j}{dt}+\left(-\frac{1}{12}u_{j+2}+\frac{2}{3}u_{j+1}-\frac{2}{3}u_{j-1}+\frac{1}{12}u_{j-2}\right)+M(u_{j+1}-2u_j+u_{j-1})-\gamma G u_j=0,$$

where

$$G=\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix},\quad H=\begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix},\quad M=G+\gamma H.$$

It is of the form (2.8) with the coefficients

$$Z_0=I,\quad L_0=-\gamma G-2M,\quad L_{\pm1}=M\pm 2I/3,\quad L_{\pm2}=\mp I/12,$$

where $I$ is the identity matrix. We will use the mapping $\Pi_h$ given by $(\Pi_h f)_{\eta,1}=(\Pi_h f)_{\eta,2}=(\Pi_h f)_{\eta,3}=f(\eta h)$ and the Euclidean norm on $\mathbb{C}^3$.

The matrix $A(\phi)$ for this scheme takes the form

$$\begin{gathered} A(\phi)=i\phi I-L(\phi)= \\ =i\phi I-\left(-\frac{I}{12}e^{2i\phi}+(M+2I/3)e^{i\phi}-\gamma G-2M+(M-2I/3)e^{-i\phi}+\frac{I}{12}e^{-2i\phi}\right)= \\ =i\left(\phi-\frac{4}{3}\sin\phi+\frac{1}{6}\sin 2\phi\right)I+4M\sin^2(\phi/2)+\gamma G= \\ =if(\phi)I+\begin{pmatrix} 0 & \gamma+g & -\gamma-g-\gamma g \\ -\gamma-g & 0 & \gamma+g+\gamma g \\ \gamma+g+\gamma g & -\gamma-g-\gamma g & 0 \end{pmatrix}, \end{gathered}$$

where

$$g=g(\phi)=4\sin^2(\phi/2),\quad f(\phi)=\phi-\frac{4}{3}\sin\phi+\frac{1}{6}\sin 2\phi=O(\phi^5).$$

Since $A^*(\phi)=-A(\phi)$, the scheme is stable with $K=1$. We have

$$v(\phi,\Pi_h)=(\Pi_1 e^{i\phi x})_0=\vec{\mathfrak{e}}=(1,1,1)^T,$$

$$\hat{\epsilon}(\phi,\Pi_h)=A(\phi)v(\phi,\Pi_h)=if(\phi)\begin{pmatrix}1\\1\\1\end{pmatrix}+\gamma g(\phi)\begin{pmatrix}-1\\1\\0\end{pmatrix},$$

so by Lemma 6.10 the optimal value of the order of the truncation error is $P_A = 1$.

Introduce the function

$$w(\phi) = \frac{1}{\gamma + g} \begin{pmatrix} \gamma + g + \gamma g \\ \gamma + g + \gamma g \\ \gamma + g \end{pmatrix} = \vec{\mathfrak{e}} + \frac{\gamma g(\phi)}{\gamma + g(\phi)} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

By construction $\gamma + g(\phi) \geqslant \gamma > 0$. There holds

$$\|w(\phi) - v(\phi, \Pi_h)\| \leqslant \sqrt{2} g(\phi) \leqslant \sqrt{2}\phi^2, \quad \|w(\phi)\| \leqslant \sqrt{3} + \sqrt{2}\phi^2.$$

Since $A(\phi)w(\phi) \equiv if(\phi)w(\phi)$, we have

$$\|A(\phi)w(\phi)\| \leqslant c|\phi|^5(\sqrt{3} + \sqrt{2}\phi^2)$$

960 with $c$ independent of $\gamma$ and $\phi$. By Lemma 8.1 the scheme possesses the 2nd order of accuracy and the 4th order of the long-time simulation, and the constants in the estimate (2.17) are uniform with respect to $\gamma$.

Now we put $\mathcal{P}_h = \Pi_h$ and find the values of $\mathfrak{C}^{(m)}$ providing the 4th order of the truncation error in the sense of $\Pi_h^{(2,4)}$. We have $L(0) = -\gamma G$ and

$$-(\epsilon_1(x^2/2, \Pi_1))_0 = -M\vec{\mathfrak{e}} = -\gamma(-1, 1, 0)^T.$$

By Lemma 10.4 the diagonal matrix $\mathfrak{C}^{(2)}$ should satisfy $-\gamma G\mathfrak{C}^{(2)}\vec{\mathfrak{e}} = -\gamma(-1, 1, 0)^T$, thus we get

$$\mathfrak{C}^{(2)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + c_2 I,$$

where $c_2 \in \mathbb{R}$ is arbitrary.

For any $\mathfrak{C}^{(2)}$ we have

$$\left(\epsilon_1\left(\frac{x^3}{6}, \Pi_1^{(2,2)}\right)\right)_0 = \left(\epsilon_1\left(\frac{x^3}{6}, \Pi_1\right)\right)_0 - Z_0\mathfrak{C}^{(2)}(\Pi_1 1)_0 + \sum_{\eta=-2}^{2} L_\eta \mathfrak{C}^{(2)}(\Pi_1 x)_\eta = \left(-Z_0 + \sum_{\eta=-2}^{2} \eta L_\eta\right)\mathfrak{C}^{(2)}\vec{\mathfrak{e}} = 0$$

and the system $-\gamma G\mathfrak{C}^{(3)}\vec{\mathfrak{e}} = 0$ yields $\mathfrak{C}^{(3)} = c_3 I$, $c_3 \in \mathbb{R}$.

Further,

$$\left(\epsilon_1\left(\frac{x^4}{24}, \Pi_1^{(2,3)}\right)\right)_0 = \left(\epsilon_1\left(\frac{x^4}{24}, \Pi_1\right)\right)_0 - Z_0\mathfrak{C}^{(3)}(\Pi_1 1)_0 + \sum_{\eta=-2}^{2} L_\eta \mathfrak{C}^{(3)}(\Pi_1 x)_\eta + \sum_{\eta=-2}^{2} L_\eta \mathfrak{C}^{(2)}\left(\Pi_1 \frac{x^2}{2}\right)_\eta =$$

$$= \frac{1}{12}M\vec{\mathfrak{e}} + \left(-Z_0 + \sum_{\eta=-2}^{2} \eta L_\eta\right)\mathfrak{C}^{(3)}\vec{\mathfrak{e}} + \left(\sum_{\eta=-2}^{2} \frac{\eta^2}{2} L_\eta\right)\mathfrak{C}^{(2)}\vec{\mathfrak{e}} = \frac{1}{12}M\vec{\mathfrak{e}} + M\mathfrak{C}^{(2)}\vec{\mathfrak{e}} =$$

$$= \frac{1}{12}\gamma \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + (1 + \gamma) \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_2\gamma \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \left(1 + \gamma\left(\frac{13}{12} + c_2\right)\right) \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}.$$

The solution of $-\gamma G\mathfrak{C}^{(4)}\vec{\mathfrak{e}} = -(\epsilon_1(x^4/24, \Pi_1^{(2,3)}))_0$ is

$$\mathfrak{C}^{(4)} = \left(\frac{1}{\gamma} + \frac{13}{12} + c_2\right) \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + c_4 I.$$

Looking on the expressions for $\mathfrak{C}^{(2)}$ and $\mathfrak{C}^{(4)}$ we see that we cannot choose $c_2$, $c_3$, and $c_4$ such that $\mathfrak{C}^{(2)}$ and $\mathfrak{C}^{(4)}$ are 965 bounded uniformly in $\gamma$ simultaneously.

Note that since $A(0) = -L(0) = \gamma G$, for each $\gamma > 0$ the scheme is "simple". For $\gamma = 0$ the scheme possesses the 4th order of accuracy and the 4th order in the long-time simulation, but it is no more "simple": $A(0) = 0$.

## 12. The bad

The crucial difference between the one-dimensional and multidimensional cases can be seen from the proof of Lemma 9.20. Consider a scheme possessing the formal order of accuracy $P$ and the long-time simulation order $Q$. The function $\mathcal{W}^{(P,Q)}(\boldsymbol{\phi})$ given by (8.5) can be componentwise represented as a ratio of the two analytical functions. This ratio is bounded as $\boldsymbol{\phi} \to 0$ (otherwise the scheme does not possess the orders $P$ and $Q$). In 1D case (and this remains true for quasi-1D case) this ratio is holomorphic. Then the Taylor expansion for $\mathcal{W}^{(P,Q)}(\boldsymbol{\phi})$ by Lemma 6.7 can be transformed to the coefficients of a local map $\Pi_h^{(P,Q)}$ that gives the $Q$-th order of the truncation error.

In the multidimensional case the ratio of two holomorphic functions can be bounded but not holomorphic. The example is $f(\phi_x, \phi_y) = \phi_x\phi_y/(\phi_x^2 + \phi_x\phi_y + \phi_y^2)$. Once the function $\mathcal{W}^{(P,Q)}(\boldsymbol{\phi})$ behaves like this, the auxiliary mapping becomes non-local, see Section 12.2. Besides, the behavior of the solution error can become surprising. Particularly, the optimal value of the long-time simulation order can be non-integer, see Section 12.4.

Let $P \in \mathbb{N}$ be the optimal value of the order of accuracy for the scheme (2.8). If one prescribes $\boldsymbol{e} \in \mathring{\Omega}$, this leads to the 1D formulation. By Proposition 8.7 for each $\boldsymbol{e} \in \mathring{\Omega}$ the $Q$-th order of the long-time simulation order on $H^q_{per,\boldsymbol{e}}(\mathbb{R}^d)$ (with $q$ big enough) is equivalent to the existence of $\Pi_{h,\boldsymbol{e}}^{(P,Q)}$ providing the $Q$-th order of the truncation error on the direction $\boldsymbol{e}$. Here we have the following fork of the possible cases.

1. For some $\boldsymbol{e} \in \mathring{\Omega}$ there exist no $\mathfrak{C}_{\boldsymbol{e}}^{(P)}$, …, $\mathfrak{C}_{\boldsymbol{e}}^{(Q)}$ such that the mapping $\Pi_{h,\boldsymbol{e}}^{(P,Q)}$ given by (2.12) provides the $Q$-th order of the truncation error on $\boldsymbol{e}$. Then the scheme does not possess the orders $P$ and $Q$. This may hold even if the resulting system for $\mathfrak{C}_{\boldsymbol{e}}^{(P)}$, …, $\mathfrak{C}_{\boldsymbol{e}}^{(Q)}$ is consistent for almost all $\boldsymbol{e} \in \Omega$, see an example in Section 12.1.
2. For each $\boldsymbol{e} \in \mathring{\Omega}$ there exist $\mathfrak{C}_{\boldsymbol{e}}^{(m)}$, $m = P, \dots, Q$, such that $\Pi_{h,\boldsymbol{e}}^{(P,Q)}$ provides the $Q$-th order of accuracy on $\boldsymbol{e}$ and $\sup\limits_{\boldsymbol{e}\in\mathring{\Omega}} \|\mathfrak{C}_{\boldsymbol{e}}^{(m)}\| < \infty$, $m = P, \dots, Q$. Then by Corollary 6.19 the scheme possesses the formal order of accuracy $P$ and long-time simulation order $Q$. See an example in Section 12.2.
3. For each $\boldsymbol{e} \in \mathring{\Omega}$ there exist $\mathfrak{C}_{\boldsymbol{e}}^{(P)}$, …, $\mathfrak{C}_{\boldsymbol{e}}^{(Q)}$ such that $\Pi_h^{(P,Q)}$ provides the $Q$-th order of accuracy on $\boldsymbol{e}$ but the previous condition does not hold. Let $P, Q'$ be optimal values of the formal order of accuracy and of the long-time simulation order. Then the following cases are possible.
   - $Q' \geqslant Q$. See Section 12.3 for example.
   - $P \leqslant Q' < Q$, and $Q'$ may be not an integer. See Section 12.4 for example.

Throughout this section we consider the Cauchy problem for the transport equation (2.6), (2.7) in $\mathbb{R}^2$ with $\boldsymbol{\omega} = 0$. We put $\boldsymbol{a}_1 = (1, 0)^T$, $\boldsymbol{a}_2 = (0, 1)^T$, so $T = I$. On $\mathbb{C}^{M^0}$ the Euclidean norm will be implied. All the schemes considered in this section are artificial and not intended to represent any scheme used in practice. They have skew-Hermitian matrix $A(\boldsymbol{\phi})$ and thus are stable with $K = 1$. Note that $\boldsymbol{\omega} = 0$ is not a limitation; to each scheme considered in this section for the case of nonzero $\boldsymbol{\omega}$ we can add the central difference approximation of $\boldsymbol{\omega} \cdot \nabla u$ of the order high enough. This will keep $A(\boldsymbol{\phi})$ skew-Hermitian and thus preserve the stability.

### *12.1. Minus-special direction with respect to the long-time simulation accuracy*

The following scheme has the following properties: on $C^3_{per,\boldsymbol{e}}(\mathbb{R}^2)$, is possesses the formal order of accuracy $P = 1$ and the long-time simulation order $Q = 1$ if $\boldsymbol{e}$ is aligned with the vertical axis, and $P = 1$, $Q = 2$ otherwise. I. e. the long-time simulation order degrades if the wave vector is aligned with the vertical axis.

Put $M^0 = \{L, R\}$. Consider the scheme

$$h\frac{du_{j,k}}{dt} + E(u_{j+1,k} - u_{j-1,k}) + W(u_{j,k+1} - 2u_{j,k} + u_{j,k-1}) = 0,$$

$$E = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}, \quad W = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

It is of the form (2.8) with the coefficients $Z_{0,0} = I$,

$$L_{0,0} = -2W, \quad L_{\pm1,0} = \pm E, \quad L_{0,\pm1} = W,$$

Unspecified coefficients are zero. The mapping $\Pi_h$ is defined by $(\Pi_h f)_{\boldsymbol{\eta},L} = (\Pi_h f)_{\boldsymbol{\eta},R} = f(\eta_1 h, \eta_2 h)$, so $v(\boldsymbol{\phi}, \Pi_h) = \vec{\mathfrak{e}} = (1,1)^T$. The matrix $A(\boldsymbol{\phi}) = -L(\boldsymbol{\phi})$ is

$$A(\phi_1, \phi_2) = \begin{pmatrix} i\sin\phi_1 & -i\sin\phi_1 - 4\sin^2(\phi_2/2) \\ -i\sin\phi_1 + 4\sin^2(\phi_2/2) & i\sin\phi_1 \end{pmatrix}. \tag{12.1}$$

We have $A(0) = 0$ and $A(\boldsymbol{\phi})\vec{\mathfrak{e}} = 4\sin^2(\phi_2/2)(-1,1)^T$, thus the optimal values of the truncation error and of the formal order of accuracy are $P_A = P = 1$.

We look for the auxiliary mapping $\Pi^{(1,2)}_{h,\boldsymbol{e}}$ of the form (2.13):

$$\Pi^{(1,2)}_{h,\boldsymbol{e}} f = \Pi_h f + h\mathfrak{C}^{(1)}_{\boldsymbol{e}} \Pi_h \frac{\partial f}{\partial \boldsymbol{e}} + h^2 \mathfrak{C}^{(2)}_{\boldsymbol{e}} \Pi_h \frac{\partial^2 f}{\partial \boldsymbol{e}^2}$$

that provides the second order of the truncation error on planar waves aligned with $\boldsymbol{e}$. Since $L(0) = 0$, the truncation error on quadratic polynomials does not depend on $\mathfrak{C}^{(2)}_{\boldsymbol{e}}$, so it is enough to consider mappings of the form

$$\Pi^{(1,1)}_{h,\boldsymbol{e}} f = \Pi_h f + h\mathfrak{C}_{\boldsymbol{e}} \Pi_h \frac{\partial f}{\partial \boldsymbol{e}}. \tag{12.2}$$

On the linear function we have $(\epsilon_1(\boldsymbol{e}\cdot\boldsymbol{r}, \Pi^{(1,1)}_{1,\boldsymbol{e}}))_0 = 0$ for each $\mathfrak{C}_{\boldsymbol{e}}$ because $L(0) = 0$. For the quadratic function the truncation error is

$$\left(\epsilon_1\left(\frac{(\boldsymbol{e}\cdot\boldsymbol{r})^2}{2}, \Pi^{(1,1)}_{1,\boldsymbol{e}}\right)\right)_0 = \sum_{\boldsymbol{\eta}} L_{\boldsymbol{\eta}} \left(\frac{(\boldsymbol{e}\cdot\boldsymbol{\eta})^2}{2}\vec{\mathfrak{e}} + (\boldsymbol{e}\cdot\boldsymbol{\eta})\mathfrak{C}_{\boldsymbol{e}}\vec{\mathfrak{e}}\right).$$

Equating to zero and substituting the values of $L_{\boldsymbol{\eta}}$ we get $e_y^2 W\vec{\mathfrak{e}} + 2e_x E\mathfrak{C}_{\boldsymbol{e}}\vec{\mathfrak{e}} = 0$. Substituting the values for $W$ and $E$ we get

$$e_x \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} (\mathfrak{C}_{\boldsymbol{e}}\vec{\mathfrak{e}}) + e_y^2 \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 0.$$

For $e_x \neq 0$ the system is consistent and has the solution

$$\mathfrak{C}_{\boldsymbol{e}} = -\frac{e_y^2}{e_x}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + c_{\boldsymbol{e}} I, \quad c_{\boldsymbol{e}} \in \mathbb{R},$$

while for $e_x = 0$ the system is inconsistent. Thus for each $\boldsymbol{e} \in \mathring{\Omega}$ with $e_x \neq 0$ the scheme possesses the first order of accuracy and the second order in the long-time simulation $C^3_{per,\boldsymbol{e}}(\mathbb{R}^2)$. The following estimate is valid:

$$\|\varepsilon_h(t, e^{i\boldsymbol{\alpha}\cdot\boldsymbol{r}}, \Pi_h)\| \leqslant \min\left\{2\frac{\alpha_y^2}{|\alpha_x|}h + C_2|\boldsymbol{\alpha}|^3 h^2 t, C_3|\boldsymbol{\alpha}|^2 ht\right\}.$$

Let us write the explicit expression for the solution error. Denote $a = \sin\phi_1$, $b = 4\sin^2(\phi_2/2)$, $v = a + bi$. Then (12.1) rewrites as

$$A(\boldsymbol{\phi}) = \begin{pmatrix} ia & -ia - b \\ -ia + b & ia \end{pmatrix} = S\Lambda S^{-1},$$

where

$$S = \frac{1}{2}\begin{pmatrix} -1 & 1 \\ v/|v| & v/|v| \end{pmatrix}, \quad \Lambda = \begin{pmatrix} ia + i|v| & 0 \\ 0 & ia - i|v| \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} -1 & \bar{v}/|v| \\ 1 & \bar{v}/|v| \end{pmatrix}.$$

Then

$$\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h) = \left(e^{\nu A(\boldsymbol{\phi})} - I\right)\vec{\mathfrak{e}} = S\begin{pmatrix} (\exp(i\nu(a + |v|)) - 1)(\bar{v}/|v| - 1) \\ (\exp(i\nu(a - |v|)) - 1)(\bar{v}/|v| + 1) \end{pmatrix}.$$

Since $A(0) = 0$, the function $\hat{\varepsilon}(\boldsymbol{\alpha}h, t/h, \Pi_h)$ is holomorphic in $h$ (see Section 9.1). Thus it can be expressed using the Taylor series in $h$. Omitting the calculations, we get

$$S^{-1}(\boldsymbol{\alpha}h)\hat{\varepsilon}(\boldsymbol{\alpha}h,t/h,\Pi_h)=i\frac{\alpha_y^2}{4\alpha_x}(1-\exp(2it\alpha_x))h\begin{pmatrix}1\\0\end{pmatrix}+O(h^2),\quad \alpha_x>0;$$
$$S^{-1}(\boldsymbol{\alpha}h)\hat{\varepsilon}(\boldsymbol{\alpha}h,t/h,\Pi_h)=i\frac{\alpha_y^2}{4\alpha_x}(1-\exp(2it\alpha_x))h\begin{pmatrix}0\\-1\end{pmatrix}+O(h^2),\quad \alpha_x<0.$$

Thus for each $\alpha_x\neq 0$ the term with $h$ is a bounded function of the time. In contrast, for $\alpha_x=0$ we have

$$\exp(\nu A(\boldsymbol{\phi}))=\begin{pmatrix}\cos(\nu b) & -\sin(\nu b)\\ \sin(\nu b) & \cos(\nu b)\end{pmatrix},\quad \hat{\varepsilon}(\boldsymbol{\alpha}h,t/h,\Pi_h)=\begin{pmatrix}\cos(\nu b)-1-\sin(\nu b)\\ \cos(\nu b)-1+\sin(\nu b)\end{pmatrix},$$

where $\nu b=4t\sin^2(\alpha_y h/2)/h$. It is clear that there is no estimate of the form $O(h+h^2t)$ for $\hat{\varepsilon}(\boldsymbol{\alpha}h,t/h,\Pi_h)$.

*12.2. The case of nonlocal auxiliary mapping*

In this section we present a scheme that possesses the first order of accuracy and the second order in the long-time simulation, however, there is no local mapping of the form (2.12) that gives the second order of the truncation error.

Consider the scheme with three degrees of freedom per cell given by

$$h\frac{du_{j,k}}{dt}+E(u_{j+1,k}-u_{j-1,k})+F(u_{j,k+1}-u_{j,k-1})+W(u_{j+1,k+1}-u_{j+1,k-1}-u_{j-1,k+1}+u_{j-1,k-1})=0,$$

$$E=\begin{pmatrix}1&0&-1\\0&-1&1\\-1&1&0\end{pmatrix},\ F=\begin{pmatrix}1&-1&0\\-1&0&1\\0&1&-1\end{pmatrix},\ W=\begin{pmatrix}0&1&0\\-1&0&1\\0&-1&0\end{pmatrix}$$

and let $\Pi_h$ be given by $(\Pi_h f)_{\boldsymbol{\eta},\xi}=f(\eta_1 h,\eta_2 h)$ for each $\xi$. This scheme is of the form (2.8) with $Z_{(0,0)}=I$,

$$\begin{gathered}L_{(-1,-1)}=W,\quad L_{(0,-1)}=-F,\quad L_{(1,-1)}=-W,\\ L_{(-1,0)}=-E,\quad L_{(0,0)}=0,\quad L_{(1,0)}=E,\\ L_{(-1,1)}=-W,\quad L_{(0,1)}=F,\quad L_{(1,1)}=W,\end{gathered}$$

(all unspecified coefficients are zero).

We start with the analysis of the truncation error. We have $(\epsilon_1(f,\Pi_1))_0=(L\Pi_1 f)_0=\sum_{\boldsymbol{\eta}}L_{\boldsymbol{\eta}}(\Pi_1 f)_{\boldsymbol{\eta}}$ and $(\Pi_1\boldsymbol{r}^{\boldsymbol{m}})_{\boldsymbol{\eta}}=\boldsymbol{\eta}^{\boldsymbol{m}}\vec{\mathfrak{e}}$, where $\vec{\mathfrak{e}}=(1,1,1)^T$,

$$\begin{gathered}\sum_{\boldsymbol{\eta}}L_{\boldsymbol{\eta}}=0,\quad \sum_{\boldsymbol{\eta}}\eta_1 L_{\boldsymbol{\eta}}=2E,\quad \sum_{\boldsymbol{\eta}}\eta_2 L_{\boldsymbol{\eta}}=2F,\\ \sum_{\boldsymbol{\eta}}\eta_1^2 L_{\boldsymbol{\eta}}=0,\quad \sum_{\boldsymbol{\eta}}\eta_1\eta_2 L_{\boldsymbol{\eta}}=4W,\quad \sum_{\boldsymbol{\eta}}\eta_2^2 L_{\boldsymbol{\eta}}=0.\end{gathered}$$

Since $E\vec{\mathfrak{e}}=F\vec{\mathfrak{e}}=0$, the scheme possesses the first order of the truncation error; since $W\vec{\mathfrak{e}}\neq 0$, it does not possess the second order of the truncation error. Since $L(0)=0$, by Theorem 1 the optimal order of accuracy is equal to $P=1$.

Now we consider this scheme on $C^3_{per,\boldsymbol{e}}(\mathbb{R}^2)$ for some $\boldsymbol{e}\in\mathring{\Omega}$. We will look for the auxiliary mapping $\Pi^{(1,1)}_{h,\boldsymbol{e}}$ given by (12.2) providing the second order of the truncation error on $\boldsymbol{e}$. Since the scheme possesses the first order of the truncation error, we need to consider the function $f(\boldsymbol{r})=(\boldsymbol{e}\cdot\boldsymbol{r})^2/2$ only. Write

$$\left(\epsilon_1(f,\Pi^{(1,1)}_{1,\boldsymbol{e}})\right)_0=\sum_{\boldsymbol{\eta}}L_{\boldsymbol{\eta}}\left(\frac{(\boldsymbol{e}\cdot\boldsymbol{\eta})^2}{2}\vec{\mathfrak{e}}+(\boldsymbol{e}\cdot\boldsymbol{\eta})\mathfrak{C}_{\boldsymbol{e}}\vec{\mathfrak{e}}\right)=4e_xe_yW\vec{\mathfrak{e}}+2(e_xE+e_yF)\mathfrak{C}_{\boldsymbol{e}}\vec{\mathfrak{e}}.$$

Substituting the expressions for $E$, $F$, and $W$ and equating to zero we get

$$\begin{pmatrix}e_x+e_y & -e_y & -e_x\\ -e_y & -e_x & e_x+e_y\\ -e_x & e_x+e_y & -e_y\end{pmatrix}\mathfrak{C}_{\boldsymbol{e}}\vec{\mathfrak{e}}=-2e_xe_y\begin{pmatrix}1\\0\\-1\end{pmatrix}.$$

The solution of this system is

$$\mathfrak{C}_{\boldsymbol{e}} = \frac{2e_xe_y}{e_x^2+e_xe_y+e_y^2}\begin{pmatrix} -e_x & 0 & 0 \\ 0 & e_y & 0 \\ 0 & 0 & 0 \end{pmatrix} + c_{\boldsymbol{e}}I.$$

We can put $c_{\boldsymbol{e}} = 0$. Since $\mathfrak{C}_{\boldsymbol{e}}$ is bounded in $\boldsymbol{e} \in \mathring{\Omega}$, by Corollary 6.19 the scheme possesses the first order of accuracy and the second order in the long-time simulation on $C^3_{per}(\mathbb{R}^2)$.

Now show that there exists no mapping $\Pi_h^{(1,2)}$ of the form (2.12) providing the second order of the truncation error. Indeed, since $L(0) = 0$, the truncation error in the sense of $\Pi_h^{(1,2)}$ on quadratic polynomials does not depend on $\mathfrak{C}^{(\boldsymbol{m})}$ for each $\boldsymbol{m}$ such that $|\boldsymbol{m}| = 2$. Thus we need to check mappings $\Pi_h^{(1,1)}$ of the form

$$\left(\Pi_h^{(1,1)}f\right)_{\boldsymbol{\eta}} = (\Pi_h f)_{\boldsymbol{\eta}} + h\mathfrak{C}^{(x)}\left(\Pi_h\frac{\partial f}{\partial x}\right)_{\boldsymbol{\eta}} + h\mathfrak{C}^{(y)}\left(\Pi_h\frac{\partial f}{\partial y}\right)_{\boldsymbol{\eta}}.$$

Considering $f = x^2/2$ we get $\mathfrak{C}^{(x)} \equiv \mathfrak{C}_{\boldsymbol{e}}$ with $\boldsymbol{e} = (1,0)^T$. Considering $f = y^2/2$ we get $\mathfrak{C}^{(y)} \equiv \mathfrak{C}_{\boldsymbol{e}}$ with $\boldsymbol{e} = (0,1)^T$. Thus we get $\mathfrak{C}^{(x)} = \kappa_x I$, $\mathfrak{C}^{(y)} = \kappa_y I$, $\kappa_x, \kappa_y \in \mathbb{R}$. It remains to see that the truncation error on $f = xy$ is equal to $4Wh$ whatever $\kappa_x$ and $\kappa_y$. Thus the mapping $\Pi_h^{(1,2)}$ of the form (2.12) providing the second order of the truncation error does not exist.

*12.3. Fake special direction*

Let a scheme of the form (2.8) possess the formal order of accuracy $P \in \mathbb{N}$ and the long-time simulation order $Q \in \mathbb{N}$, $Q > P$. By Lemma 8.7 for each $\boldsymbol{e} \in \mathring{\Omega}$ there exist coefficients $\mathfrak{C}_{\boldsymbol{e}}^{(m)}$, $m = P, \dots, Q$ such that the scheme possesses the truncation error of order $Q$ in the sense of $\Pi_{h,\boldsymbol{e}}^{(P,Q)}$ given by (2.13) on $\boldsymbol{e}$. But it is generally impossible to define these coefficients such that they satisfy $\max\limits_{m=P,\dots,Q}\sup\limits_{\boldsymbol{e}\in\mathring{\Omega}}\|\mathfrak{C}_{\boldsymbol{e}}^{(m)}\| < \infty$. To show this, we adapt the example of Section 11.4.

Consider the scheme

$$h\frac{du_{j,k}}{dt} + \sum_{\eta_1=-2}^{2}\sum_{\eta_2=-2}^{2} L_{\eta_1,\eta_2}u_{j+\eta_1,k+\eta_2} = 0$$

with

$$L_{\eta_1,\eta_2} = \left(-c^{(2)}_{\eta_2} + c^{(4)}_{\eta_1} + c^{(4)}_{\eta_2}\right)W + \left(c^{(2)}_{\eta_1}c^{(2)}_{\eta_2} + c^{(4)}_{\eta_2}\right)G,$$

$$W = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix},$$

and $c_\eta^{(m)}$ be the coefficients of $(5-m)$-th order finite difference approximations of the $m$-th derivative on the uniform mesh with the unit step:

$$c_0^{(2)} = -5/2; \quad c_{\pm1}^{(2)} = 4/3; \quad c_{\pm2}^{(2)} = -1/12;$$

$$c_0^{(4)} = 6; \quad c_{\pm1}^{(4)} = -4; \quad c_{\pm2}^{(4)} = 1.$$

It is of the form (2.8) with $Z_0 = I$, $Z_{\boldsymbol{\eta}} = 0$ for $\boldsymbol{\eta} \neq 0$ and $L_{\boldsymbol{\eta}}$ given above. Define $\Pi_h$ is by $(\Pi_h f)_{\boldsymbol{\eta},\xi} = f(\eta_1 h, \eta_2 h)$ for each $\xi$, so $v(\boldsymbol{\phi}, \Pi_h) = \vec{\mathfrak{e}} = (1,1,1)^T$. The matrix $L(\boldsymbol{\phi})$ is

$$L(\boldsymbol{\phi}) = (F^{(2)}(\phi_2) + F^{(4)}(\phi_1) + F^{(4)}(\phi_2))W + (F^{(2)}(\phi_1)F^{(2)}(\phi_2) + F^{(4)}(\phi_2))G,$$

where

$$F^{(m)}(\psi) = (-i)^m\sum_\eta c_\eta^{(m)}\exp(i\eta\psi) = \psi^m + O(|\psi|^{m+1}). \tag{12.3}$$

Then

$$A(\boldsymbol{\phi}) = -L(\boldsymbol{\phi}) = \begin{pmatrix} 0 & -x & x+y \\ x & 0 & -x-y \\ -x-y & x+y & 0 \end{pmatrix},$$

where

$$x \equiv x(\boldsymbol{\phi}) = F^{(2)}(\phi_2) + F^{(4)}(\phi_2) + F^{(4)}(\phi_1), \quad y \equiv y(\boldsymbol{\phi}) = F^{(2)}(\phi_1)F^{(2)}(\phi_2) + F^{(4)}(\phi_2). \tag{12.4}$$

For $\boldsymbol{\phi}$ in a neighborhood of $\boldsymbol{\phi} = 0$ introduce the vector

$$w(\boldsymbol{\phi}) = \frac{1}{x(\boldsymbol{\phi})} \begin{pmatrix} x(\boldsymbol{\phi}) + y(\boldsymbol{\phi}) \\ x(\boldsymbol{\phi}) + y(\boldsymbol{\phi}) \\ x(\boldsymbol{\phi}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \frac{y(\boldsymbol{\phi})}{x(\boldsymbol{\phi})} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\phi} \neq 0; \quad w(0) = 0.$$

Obviously, $A(\boldsymbol{\phi})w(\boldsymbol{\phi}) \equiv 0$. Using (12.3) and $v(\boldsymbol{\phi}, \Pi_h) = (1,1,1)^T$ in a neighborhood of $\boldsymbol{\phi} = 0$ we get

$$\|w(\boldsymbol{\phi}) - v(\boldsymbol{\phi}, \Pi_h)\| = \left|\frac{y(\boldsymbol{\phi})}{x(\boldsymbol{\phi})}\right| \left\| \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\| = \sqrt{2} \left| \frac{F^{(2)}(\phi_1)F^{(2)}(\phi_2) + F^{(4)}(\phi_2)}{F^{(2)}(\phi_2) + F^{(4)}(\phi_2) + F^{(4)}(\phi_1)} \right| \leqslant 2|\boldsymbol{\phi}|^2.$$

By Lemma 8.1 the scheme possesses the second order of accuracy and the infinite order in the long-time simulation. These values are not optimal ($\|\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h)\| \sim \sqrt{2}|y(\boldsymbol{\phi})| = O(|\boldsymbol{\phi}|^4)$, so the optimal value of the solution error is equal to 3) but the scheme may be easily modified by adding a disjoint spurious component that will decrease the formal order of accuracy.

**Proposition 12.1.** *Let $\{\mathfrak{C}_{\boldsymbol{e}}^{(m)}, m = 2, \dots, 9, \boldsymbol{e} \in \mathring{\Omega}\}$ be a set of diagonal matrices such that for each $\boldsymbol{e} \in \mathring{\Omega}$ the scheme possesses the truncation error of order 9 on the direction $\boldsymbol{e}$ in the sense of $\Pi_{h,\boldsymbol{e}}^{(2,9)}$ given by* (2.13)*. Then*

$$\max_{m=2,\dots,9} \sup_{\boldsymbol{e} \in \mathring{\Omega}} \|\mathfrak{C}_{\boldsymbol{e}}^{(m)}\| = \infty.$$

*Proof.* Assume the converse, then for each $\boldsymbol{e} = (e_1, e_2) \in \mathring{\Omega}$ for the function

$$w_{\boldsymbol{e}}(\psi) = v(\psi \boldsymbol{e}, \Pi_{h,\boldsymbol{e}}^{(2,9)}) = \left( I + \sum_{m=2}^{9} \psi^m \mathfrak{C}_{\boldsymbol{e}}^{(m)} \right) \vec{\mathfrak{e}}$$

in a neighborhood of $\psi = 0$ there holds

$$\|A(\boldsymbol{e}\psi) w_{\boldsymbol{e}}(\psi)\| \leqslant c_{\boldsymbol{e}} |\psi|^{10}, \quad \|v(\boldsymbol{e}\psi, \Pi_h) - w_{\boldsymbol{e}}(\psi)\| \leqslant c_{\boldsymbol{e}} |\psi|^2, \tag{12.5}$$

$$\max_{m=2,\dots,9} \sup_{\boldsymbol{e} \in \mathring{\Omega}} \left\| \frac{d^m w_{\boldsymbol{e}}}{d\psi^m}(0) \right\| < \infty. \tag{12.6}$$

The eigenvalues of $A(\boldsymbol{\phi})$ are 0 and $\pm i(x^2 + 2(x+y)^2)^{1/2}$ where $x$ and $y$ are given by (12.4). The last two eigenvalues satisfy $|\lambda| \geqslant |\boldsymbol{\phi}|^4/2$ in a neighborhood of $\boldsymbol{\phi} = 0$. Thus the inequality $\|A(\boldsymbol{e}\psi) w_{\boldsymbol{e}}(\psi)\| = O(|\psi|^{10})$ is possible only if there exists $\gamma_{\boldsymbol{e}}(\psi) > 0$ such that $w_{\boldsymbol{e}}(\psi) = \gamma_{\boldsymbol{e}}(\psi) w(\boldsymbol{e}\psi) + O(|\psi|^6)$. Taking $\psi = 0$ we get $\gamma_{\boldsymbol{e}}(0) = 1$. This means that for $\boldsymbol{\phi} = \psi \boldsymbol{e}$ we have

$$w_{\boldsymbol{e}}(\psi) = \gamma_{\boldsymbol{e}}(\psi) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \frac{\gamma_{\boldsymbol{e}}(\psi) y(\boldsymbol{e}\psi)}{x(\boldsymbol{e}\psi)} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + O(|\psi|^6). \tag{12.7}$$

By assumption $d^n w_{\boldsymbol{e}}/d\psi^n(0)$ is bounded in $\boldsymbol{e}$ for $n = 0, \dots, 4$. Taking the last component of (12.7) we see that this also holds for the derivatives of $\gamma_{\boldsymbol{e}}(\psi)$; thus this holds for the derivatives of $\gamma_{\boldsymbol{e}}(\psi) y(\boldsymbol{e}\psi)/x(\boldsymbol{e}\psi)$ and $y(\boldsymbol{e}\psi)/x(\boldsymbol{e}\psi)$. But the Taylor expansion

$$\frac{y(\psi e_1, \psi e_2)}{x(\psi e_1, \psi e_2)} \approx \frac{\psi^4 e_2^2}{\psi^2 e_2^2 + \psi^4(1 - 2e_1^2 e_2^2)} = \psi^2 \frac{1}{1 + \psi^2(e_2^{-2} - 2e_1^2)} = \psi^2 - \psi^4(e_2^{-2} - 2e_1^2) + O(\psi^6)$$

shows that the 4-th derivative by $\psi$ of this expression is of order $e_2^{-2}$. This contradiction proves the proposition. □

Note that there exist no mapping $\Pi_h^{(2,9)}$ that gives the 9-th order of the truncation error. Assuming the converse we get that $v(\boldsymbol{\phi}, \Pi_h^{(2,9)})$ is a holomorphic function satisfying $\|v(\boldsymbol{\phi}, \Pi_h^{(2,9)}) - v(\boldsymbol{\phi}, \Pi_h)\| = O(|\boldsymbol{\phi}|^2)$ and $A(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h^{(2,9)}) = O(|\boldsymbol{\phi}|^{10})$. Thus $w_{\boldsymbol{e}}(\psi) = v(\boldsymbol{e}\psi, \Pi_h^{(2,9)})$ satisfies (12.5) and (12.6), which leads to the same contradiction as above.

*12.4. Strange special direction*

In this section we consider a family of schemes with the following properties. The optimal value of the order of accuracy is equal to $P = 1$. For some $q \in \mathbb{N}$ and each $\boldsymbol{e} \in \mathring{\Omega}$ there exist diagonal matrices $\mathfrak{C}_{\boldsymbol{e}}^{(m)}$, $m = 1, \dots, q$, such that the scheme possesses the truncation error of order $q$ in the sense of $\Pi_{h,\boldsymbol{e}}^{(1,q)}$ on the direction $\boldsymbol{e}$. However, the scheme does not possess the formal order $P$ and the long-time simulation order $q$. For instance, the optimal value $Q$ of the long-time simulation order may be equal to $q - 1$ or $q - 1/2$.

Prior to construct the schemes we prove the following lemma. Denote $\mathbb{R}^d_+ = \{(x_1, \dots, x_d) \,:\, x_1, \dots, x_d \geqslant 0\}$ and for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d_+$ put $\boldsymbol{x}^{\boldsymbol{y}} = x_1^{y_1} \cdot \ldots \cdot x_d^{y_d}$.

**Lemma 12.2.** *Let $\boldsymbol{m}, \boldsymbol{n}_j \in \mathbb{R}^d_+$, $j = 1, \dots, N$. The following statements are equivalent.*

1. *The ratio of $\boldsymbol{\phi}^{\boldsymbol{m}}$ and $\sum\limits_j \boldsymbol{\phi}^{\boldsymbol{n}_j}$ is bounded on $[0,1]^d$;*
2. *There exist $\delta_1, \dots, \delta_N \geqslant 0$ with $\sum \delta_j = 1$ such that $\sum \delta_j \boldsymbol{n}_j \leqslant \boldsymbol{m}$.*

*Proof.* Assume the second statement. Then for each $\boldsymbol{\phi} \in [0,1]^d$ there holds

$$\boldsymbol{\phi}^{\boldsymbol{m}} \leqslant \boldsymbol{\phi}^{\sum \delta_j \boldsymbol{n}_j} = \prod_{j=1}^{N} (\boldsymbol{\phi}^{\boldsymbol{n}_j})^{\delta_j} \leqslant \sum_{j=1}^{N} \boldsymbol{\phi}^{\boldsymbol{n}_j}.$$

The last inequality is due to the Young inequality. Thus we get the first statement.

Assume the first statement. Let $\boldsymbol{q} \in \mathbb{R}^d_+$. Put $\phi_1 = \varepsilon^{q_1}$, …, $\phi_d = \varepsilon^{q_d}$ and send $\varepsilon$ to zero. Since $\boldsymbol{\phi}^{\boldsymbol{m}} = \varepsilon^{\boldsymbol{q}\cdot\boldsymbol{m}}$ and $\sum\limits_j \boldsymbol{\phi}^{\boldsymbol{n}_j} = \sum\limits_j \varepsilon^{\boldsymbol{q}\cdot\boldsymbol{n}_j}$ we get

$$\boldsymbol{q} \cdot \boldsymbol{m} \geqslant \min_j \{\boldsymbol{q} \cdot \boldsymbol{n}_j\}. \tag{12.8}$$

Assume that the second statement does not hold. Let $M = \operatorname{Conv}\{\boldsymbol{n}_j\} + \mathbb{R}^d_+$ where the plus sign means the Minkowski addition. Obviously, $M$ is convex and $\boldsymbol{m} \notin M$. By the hyperplane separation theorem there exists $\boldsymbol{q} \in \mathbb{R}^d$ such that there holds $\boldsymbol{q} \cdot \boldsymbol{a} > C > \boldsymbol{q} \cdot \boldsymbol{m}$ for each $\boldsymbol{a} \in M$. Therefore

$$\min_j \{\boldsymbol{q} \cdot \boldsymbol{n}_j\} > C > \boldsymbol{q} \cdot \boldsymbol{m}. \tag{12.9}$$

It remains to see that $\boldsymbol{q} \in \mathbb{R}^d_+$. Assume the converse, i. e. $q_k < 0$ for some $k$. Let $\boldsymbol{a} \in M$. Consider the sequence

$$\boldsymbol{a}_l = \boldsymbol{a} + l(0, \dots, 1, \dots, 0)^T,$$

where unit stays in the $k$-th position. This vector belongs to $M$ by construction, but $\boldsymbol{q} \cdot \boldsymbol{a}_l \to -\infty$ and thus can't be greater than $C$ for each $l$. This contradiction proves that $\boldsymbol{q} \in \mathbb{R}^d_+$. So (12.9) contradicts (12.8), which proves the lemma. □

Now we construct the schemes. Put $M^0 = \{1, 2, 3, 4\}$ and consider the schemes of the form (2.8) with the coefficients $Z_0 = I$,

$$L_{\boldsymbol{\eta}} = c^{(\alpha)}_{\eta_1} c^{(\beta)}_{\eta_2} W + c^{(\gamma)}_{\eta_1} c^{(\delta)}_{\eta_2} E + c^{(1)}_{\eta_1} R,$$

where $\alpha$, $\beta$, $\gamma$ are odd natural numbers and $\delta$ is an even natural number,

$$W = \begin{pmatrix} 0 & 1 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad E = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and $c_\eta^{(m)}$ are the coefficients of high-order finite difference approximations for the $m$-th derivative on the uniform mesh with the unit step such that $c_{-\eta}^{(m)} = (-1)^m c_\eta^{(m)}$. The order of the approximation should be not less than $\max\{\alpha,\beta,\gamma,\delta\}+1-m$. For example, for $\alpha=5$, $\beta=\gamma=1$, $\delta=8$ we can put

$$c_{\pm1}^{(1)} = \pm 4/5; \quad c_{\pm2}^{(1)} = \mp 1/5; \quad c_{\pm3}^{(1)} = \pm 4/105; \quad c_{\pm4}^{(1)} = \mp 1/280;$$

$$c_{\pm1}^{(5)} = \pm 29/6; \quad c_{\pm2}^{(5)} = \mp 13/3; \quad c_{\pm3}^{(5)} = \pm 3/2; \quad c_{\pm4}^{(5)} = \mp 1/6;$$

$$c_0^{(8)} = 70; \quad c_{\pm1}^{(8)} = -56; \quad c_{\pm2}^{(8)} = 28; \quad c_{\pm3}^{(8)} = -8; \quad c_{\pm4}^{(8)} = 1$$

(unspecified coefficients are zeros). The mapping $\Pi_h$ is defined by

$$(\Pi_h f)_{\boldsymbol{\eta},\xi} = f(\boldsymbol{\eta}h), \quad \xi = 1,\dots,3; \quad (\Pi_h f)_{\boldsymbol{\eta},4} = h\frac{\partial f}{\partial x}(\boldsymbol{\eta}h),$$

so $v(\boldsymbol{\phi},\Pi_h) = (1,1,1,i\phi_x)^T$. The matrix $A(\boldsymbol{\phi}) = -L(\boldsymbol{\phi})$ is

$$A(\boldsymbol{\phi}) = f(\boldsymbol{\phi})W - ig(\boldsymbol{\phi})E - iF^{(1)}(\phi_x)R, \tag{12.10}$$

where

$$f(\boldsymbol{\phi}) = i^{\alpha+\beta+2}F^{(\alpha)}(\phi_x)F^{(\beta)}(\phi_y), \quad g(\boldsymbol{\phi}) = i^{\gamma+\delta-1}F^{(\gamma)}(\phi_x)F^{(\delta)}(\phi_y), \tag{12.11}$$

and

$$F^{(m)}(\psi) = (-i)^m \sum_\eta c_\eta^{(m)} \exp(i\eta\psi) = \psi^m + O(|\psi|^{\max\{\alpha,\beta,\gamma,\delta\}+1}).$$

Since $(A(\boldsymbol{\phi}))^* = -(A(\boldsymbol{\phi}))$ the scheme is stable. Note that $i^{\alpha+\beta+2} = \pm 1$ and $i^{\gamma+\delta-1} = \pm 1$. By construction $F^{(m)}(\psi)$, $f(\boldsymbol{\phi})$, and $g(\boldsymbol{\phi})$ are real-valued for real-valued $\psi$ and $\boldsymbol{\phi}$.

Determine the order of the truncation error first. We have

$$\hat\epsilon(\boldsymbol{\phi},\Pi_h) = A(\boldsymbol{\phi})v(\boldsymbol{\phi},\Pi_h) = (-ig(\boldsymbol{\phi}),0,ig(\boldsymbol{\phi}),\phi_x F^{(1)}(\phi_x))^T.$$

By assumption $\gamma+\delta \geqslant 3$, then $\hat\epsilon(\boldsymbol{\phi},\Pi_h) \leqslant c|\boldsymbol{\phi}|^2$ and the scheme possesses the truncation error of the order $P_A = 1$. Since the last component is of the second order as $\phi_x \to 0$, the value $P_A = 1$ is optimal. Since $A(0) = 0$, by
1055 Theorem 1 the optimal value of the order of accuracy is $P = 1$.

The optimal value of the long-time simulation order will be obtained with the use of Theorem 8.5. Since $A^*(\boldsymbol{\phi}) = -A(\boldsymbol{\phi})$, the functional $\mathcal{F}(\boldsymbol{\phi})$ defined by (8.9) can be rewritten as

$$\mathcal{F}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2P}\,(\hat\epsilon(\boldsymbol{\phi},\Pi_h))^*\left(A^*(\boldsymbol{\phi})A(\boldsymbol{\phi}) + |\boldsymbol{\phi}|^{2(Q+1-P)}\right)^{-1}\hat\epsilon(\boldsymbol{\phi},\Pi_h).$$

Denote by $\mathcal{A}(\phi_x,\phi_y)$ the restriction of $A(\phi_x,\phi_y)$ to the first three components. Taking into account that $P=1$ we get

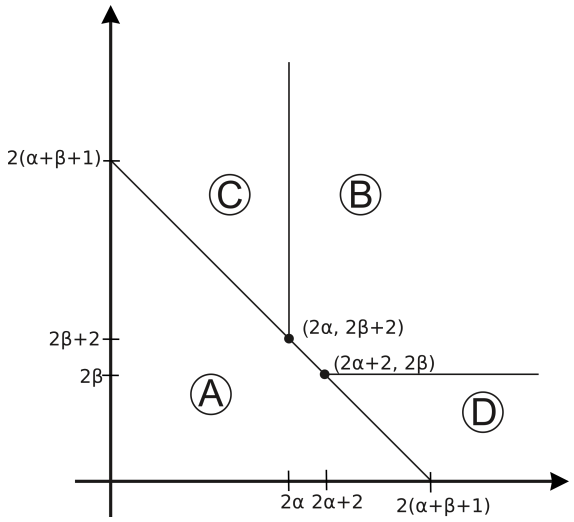
$$\mathcal{F}(\boldsymbol{\phi}) = |\boldsymbol{\phi}|^{-2}(g(\boldsymbol{\phi}))^2\,(1,0,-1)\left(\mathcal{A}^*(\boldsymbol{\phi})\mathcal{A}(\boldsymbol{\phi}) + |\boldsymbol{\phi}|^{2Q}\right)^{-1}\begin{pmatrix}1\\0\\-1\end{pmatrix} + \frac{|\phi_x|^2}{|\boldsymbol{\phi}|^2}\frac{|F^{(1)}(\phi_x)|^2}{|F^{(1)}(\phi_x)|^2+|\boldsymbol{\phi}|^{2Q}}.$$

The last term (corresponding to the last component of the solution error) is bounded for each $Q$. Now write

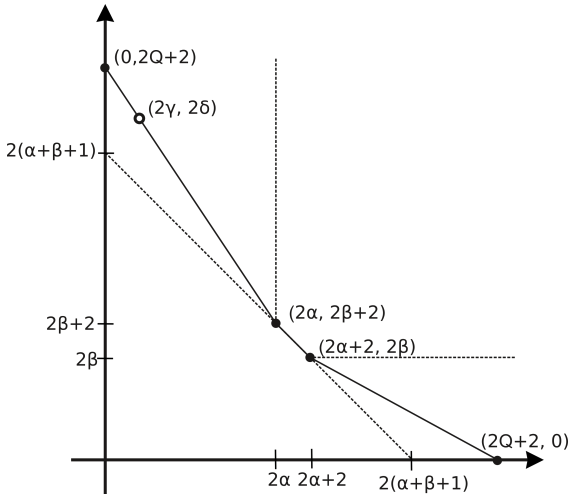
$$\mathcal{A} \equiv \mathcal{A}(\boldsymbol{\phi}) = \begin{pmatrix} -ig & f & -f \\ -f & 0 & f \\ f & -f & ig \end{pmatrix}, \quad \mathcal{A}^*\mathcal{A} = -\mathcal{A}^2 = \begin{pmatrix} g^2+2f^2 & ifg-f^2 & -f^2 \\ -igf-f^2 & 2f^2 & -igf-f^2 \\ -f^2 & igf-f^2 & g^2+2f^2 \end{pmatrix},$$

where $f \equiv f(\boldsymbol{\phi})$ and $g \equiv g(\boldsymbol{\phi})$ are defined by (12.11). Now see that $(1,0,-1)^T$ is a right eigenvector of $\mathcal{A}^*\mathcal{A}$ corresponding to the eigenvalue $3f^2+g^2$. Thus

$$\mathcal{F}(\boldsymbol{\phi}) = 2g^2|\boldsymbol{\phi}|^{-2}(3f^2+g^2+|\boldsymbol{\phi}|^{2Q})^{-1} + < bounded\ term > .$$




(a) zones for $(2\gamma, 2\delta)$ for the cases (A), (B), (C), (D). Each inner boundary belongs to the zone upper or right to it




(b) how to obtain the optimal value for $Q$ for the case (C). Solid line is the left-bottom boundary of the Newton polygon

Figure 4: Illustration to the example of Section 12.4

The optimal value of the long-time simulation order is the highest value $Q$ such that $\mathcal{F}(\boldsymbol{\phi})$ is bounded at $\boldsymbol{\phi} = 0$, i. e. there exist $c > 0$ and a neighborhood of $\boldsymbol{\phi} = 0$ such that

$$\exists\, c\ :\quad (g(\boldsymbol{\phi}))^2 \leqslant c|\boldsymbol{\phi}|^2(3(f(\boldsymbol{\phi}))^2 + (g(\boldsymbol{\phi}))^2 + |\boldsymbol{\phi}|^{2Q}).$$

Now we substitute the expressions (12.11) for $f(\boldsymbol{\phi})$ and $g(\boldsymbol{\phi})$. Obviously, replacing a function $F^{(m)}(\phi_j)$ by $\phi_j^m$ does not affect the condition to check. The multiplier 3 can be omitted also. Thus we need to check that

$$\exists\, c\ :\quad \phi_x^{2\gamma}\phi_y^{2\delta} \leqslant c\Big[(\phi_x^2+\phi_y^2)(\phi_x^{2\alpha}\phi_y^{2\beta} + \phi_x^{2\gamma}\phi_y^{2\delta}) + (\phi_x^2+\phi_y^2)^{Q+1}\Big].$$

The expression $(\phi_x^2+\phi_y^2)^{Q+1}$ is equivalent to $|\phi_x|^{2Q+2} + |\phi_y|^{2Q+2}$. The term $(\phi_x^2+\phi_y^2)\phi_x^{2\gamma}\phi_y^{2\delta}$ does not affect the condition to check. Thus we get

$$\exists\, c\ :\quad \phi_x^{2\gamma}\phi_y^{2\delta} \leqslant c\Big[\phi_x^{2\alpha+2}\phi_y^{2\beta} + \phi_x^{2\alpha}\phi_y^{2\beta+2} + \phi_x^{2Q+2} + \phi_y^{2Q+2}\Big]. \tag{12.12}$$

We need the biggest value of $Q$ for which this condition is satisfied. Use Lemma 12.2 to find it. At the left-hand side we have $\boldsymbol{\phi}^{\boldsymbol{m}}$, where $\boldsymbol{m} = (2\gamma, 2\delta)$. At the right-hand side we have $\sum_{j=1}^4 \boldsymbol{\phi}^{\boldsymbol{n}_j}$, where

$$\boldsymbol{n}_1 = (2\alpha+2, 2\beta),\quad \boldsymbol{n}_2 = (2\alpha, 2\beta+2),\quad \boldsymbol{n}_3 = (2Q+2, 0),\quad \boldsymbol{n}_4 = (0, 2+2Q).$$

We have four possible cases here.

(A) $\gamma + \delta < \alpha + \beta + 1$. Then $Q = \gamma + \delta - 1$, i. e. the long-time simulation order is defined by the truncation error for the first three components of the solution.
(B) $\gamma \geqslant \alpha$ and $\delta \geqslant \beta$, and $\gamma + \delta > \alpha + \beta$. Then $Q = \infty$.
(C) $\gamma < \alpha$, $\gamma + \delta \geqslant \alpha + \beta + 1$. Then $Q = \beta + \alpha(\delta - \beta - 1)/(\alpha - \gamma)$.
(D) $\delta < \beta$, $\gamma + \delta \geqslant \alpha + \beta + 1$. Then $Q = \alpha + \beta(\gamma - \alpha - 1)/(\beta - \delta)$.

In Fig. 4(a) for a fixed pair $(\alpha, \beta)$ the zones for $(2\gamma, 2\delta)$ corresponding to the cases (A), (B), (C), and (D) are plotted. In Fig. 4(b) for the case (C) we illustrate the location of $\boldsymbol{m}$ and $\boldsymbol{n}_j$. For example, for $\alpha = 5$, $\beta = \gamma = 1$, $\delta = 8$ the optimal value for $Q$ can be found from the condition that the points $(10, 4)$, $(2, 16)$, and $(0, 2Q+2)$ can be

connected by a straight line. Thus we get $Q = 17/2$. The solution error possesses the estimate $O(h + th^{17/2})$, and the values $P = 1$, $Q = 17/2$ are optimal.

Compare this result with the accuracy on $C^1_{per,\boldsymbol{e}}(\mathbb{R}^2)$ for a fixed $\boldsymbol{e}$. If $\boldsymbol{e} = (\pm 1, 0)$ or $\boldsymbol{e} = (0, \pm 1)$ the first three components of the solution error are zero. Otherwise we can write $\phi_x = e_1\psi$ and $\phi_y = e_2\psi$, and the criterion (12.12) reduces to

$$\exists\, c \ : \quad \psi^{2(\gamma+\delta)} \leqslant c\Big[\psi^{2(\alpha+\beta+1)} + \psi^{2Q+2}\Big].$$

So in the case (A) we still have $Q = \gamma + \delta - 1$. But in the cases (B), (C), (D) we have $Q = \infty$, i. e. the scheme possesses the error estimate

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1(\boldsymbol{e})h\|\nabla v_0\| \tag{12.13}$$

on $C^1_{per,\boldsymbol{e}}(\mathbb{R}^2)$ for each $\boldsymbol{e} \in \mathring{\Omega}$.

This example also reveals the difference between the long-time simulation accuracy in the strong sense and in the weak sense. To show this, we need an estimate for the multiplier $C_1(\boldsymbol{e})$ in (12.13). Theorem 8.5 does not provide this estimate (one can obtain an estimate for $C_1(\boldsymbol{e})$ this way, but this is difficult due to the need of explicit control of the neighborhood of $\boldsymbol{\phi} = 0$ where intermediate estimates are valid). We will use the direct spectral analysis instead.

Now we put $\alpha = 3$, $\beta = \gamma = 1$, $\delta = 8$.

The equation for the last component of the solution is separate from others, so we will consider only the first three components. The eigenvalues of $\mathcal{A}(\boldsymbol{\phi})$ are $\pm ir$ and 0 where $r = \sqrt{3f^2 + g^2}$ with $f \equiv f(\boldsymbol{\phi})$ and $g \equiv g(\boldsymbol{\phi})$ given by (12.11). If $f = g = 0$, then $\mathcal{A} = 0$ and the error is exactly zero; below assume $r \neq 0$. We don't need to provide an explicit expression of the eigenvectors; it is enough to see that

$$A(\boldsymbol{\phi}) = S(\boldsymbol{\phi})M(\boldsymbol{\phi})S^{-1}(\boldsymbol{\phi}),$$

where

$$S(\boldsymbol{\phi}) = \frac{1}{r}\begin{pmatrix} -ig & r & f & 0 \\ -f & 0 & f+ig & 0 \\ f & -r & f & 0 \\ 0 & 0 & 0 & r \end{pmatrix}, \quad M(\boldsymbol{\phi}) = \begin{pmatrix} -f-ig & 2r & 0 & 0 \\ -f(2f+ig)/r & f+ig & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -iF^{(1)}(\phi_x) \end{pmatrix},$$

$$S^{-1}(\boldsymbol{\phi}) = \frac{1}{r^2}\begin{pmatrix} r(f+ig) & -2fr & r(f+ig) & 0 \\ f(2f+ig) & -f(f+ig) & ifg - g^2 - f^2 & 0 \\ rf & r(f-ig) & rf & 0 \\ 0 & 0 & 0 & r^2 \end{pmatrix}.$$

Obviously, $\|S(\boldsymbol{\phi})\| \leqslant \tilde{c}$, $\|S^{-1}(\boldsymbol{\phi})\| \leqslant \tilde{c}$, where $\tilde{c}$ does not depend on $\boldsymbol{\phi}$, and by stability for each $\nu \geqslant 0$ there holds $\|\exp(\nu M(\boldsymbol{\phi}))\| \leqslant \|S(\boldsymbol{\phi})\|\,\|\exp(\nu A(\boldsymbol{\phi}))\|\,\|S(\boldsymbol{\phi})\| \leqslant \tilde{c}^2$. Then

$$\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h) = \left(e^{\nu A(\boldsymbol{\phi})} - I\right) v(\boldsymbol{\phi}, \Pi_h) = S(\boldsymbol{\phi})\left(e^{\nu M(\boldsymbol{\phi})} - I\right) S^{-1}(\boldsymbol{\phi}) v(\boldsymbol{\phi}, \Pi_h).$$

Denote by $\|\ \cdot\ \|_{[1,2]}$ and by $\|\ \cdot\ \|_{[1,2,3]}$ the seminorms on $\mathbb{C}^4$ defined as $\|(a,b,c,d)^T\|_{[1,2]} = \sqrt{|a|^2 + |b|^2}$, and $\|(a,b,c,d)^T\|_{[1,2,3]} = \sqrt{|a|^2 + |b|^2 + |c|^2}$. Then

$$S^{-1}(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h) = \begin{pmatrix} 2ig/r \\ (ifg - g^2)/r^2 \\ (3f - ig)/r \\ i\phi_x \end{pmatrix}, \quad \|S^{-1}(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h)\|_{[1,2]} = \sqrt{4r^2 + |g - if|^2}\frac{|g|}{r^2} \leqslant \sqrt{5}\frac{|g|}{r},$$

$$\|\hat{\varepsilon}(\boldsymbol{\phi}, \nu, \Pi_h)\|_{[1,2,3]} \leqslant \|S(\boldsymbol{\phi})\|\left(1 + \sup_{\nu \geqslant 0}\|e^{\nu M(\boldsymbol{\phi})}\|_{[1,2]}\right)\|S^{-1}(\boldsymbol{\phi})v(\boldsymbol{\phi}, \Pi_h)\|_{[1,2]} \leqslant \sqrt{5}(\tilde{c} + \tilde{c}^3)\frac{|g|}{r}.$$

Neglecting high order terms we can write $f \approx i^{\alpha+\beta+2}\phi_x^\alpha\phi_y^\beta$ and $g \approx i^{\gamma+\delta-1}\phi_x^\gamma\phi_y^\delta$ and, for the chosen parameters, $f \approx -\phi_x^3\phi_y$ and $g \approx \phi_x\phi_y^8$. Thus

$$\|\hat{\varepsilon}(\boldsymbol{\phi},\nu,\Pi_h)\|_{[1,2,3]} \leqslant c\frac{|\phi_y|^7}{|\phi_x|^2+|\phi_y|^7} = c\frac{|\phi_y|^{7/2}}{|\phi_x|}\frac{|\phi_x|\,|\phi_y|^{7/2}}{|\phi_x|^2+|\phi_y|^7} \leqslant c\frac{|\phi_y|}{|\phi_x|}|\boldsymbol{\phi}|^{5/2}. \tag{12.14}$$

Let $v_0 \in H^3_{per}(\mathbb{R}^2)$, $N_0$ be its period. Then $w = \Pi_h v_0$ has the period $N_0/h$, so its Fourier series $F[w](\boldsymbol{\phi})$ may be nonzero only for $\boldsymbol{\phi} = 2\pi\boldsymbol{k}h/N_0$, $\boldsymbol{k} \in (\mathbb{Z}\cap[-N_0/(2h), N_0/(2h)))^2$, see (6.1). For such $\boldsymbol{\phi}$, we have either $\phi_x = 0$ or

$$\frac{|\phi_y|}{|\phi_x|} = \frac{|k_2|}{|k_1|} \leqslant |k_2| = \frac{N_0}{2\pi h}|\phi_y|.$$

If $\phi_x \neq 0$, (12.14) yields

$$\|\hat{\varepsilon}(\boldsymbol{\phi},\nu,\Pi_h)\|_{[1,2,3]} \leqslant c\frac{N_0}{2\pi h}|\boldsymbol{\phi}|^{7/2} \leqslant c\frac{N_0}{\sqrt{2\pi}h}|\boldsymbol{\phi}|^3. \tag{12.15}$$

If $\phi_x = 0$ then $A(\boldsymbol{\phi}) = 0$ and (12.15) also holds. Adding the disjoined component, we finally obtain

$$\|\hat{\varepsilon}(\boldsymbol{\phi},\nu,\Pi_h)\| \leqslant cN_0h^{-1}|\boldsymbol{\phi}|^3 + 2|\boldsymbol{\phi}|.$$

Note that $H^3_{per}(\mathbb{R}^2) \subset\subset C_{per}(\mathbb{R}^2)$, so $\Pi_h$ is well defined on $H^3_{per}(\mathbb{R}^2)$ and is a bounded homogeneous mapping of $H^3_{per}(\mathbb{R}^2)$ to $V_{per}$. Let $F_h$ be the operator that takes $v_0 \in H^3_{per}(\mathbb{R}^2)$ to $F_hv_0 = \varepsilon_h(t, v_0, \Pi_h)$. By Lemma 6.9 we have $\|F_h\|_{(3,h)} \leqslant 2\|\Pi_h\|_{(3,h)}$. Thus the conditions of Lemma 6.14 are satisfied with $p(|\boldsymbol{\alpha}|) = cN_0h^2|\boldsymbol{\alpha}|^3 + 2h|\boldsymbol{\alpha}|$. For each $v_0 \in H^3_{per}(\mathbb{R}^2)$ with period $N_0$ Lemma 6.14 yields

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \equiv \|F_hv_0\| \leqslant \tilde{C}(h\|\nabla v_0\| + N_0h^2\|\nabla^3 v_0\|).$$

This means that the scheme possesses the first formal order of accuracy and the infinite long-time simulation order on $H^3_{per}(\mathbb{R}^2)$ in sense of $\Pi_h$ in the weak sense.

Note that the set of values $\alpha = 3$, $\beta = \gamma = 1$, $\delta = 8$ falls in the case (C) and the optimal values of the formal order of accuracy and of the long-time simulation order in the strong sense are $P = 1$, $Q = 10$.

## 13. Order of accuracy depending on the transport velocity

Up to this point, we studied properties of a fixed scheme (2.8) for the transport equation (2.6) with a fixed the transport velocity $\boldsymbol{\omega}$ in (2.6). In this section we assume that the scheme coefficients linearly depend on $\boldsymbol{\omega} = (\omega_x, \omega_y)$, which is constant in space and time.

Consider the following scheme with three degrees of freedom per cell:

$$h\frac{du_{j,k}}{dt} + \omega_x\frac{u_{j+1,k} - u_{j-1,k}}{2} + \omega_y\frac{u_{j,k+1} - u_{j,k-1}}{2} + \omega_xF(u_{j,k+1} - 2u_{j,k} + u_{j,k-1}) + \omega_yGu_{j,k} = 0,$$

$$F = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

It is of the form (2.8) with $L_{\boldsymbol{\eta}} = L^{(x)}_{\boldsymbol{\eta}}\omega_x + L^{(y)}_{\boldsymbol{\eta}}\omega_y$, where

$$Z_{0,0} = I, L^{(x)}_{\pm1,0} = \pm I/2, L^{(x)}_{0,\pm1} = F, L^{(x)}_{0,0} = -2F, L^{(y)}_{0,\pm1} = \pm I/2, L^{(y)}_{0,0} = G,$$

and other coefficients are zero. Let $\Pi_h$ be given by

$$(\Pi_hf)_{\boldsymbol{\eta},1} = (\Pi_hf)_{\boldsymbol{\eta},2} = -h^2\frac{\partial^2 f}{\partial x^2}(\boldsymbol{\eta}h), \quad (\Pi_hf)_{\boldsymbol{\eta},3} = f(\boldsymbol{\eta}h).$$

Then $v(\boldsymbol{\phi}, \Pi_h) = (\phi_x^2, \phi_x^2, 1)^T$ and $A(\boldsymbol{\phi}) = i\boldsymbol{\phi}\cdot\boldsymbol{\omega} - L(\boldsymbol{\phi})$ is

$$A(\boldsymbol{\phi}) = i\omega_x(\phi_x - \sin\phi_x)I + i\omega_y(\phi_y - \sin\phi_y)I + \begin{pmatrix} 0 & \omega_y & \omega_x g(\phi_y) \\ -\omega_y & 0 & -\omega_x g(\phi_y) \\ -\omega_x g(\phi_y) & \omega_x g(\phi_y) & 0 \end{pmatrix},$$

where $g(\phi_y) = 4\sin^2(\phi_y/2)$. Since $(A(\boldsymbol{\phi}))^* = -A(\boldsymbol{\phi})$, the scheme is stable with the stability constant $K = 1$ for each $\boldsymbol{\omega}$. We have

$$\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = i\left[\omega_x(\phi_x - \sin\phi_x) + \omega_y(\phi_y - \sin\phi_y)\right]\begin{pmatrix} \phi_x^2 \\ \phi_x^2 \\ 1 \end{pmatrix} + \left[\omega_y\phi_x^2 + 4\omega_x\sin^2(\phi_y/2)\right]\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

Thus for $\boldsymbol{\omega} \neq 0$ we have $\hat{\epsilon}(\boldsymbol{\phi}, \Pi_h) = O(|\boldsymbol{\phi}|^2)$ and the optimal value of the truncation error is $P_A = 1$.

Now we use Theorem 1 to obtain the order of the solution error. We have

$$(\Pi_1 x^2/2)_{(\eta_x,\eta_y)} = (-1, -1, \eta_x^2/2)^T,$$
$$(\Pi_1 xy)_{(\eta_x,\eta_y)} = (0, 0, \eta_x\eta_y)^T,$$
$$(\Pi_1 y^2/2)_{(\eta_x,\eta_y)} = (0, 0, \eta_y^2/2)^T,$$

so

$$(\epsilon_1(x^2/2, \Pi_h))_0 = -\omega_y G(-1, -1, 0)^T = \omega_y(-1, 1, 0)^T,$$
$$(\epsilon_1(xy, \Pi_h))_0 = 0,$$
$$(\epsilon_1(y^2/2, \Pi_h))_0 = \omega_x F(0, 0, 1)^T = \omega_x(-1, 1, 0)^T.$$

Note that $L(0) = \omega_y G$. If $\omega_y \neq 0$, then all the vectors $(\epsilon_1(x^2/2, \Pi_h))_0$, $(\epsilon_1(xy, \Pi_h))_0$, $(\epsilon_1(y^2/2, \Pi_h))_0$ belong to the image of $L(0)$ and thus by Theorem 1 the scheme possesses the second order of accuracy. For $\omega_y = 0$, $\omega_x \neq 0$ the vector $(\epsilon_1(y^2/2, \Pi_h))_0$ does not belong to $\operatorname{Im} L(0) = \{0\}$, thus the scheme does not possess the second order of accuracy.

## 14. Conclusion

The Lax – Ryabeknii theorem states that a stable scheme possessing the truncation error $O(h^{P_A})$ provides the solution error with an estimate $O(h^{P_A}t)$. For the classical finite-difference schemes these estimates are optimal, however, for some schemes they can be very rough. The method of auxiliary mapping is a powerful tool to obtain better estimates, which was successfully used for the DG scheme by several authors. However, up to now this method has been applied on ad hoc basis.

We consider a linear $L_2$-stable scheme with several degrees of freedom per cell on a uniform mesh and a local mapping $\Pi_h$ to the mesh space. Based on the method of auxiliary mapping we present a unified approach to find the largest possible $P$ and, for this $P$, the largest possible $Q$ such that the scheme possesses a solution error estimate of the form $O(h^P + th^Q)$. The outline of our results is below.

To analyze the accuracy of the scheme one first finds the optimal order $P_A$ of the truncation error. The optimal value $P$ of the formal order of accuracy can be either $P_A$ or $P_A + 1$. To check which of these two cases holds, one considers auxiliary mappings $\Pi_h^{(P_A+1,P_A+1)}$ of the form (2.12) with undetermined coefficients $\mathfrak{C}^{(\boldsymbol{m})}$. The condition that the truncation error is of order $P_A + 1$ in the sense of $\Pi_h^{(P_A+1,P_A+1)}$ forms linear systems for $\mathfrak{C}^{(\boldsymbol{m})}$, $|\boldsymbol{m}| = P_A + 1$. The scheme possesses the order of accuracy $P = P_A + 1$ if and only if each of these systems is consistent, see Algorithm 2 for details. Below we assume that $P > 0$.

For a quasi-1D or a "simple" scheme (see Definitions 6 and 7), the optimal value of $Q$ in an estimate of the form $O(h^P + th^Q)$ is an integer (or it can be infinity, if the transport velocity is zero). In order to check whether the scheme possesses this estimate, one considers auxiliary mappings $\Pi_h^{(P,Q)}$ of the form (2.12) with undetermined coefficients $\mathfrak{C}^{(\boldsymbol{m})}$. The condition that the truncation error is of order $Q$ in the sense of $\Pi_h^{(P,Q)}$ forms a linear system for $\mathfrak{C}^{(\boldsymbol{m})}$, $P \leqslant |\boldsymbol{m}| \leqslant Q$. If this system is consistent, then the scheme possesses a solution error estimate of the form

$O(h^P + th^Q)$, otherwise it does not. See Algorithm 3 for details. For "simple" schemes the diagonal matrices $\mathfrak{C}^{(\boldsymbol{m})}$ can be found successively, and thus Algorithm 3 simplifies to Algorithm 4.

If a scheme is neither quasi-1D nor simple, the situation may be tricky. The definitions of the long-time simulation order in the weak and strong sense, namely Definition 3 and Definition 4, are not equivalent to each other. The optimal value of the long-time simulation order $Q$ (with the formal order $P$) in the sense of Definition 4 is generally *not* a natural number. If the scheme possesses the formal order of accuracy $P$ and the long-time simulation order $Q$ in the sense of Definition 4, then there exists a homogeneous mapping $\tilde{\Pi}_h$ such that $\|\tilde{\Pi}_h f - \Pi_h f\| = O(h^P)$ and the scheme possesses the truncation error of order $Q$ in the sense of $\tilde{\Pi}_h$. However, even if $Q$ is a natural number, there can be no such local mapping and thus Algorithm 3 generally fails. A general criterion of the $Q$-th order in the long-time simulation is given by Theorem 8.5, however we do not have an algorithm to check it.

In the general case one can specify a unit vector $\boldsymbol{e}$ and restrict the analysis to solutions of the form $f(t, \boldsymbol{e}\cdot\boldsymbol{r})$ and write a system for diagonal matrices $\mathfrak{C}^{(m)}_{\boldsymbol{e}}$, $m = P, \dots, Q$, such that the scheme possesses the long-time simulation order $Q$ in the sense of the mapping $\Pi^{(P,Q)}_{\boldsymbol{e}}$ of the form (2.13). If these matrices are bounded over $\boldsymbol{e}$ on the unit sphere then the scheme possesses the long-time simulation order $Q$. If for some $\boldsymbol{e} \in \mathring{\Omega}$ these matrices do not exist then the scheme does not possess the long-time simulation order $Q$. If these matrices exist but there is no way to specify them boundedly over the unit sphere then this tells us nothing.

All cases and methods mentioned above were demonstrated on the artificial examples of numerical schemes.

## A. Appendix

**Lemma A.1.** *Let $n \in \mathbb{N} \cup \{0\}$ and $w \in H^n_{per}(\mathbb{R}^d)$. Then the values $\|\nabla^n w\|^2$ defined by* (2.2) *and* (2.5) *coincide.*

*Proof.* Let $w \in H^n_{per}(\mathbb{R}^d)$ have the Fourier series (2.4). For a multiindex $\boldsymbol{n} = (n_1, \dots, n_d)$, $|\boldsymbol{n}| = n$, the Fourier coefficients for the function $D^{\boldsymbol{n}} w$ are equal to $(i\alpha_1)^{n_1} \dots (i\alpha_d)^{n_d} w_{\boldsymbol{\alpha}}$. Hence, by Parseval's identity

$$\|D^{\boldsymbol{n}} w\|^2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} |\alpha_1|^{2n_1} \dots |\alpha_d|^{2n_d} |w_{\boldsymbol{\alpha}}|^2.$$

Multiply this by $n!/\boldsymbol{n}!$ and take the sum over all multiindexes $\boldsymbol{n}$ such that $|\boldsymbol{n}| = n$. We have

$$\|\nabla^n w\|^2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 \sum_{\substack{n_1\geqslant 0,\dots,n_d\geqslant 0,\\ n_1+\dots+n_d=n}} \frac{n!}{\boldsymbol{n}!} \left|\alpha_1^2\right|^{n_1} \dots \left|\alpha_d^2\right|^{n_d} =$$

$$= \sum_{\boldsymbol{\alpha}\in\mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 \left(\alpha_1^2 + \dots + \alpha_d^2\right)^n = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} |w_{\boldsymbol{\alpha}}|^2 |\boldsymbol{\alpha}|^{2n}.$$

The proof of the lemma is complete. □

**Lemma A.2.** *For $\boldsymbol{e} \in \mathring{\Omega}$ the spaces $H^q_{per,\boldsymbol{e}}(\mathbb{R}^d)$ and $C^q_{per,\boldsymbol{e}}(\mathbb{R}^d)$ are infinite dimensional for each $q \in \mathbb{N} \cup \{0\}$.*

*Proof.* By definition there exists $\lambda > 0$ such that $\lambda \boldsymbol{e}\cdot\boldsymbol{a}_j = n_j/d_j$, where $n_j \in \mathbb{Z}$ and $d_j \in \mathbb{N}$, for all $j = 1, \dots, d$. Consider $m \in \mathbb{Z}$, the functions $g(x) = \exp(2\pi i m\lambda x)$ and $f(\boldsymbol{r}) = g(\boldsymbol{e}\cdot\boldsymbol{r})$. Then we have

$$f(\boldsymbol{r} + N_0\boldsymbol{a}_j) = g(\boldsymbol{e}\cdot\boldsymbol{r} + \boldsymbol{e}\cdot\boldsymbol{a}_j N_0) = g(\boldsymbol{e}\cdot\boldsymbol{r}) \exp(2\pi i m\lambda \boldsymbol{e}\cdot\boldsymbol{a}_j N_0) = f(\boldsymbol{r}) \exp(2\pi i m\lambda \boldsymbol{e}\cdot\boldsymbol{a}_j N_0).$$

Assuming $N_0 = \prod_j d_j$ we get $f(\boldsymbol{r} + N_0\boldsymbol{a}_j) = f(\boldsymbol{r})$ for each $j = 1, \dots, d$, then $N_0$ is a period of $f$. Hence, $f \in C^q_{per,\boldsymbol{e}}(\mathbb{R}^d) \subseteq H^q_{per,\boldsymbol{e}}(\mathbb{R}^d)$. □

**Lemma A.3.** *Let $G$ be a bounded domain in $\mathbb{R}^d$. For each function $f \in L_{2,per}(\mathbb{R}^d)$ with period $N_0$ and each $h$ such that $1/h \in \mathbb{N}$ there holds*

$$Y(f) := \frac{1}{(N_0/h)^d} \sum_{\boldsymbol{\eta}\in\{0,\dots,N_0/h-1\}^d} \int_G |f(h\boldsymbol{r} + hT\boldsymbol{\eta})|^2 \, d\boldsymbol{r} \leqslant C\|f\|^2 \tag{A.1}$$

*with $C > 0$ independent of $f$, $N_0$, $h$.*

*Proof.* Let $\square$ be the parallelepiped generated by vectors $N_0\boldsymbol{a}_1, \ldots, N_0\boldsymbol{a}_d$ and $\square_1$ be the parallelepiped generated by the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$. If $G$ is a translation of $\square_1$ then we have

$$\frac{1}{(N_0/h)^d} \sum_{\boldsymbol{\eta}\in\{0,\ldots,N_0/h-1\}^d} \int\limits_G |f(h\boldsymbol{r} + hT\boldsymbol{\eta})|^2 \, d\boldsymbol{r} = \frac{1}{N_0^d} \int\limits_{\square} |f(\boldsymbol{r})|^2 \, d\boldsymbol{r} = |\square_1| \cdot \|f\|^2.$$

Since $G$ is bounded, $G$ belongs to the union of $m \in \mathbb{N}$ translations of $\square_1$. Then (A.1) holds with $C = m|\square_1|$. $\square$

**Lemma A.4.** *A local mapping with kernel $\mu \in (W_2^q(G))^*$ is a bounded homogeneous mapping of $H^q_{per}(\mathbb{R}^d)$ to $V_{per}$.*

*Proof.* We need to show that for a local mapping $\Pi_h$ with $\mu \in (W_2^q(G))^*$ the norm $\|\Pi_h\| = \sup \|\Pi_h f\|/\|f\|_{(q,h)}$ is bounded. First equip $W_2^q(G)$ with the norm

$$\|f\|^2_{W_2^q} = \sum_{|\boldsymbol{m}|\leqslant q} \frac{|\boldsymbol{m}|!}{\boldsymbol{m}!} \int\limits_G |D^{\boldsymbol{m}} f(\boldsymbol{r})|^2 d\boldsymbol{r}.$$

Let $f \in H^q_{per}(\mathbb{R}^d)$ have a period $N_0$. Put $N = N_0/h$. By the norm equivalence on the finite-dimensional space $\mathbb{C}^{M^0}$ we get

$$\|\Pi_h f\|^2 = \frac{1}{N^d} \sum_{\boldsymbol{\eta}\in\{0,\ldots,N-1\}^d} \|(\Pi_h f)_{\boldsymbol{\eta}}\|^2 \leqslant C \frac{1}{N^d} \sum_{\boldsymbol{\eta}\in\{0,\ldots,N-1\}^d} \sum_{\xi\in M^0} |(\Pi_h f)_{\boldsymbol{\eta},\xi}|^2 .$$

The value of $(\Pi_h f)_{\boldsymbol{\eta},\xi}$ results from the application of the functional $\mu_\xi \in (W_2^q(G))^*$ to the function $g(\boldsymbol{r}) = f(h(\boldsymbol{r} + T\boldsymbol{\eta}))$ so $|(\Pi_h f)_{\boldsymbol{\eta},\xi}| \leqslant \|\mu_\xi\| \, \|g\|_{W_2^q(G)}$. Thus

$$\|\Pi_h f\|^2 \leqslant C|M^0| \left(\max_{\xi\in M^0} \|\mu_\xi\|\right)^2 \frac{1}{N^d} \sum_{\boldsymbol{\eta}\in\{0,\ldots,N-1\}^d} \sum_{|\boldsymbol{m}|\leqslant q} h^{2|\boldsymbol{m}|} \frac{|\boldsymbol{m}|!}{\boldsymbol{m}!} \int\limits_G |(D^{\boldsymbol{m}} f)(h(\boldsymbol{r} + T\boldsymbol{\eta}))|^2 \, d\boldsymbol{r} =$$

$$= C|M^0| \left(\max_{\xi\in M^0} \|\mu_\xi\|\right)^2 \sum_{|\boldsymbol{m}|\leqslant q} h^{2|\boldsymbol{m}|} \frac{|\boldsymbol{m}|!}{\boldsymbol{m}!} Y(D^{\boldsymbol{m}} f),$$

where $Y$ is given by (A.1). By Lemma A.3 we get

$$\|\Pi_h f\|^2 \leqslant \tilde{C} \sum_{|\boldsymbol{m}|\leqslant q} h^{2|\boldsymbol{m}|} \frac{|\boldsymbol{m}|!}{\boldsymbol{m}!} \|D^{\boldsymbol{m}} f\|^2 = \tilde{C} \|f\|^2_{(q,h)},$$

thus $\|\Pi_h\|_{(q,h)} \leqslant \tilde{C} < \infty$. $\square$

**Lemma A.5.** *Let $\{(\boldsymbol{e}_1 \cdot \boldsymbol{r})^m, \ldots, (\boldsymbol{e}_J \cdot \boldsymbol{r})^m\}$ be a complete system in the space of homogeneous polynomials of order $m$. Then $\{(\boldsymbol{e}_1 \cdot \boldsymbol{r})^n, \ldots, (\boldsymbol{e}_J \cdot \boldsymbol{r})^n\}$ is a complete system in the space of homogeneous polynomials of order $n \leqslant m$.*

*Proof.* For $m = n$ this is obvious. If $m > n$ choose any $\boldsymbol{e}_0 \in \Omega$. Let $u$ be a homogeneous polynomial of order $n$. Then there exists a homogeneous polynomial $U$ of order $m$ such that $u = (\boldsymbol{e}_0 \cdot \nabla)^{m-n} U$. For instance, one can construct $U$ using an $(m-n)$-th primitive of $u(t(\boldsymbol{e}_0 \cdot \boldsymbol{r})\boldsymbol{e}_0 + \boldsymbol{r} - (\boldsymbol{e}_0 \cdot \boldsymbol{r})\boldsymbol{e}_0)$. By assumption $U = \sum\limits_{j=1}^{J} U_j(\boldsymbol{e}_j \cdot \boldsymbol{r})^m$, thus

$$u(\boldsymbol{r}) = \sum_{j=1}^{J} U_j(\boldsymbol{e}_0 \cdot \nabla)^{m-n} (\boldsymbol{e}_j \cdot \boldsymbol{r})^m = \frac{m!}{n!} \sum_{j=1}^{J} U_j(\boldsymbol{e}_0 \cdot \boldsymbol{e}_j)^{m-n} (\boldsymbol{e}_j \cdot \boldsymbol{r})^n.$$

Thus $u \in \operatorname{span}\{(\boldsymbol{e}_j \cdot \boldsymbol{r})^n, j = 1, \ldots, J\}$. $\square$

**Lemma A.6.** *Let $n, d \in \mathbb{N}$, $\Omega_d$ be the unit sphere in $\mathbb{R}^d$ and $\breve{C}_d$ be an open cone in $\mathbb{R}^d$. Then there exists a set of vectors $\{\boldsymbol{e}_k \in \mathring{\Omega}_d \cap \breve{C}_d,\ k = 1, \dots, C_{n+d-1}^{d-1}\}$ such that $\{(\boldsymbol{e}_k \cdot \boldsymbol{r})^n\}$ form a basis in the space of homogeneous polynomials of order $n$ of $d$ variables.*

*Proof.* First prove that there exists a set of vectors $\{\boldsymbol{e}_k \in \Omega_d \cap \breve{C}_d,\ k = 1, \dots, C_{n+d-1}^{d-1}\}$ such that $\{(\boldsymbol{e}_k \cdot \boldsymbol{r})^n\}$ form a basis in the space of homogeneous polynomials of order $n$ of $d$ variables.

Assume without loss that $(1, 0, \dots, 0) \in \breve{C}_d$.

The proof is by induction in $d$. For $d = 1$ the statement is obvious. Let us prove this for $d = 2$. Let $k \in \{0, \dots, n\}$. Let $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_n)$ be a set of pairwise different numbers. Find $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_n)$ such that

$$x^k y^{n-k} = \sum_{j=0}^{n} \gamma_j (x + \alpha_j y)^n.$$

This is equivalent to the system of equations

$$\sum_{j=0}^{n} \gamma_j \alpha_j^l = \frac{\delta_{l,n-k}}{C_n^l}, \quad l = 0, \dots, n.$$

This is a system of linear equations for $\gamma_j$ with the Wandermond determinant and for any set of pairwise different $\alpha_j$ this system has a unique solution. Thus for $d = 2$ and any system of vectors $\boldsymbol{e}_j = (1, \alpha_j)$, $j = 0, \dots, n$ with pairwise different $\alpha_j$ the polynomials $(\boldsymbol{e}_j \cdot \boldsymbol{r})^n$ form a basis in the space of homogeneous polynomials of order $n$. Clearly we can choose $\boldsymbol{\alpha}$ so that the vectors $(1, \alpha_j)$ belong to $\breve{C}_d$.

Now let the statement hold for the space dimension $d \geqslant 2$, we will prove it for $d + 1$. Let $\{\boldsymbol{e}_j \in \Omega_d,\ j = 1, \dots, C_{n+d-1}^{d-1}\}$, be the set of vectors given by the induction assumption. Let $\{\boldsymbol{g}_j, j = 1, \dots, d+1\}$ be the standard basis in $\mathbb{R}^{d+1}$. For $\boldsymbol{r} \in \mathbb{R}^{d+1}$ denote $\boldsymbol{r} = (\boldsymbol{r}', r_{d+1})$, $\boldsymbol{r}' \in \mathbb{R}^d$.

Using the argument for $d = 2$, given any set of pairwise different numbers $\alpha_j$, $j = 0, \dots, n$, we can find $\gamma_{k,j}$, $k, j = 0, \dots, n$, such that

$$(\boldsymbol{r}' \cdot \boldsymbol{e}')^k (r_{d+1})^{n-k} = \sum_{j=0}^{n} \gamma_{k,j} ((\boldsymbol{r}' \cdot \boldsymbol{e}') + \alpha_j r_{d+1})^n, \quad \boldsymbol{e}', \boldsymbol{r}' \in \mathbb{R}^d, \quad r_{d+1} \in \mathbb{R}.$$

By the induction assumption and Lemma A.5, for any multiindex $\boldsymbol{k} = (k_1, \dots, k_d)$ with $|\boldsymbol{k}| = k$ we can find coefficients $\beta_{\boldsymbol{k},l}$, $l = 1, \dots, J$ where $J = C_{n+d-1}^{d-1}$, such that

$$(\boldsymbol{r}')^{\boldsymbol{k}} = \sum_{l=1}^{J} \beta_{\boldsymbol{k},l} (\boldsymbol{r}' \cdot \boldsymbol{e}_l)^k$$

and thus

$$(\boldsymbol{r}')^{\boldsymbol{k}} (r_{d+1})^{n-k} = \sum_{l=1}^{J} \beta_{\boldsymbol{k},l} \sum_{j=0}^{n} \gamma_{k,j} ((\boldsymbol{r}' \cdot \boldsymbol{e}_l) + \alpha_j r_{d+1})^n.$$

So for any set of pairwise different numbers $\{\alpha_j\}_{j=0}^n$, for the system of vectors

$$\boldsymbol{e}_{l,j} = (1 + \alpha_j^2)^{-1/2} (\boldsymbol{e}_l, \alpha_j) \in \mathbb{R}^{d+1}, \quad l = 1, \dots, C_{n+d-1}^{d-1}, \quad j = 0, \dots, n,$$

the set of polynomials $(\boldsymbol{e}_{l,j} \cdot \boldsymbol{r})^n$ forms a complete system in the space of homogeneous polynomials of order $n$ in $d + 1$ variables. By construction all these vectors have the form $t(1, \alpha_{i_1}, \dots, \alpha_{i_d})$, $i_k \in \{0, \dots, n\}$, $t > 0$. Choosing sufficiently small $\boldsymbol{\alpha}$ we can guarantee that these vectors belong to $\Omega_{d+1} \cap \breve{C}_d$. Extracting a basis from this complete system we prove the induction statement.

Now let us show that we can find the required system of vectors in $\mathring{\Omega}_d \cap \breve{C}_d$. Let $\boldsymbol{e}_l \in \Omega_d \cap \breve{C}_d$, $l = 1, \dots, C_{n+d-1}^{d-1}$ be the output of the argument above. It is clear that any set of vectors sufficiently close to a basis also form a basis.

Given $\varepsilon > 0$, for each $\boldsymbol{e}_l$ we find a vector $\boldsymbol{g}_l \in \mathbb{Q}^d$ such that $\|U^{-1}\boldsymbol{e}_l - \boldsymbol{g}_l\| < \varepsilon$. Then $\|\boldsymbol{e}_l - U\boldsymbol{g}_l\| \leqslant \varepsilon\|U\|$, thus $|\|U\boldsymbol{g}_l\| - 1| = |\|U\boldsymbol{g}_l\| - \|\boldsymbol{e}_l\|| \leqslant \varepsilon\|U\|$ and

$$\left\|\boldsymbol{e}_l - \frac{U\boldsymbol{g}_l}{\|U\boldsymbol{g}_l\|}\right\| \leqslant \|\boldsymbol{e}_l - U\boldsymbol{g}_l\| + \left\|U\boldsymbol{g}_l - \frac{U\boldsymbol{g}_l}{\|U\boldsymbol{g}_l\|}\right\| = \|\boldsymbol{e}_l - U\boldsymbol{g}_l\| + |\|U\boldsymbol{g}_l\| - 1| \leqslant 2\varepsilon\|U\|.$$

Thus for sufficienty small $\varepsilon > 0$ the system $U\boldsymbol{g}_l/\|U\boldsymbol{g}_l\|$ is the required set of vectors in $\mathring{\Omega} \cap \check{C}_d$. □

**Acknowledgements**

The research is funded by the Russian Science Foundation, project 22-11-00199.

**References**

[1] Cockburn B., Luskin M., Shu C.-W., Süli E., Enhanced accuracy by post-processing for finite element methods for hyperbolic equations, Mathematics of Computation 72 (2003) 577–606.

[2] Vichnevetsky R., Bowles J. B., Fourier analysis of numerical approximations of hyperbolic equations, SIAM, 1982.

[3] Lowrie R., Compact higher-order numerical methods for hyperbolic conservation laws, Ph.D. thesis, The University of Michigan (1996).

[4] Hu F., Hussaini M., Rasetarinera P., An Analysis of the Discontinuous Galerkin Method for Wave Propagation Problems, J. Comput. Phys. 151 (1999) 921–946.

[5] Zhang M., Shu C.-W., An analysis of and a comparison between the discontinuous Galerkin and the spectral finite volume methods, Computers and Fluids 34 (2005) 581–592.

[6] Guo W., Zhong X., Qui C.-M., Superconvergence of discontinuous Galerkin and local discontinuous Galerkin methods: Eigen-structure analysis based on Fourier approach, J. Comput. Phys. 235 (2013) 458–485.

[7] Balan A., May G., Schöberl J., A stable high-order Spectral Difference method for hyperbolic conservation laws on triangular elements, J. Comput. Phys. 231 (2012) 2359–2375.

[8] Vanharen J., Puigt G., Vasseur X., Boussuge J.-F., Sagaut P., Revisiting the spectral analysis for high-order spectral discontinuous methods, J. Comput. Phys. 337 (2017) 379–402.

[9] Huynh H. T., A flux reconstruction approach to high-order schemes including discontinuous Galerkin methods, AIAA paper No 2007-4079.

[10] Vincent P. E., Castonguay P., Jameson A., Insights from von Neumann analysis of high-order flux reconstruction schemes, J. Comput. Phys. 230 (2011) 8134–8154.

[11] Cheng Y., Shu C.-W., Superconvergence and time evolution of discontinuous Galerkin finite element solutions, J. Comput. Phys. 227 (22) (2008) 9612–9627.

[12] Cheng Y., Shu C.-W., Superconvergence of discontinuous Galerkin and local discontinuous Galerkin schemes for linear hyperbolic and convection-diffusion equations in one space dimension, SIAM Journal on Numerical Analysis 47 (2010) 4044–4072.

[13] Zhang Q., Shu C.-W., Stability analysis and a priori error estimates of the third order explicit Runge–Kutta discontinuous Galerkin method for scalar conservation laws, SIAM Journal on Numerical Analysis 48 (3) (2010) 1038–1063.

[14] Yang Y., Shu C.-W., Analysis of optimal superconvergence of discontinuous Galerkin method for linear hyperbolic equations, SIAM Journal on Numerical Analysis 50 (6) (2012) 3110–3133.

[15] Cao W., Zhang Z., Zou Q., Superconvergence of discontinuous Galerkin methods for linear hyperbolic equations, SIAM Journal on Numerical Analysis 52 (5) (2014) 2555–2573.

[16] Bouche D., Ghidaglia J.-M., Pascal F., Error estimate and the geometric corrector for the upwind finite volume method applied to the linear advection equation, SIAM Journal on Numerical Analysis 43 (2006) 557–603.

[17] Cao W., Shu C.-W., Yang Y., Zhang Z., Superconvergence of discontinuous Galerkin methods for two-dimensional hyperbolic equations, SIAM Journal on Numerical Analysis 53 (4) (2015) 1651–1671.

[18] Cao W., Shu C.-W., Zhang Z., Superconvergence of discontinuous Galerkin methods for 1-D linear hyperbolic equations with degenerate variable coefficients, ESAIM: M2AN 51 (6) (2017) 2213–2235.

[19] Kučera V., Shu C.-W., On the time growth of the error of the DG method for advective problems, IMA Journal of Numerical Analysis 39 (2019) 687–712.

[20] Liu Y., Shu C.-W., Zhang M., Optimal error estimates of the semidiscrete discontinuous Galerkin methods for two dimensional hyperbolic equations on Cartesian meshes using Pk elements, ESAIM: M2AN 54 (2020) 705–726.

[21] Liu Y., Shu C.-W., Zhang M., Superconvergence of energy-conserving discontinuous Galerkin methods for linear hyperbolic equations, Communications on Applied Mathematics and Computation 1 (2019) 101–116.

[22] Xu Y., Meng X., Shu C.-W., Zhang Q., Superconvergence analysis of the Runge–Kutta discontinuous Galerkin methods for a linear hyperbolic equation, Journal of Scientific Computing 84 (2020) article number 23.

[23] Bouche D., Ghidaglia J.-M., Pascal F., Error estimate for the upwind finite volume method for the nonlinear scalar conservation law, Journal of Computational and Applied Mathematics 235 (2011) 5394–5410.

[24] Work C. D., Katz A. J., Aspects of the flux correction method for solving the Navier – Stokes equations on unstructured meshes, AIAA paper No. 2015-0834.

[25] Kolmogorov A. N., On inequalities between the upper bounds of the successive derivatives of an arbitrary function on an infinite integral, Amer. Math. Soc. Translations, Ser. 1 2 (1962) 233–243.

[26] Kato T., Perturbation theory for linear operators, Grund. math. Wiss., B. 132, Springer, 1966.

[27] Kreiss H. O., Über Matrizen die beshränkte Halbgruppen erzeugen, Mathematica Scandinavica (1959) 71–80.

[28] Bakhvalov P. A., Surnachev M. D., Transformation to a block-diagonal form of matrices generating bounded semigroups, Linear Algebra and Its Applications 603 (2020) 275–288.

[29] Bakhvalov P. A., Surnachev M. D., On analytical families of matrices generating bounded semigroups, Numerical Analysis and Applications 14 (2021) 1–12.

[30] Baumgartel H., Analytic perturbation theory for matrices and operators, Basel; Boston; Stuttgart: Birkhäuser Verlag, 1985.

[31] Shu C.-W., Discontinuous Galerkin method for time-dependent problems: survey and recent developments, Springer, Cham, 2014.