IMA Journal of Numerical Analysis (2006) **26**, 686−722
doi:10.1093/imanum/drl005
Advance Access publication on March 21, 2006

Stability and consistency of kinetic upwinding for advection–diffusion equations

MANUEL TORRILHON[†] AND KUN XU[‡]
Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

[Received on 1 November 2005]

Numerical methods based on kinetic models of fluid flows, like the so-called BGK scheme, are becoming increasingly popular for the solution of convection-dominated viscous fluid equations in a finite-volume approach due to their accuracy and robustness. Based on kinetic-gas theory, the BGK scheme approximately solves the BGK kinetic model of the Boltzmann equation at each cell interface and obtains a numerical flux from integration of the distribution function. This paper provides the first analytical investigations of the BGK-scheme and its stability and consistency applied to a linear advection–diffusion equation. The structure of the method and its limiting cases are discussed. The stability results concern explicit time marching and demonstrate the upwinding ability of the kinetic method. Furthermore, its stability domain is larger than that of common finite-volume methods in the under-resolved case, i.e. where the grid Reynolds number is large. In this regime, the BGK scheme is shown to allow the time step to be controlled from the advection alone. We show the existence of a third-order 'super-convergence' on coarse grids independent of the initial condition. We also prove a limiting order for the local consistency error and show the error of the BGK scheme to be asymptotically first order on very fine grids. However, in advection-dominated regimes super-convergence is responsible for the high accuracy of the method.

Keywords: advection–diffusion; finite difference methods; kinetic schemes; stability; BGK model.

1. Introduction

Typically, fluid flow processes in physics and engineering consist of a hyperbolic convection or advection part and a parabolic dissipative or diffusive part, see e.g. Morton (1996). They are described by the compressible or incompressible system of balance laws of continuum physics which are closed by the constitutive relations of Navier–Stokes and Fourier for viscosity and heat conduction. The development of numerical methods for those processes often focuses on the physical ingredients, advection and diffusion, separately. For the diffusive part representing an elliptic operator, the use of highly developed finite-element methods is popular, while the hyperbolic advection part is best approximated using high resolution finite-volume wave propagation methods.

The so-called kinetic schemes were originally developed for hyperbolic equation. These schemes use the microscopic kinetic background of the physical equations to derive macroscopic numerical methods. Initially described by Pullin (1980), Deshpande (1986) and Prendergast & Xu (1993), kinetic schemes have been used, modified and further developed for hyperbolic equations in many papers, see e.g. the work of Aregba-Driollet & Natalini (2000), Kim *et al.* (1997) and Perthame (1992). This paper will use

[†]Email: matorril@ust.hk
[‡]Email: makxu@ust.hk

the gas-kinetic BGK scheme in which the BGK model (after Bhatnagar, Gross & Krook, 1954) of the Boltzmann equation is solved at each cell interface in order to obtain a numerical flux, see Sections 2 and 3 for a description. Xu (2001) proposed a procedure for applying the framework of BGK kinetic schemes to the full compressible 'viscous' fluid equations. Similar ideas can be found in the works of Chou & Baganoff (1997) and Junk & Rao (1999) in different settings. The common idea is to formulate a numerical method which includes the physical phenomena of advection and diffusion in a unified framework given by kinetic-gas theory. In these methods, the numerical approximation of the diffusive part of the equations is strongly coupled with the approximation of the advection and this coupling is realized according to physical, i.e. gas-kinetical requirements. In this sense, the gas-kinetic BGK scheme for viscous equations has to be distinguished from standard methods.

The BGK scheme for the full convection-dominated compressible Navier–Stokes equations turned out to be an accurate and robust solver. It has been studied by Ohwada (2002) and Ohwada & Kobayashi (2004) . Extensions to include scalar transport and multi-dimensional effects have been done by Li *et al.* (2004) and Xu *et al.* (2005). Multi-fluid applications have been considered by Li & Fu (2003) and Song & Ni (2004). Recently, May, Srinivasan and Jameson applied the method to practical engineering 3D problems like aircraft flow in May *et al.* (2005).

The aim of this paper is to provide some analytical results about the numerical behavior of the BGK scheme in a simplified setting. To our knowledge, no analytical investigations of the method are available so far. Due to the setting, the approach of the paper is of fundamental nature and we will discuss thoroughly different aspects and variants of the method. We are not concerned with the competitiveness of the method for the full compressible gas dynamics, however, our results may provide some explanations for robustness and accuracy of the method. We apply the gas-kinetic BGK method to a model equation of viscous flow given by the scalar, linear advection–diffusion equation

$$\partial_t u + a\partial_x u = \nu \partial_{xx} u, \tag{1.1}$$

where $a \in \mathbb{R}$ is the advection velocity and $\nu \in \mathbb{R}$ is a positive viscosity or diffusion coefficient. This equation allows to uncover and investigate the structure of the BGK scheme in a most explicit setting which will provide a thorough understanding of the kinetic mechanisms in the method. The paper concentrates on results for L^2-stability and asymptotic error consistency of the method. The results serve as a first step towards an analysis of the full gas-dynamic case.

The section on von-Neumann stability will consider the explicit time marching scheme as employed in most of the applications. We prove and discuss the stability domains of limiting inviscid methods and the full BGK method. In order to compare the BGK results, we discuss the stability domains of classical methods like Lax–Wendroff and Upwinding where the diffusion part is included via central differences. These classical methods mimic the procedure for the full gas-dynamic equations where a hyperbolic flow solver is supplemented by central differences for the viscous terms in an ad hoc way. The stability of the BGK scheme is superior over these methods for a wide range of parameters. The results for the inviscid schemes show the upwinding ability of the kinetic approach and its limitations in the case small values of a. We also find that the scheme exhibits an enlarged stability domain CFL $\lesssim 1.3$ in special cases of parameter choices. The results for the viscous method also show an enlarged stability domain for large values of the grid Reynolds number $|a|\Delta x/\nu$. In this under-resolved regime, the stability of the BGK scheme is fully controlled by the advection. The reason is an upwind mechanism which is extended to include the expressions for the diffusion. A similar stability gain was observed in Morton & Sobey (1993) also due to upwinding of the diffusive gradients.

The investigation of the local asymptotic consistency order leads to the observation of a third-order convergence behavior for coarse grids. In case of the advection–diffusion equation, we prove that this behavior is due to a strong non-uniform convergence on coarse grids which depends on the parameters of the equation and the scheme but is independent of the initial conditions. The reason is a special shape of the error constant which vanishes due to its non-linear behavior. This introduces pronounced grooves into the error landscape which results in locally high order of convergence. Due to this super-convergence, the BGK method presents itself as a high-order method in most simulations. Indeed, a related behavior has recently been described by Jameson (2004) for the BGK scheme applied to the full Navier–Stokes equations.

However, asymptotically for fine grids the method is proven to be first order in space and time. We also prove a general limit for the order of consistency of the BGK scheme which is due to the asymptotic behavior of the underlying kinetic equation. The superior error performance of the BGK scheme due to super-convergence is demonstrated in empirical investigations of the order of convergence and comparison with other methods. To our knowledge, this is the first time that empirical error curves of a convergence study are presented for the BGK scheme.

Beside the investigations on stability and consistency of the BGK scheme, another theme of this paper is the relevance of physically constructed numerical methods. The gas-kinetic BGK scheme is one example of a paradigm of computational science which states that the ‘physical’ requirements and processes should guide the construction of an efficient and accurate numerical scheme. Clearly, in the case of non-linear systems of partial differential equations where a rigorous mathematical theory is missing, the consideration of physical requirements is of great help. However, in this paper we want to go back to a simpler model in order to prove that the physical guidance which leads to the BGK scheme, indeed, helped to construct an accurate and robust numerical method. The result is a method with enlarged stability domain and minimized error constant.

The paper is organized as follows: Section 2 introduces the kinetic framework for the advection–diffusion equation. It is obtained according to the kinetic framework of the full gas-dynamic equations. In Section 3, this framework is used to construct the gas-kinetic BGK method. We will discuss the general structure and limiting cases. The stability of kinetic upwinding and viscous effect treatment in the BGK framework are investigated in Section 4. Section 5 on consistency starts with an example simulation and an empirical investigation of the error and order of convergence. The general limit and actual order of consistency of the BGK method are given and the existence of high-order error grooves is proven. At the end of this section, empirical error plots are given for BGK and classical methods.

2. Kinetic framework

Inspired by kinetic-gas theory, see e.g. Vincenti & Kruger (1965), the fundamental idea of kinetic schemes is to consider a kinetic model for (1.1) based on a distribution function f in a microscopic velocity space and its evolution equation. Here, we consider a continuous model with 1D for both the macroscopic space $\Omega \subset \mathbb{R}$ and velocity space $\mathbb{R}$, so we have

$$f\colon \mathbb{R} \times \Omega \times \mathbb{R}^{+} \to \mathbb{R}, \quad (c, x, t) \mapsto f(c, x, t), \tag{2.1}$$

where c denotes the microscopic velocity and (x, t) are space and time. The distribution function is assumed to be positive and integrable $f(\cdot, x, t) \in L^{1}(\mathbb{R})$ with the property

$$u(x, t) = \int_{\mathbb{R}} f(c, x, t)\mathrm{d}c, \tag{2.2}$$

which links the kinetic description to the macroscopic variable u.

2.1 *BGK equation*

In kinetic-gas theory, the evolution of the distribution function is described by the Boltzmann equation (Vincenti & Kruger, 1965). Here, we use the simplified BGK model (Bhatnagar *et al.*, 1954) as evolution equation for f which is given by

$$\partial_t f + c\partial_x f = \frac{1}{\tau}(g - f) \tag{2.3}$$

with a positive relaxation time $\tau \in \mathbb{R}$, $\tau > 0$. The relaxational right-hand side of the equation assures the relaxation towards an equilibrium distribution function g. Different choices for the equilibrium g lead to models for various physical equations, like gas dynamics or shallow-water equations. For the full gas dynamic equation the equilibrium is given by a Maxwell distribution. For the present case of the scalar advection–diffusion equation (1.1), we will use

$$g[u](c) = u\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}} \tag{2.4}$$

as in Kim *et al.* (1997). The equilibrium distribution depends on the macroscopic variable u and contains the new parameter $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$. The value ε corresponds to the temperature in energy density units in the full gas dynamic case. In the exponent the constant advection velocity a appears. In comparison with the Maxwell distribution, u corresponds to the density and a to the mean fluid velocity. The specific form of (2.4) is justified by the fact that it will reproduce the macroscopic equation (1.1). This choice is not unique but resembles the case of the full gas dynamic settings.

2.2 *Kinetic model*

The BGK equation (2.3) has to be connected to the macroscopic equation (1.1). This is done by integrating the equation over the velocity space. By construction, the equilibrium distribution satisfies

$$\int_{\mathbb{R}} g[u](c)\mathrm{d}c \equiv u, \tag{2.5}$$

so it reproduces the first moment of f and the first moment of the right-hand side of (2.3) will vanish. Integration of the left-hand side yields the transport equation

$$\partial_t u + \partial_x F = 0 \tag{2.6}$$

containing the flux

$$F(x, t) = \int_{\mathbb{R}} cf(c, x, t)\mathrm{d}c. \tag{2.7}$$

Of course, this equation is not closed, since we need to specify a relation between F and u. This relation will be obtained by means of the Chapman–Enskog expansion.

The Chapman–Enskog expansion was developed for the Boltzmann equation in order to derive explicit relations for transport coefficients of heat conduction and viscous stresses. See Vincenti & Kruger (1965) for a physical introduction. The approach of Chapman–Enskog considers an asymptotic expansion

$$f^{(N)}(c, x, t) = \sum_{n=0}^{N} \tau^n f_n(c, x, t) \tag{2.8}$$

of the distribution function in terms of the relaxation time τ up to an order N. The parameter τ is assumed to be small in an appropriate dimensionless scaling. The coefficients f_n will include the function u and its derivatives. The expansion has to satisfy the compatibility condition $u = \int f^{(N)}\,\mathrm{d}c$.

The first non-equilibrium correction is given in the case $N = 1$ with

$$f^{(1)} = (u - \tau(c-a)\partial_x u)\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}} \tag{2.9}$$

which includes the gradient of u. While the quantity u is always reproduced by the expanded distribution function we obtain different approximations results for the flux (2.7). The flux reads

$$F^{(1)}[u] = au - \frac{\varepsilon\tau}{2}\partial_x u \tag{2.10}$$

and leads to the closed transport equation

$$\partial_t u + a\partial_x u = \frac{\varepsilon\tau}{2}\partial_{xx} u. \tag{2.11}$$

By comparison with (1.1) we identify the transport coefficient

$$\nu = \frac{\varepsilon\tau}{2}. \tag{2.12}$$

This relation will be used from now on in the entire paper.

The above derivation of the advection–diffusion equation (1.1) demonstrates that the BGK equation (2.3) together with the equilibrium distribution (2.4) represent a valid kinetic model of the macroscopic equation. That is, for small values of τ the solution of the kinetic BGK equation will be close to the solution of the advection–diffusion equation. This link will be used to construct a numerical method for the advection–diffusion equation by solving the kinetic model.

The full compressible Navier–Stokes system consists of several equations while the kinetic model is still a scalar equation. By considering the scalar kinetic equation numerical methods for the full system are easier to construct.

3. BGK-kinetic numerical method

In this section, we will derive the gas-kinetic BGK numerical method for the scalar equation

$$\partial_t u + \partial_x F[u] = 0 \tag{3.1}$$

with the advection–diffusion flux function $F[u] = au - \frac{\varepsilon\tau}{2}\partial_x u$. We will follow the presentation of the method for the full compressible Navier–Stokes equations in Xu (2001). However, due to the simpler structure of the present equation we will be able to present the scheme more explicitly and uncover more structure. In most cases, this structure remains present in the full gas dynamic case, though it is not easily seen.

3.1 *Finite-volume update*

We consider a discretization of the space-time $\Omega \times [0, T]$ with constant spatial cell size Δx and time step Δt. The cell centers are given by $x_i = i\,\Delta x$ and the time instances by $t_n = n\,\Delta t$. Following a finite-volume approach, see e.g. Godlewski & Raviart (1996), we consider the cell mean values $\bar{u}_i^n$ omitting the bar in the sequel. The basic numerical method is an explicit finite-volume update

$$\bar{u}_i^{n+1} = \bar{u}_i^n + \frac{\Delta t}{\Delta x}(\tilde{F}^n_{i-\frac{1}{2}} - \tilde{F}^n_{i+\frac{1}{2}}) \tag{3.2}$$

based on numerical intercell fluxes $\tilde{F}^n_{i+\frac{1}{2}}$ which have to be modeled in a consistent and stable way, see e.g. LeVeque (2002). This paper will only consider explicit schemes. Implicit versions of the BGK scheme are discussed, e.g. in Xu *et al.* (2005).

Generally, the numerical flux is defined by $\tilde{F}^n_{i+\frac{1}{2}} = \frac{1}{\Delta t}\int_0^{\Delta t} F[u]^n_{i+\frac{1}{2}}\,\mathrm{d}t$. The idea of the BGK method is to use the (approximate) solution of the BGK equation (2.3) and the definition of the flux in (2.7) to define a numerical flux function by

$$\tilde{F}^n_{i+\frac{1}{2}} := \frac{1}{\Delta t}\int_0^{\Delta t}\int_{-\infty}^{\infty} cf(c, x_{i+\frac{1}{2}}, t^n + t)\mathrm{d}c\,\mathrm{d}t. \tag{3.3}$$

This approach follows the spirit of the Godunov method and its refinement in methods using generalized Riemann problems, see Ben-Artzi & Falcovitz (2003). The main difference of the BGK method is that it considers a kinetic model as auxiliary equation in order to obtain a simple solution of the generalized Riemann problem and that it formulates the Riemann problem and the resulting flux for the entire viscous equation.

Given a distribution $f(c, x, 0)$ the solution of the BGK equation (2.3) can be formally expressed by

$$f(c, x, t) = \frac{1}{\tau}\int_0^t g(c, x - cs, t - s)\mathrm{e}^{-\frac{s}{\tau}}\,\mathrm{d}s + \mathrm{e}^{-\frac{t}{\tau}} f(c, x - ct, 0) \tag{3.4}$$

which will be used below. The BGK scheme assumes that the values u_i^n are given and approximations to the cell-wise gradients $(\delta_x u)_i^n$ are obtained from a reconstruction procedure. Based on these macroscopic quantities, an appropriate distribution function is built and furnishes as initial condition for the solution (3.4). Note, that this solution is not explicit, since the equilibrium distribution g given by (2.4) depends on the solution f through u.

3.2 *Time-averaged interface distribution*

The time-averaged interface distribution at an interface $x_{i+\frac{1}{2}}$ for a time step Δt is calculated by

$$\begin{aligned} \tilde{f}\left(c, x_{i+\frac{1}{2}}\right) &= \frac{1}{\Delta t}\int_0^{\Delta t} f(c, x_{i+\frac{1}{2}}, t^n + t)\mathrm{d}t \\ &= \frac{1}{\Delta t}\int_0^{\Delta t}\left(\frac{1}{\tau}\int_0^t g(c, x_{i+\frac{1}{2}} - cs, t_n + t - s)\mathrm{e}^{-\frac{s}{\tau}}\,\mathrm{d}s + \mathrm{e}^{-\frac{t}{\tau}} f(c, x_{i+\frac{1}{2}} - ct, t_n)\right)\mathrm{d}t, \end{aligned} \tag{3.5}$$

where the solution (3.4) was used. In the BGK gas-kinetic approach, the functions $g(c, x, t)$ and $f(c, x, 0)$ are approximated by Taylor expansions around the point $(x_{i+\frac{1}{2}}, t_n)$. We write

$$g(c, x, t) = g(c, x_{i+\frac{1}{2}}, t_n) + (x - x_{i+\frac{1}{2}})\partial_x g(c, x_{i+\frac{1}{2}}, t_n) + (t - t^n)\partial_t g(c, x_{i+\frac{1}{2}}, t_n) \tag{3.6}$$

for the equilibrium distribution and

$$f(c, x, t_n) = f(c, x_{i+\frac{1}{2}}, t_n) + (x - x_{i+\frac{1}{2}})\partial_x f(c, x_{i+\frac{1}{2}}, t_n) \tag{3.7}$$

for the initial distribution. The values $g^n_{i+\frac{1}{2}}$ and $f^n_{i+\frac{1}{2}}$ and the derivatives $(\partial_x g)^n_{i+\frac{1}{2}}$, $(\partial_t g)^n_{i+\frac{1}{2}}$ and $(\partial_x f)^n_{i+\frac{1}{2}}$ will be given below. At this stage, it is tempting to introduce further approximations by evaluating the integrals according to quadrature rules since Δt is a small number. Such an approach is discussed in Ohwada (2002). The resulting schemes loose some of the qualities of the fully integrated scheme. The reason is that the quadrature rules give reasonable approximations only if the ratio $\Delta t/\tau$ is small, which might not be the case in all simulations.

By use of the Taylor expansions (3.6)/(3.7), the integrals in (3.5) can be explicitly evaluated to give the expression

$$\tilde{f}(c, x_{i+\frac{1}{2}}) = g^n_{i+\frac{1}{2}}\left(1 - W_1\left(\frac{\Delta t}{\tau}\right)\right) + f^n_{i+\frac{1}{2}} W_1\left(\frac{\Delta t}{\tau}\right) - \tau c(\partial_x g)^n_{i+\frac{1}{2}} W_2\left(\frac{\Delta t}{\tau}\right) - \tau c(\partial_x f)^n_{i+\frac{1}{2}} W_3\left(\frac{\Delta t}{\tau}\right) + \Delta t(\partial_t g)^n_{i+\frac{1}{2}} W_4\left(\frac{\Delta t}{\tau}\right), \tag{3.8}$$

where weight functions W_i depending on the ratio $\Delta t/\tau$ are introduced. The definition of the weight functions and some properties are summarized in the following lemma.

LEMMA 3.1 (Weight functions) Let $\omega = \Delta t/\tau$. The functions W_i, $i = 1, 2, 3, 4$, building the time-averaged interface distribution are given by

$$W_1(\omega) = \frac{1 - \mathrm{e}^{-\omega}}{\omega}, \quad W_2(\omega) = \frac{\omega - 2 + (\omega + 2)\mathrm{e}^{-\omega}}{\omega},$$
$$W_3(\omega) = \frac{1 - (1 + \omega)\mathrm{e}^{-\omega}}{\omega}, \quad W_4(\omega) = \frac{1 - \omega + \frac{1}{2}\omega^2 - \mathrm{e}^{-\omega}}{\omega^2}, \tag{3.9}$$

for $0 < \omega < \infty$. They form weight functions in the sense that $0 < W_i(\omega) < 1$ and $W_i \in C^\infty(\mathbb{R}^+)$. They satisfy the relations

$$W_1(\omega) + W_2(\omega) + W_3(\omega) = 1 \tag{3.10}$$

and

$$W_4(\omega) = \frac{1}{2} - \frac{1 - W_1(\omega)}{\omega}. \tag{3.11}$$

We skip the proof which consists of evaluating (3.5) and elementary calculations. For later use, we also introduce an additional weight function

$$W_5(\omega) = \frac{W_3(\omega)}{1 - W_1(\omega)}, \tag{3.12}$$

for which we also have $0 < W_5(\omega) < 1$. The shape of the five weight functions is shown in Fig. 1.

Inspection of the result (3.8) gives some interpretation of the role of the weight functions. They balance the influence of different components of the BGK solution. Namely, the collision-less free flight which solely convects $f^n_{i+\frac{1}{2}}$ and the dissipative mechanism of the relaxational part which drives the solution to the equilibrium $g^n_{i+\frac{1}{2}}$. For large values of ω, i.e. $\Delta t \gg \tau$, the equilibrium will dominate, while for $\Delta t \ll \tau$, that is small ω, the relaxational part has almost no influence. The interplay of the gradients is most involved, since the weight W_3 vanishes for $\omega \to 0$ as well as for $\omega \to \infty$.

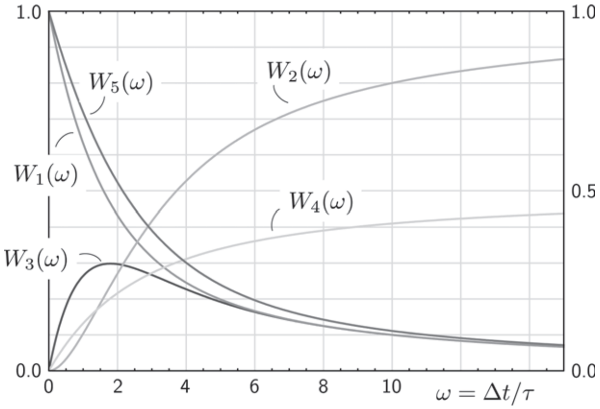

FIG. 1. Weight functions building the time-averaged interface distribution of a gas-kinetic BGK scheme. They balance the influence of free flight and collisonal dissipation depending on the value of $\Delta t/\tau$.

The definition of the general flux (3.3) is valid also for the BGK schemes for gas dynamics. Since we have not used more than the general BGK equation and its equilibrium, the time-averaged interface distribution (3.8) is also valid in the case of the Navier–Stokes system. Indeed, related expressions can be found in, e.g. Kim *et al.* (1997), Ohwada (2002) and Xu (2001). However, in these works the weight functions have not been explicitly identified.

3.3 *Kinetic evaluations*

Given a cell-wise linear reconstruction, the straight-forward evaluation of the numerical flux at the interface is difficult due to the discontinuities at the cell interfaces. This fact gave rise to the idea of a Riemann solver which handles the interface flux evaluation. The use of a underlying kinetic model also simplifies the evaluations at discontinuities considerably. In fact, due to the following definition the evaluation of the distribution function is well-defined even for spatial discontinuities. This makes the evaluation of the intercell flux straight-forward once the (eventually discontinuous) distribution function is found.

DEFINITION 3.1 (Discontinuous evaluation) Let the distribution function be integrable $f(\cdot, x, t) \in L^1(\mathbb{R})$ for almost all $x \in \Omega$, while $f(c, \cdot, t)$ be a piecewise C^k-function with respect to the grid introduced in Section 3.1. Let x_0 be the position of a discontinuity. The evaluation of f at x_0 is defined by

$$f(c, x_0, t) := \begin{cases} \lim\limits_{x \to (x_0)^-} f(c, x, t) & c < 0 \\ \lim\limits_{x \to (x_0)^+} f(c, x, t) & c > 0. \end{cases} \tag{3.13}$$

This definition is physically justified through the picture of particles approaching the discontinuity from both sides carrying the respective value of the distribution function. It also follows from the limit $t \to 0$ of the solution (3.4).

The procedure of discontinuous evaluations is the basis for all kinetic schemes. It enables us to evaluate the moments of the distribution function at discontinuities in an upwinding manner since the particle velocities are scanned in the moment integration.

3.4 *BGK scheme*

The actual BGK numerical method follows from (3.8) by specifying the equilibrium and initial distribution and their derivatives at the interface $x_{i+\frac{1}{2}}$. The distribution functions are completely defined once the function u and its gradient is specified. The following definitions of $f(c,x,t^n)$ and $g(c,x,t^n)$ correspond to the choices given in Xu (2001), which are also used in Li *et al.* (2004), May *et al.* (2005) and Xu *et al.* (2005).

DEFINITION 3.2 (BGK scheme construction)

(i) The initial distribution function $f(c,x,t^n)$ is defined cell-wise using the values u_i^n with gradient values $(\delta_x u)_i^n$ and the first-order Chapman–Enskog distribution (2.9). For the cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ we have

$$f(c,x,t^n)|_{x\in[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]} = (u_i^n + (x-x_i)(\delta_x u)_i^n - \tau(c-a)(\delta_x u)_i^n)\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}} \tag{3.14}$$

which assembles to a function in Ω which is cell-wise linear.

(ii) The equilibrium distribution is based on the interface values $u^n_{i+\frac{1}{2}}$ with one-sided equilibrium gradient values $(\widetilde{\delta_x u})^n_{i+\frac{1}{2},L/R}$ and the equilibrium distribution. For the interval $[x_i, x_{i+1}]$ we have

$$g(c,x,t)|_{x\in[x_i,x_{i+1}]} = \begin{cases} (u^n_{i+\frac{1}{2}} + (x-x_{i+\frac{1}{2}})(\widetilde{\delta_x u})^n_{i+\frac{1}{2},L} + (t-t^n)A^n_{i+\frac{1}{2}})\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}} & x < x_{i+\frac{1}{2}} \\ (u^n_{i+\frac{1}{2}} + (x-x_{i+\frac{1}{2}})(\widetilde{\delta_x u})^n_{i+\frac{1}{2},R} + (t-t^n)A^n_{i+\frac{1}{2}})\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}} & x > x_{i+\frac{1}{2}} \end{cases} \tag{3.15}$$

which leads to a function that is half-cell-wise linear and continuous in space.

(iii) The interface value is computed by

$$u^n_{i+\frac{1}{2}} := \int_{\mathbb{R}} f(c, x_{i+\frac{1}{2}}, t_n)\mathrm{d}c \tag{3.16}$$

using (3.14) while the gradient values follow from

$$(\delta_x u)_i^n = \frac{u^n_{i+1} - u^n_{i-1}}{2\varDelta x} \quad \text{and} \quad (\widetilde{\delta_x u})^n_{i+\frac{1}{2},L} = \frac{u^n_{i+\frac{1}{2}} - u^n_i}{\varDelta x/2}, \quad (\widetilde{\delta_x u})^n_{i+\frac{1}{2},R} = \frac{u^n_{i+1} - u^n_{i+\frac{1}{2}}}{\varDelta x/2}. \tag{3.17}$$

(iv) The unknown $A^n_{i+\frac{1}{2}}$ specifies the time derivative of g. It is obtained from the conservation condition

$$\int_{\mathbb{R}} \tilde{f}(c, x_{i+\frac{1}{2}})\mathrm{d}c = \int_{\mathbb{R}} \left(g^n_{i+\frac{1}{2}} + \frac{\varDelta t}{2}(\partial_t g)^n_{i+\frac{1}{2}} \right)\mathrm{d}c \tag{3.18}$$

with use of (3.8).

The idea of this definition is to use two different reconstructed functions of u for the initial distribution f and the equilibrium distribution g. Both reconstructions are shown in Fig. 2. The reconstruction

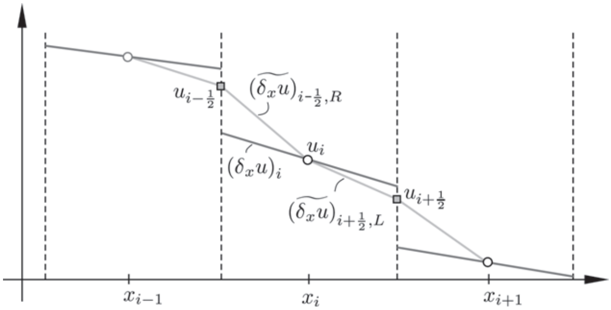
$(\widetilde{\delta_x u})_{i-\frac{1}{2},R}$
$u_{i-\frac{1}{2}}$
u_i
$(\delta_x u)_i$
$u_{i+\frac{1}{2}}$
$(\widetilde{\delta_x u})_{i+\frac{1}{2},L}$
x_{i-1}
x_i
x_{i+1}

FIG. 2. Two different reconstrucions used to formulate the initial and equilibrium distribution function for the BGK scheme. One reconstruction is discontinuous and based on cell-wise gradients $(\delta u)_i$. The other is continuous and uses interface equilibrium values and half-cell-wise gradients $(\widetilde{\delta u})_{i\pm\frac{1}{2},R/L}$.

for f is based on the values u_i^n and cell-wise gradients $(\delta_x u)_i^n$ which leads to a discontinuous cell-wise linear function. In full gas-dynamic calculations the gradients will need to be limited. From f follow interface values $u^n_{i+\frac{1}{2}}$, see (A.6), which are used to formulate an equilibrium reconstruction used for calculating g. It is a continuous half-cell-wise linear reconstruction through the cell and interface values.

The gradients $(\partial_x f)^n_{i+\frac{1}{2}}$ as well as $(\partial_x g)^n_{i+\frac{1}{2}}$ and $(\partial_t g)^n_{i+\frac{1}{2}}$ follow by cell-wise differentiation of (3.14) and (3.15). For the evaluation of these values and for the calculation of the interface value (3.16), the application of Definition 1 is required.

In principle, the time derivative $(\partial_t g)^n_{i+\frac{1}{2}}$ specified by $A^n_{i+\frac{1}{2}}$ could be transformed into an expression containing time derivatives of u which then could be expressed by spatial derivatives using the evolution equation (1.1) itself, see Ohwada (2002). However, this is quite cumbersome especially for the full compressible Navier–Stokes system. The relation (3.18) is used in Xu (2001). It links the coupled quantities f and g at least by the first moment.

For a more detailed discussion of the construction of the BGK gas-kinetic numerical method we refer to the paper Xu (2001).

Appendix Appendix A gives the details about the final integration to obtain the numerical flux $\tilde{F}_{i+\frac{1}{2}}$ from the interface distribution $\tilde{f}_{i+\frac{1}{2}}$. Here, we only give the result

$$\begin{aligned}\tilde{F}^{(\mathrm{BGK})}_{i+\frac{1}{2}} = {} & a u^n_{i+\frac{1}{2}}\left(1 - W_1\left(\frac{\Delta t}{\tau}\right)\right) + a\left(u^n_{i+\frac{1}{2},L} Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right) + u^n_{i+\frac{1}{2},R}\left(1 - Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right) W_1\left(\frac{\Delta t}{\tau}\right)\\ & - \frac{\varepsilon\tau}{2}\left[\left((\delta_x u)^n_i Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right) + (\delta_x u)^n_{i+1}\left(1 - Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right)\left(1 - W_2\left(\frac{\Delta t}{\tau}\right)\right)\right.\\ & + \left.\left((\widetilde{\delta_x u})^n_{i+\frac{1}{2},L} Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right) + (\widetilde{\delta_x u})^n_{i+\frac{1}{2},R}\left(1 - Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right) W_2\left(\frac{\Delta t}{\tau}\right)\right]\\ & - \frac{a^2\Delta t}{2}\left[\left((\delta_x u)^n_i Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right) + (\delta_x u)^n_{i+1}\left(1 - Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right) W_5\left(\frac{\Delta t}{\tau}\right)\right.\\ & + \left.\left((\widetilde{\delta_x u})^n_{i+\frac{1}{2},L} Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right) + (\widetilde{\delta_x u})^n_{i+\frac{1}{2},R}\left(1 - Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right)\left(1 - W_5\left(\frac{\Delta t}{\tau}\right)\right)\right], \qquad (3.19)\end{aligned}$$

where the ‘kinetic upwind or averaging’ weights $Z_{0,1}$ are given by

$$Z_0(\alpha) = \frac{1}{2}(1 + \operatorname{erf}(\alpha)), \tag{3.20}$$

$$Z_1(\alpha) = \frac{1}{2}\left(1 + \operatorname{erf}(\alpha) + \frac{1}{\alpha\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}\right) \tag{3.21}$$

with $\alpha = a/\sqrt{\varepsilon}$. The flux also uses the one-sided interface values

$$u^n_{i+\frac{1}{2},L} = u_i + \frac{\Delta x}{2}(\delta_x u)^n_i, \quad u^n_{i+\frac{1}{2},R} = u_{i+1} - \frac{\Delta x}{2}(\delta_x u)^n_{i+1}, \tag{3.22}$$

which follow from the reconstructed function u alone.

The first line in the flux corresponds to the advection part au of the flux, while the second and third line represent the dissipative gradient $\frac{\varepsilon\tau}{2}\partial_x u = \nu\partial_x u$. The last two lines introduce the second-order Lax–Wendroff type correction for the advection part $\frac{a^2\Delta t}{2}\partial_x u$. The final flux consists of an interplay of upwinding accounted for by the weight functions $Z_{0,1}(a/\sqrt{\varepsilon})$ and kinetic dissipation/transport mechanisms controlled by the weight functions $W_i(\Delta t/\tau)$. The quantities $u^n_{i+\frac{1}{2},L/R}$ and $(\delta_x u)^n_i$ are attributed with the transported initial distribution function and are activated for $\Delta t \ll \tau$. On the other hand, the quantities $u^n_{i+\frac{1}{2}}$ and $\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L/R}$ are introduced by the equilibrium distribution and control the flux for $\Delta t \gg \tau$. The kinetic upwind functions $Z_{0,1}(a/\sqrt{\varepsilon})$ choose the value $u^n_{i+\frac{1}{2},L/R}$ or $(\delta_x u)^n_{i/i+1}$ according to the direction of the advection velocity a.

The derived BGK numerical flux is valid for the advection–diffusion equation (1.1). However, the structure of the BGK flux will also be present in the numerical flux functions of the BGK scheme for the full gas-dynamic equations. Note that the BGK approach treats the hyperbolic advection and dissipative diffusion simultaneously in one framework. This is an interesting advantage of the BGK method. In many standard approaches a specialized numerical method is applied to either the hyperbolic or the parabolic part and the respective other mechanism is simply added in a more or less ad hoc way.

3.5 *Limiting cases*

It is suggestive to discuss limiting cases of the BGK scheme. The limiting cases will provide insight into the characteristics of kinetic schemes and will also be considered in the later sections on stability and consistency.

A major ingredient of the BGK scheme is the occurrence of different forms of kinetic upwinding through the weights $Z_{0,1}$. In order to study the mechanism it suffices to consider the pure advection case. Neglecting all gradients $\delta_x u \equiv 0$ makes the diffusive part vanish and only advection remains. The two limits $\Delta t/\tau \to 0$

$$\tilde{F}^{(\text{KIN1})}_{i+\frac{1}{2}} = a\frac{u^n_i + u^n_{i+1}}{2} + a\frac{u^n_i - u^n_{i+1}}{2}\operatorname{erf}\left(\frac{a}{\sqrt{\varepsilon}}\right) \tag{3.23}$$

and $\Delta t/\tau \to \infty$

$$\tilde{F}^{(\text{KIN2})}_{i+\frac{1}{2}} = a\frac{u^n_i + u^n_{i+1}}{2} + a\frac{u^n_i - u^n_{i+1}}{2}\left(\operatorname{erf}\left(\frac{a}{\sqrt{\varepsilon}}\right) + \frac{1}{a}\sqrt{\frac{\varepsilon}{\pi}}\,\mathrm{e}^{-\frac{a^2}{\varepsilon}}\right) \tag{3.24}$$

provide two examples of kinetic upwinding. These methods correspond to a classical upwind method which chooses the right or left-hand state as the upwind state according to the advection velocity. In the above case the switch between left and right is regularized by the use of erf-functions which stem from the weights $Z_{0,1}$. For $\varepsilon \to 0$ the classical upwind method is recovered.

A third method for advection is obtained from the full BGK equation with $\delta_x u_i \neq 0$ in the limit $\Delta t/\tau \to \infty$ which means essentially $\tau \to 0$. In this limit the inviscid equations are solved, due to the dominating dissipative part of the BGK evolution. As resulting kinetic flux we obtain

$$\begin{aligned}\tilde{F}^{(\textsc{Kin3})}_{i+\frac{1}{2}} &= a\left(u^n_{i+\frac{1}{2},L} Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right) + u^n_{i+\frac{1}{2},R}\left(1 - Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right)\\ &\quad - \frac{a^2 \Delta t}{2}\left(\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L} Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right) + \widetilde{(\delta_x u)}^n_{i+\frac{1}{2},R}\left(1 - Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right) \qquad (3.25)\end{aligned}$$

as flux for the inviscid advection equation (1.1) with $\nu \equiv 0$. Here, not only the value of u is upwinded through Z_0 but also the gradient of the second-order correction part. Though only for the inviscid case, the fluxes $\tilde{F}^{(\textsc{Kin1},\textsc{Kin2},\textsc{Kin3})}$ enable us to study the behavior of the kinetic upwinding mechanism in detail in Section 4.

To understand the treatment of the viscous part in the BGK scheme two limiting methods are of interest. For $\frac{a}{\sqrt{\varepsilon}} \gg 1$, we have for the interface value $u^n_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2},L}$, see (A.6) and for the equilibrium gradient $\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L} = (\delta_x u)^n_i$. In addition, the kinetic upwinding weights reduce to $Z_{0,1}(a/\sqrt{\varepsilon}) \to 1$. The resulting flux has the form

$$\tilde{F}^{(\textsc{FullUp})}_{i+\frac{1}{2}} = a u^n_{i+\frac{1}{2},L} - \frac{\varepsilon\tau}{2}(\delta_x u)^n_i - \frac{a^2 \Delta t}{2}(\delta_x u)^n_i \qquad (3.26)$$

and is called fully upwinded flux, since all quantities, values and gradients, are taken from the upwind side. It represents a natural and consequent extension of the upwind idea to the viscous case.

Another interesting case results from the limit $\omega \to 0$. In this case, the influence of the dissipation part of the BGK solution vanishes. The flux reads

$$\begin{aligned}\tilde{F}^{(\textsc{KinUp})}_{i+\frac{1}{2}} &= a\left(u^n_{i+\frac{1}{2},L} Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right) + u^n_{i+\frac{1}{2},R}\left(1 - Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right)\\ &\quad - \frac{\varepsilon\tau}{2}\left((\delta_x u)^n_i Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right) + (\delta_x u)^n_{i+1}\left(1 - Z_0\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right)\\ &\quad - \frac{a^2 \Delta t}{2}\left((\delta_x u)^n_i Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right) + (\delta_x u)^n_{i+1}\left(1 - Z_1\left(\frac{a}{\sqrt{\varepsilon}}\right)\right)\right) \qquad (3.27)\end{aligned}$$

and is named kinetically upwinded flux. In it all quantities are upwinded according to the kinetic weighting formulas $Z_{0,1}$. Except for the last second-order correction term, this method corresponds to the so-called kinetic flux vector splitting scheme which considers the collision-less kinetic equation for obtaining the numerical flux Deshpande (1986).

4. Stability

Kinetic schemes have been proven to satisfy stability properties in various settings. Mostly L^1-stability is considered, see e.g. Aregba-Driollet & Natalini (2000) and Perthame (1992). In Aregba-Driollet & Natalini (2000) discrete kinetic schemes are considered and in Perthame (1992) the distribution function is assumed to be compactly supported, which is not the case in the present BGK method. Indeed, in view of the CFL condition, the kinetic approach seems to be questionable. Due to the infinite integration over the velocity space in (3.3) and the transport behavior of the kinetic solution, informations are taken from locally linear reconstructed data where it is not valid anymore. Hence, the support of f in velocity space is a crucial quantity for stability. We recall that the parameter ε, which corresponds to the energy in gas dynamics, is responsible for the shape of the Gaussian (2.4). For small values of ε the distribution function tends to a Dirac delta exhibiting a smaller essential width. We expect the stability to be affected by the value of ε.

In the following, we will investigate the stability properties in a L^2-setting following the stability analysis of von-Neumann for linear problems, see e.g. Godlewski & Raviart (1996). The generic evolution function is given by

$$u_i^{n+1} = \mathcal{H}(u^n; i, \Delta t) \tag{4.1}$$

following the notation of LeVeque (2002). We investigate the effect of the method on harmonic waves with wave number k and amplitude $\hat{u}_k^n$

$$u_i^n = \hat{u}_k^n \, \mathrm{e}^{\mathrm{i}k\pi i \,\Delta x} \tag{4.2}$$

where i denotes the imaginary unit. Introducing this ansatz into the scheme we obtain

$$\hat{u}_k^{n+1} = G(\xi)\hat{u}_k^n \tag{4.3}$$

due to linearity. The amplification function $G(\xi) \in \mathbb{C}$ depends on $\xi = k\pi \,\Delta x$ and follows from $G(\xi) = \mathrm{e}^{-\mathrm{i}\xi i}\mathcal{H}(\mathrm{e}^{\mathrm{i}\xi i}; i, \Delta t)$. For stability of the method $\mathcal{H}$ we will consider the condition

$$\max_{\xi \in [-\pi,\pi]} |G(\xi)| \leqslant 1. \tag{4.4}$$

It would be possible to allow the spectral radius to be smaller than $1 + \mathrm{O}(\Delta t)$, however, this is not considered in present considerations.

4.1 *Characteristic parameters*

The advection–diffusion equation and the BGK scheme depend on the physical parameters a, ν, ε, τ, Δt and Δx. However, these parameters only occur in typical combinations and essentially only three numbers describe a certain setting completely. The following parameter combinations are frequently used throughout the paper.

Since we are concerned with explicit methods for convection-dominated flow the most important quantity is the Courant number

$$\lambda = \frac{a\,\Delta t}{\Delta x} \tag{4.5}$$

with the advection velocity a. Stability results are formulated as restrictions for λ from which the possible time step can be chosen. The influence of diffusion is measured by the quantity

$$\kappa = \frac{2\nu}{a\,\Delta x} = \frac{\varepsilon\tau}{a\,\Delta x} \tag{4.6}$$

which corresponds to an inverse grid Reynolds number

$$\mathrm{Re}_{\Delta x} = \frac{a\,\Delta x}{\varepsilon\tau}. \tag{4.7}$$

Restriction on the value of λ will typically depend on κ, i.e. on the value of diffusion, but also on the grid size and advection velocity. If physical units are considered κ is dimensionless.

In the BGK scheme, the diffusion ν is split into two parameters ε and τ which leads to a precise control of dissipation and transport mechanisms in the numerical method. Since ε is the thickness or variance of the equilibrium distribution function, $\sqrt{\varepsilon}$ is an additional velocity scale from the kinetic model which describes in average how fast the kinetic random walk particles are. In the numerical method, this quantity only appears in relation to the advection velocity

$$\alpha = \frac{a}{\sqrt{\varepsilon}}. \tag{4.8}$$

The quantity ε seems to be artificial for the advection case, but it connects the model to the full gas-dynamic case where ε represents the internal energy density. The parameter τ controls the influence of the kinetic dissipation. As a relaxation time it only appears in relation to the time step

$$\omega = \frac{\Delta t}{\tau} \tag{4.9}$$

in the numerical method. Note that also the physical viscosity of fluids is the product of a energy density and a relaxation time. The BGK scheme distinguishes between high diffusion due to the energy scale or relaxation time and chooses a specific numerical method for both cases.

The four parameters λ, κ, α and ω are related by $\lambda = \kappa\omega\alpha^2$ hence, only three of them control a certain setting.

4.2 *Classical explicit methods*

In this and the next subsections we will consider numerical methods for the full advection–diffusion equation. The investigations include the full BGK method (3.19) and its limiting cases, the fully upwinded method FULLUP (3.26) and the kinetically upwinded method KINUP (3.27). In order to compare the results with the results of classical methods we briefly introduce two classical methods and discuss their stability conditions.

In many simulations of viscous flow a standard hyperbolic upwind scheme is utilized for the convection part of the system. The dissipative second-order derivatives are then simply added to the finite-volume flux by central differences. We mimic this procedure for the advection–diffusion equation by adding a central difference to an upwind flux. The flux reads

$$\tilde{F}^{(\mathrm{UpCen})}_{i+\frac{1}{2}} = au_i - \frac{\varepsilon\tau}{2\Delta x}(u_{i+1} - u_i) \tag{4.10}$$

and is used in the finite-volume update (3.2). A related method uses the standard Lax–Wendroff flux for advection equations and adds a central difference similarly, yielding

$$\tilde{F}^{(\mathrm{LW})}_{i+\frac{1}{2}} = \frac{a}{2}(u_i + u_{i+1}) - \left(\frac{a^2\Delta t}{2\Delta x} + \frac{\varepsilon\tau}{2\Delta x}\right)(u_{i+1} - u_i) \tag{4.11}$$

as viscous Lax–Wendroff method.

The stability conditions for the UpCen method are given by

$$0 \leqslant \frac{a\,\Delta t}{\Delta x} \leqslant \frac{1}{1+\frac{\varepsilon\tau}{|a|\,\Delta x}} \tag{4.12}$$

which can be extended to negative values of a by introducing the case decision between au_i and au_{i+1} depending on the sign of a. For vanishing diffusion coefficient $\varepsilon\tau \rightarrow 0$ the classical result for the advection equation is recovered. For increasing values of $\varepsilon\tau$ the stability domain decreases. Assuming a case-decisive upwinding the stability condition can be written

$$\Delta t \leqslant \frac{1}{|a|+\frac{\varepsilon\tau}{\Delta x}}\,\Delta x \tag{4.13}$$

directly for the time step Δt. It is interesting to see that the diffusion enters the condition by means of a 'grid diffusion velocity' $\frac{\varepsilon\tau}{\Delta x}$ which increases the signal speed given by the advection. For small grid sizes the signal speed tends to infinity accounting for the parabolic nature of the equation.

For the analogous stability result of the LW method we obtain

$$\frac{|a|\,\Delta t}{\Delta x} \leqslant \sqrt{\left(\frac{\varepsilon\tau}{2a\,\Delta x}\right)^2+1} - \frac{\varepsilon\tau}{2|a|\,\Delta x} \tag{4.14}$$

which has the same qualitative properties as the UpCen result, but is a little less restrictive. For $\frac{\varepsilon\tau}{|a|\,\Delta x} \gg 1$ the stability condition in both cases, UpCen and LW, reduces to

$$\Delta t \leqslant \frac{\Delta x^2}{\varepsilon\tau} = \frac{\Delta x^2}{2\,\nu} \tag{4.15}$$

which is the standard result for explicit methods for diffusion equations.

The stability domains of UpCen and LW as functions of κ are depicted in Fig. 4. The less restrictive shape for the LW scheme is visible.

4.3 *Kinetic upwinding*

First, the effect of kinetic upwinding is investigated for the inviscid case $\tau = 0$, i.e. $\nu = 0$ only.

Theorem 4.1 (Stability of kinetic upwinding) Consider the kinetic fluxes Kin1 (3.23), Kin2 (3.24) and Kin3 (3.25) for the finite-volume update (3.2) in order to solve the inviscid ($\nu = 0$) advection equation (1.1) with $a, \varepsilon, \Delta t, \Delta x \in \mathbb{R}$ and $\varepsilon, \Delta t, \Delta x > 0$. The conditions

(Kin1)	$\frac{\|a\|\,\Delta t}{\Delta x} \leqslant \operatorname{erf}\left(\frac{\|a\|}{\sqrt{\varepsilon}}\right)$	$\frac{\sqrt{\varepsilon}\,\Delta t}{\Delta x} \leqslant \frac{2}{\sqrt{\pi}}$
(Kin2)	$\frac{\|a\|\,\Delta t}{\Delta x} \leqslant \frac{\|a\|}{a\operatorname{erf}\left(\frac{a}{\sqrt{\varepsilon}}\right)+\sqrt{\frac{\varepsilon}{\pi}}\,\mathrm{e}^{-\frac{a^2}{\varepsilon}}}$	$\frac{\sqrt{\varepsilon}\,\Delta t}{\Delta x} \leqslant \sqrt{\pi}$
(Kin3)	$\frac{\|a\|\,\Delta t}{\Delta x} \leqslant \lambda^{(\star)}\left(\frac{a}{\sqrt{\varepsilon}}\right)$	$\frac{\sqrt{\varepsilon}\,\Delta t}{\Delta x} \leqslant 7.84\ldots$

(4.16)

are necessary and sufficient (first column) and necessary (second column) for stability of the three respective methods in the sense of von-Neumann. $\lambda^{(\star)}$ is given in (B.18).

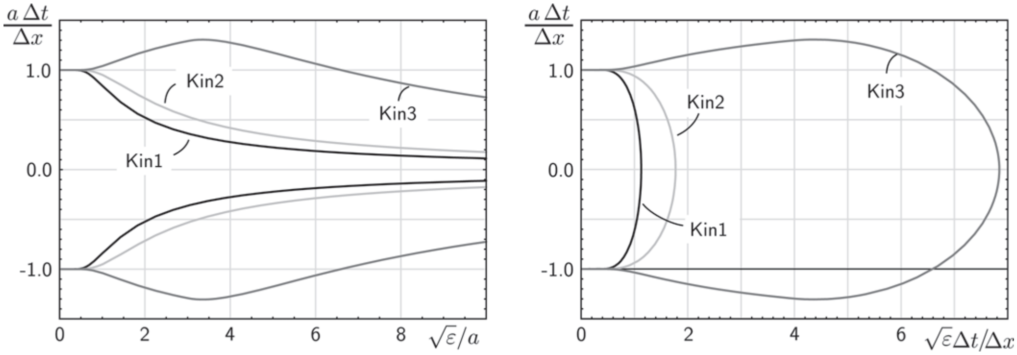

FIG. 3. Stability domains for kinetic upwinding methods KIN1, KIN2, and KIN3 for pure advection. Both plots show the same functions with different independent variables. Due to the increased width of the distribution function, the stability is reduced for large values of ε. The method KIN3 shows a non-monotone behavior and admitts values $|\lambda| > 1$.

The proof can be found in Appendix B.1.

The results reflect the influence of the parameter $\sqrt{\varepsilon}$ on the stability of the kinetic schemes. For large values of $\sqrt{\varepsilon}$ compared to the advection velocity a, the stability domain decreases. This corresponds to the fact that the distribution function has a large width and, hence, the kinetic flux has a large domain of dependence. This behavior becomes most obvious when the kinetic Courant number $\sqrt{\varepsilon}\,\Delta t/\Delta x$ is considered: The values in the right column of (4.16) give explicit limitations on the width of the distribution function beyond which stability is not anymore assured. In the limit of small values of ε all considered methods reduce to the case decisive upwind method with stability condition $|a|\,\Delta t/\Delta x \leqslant 1$. In the case of KIN1 and KIN2 the kinetic upwinding monotonically reduces the stability for increasing values of ε, hence, the CFL condition $|\lambda| < 1$ represents a necessary condition. However, the relation $\lambda^{(\star)}\left(\frac{a}{\sqrt{\varepsilon}}\right)$ for the method KIN3 shows a non-monotone behavior and admits Courant numbers $|\lambda| > 1$ and still guaranteeing L^2-stability. Note that this does not contradict the general CFL condition, since the KIN3 method works with a five-point stencil.

In the case of full gas dynamics the parameter ε corresponds to the energy and $\alpha = a/\sqrt{\varepsilon}$ is the Mach number of the flow. The present result suggests that the KIN3 method shows improved stability properties for medium to low Mach number flows. However, the extrapolation is difficult in this case since the sound wave modes of the full gas dynamics are not captured in the present analysis.

Figure 3 displays the stability domains of the three kinetic upwinding methods KIN1, KIN2 and KIN3. The left-hand side shows the marginal value of λ depending on the inverse Mach number $\sqrt{\varepsilon}/a$. The asymptotically decreasing curves are clearly visible. The right-hand side shows the same function but with the kinetic Courant number $\sqrt{\varepsilon}\,\Delta t/\Delta x$ as independent variable. Here, the maximal value of λ_ε is visible at which the stability interval for λ shrinks to zero. Both plots also clearly show the maximal stability range $|\lambda| \lesssim 1.3$ for the KIN3 flux for values $\sqrt{\varepsilon}/a \approx 3$.

4.4 *Full viscous upwinding*

The third curve in Fig. 4 corresponds to the stability domain of the fully upwinded scheme (3.26). For this method we have the following result.

THEOREM 4.2 (Stability of full upwinding) Consider the finite-volume update (3.2) for the full advection–diffusion equation with the fully upwinded flux (3.26) and $a, \tau, \varepsilon, \Delta t, \Delta x \in \mathbb{R}$ as well as ε,

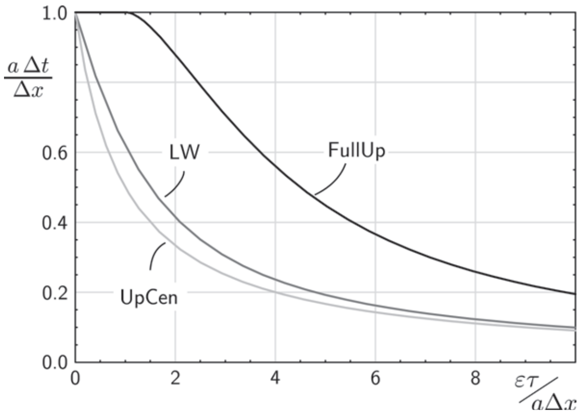

FIG. 4. Comparison of the stability domains of the classical methods for viscous equations: upwind with viscous central differences (UPCEN) and Lax–Wendroff with viscous central differences (LW), and the fully upwinded scheme (FULLUP). The FULLUP method is a limiting case of the BGK scheme and shows superior stability properties.

$\varDelta t, \varDelta x > 0$. The conditions

$$\begin{aligned} &\text{if } \frac{\varepsilon\tau}{|a|\varDelta x} \leqslant 1\text{: } 0 \leqslant \frac{a\varDelta t}{\varDelta x} \leqslant 1 \\ &\text{if } \frac{\varepsilon\tau}{|a|\varDelta x} > 1\text{: } 0 \leqslant \frac{a\varDelta t}{\varDelta x} \leqslant \lambda^{(\text{crit})}\left(\frac{\varepsilon\tau}{|a|\varDelta x}\right) < 1 \end{aligned} \tag{4.17}$$

together with $\tau > 0$ are necessary and sufficient for stability in the sense of von-Neumann. $\lambda^{(\text{crit})}$ is found by inverting (B.28) given in the proof.

See Appendix B.2 for the proof.

The results for FULLUP show a considerable improvement of stability in comparison with the standard methods UPCEN and LW. Figure 4 displays the marginal values for λ obtained numerically from (B.28). Most impressive is the independence of λ from the value of $\kappa = \frac{\varepsilon\tau}{|a|\varDelta x}$, for $\kappa < 1$. The remaining condition $0 \leqslant \lambda \leqslant 1$ in this regime shows that stability is solely controlled from the advection part with no influence from the viscous term. Hence, for small values of the diffusion and/or large values of the grid size $\varDelta x$, i.e. large values of the grid Reynolds number (4.7), the computation can be conducted as if there was no diffusion expression. Even in the range $\kappa > 1$ stability is improved due to a slower decay rate of $\lambda^{(\text{crit})}$. However, in the limit $\kappa \to \infty$ the diffusion condition (4.15) is recovered for the fully upwinded method as well.

The flux of the FULLUP scheme (3.26) can be generalized to quasi-linear systems for which a numerical method based on locally linearized equations, like the Riemann solver of Roe, is available. In those cases the upwinding of the diffusive gradients can be realized by a characteristic decomposition according to the Jacobian of the flux function. The formulation and investigation of such a method is left for future work.

4.5 *BGK method*

The amplification function of the full BGK method (3.19) is highly involved and a detailed analytical investigation appears to be forbiddingly complicated. Thus, we will only state a non-explicit stability statement and present and discuss a numerical evaluation of the stability domain afterward.

THEOREM 4.3 (Stability of the BGK method) Consider the finite-volume update (3.2) with the gas-kinetic BGK flux (3.19) for solving the advection–diffusion equation (1.1). Assume $a, \tau, \varepsilon, \Delta x, \Delta t \in \mathbb{R}$ and $\varepsilon, \Delta x, \Delta t > 0$. Then, for sufficiently small values of Δt the resulting BGK scheme is stable in the sense of von-Neumann for all a, ε and for all $\tau \geqslant 0$.

The proof is given in Appendix B.3.

The amplification function of the BGK scheme depends on the parameter $\{\lambda, \kappa, \alpha\}$ or equivalently $\{\lambda, \omega, \alpha\}$. Hence, the marginal value of λ will depend on κ and α and can be found numerically by searching for the maximal value of $|G(\xi)|$. The result is plotted over the variable $\log(\kappa)$ in different view graphs varying the parameter $\sqrt{\varepsilon}/a$, see Fig. 5. In the figure, the BGK stability domain is compared with that of the classical methods UPCEN (4.10) and LW (4.11), as well as with that of the limiting cases FULLUP (3.26) and KINUP (3.27). The logarithmic scale is chosen to focus on the regime of small values of κ which corresponds to large values of $\mathrm{Re}_{\Delta x}$. Due to the kinetic background of the BGK method the diffusion coefficient $\nu = \frac{\varepsilon\tau}{2}$ is split into two independent parameters, the relaxation time τ and the energy scale ε. Since the stability domains of the classical methods UPCEN and LW, as well as the limiting scheme FULLUP, depend solely on κ as a fixed combination of τ and ε, these curves do not vary among the different plots of Fig. 5. However, for the BGK scheme τ and ε can be chosen independently.

For large values of α or small values of $\sqrt{\varepsilon}/a$ the BGK scheme (3.19) and the kinetically upwinded scheme (3.27) reduce to the fully upwinded case FULLUP. This can be observed in the upper left plot of Fig. 5, where the curves of these methods coincide. In addition, the full BGK scheme reduces to the KINUP scheme for small values of $\omega = \Delta t/\tau$, or equivalently, large values of κ. Indeed, from the plots of Fig. 5 we see the curves meeting beyond a certain value of κ which is varying for different choices of $\sqrt{\varepsilon}/a$. In contrast to KINUP the full BGK scheme includes the dissipative part of the solution of the BGK equation which is activated for small values of κ. The plots in the figure exhibit the stability gain of the dissipative mechanism. While the stability domain for small κ and increasing $\sqrt{\varepsilon}/a$ is reduced for KINUP, it is enlarged in the case of the full BGK scheme.

To some extent the full BGK scheme combines the stability properties of the inviscid kinetic scheme KIN3 and the viscous scheme FULLUP into a numerical method whose stability domain is clearly superior over classical schemes like UPCEN and LW. It partially assures stability beyond $|\lambda| \leqslant 1$ and allows a advection controlled time step decoupled from diffusion for small κ. However, the stability domain collapses for large values of $\sqrt{\varepsilon}/a$, reflecting the behavior found in case of the inviscid kinetic methods KIN1, KIN2 and KIN3. Numerical evaluation shows that the barrier for the kinetic Courant number $\lambda_\varepsilon \leqslant 7.84\ \ldots$ is also present for the BGK if $\kappa \to 0$. However, there seems to be no ultimate restriction for λ_ε in the case $\kappa > 0$.

Extrapolated to the full gas-dynamic case the result suggests the BGK scheme to be a very robust numerical method for high values of the grid Reynolds number and not too small Mach number. Stability might be reduced for low Mach number flow, which is also observed in May *et al.* (2005) for a full gas-dynamic viscous BGK method.

5. Consistency

While the order of consistency of the classical methods UPCEN or LW is more or less obvious (first and second order in space, respectively, first order in time), it is relatively difficult to state for the full BGK method (3.19) due to the different use of gradients and implicit dependence on Δt through the weights $W_i(\omega)$. Some authors claim that it is a ‘second-order method’, which is in fact wrong or at least not very precisely stated. Furthermore, empirical investigations on the order of convergence of the BGK scheme cannot be found in the literature.

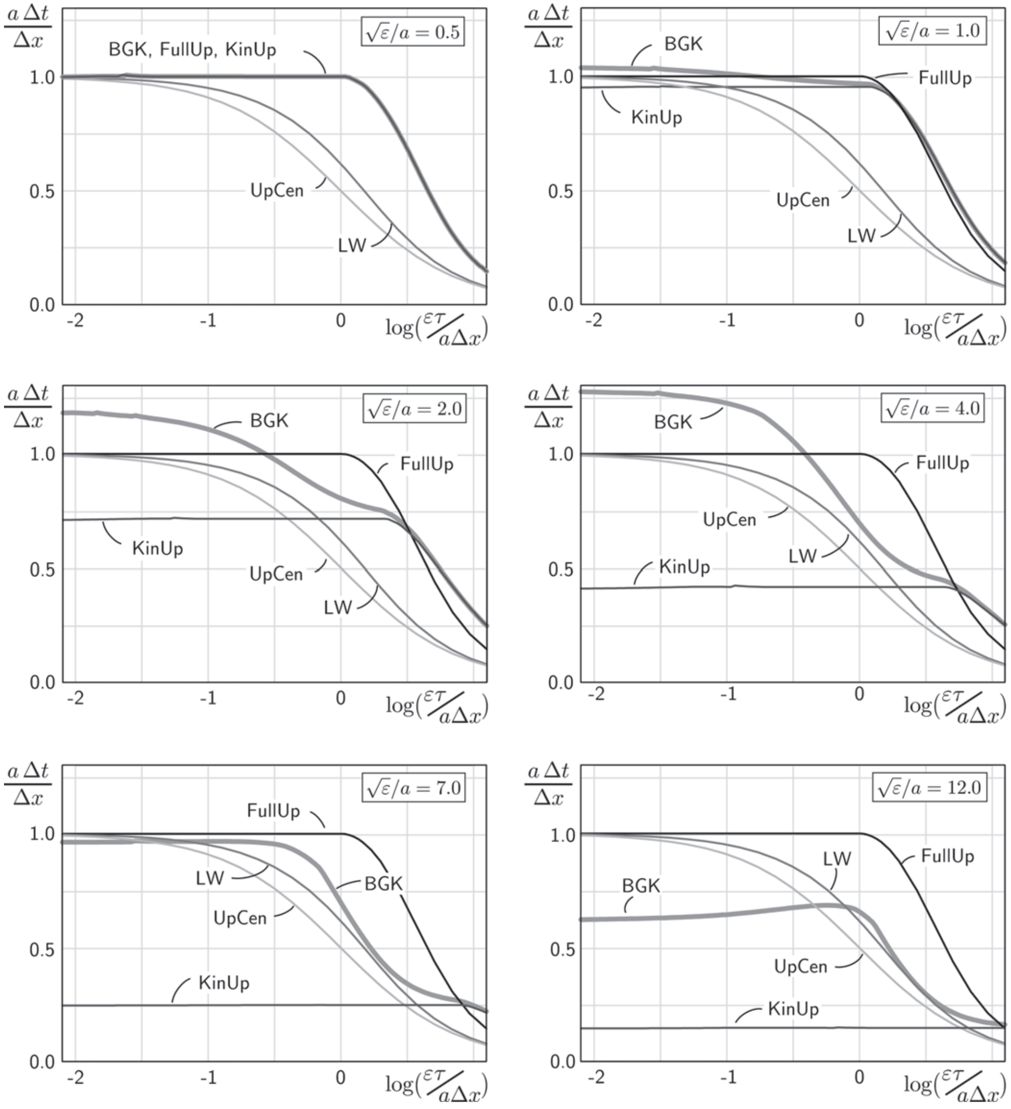

FIG. 5. The stability domain for the BGK scheme compared to classical methods and limiting cases. UPCEN and LW are classical methods where a standard advection scheme is supplemented with central differences for the diffusion. The FULLUP scheme results from BGK in the case $a/\sqrt{\varepsilon} \gg 1$, while the KINUP method represents the case $\Delta t/\tau \to 0$. Up to moderate values of $\sqrt{\varepsilon}/a$ the BGK scheme exhibits a superior stability domain. The curves show symmetric behavior in the range $a < 0$.

In this section, we will present the order of consistency in space and time including error constants and empirical investigations. We also prove the phenomenon of super-convergence for the BGK method.

5.1 *Example*

To give an impression of the convergence and consistency behavior of the BGK scheme, we start with an example. Consider the advection–diffusion equation (1.1) and the numerical method (3.2) with the

BGK flux (3.19). We apply the method in the domain $x \in [-1, 3]$ with periodic boundary conditions to the the initial conditions

$$u_0(x) = 4 + \frac{8}{\pi} \sin\left(\frac{\pi}{L} x\right) + \frac{16}{3\pi} \sin\left(3\frac{\pi}{L} x\right) \tag{5.1}$$

with $L = 2$. The exact solution of the advection–diffusion equation is given by

$$u(x,t) = 4 + \frac{8}{\pi}\left(\mathrm{e}^{-\frac{\pi^2}{L^2}\nu t} \sin\left(\frac{\pi}{L}(x - at)\right) + \frac{2}{3}\,\mathrm{e}^{-\frac{9\pi^2}{L^2}\nu t} \sin\left(3\frac{\pi}{L}(x - at)\right)\right), \tag{5.2}$$

where we have $\nu = \frac{\varepsilon\tau}{2}$ for the viscosity coefficient according to the kinetic model. In this numerical experiment we use $\varepsilon = 1$ and $\tau = 0.2$, hence $\nu = 0.1$, and $a = 2$ for the advection velocity. The result of the BGK method at time $t = 0.7$ for two different grids $\Delta x = 0.1$ (triangles) and $\Delta x = 0.04$ (squares) is shown at the left-hand side of Fig. 6. The right-hand side of the figure shows the corresponding result of the Lax–Wendroff scheme (4.11). The time step was chosen to be $\Delta t = 0.0225$ and $\Delta t = 0.00514$, respectively, and the figure shows only the section $x \in [1.8, 3.0]$. Interestingly, the BGK solution approximates the exact curve (solid line) much better on the coarse grid than on the fine grid. The coarse BGK solution is even better than the fine LW result which shows a monotone convergence behavior. It seems that this non-monotone behavior of the BGK scheme can also be found in the full gas-dynamic case, see Jameson (2004).

To investigate the behavior further we conduct several calculations with $\varepsilon = 1$, $\tau = 0.01$ and $a = 0.5$ and plot the error against the number of grid points or the value of the time step used in the simulations. The time step and grid size are coupled according to

$$\Delta t = \frac{\mathrm{CFL}}{a + \frac{\varepsilon\tau}{\Delta x}} \Delta x, \tag{5.3}$$

where CFL > 0 has to be chosen. For CFL $\leqslant 1$ this coupling realizes the stability condition of the UPCEN method (4.10) which is the most restrictive among the presented methods for $a/\sqrt{\varepsilon} \gtrsim 0.2$. This procedure assures that the same time step can be used for all methods in this investigation without loss of stability. The error is calculated using

$$\mathrm{err} = \sum_{i=1}^{N} \Delta x |u_i^{(\mathrm{num})} - u^{(\mathrm{ex})}(x_i)| \tag{5.4}$$

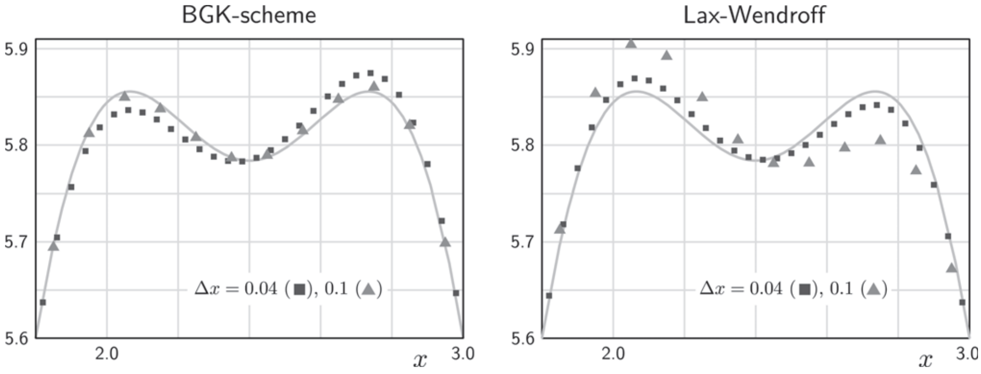

FIG. 6. Sketch of the approximation ability of the BGK and LW method for a sine-wave-type solution (solid line). The BGK method gives a better approximation on the coarse grid (triangles) than on the fine grid (squares).

which corresponds to the L^1-norm and can be viewed as a function of Δt, Δx or N, the number of grid points. The empirical order of convergence is defined by

$$\mathrm{EOC}_N = \frac{\log\left(\frac{\mathrm{err}_{N_1}}{\mathrm{err}_{N_2}}\right)}{\log\left(\frac{N_2}{N_1}\right)}, \quad \text{or} \quad \mathrm{EOC}_{\Delta t} = \frac{\log\left(\frac{\mathrm{err}_{\Delta t_1}}{\mathrm{err}_{\Delta t_2}}\right)}{\log\left(\frac{\Delta t_1}{\Delta t_2}\right)} \tag{5.5}$$

for convergence with respect to the time step Δt or with respect to the number of grid points N, i.e. the grid size Δx. Since Δt and Δx are coupled in a non-linear way the order of convergence is expected to be different with respect to space and time. For that reason we display the error curves separately with respect to N and with respect to Δt. The upper row of Fig. 7 shows these two plots with error curves for different values of CFL $= 0.25$, 0.5, 0.6, 0.7, 0.8, 0.9 in (5.3) and the grid sizes between $N = 20$ and 1000. The dots and curves in both plots belong to the same simulations but are shown with respect to N and Δt, respectively. The error curves exhibit a strong non-monotone behavior in correspondence to the findings in Fig. 6. Some coarse grid and large time step calculations show considerably smaller errors than finer calculations. Most of the curves even show two minima, most pronounced in the curve for CFL $= 0.8$.

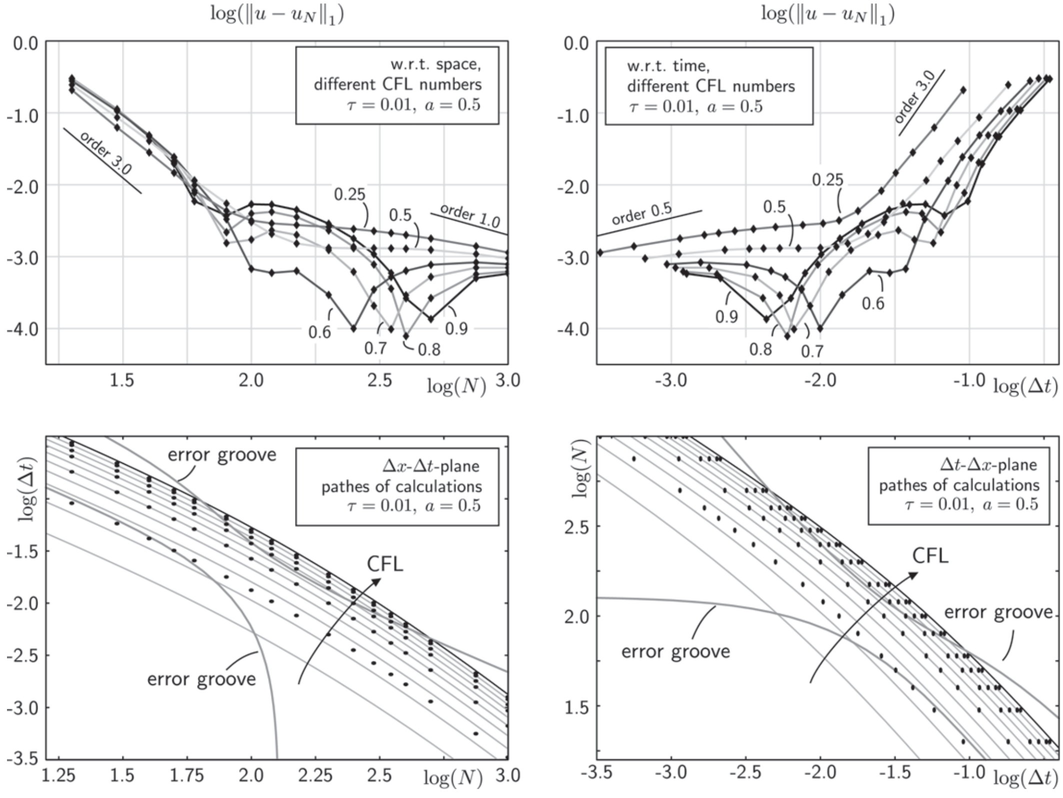

FIG. 7. L^1-error curves for the BGK scheme with respect to grid size N (upper left) and time step Δt (upper right) for various CFL numbers. The strong non-monotone behavior can be explained by pronounced grooves in the error landscape which are indicated in the lower row.

A reliable empirical order of convergence is difficult to obtain from those error curves. Asymptotically for fine grids and small time steps the error tends to reduce with first order in space and order 0.5 in time. However, for coarse grids and large time steps the BGK method performs like a third-order method both with respect to space and time discretization. In the next subsections we will prove this behavior and then return to Fig. 7 giving a detailed explanation using the lower row of the figure.

We will use the local error of consistency of a numerical method defined by

$$e(\varDelta t, \varDelta x, u^{(\mathrm{ex})}) = \left\| \frac{1}{\varDelta t}(u^{(\mathrm{num})} - u^{(\mathrm{ex})})\Big|_{t=\varDelta t} \right\|. \tag{5.6}$$

Together with appropriate stability properties of the scheme the boundedness of the local error can imply convergence of the method, see Godlewski & Raviart (1996) or LeVeque (2002). In those cases the order of the local error e as defined above with respect to $\varDelta t$ or $\varDelta x$ will also be the order of convergence. Sometimes the factor $\frac{1}{\varDelta t}$ is not used in the definition of the local error in other works. In the present case considering a linear advection–diffusion equation and a linear BGK scheme, the Lax-Equivalence Theorem applies and the stability results in the previous section imply convergence of the BGK method once consistency is shown.

5.2 *General limit*

The question for consistency of the gas-kinetic BGK method is reasonable, since its flux is based on a different equation than that which is subject to solve, namely the BGK equation. It is not a priori obvious that the flux based on the solution of the BGK model of Boltzmann's equation (3.3) with (2.3)/(2.4) is consistent with the advection–diffusion equation (1.1). Of course, the Chapman–Enskog expansion clarifies the situation and the result (C.4) in Section C.1 gives the quantitative relation between the two models. However, this very result also gives a limit in the accuracy of BGK schemes.

PROPOSITION 5.1 (Consistency limit of the BGK scheme) Consider a hypothetical BGK scheme (3.2) with (3.3) in which the 'full exact' solution of the kinetic equation is used to calculate the flux. Assume the numerical solution of this scheme $u^{(\mathrm{BGK})}$ and the exact solution of the advection–diffusion equation $u^{(\mathrm{ex})}$ to be smooth and $L^1(\mathbb{R})$-integrable. Then, the local consistency error is given by

$$\left\| \frac{1}{\varDelta t}(u^{(\mathrm{BGK})} - u^{(\mathrm{ex})})\Big|_{t=\varDelta t} \right\|_{L^1(\mathbb{R})} \leqslant C_1 \varDelta x^p + C_2 \tau \varDelta x^q + C_3 \tau^2 \varDelta t, \tag{5.7}$$

for small τ with $C_{1,2,3} > 0$. Here, p and q are the spatial approximation orders of the initial reconstruction of $u^{(\mathrm{BGK})}$ and its gradient.

The proof can be found in Appendix C.1.

A similar result is known for the approximation of inviscid equations with the kinetic flux vector splitting method which considers a non-dissipative kinetic equation to formulate the flux, see Deshpande (1986). However, here we consider the dissipative kinetic model and also the viscous equation (1.1). The result reflects the fact that there is indeed a difference between the solution of the BGK model and the actual evolution equation.

The result (5.7) considers the direct update after one time step. In principle the flux evaluation can be included into a Runge–Kutta time integration which then yields high time accuracy. In these cases also high accuracy of the spatial approximation is needed, since the error is swapped in case of a coupling of time step and grid size.

5.3 *Asymptotic error*

If the solution is assumed to be sufficiently smooth, the local consistency error (5.6) can be directly calculated for a numerical method using Taylor expansion. The proofs of the following results are skipped, since they were found mostly by automated series expansion provided by the software Mathematica. We proceed with presenting the results for the kinetic methods for the inviscid equation derived from the full BGK scheme.

LEMMA 5.1 (Consistency of inviscid kinetic schemes) Consider the numerical fluxes KIN1 (3.23), KIN2 (3.24) and KIN3 (3.25) for the finite-volume update (3.2) in order to solve the inviscid advection equation (1.1) with $\nu \equiv 0$ and $a \in \mathbb{R}$. The exact solution u is assumed to be smooth and let Δx, Δt, $\varepsilon \in \mathbb{R}$ be positive. Then, the local consistency error of the three methods is given by

$$\left\| \frac{1}{\Delta t}\left(u^{(\mathrm{KIN1},\mathrm{KIN2})} - u\right)\Big|_{t=\Delta t} \right\|_\infty \leqslant \left(\frac{a^2}{2}\Delta t + \frac{a}{2} h_{1,2}\left(\frac{a}{\sqrt{\varepsilon}}\right)\Delta x \right) \left\| u''\big|_{t=0} \right\|_\infty \tag{5.8}$$

and

$$\left\| \frac{1}{\Delta t}\left(u^{(\mathrm{KIN3})} - u\right)\Big|_{t=\Delta t} \right\|_\infty \leqslant |a| \left(\frac{a^2}{6}\Delta t^2 + \frac{a}{4} h_2\left(\frac{a}{\sqrt{\varepsilon}}\right)\Delta x\, \Delta t + \frac{1}{12}\Delta x^2 \right) \left\| u'''\big|_{t=0} \right\|_\infty, \tag{5.9}$$

respectively, for small Δt and Δx. The functions $h_{1,2}$ are defined by

$$h_1(\alpha) = \operatorname{erf}\alpha \quad \text{and} \quad h_2(\alpha) = \operatorname{erf}\alpha + \frac{1}{\alpha\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}. \tag{5.10}$$

As expected, the KIN1 and KIN2 methods are only first order in space and time, while the method KIN3 which includes the second-order gradient term shows a second-order error constant. According to the stability result for the inviscid methods, the time step can be taken essentially proportional to the grid size. Hence, in empirical investigations of the order of convergence KIN1 and KIN2 will perform as asymptotically first order, and KIN3 as asymptotically second order. Together with the stability results for the KIN3 method we conclude that this method (gas-kinetic BGK scheme with $\tau \to 0$) appears to be an accurate and robust solver for inviscid equations.

For the full gas-kinetic BGK scheme for the viscous equation we have the following lemma.

LEMMA 5.2 (Consistency of the BGK scheme) Consider the BGK numerical flux (3.19) for the finite-volume update (3.2) in order to solve the viscous equation (1.1) with $\nu = \frac{\varepsilon\tau}{2}$ and $\varepsilon, \tau, a \in \mathbb{R}$. The exact solution u is assumed to be smooth and let Δx, Δt, ε, $\tau \in \mathbb{R}$ be positive. Then, the local consistency error of the BGK scheme is given by

$$\begin{aligned} &\left\| \frac{1}{\Delta t}\left(u^{(\mathrm{BGK})} - u\right)\Big|_{t=\Delta t} \right\|_\infty \\ &\quad \leqslant \frac{\varepsilon\tau}{4} \operatorname{erf}\left(\frac{|a|}{\sqrt{\varepsilon}}\right) \Delta x \left\| u'''\big|_{t=0} \right\|_\infty + \frac{\varepsilon\tau}{8}\Delta t \left(4|a| \left\| u'''\big|_{t=0} \right\|_\infty + \varepsilon\tau \left\| u^{IV}\big|_{t=0} \right\|_\infty \right), \end{aligned} \tag{5.11}$$

for small Δt and Δx.

This result shows first order in time and space. In order to find first order of convergence in the empirical investigations of Fig. 7 we have to recall that for small time steps and fine grids the grid size is coupled with the time step by $\Delta t \sim \Delta x^2$ or $\Delta x \sim \sqrt{\Delta t}$ due to the time step selection in (5.3). Accordingly, if we look at the error with varying grid size as in the upper left of Fig. 7, Δt is replaced by Δx^2 in (5.11) and the asymptotic error is of first order in space as observed. However, looking at the error with varying time step requires to substitute Δx by $\sqrt{\Delta t}$ which results in the order of 0.5 visible in the figure.

The result in (5.11) can also be compared with the general limit (5.7). In the construction of the BGK scheme a linear reconstruction of the variable u has been used which introduces the orders $p = 2$ and $q = 1$. The second-order approximation of the function is left out in (5.11), but the first-order approximation of the gradient is present with its factor τ. The time error consists of two parts, one proportional to $\tau\,\Delta t$ and the other proportional to $\tau^2\,\Delta t$. The latter corresponds to the optimal expression in (5.7) and cannot be expected to be improved. The first corresponds to a simplified argument when the Taylor expansion of the Chapman–Enskog distribution was presented in (3.14). This term vanishes if the Taylor expansion is extended to the gradient of the Chapman–Enskog distribution in the derivation of the scheme.

The somewhat low order of the BGK scheme may be disappointing. However, the investigations in Fig. 7 show that the asymptotically low order is accompanied by an initial high order of convergence, which may be described as super-convergence.

5.4 *Super-convergence*

The high-order convergence of the BGK scheme for coarse grids is caused by a special shape of the consistency error. This results in the vanishing of the leading error constant along several lines in the parameter space independent of the solution itself.

THEOREM 5.1 (BGK Error Grooves) The local consistency error of the BGK scheme (3.2) with (3.19) for the advection–diffusion equation (1.1) satisfies

$$\left\| \frac{1}{\Delta t}\left(u^{(\mathrm{BGK})} - u\right)\Big|_{t=\Delta t} \right\|_{\infty} \leqslant C\,\frac{\Delta x^4}{\Delta t}, \tag{5.12}$$

for small $\Delta t, \Delta x > 0$ with $C = \mathrm{O}(1)$ along lines given by $G(\Delta x, \Delta t) = 0$, see (C.17). These error grooves are independent of the solution u.

The proof is given in Appendix C.2.

The existence of so-called error grooves changes the shape of the error landscape of the Δt-Δx-plane considerably. In order to visualize the phenomenon Fig. 8 shows two plots of an error landscape over the Δt-Δx-plane. Imagine a function $\mathrm{err}(\Delta t, \Delta x)$ which represents the error of a numerical method depending on the time step and the grid size. For a smooth solution this function is expected to depend smoothly on Δt and Δx. If we have $\mathrm{err} \sim C_1\,\Delta t + C_2\,\Delta x + \mathrm{O}(\Delta x^2)$ with fixed numbers for C_1 and C_2, the error will monotonically decrease for smaller values of Δx and Δt. Using a logarithmic scale this situation is displayed in the left-hand plot of Fig. 8 with arbitrary but constant values for $C_{1,2}$. In the case of the BGK, as shown above, these constants themselves must be viewed as varying with the time step and grid size and, in fact, vanishing eventually. Along these error grooves the higher-order part of the error is dominant. However, since this perturbation is itself smooth the vicinity of the optimal lines benefits for the locally reduced error. This is demonstrated in the right plot of Fig. 8, where the function of the left-hand side is used but the constants are replaced such that they vanish along two distinct lines. The strong grooves cutting the error landscape are clearly visible. The almost singular behavior of these grooves is due to the strong decrease of the error by one or two orders of magnitude. Any path following

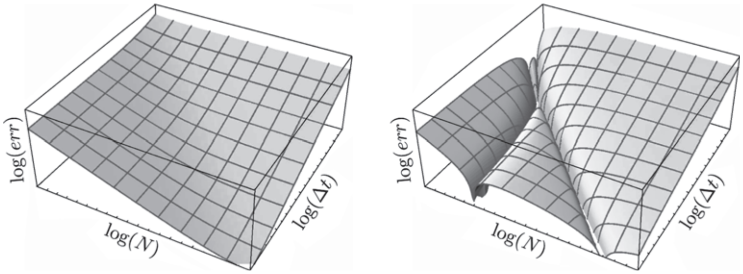

FIG. 8. Effect of a vanishing error constant on the error landscape of a numerical method. The results are strong error grooves (right plot) which introduce regions with locally high order of convergence into a non-perturbed error landscape (left plot).

the error into or in the vicinity of the grooves will exhibit locally high order of convergence. Hence, the distortion of the functional dependence leads to strong error reduction also in vicinity of the grooves.

Any investigation which conducts calculations with the BGK scheme along certain paths of the Δt-Δx-plane will occasionally become trapped in the grooves or their vicinity subsequently exhibiting a locally high order of convergence. In this light we re-consider the empirical results of Fig. 7. The lower row of the figure shows the Δt-Δx-plane. Both plots are identical except that abscissa and ordinate are exchanged such that each plot considers the same variable on the x-axis as the error plot above it. The thin light parallel lines in the Δt-Δx-plane correspond to the time step relation (5.3) for different values of CFL $= 1, 0.9, 0.8, 0.7, \ldots 0.1$. The dots in the plot following these lines represent the calculations done with the particular value of $(\Delta t, \Delta x)$. The same dots can be found in the respective plot above in Fig. 7 where the corresponding error of the calculation is displayed.

The paths of the error grooves are indicated in Fig. 7 as curved dark lines. The minimal errors of the calculations can now be exactly predicted at those points where the dots of the calculations along the line CFL $=$ const are closest to an error groove or cross a groove. Indeed, the calculations for CFL $= 0.8$ crosses an error groove twice for $\log(N) \approx 1.9$ and $\log(N) \approx 2.6$ which corresponds to the values of N exhibiting the minimal errors. Similarly, we find the points of minimal errors in the error plot with respect to the time step Δt.

The explicit expression for the error grooves is highly involved. In the limit $\alpha \gg 1$ the BGK method reduces to the FULLUP scheme (3.26). The two lines $\kappa_{1,2}^{(\mathrm{opt})}$ reduce to

$$\kappa_1^{(\mathrm{opt})}(\gamma) = \frac{1}{3+\gamma}, \qquad \kappa_2^{(\mathrm{opt})}(\gamma) = \frac{1}{2\gamma} \tag{5.13}$$

which can be written in the form

$$\Delta x_1^{(\mathrm{opt})}(\Delta t, a, \varepsilon, \tau) = 3\frac{\varepsilon\tau}{a} + a\,\Delta t, \quad \Delta x_2^{(\mathrm{opt})}(\Delta t, a, \varepsilon, \tau) = 2a\,\Delta t. \tag{5.14}$$

These curves correspond to the optimal lines in Δt-Δx-plane where the leading error constant of the FULLUP scheme vanishes.

5.5 *Comparison*

We conclude this section with a comparison of the empirical order of convergence of the BGK method with the classical methods UPCEN (4.10) and LW (4.11) for the advection–diffusion equation (1.1).

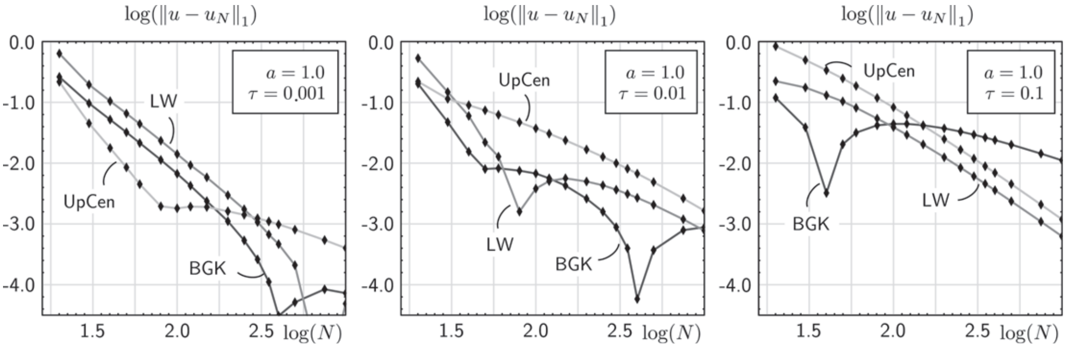

FIG. 9. Comparison of empirical error curves of the BGK method with two classical methods given by the Lax–Wendroff and the upwind method supplemented with central differences for the viscous part. Due to the super-convergence property of the BGK scheme it exhibits fast convergence and small errors on coarse grids.

We consider the initial condition (5.1) with the exact solution (5.2) with $L = 2$ in the periodic domain $x \in [-1, 3]$. The parameters are chosen to be $a = 1$, $\varepsilon = 1$, and three different values for the relaxation time $\tau = 0.001, 0.01, 0.1$. The numerical error is computed as in (5.4). For the time step the relation (5.3) is used with CFL $= 0.9$ and the computations are conducted with $N = 20, 30, 40, 50, 60, 80, 100, 120, 150, 200, 250, 300, 350, 400, 500, 750, 1000$. Figure 9 shows the error plots of the three methods for the three increasing values of τ.

Only in the case $\tau = 0.1$ the poor asymptotic order of convergence of the BGK scheme becomes visible for fine grids. In the other cases the super-convergence is dominant which turns the scheme into a competitive numerical method. In all cases the BGK scheme reaches the smallest values for the error among the methods shown, however, in a non-monotone way. Interestingly, the other two methods show some non-monotone behavior as well, most pronounced for the LW method in the case $\tau = 0.01$. In the appendix we give explicit expressions of the error groove for the Lax–Wendroff method.

High values of τ produce very smooth solutions. The BGK scheme is not especially designed for this regime, since it is based on a discontinuous reconstruction which assumes strong gradients. For very smooth and highly resolved situations this might not be an appropriate approach. Indeed, the LW and upwind scheme with central differences which can be viewed to be based on continuously reconstructed data show the better performances for high values of τ. These results correspond to recent findings in May *et al.* (2005) in the full gas-dynamic case. A remedy could be introduced into the BGK scheme by a non-linear coupling to a continuous reconstruction for high values of τ.

The gas-kinetic BGK scheme is computationally expensive due to the evaluation of error functions and exponentials, especially in the case of full viscous gas dynamics. A fast implementation needs to thoroughly collect the evaluation in order to avoid redundant function calls. A comparison of the efficiency of the presented methods is skipped in this paper.

6. Conclusions

The gas-kinetic BGK method provides an accurate and robust way to solve convection-dominated compressible Navier–Stokes problems. In order to provide fundamental analytical results, this paper applied the BGK method to the scalar advection–diffusion equation. The construction of the scheme provides a framework to model the advection and diffusion of the equation in a unifying way. The detailed structure and limiting cases of the scheme have been discussed and stability and consistency results have been presented. We identified several weight functions which control the influence of the different mechanisms of dissipation and advection in the scheme.

The stability results demonstrate the upwinding ability of the kinetic scheme which leads to enlarged stability domains. In regions of large values of the grid Reynolds number the stability of the BGK scheme shows additional improvement over classical methods. In this under-resolved case, the time step restriction is solely determined by the stability range of the advective part of the method. Furthermore, the method allows for Courant numbers larger than unity in regions of the parameters, which correspond to large Mach numbers in the full gas-dynamic case.

The BGK method was shown to behave as a third-order method due to a kind of super-convergence on coarse grids. We explained this behavior by demonstrating the existence of grooves in the error landscapes where the leading error constant vanishes. In addition, we proved a general limit of the order of consistency for the BGK scheme which is $\mathrm{O}(\tau^2 \Delta t)$ for the time accuracy, where τ is the relaxation time of the BGK model. This limit is visible only for very fine grids where the scheme is proven to be asymptotically first order in space and time. Empirical investigations on the error and the order of convergence together with comparison to classical methods have been presented.

To some extent, most of the results can be extrapolated to the full gas-dynamic case of the BGK scheme in a qualitative way. The results of this paper show that the physically motivated BGK method indeed may lead to an accurate and robust numerical scheme.

Acknowledgements

This work was supported by the German National Science Foundation (Deutsche Forschungsgesellschaft) and the Research Grants Council of Hong Kong.

References

Aregba-Driollet, D. & Natalini, R. (2000) Discrete kinetic schemes for multidimensional systems of conservation laws. *SIAM J. Numer. Anal.*, **37**, 1973–2004.

Ben-Artzi, M. & Falcovitz, J. (2003) Generalized Riemann problems in computational fluid dynamics. *Cambridge Monographs on Applied and Computational Mathematics*, vol. 11. Cambridge: Cambridge University Press.

Bhatnagar, P. L., Gross, E. P. & Krook, M. (1954) A model for collision processes in gases I: small amplitude processes in charged and neutral one-component systems. *Phys. Rev.*, **94**, 511–525.

Chou, S. Y. & Baganoff, D. (1997) Kinetic flux vector splitting for the Navier–Stokes equations. *J. Comput. Phys.*, **130**, 217.

Deshpande, S. M. (1986) A second order accurate, kinetic-theory based, method for inviscid compressible flows. *Technical Paper 2613*. Langley: NASA.

Godlewski, E. & Raviart, P.-A. (1996) *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. New York: Springer.

Jameson, A. (2004) *Solution Algorithms for Viscous Flow*. Talk at the Indian Institute of Science, Bangalore.

Junk, M. & Rao, S. V. R. (1999) A new discrete velocity method for Navier-Stokes equation. *J. Comput. Phys.*, **155**, 178.

Kim, C., Xu, K., Martinelli, L. & Jameson, A. (1997) Analysis and implementation of the gas-kinetic BGK scheme for computational gas dynamics. *Int. J. Numer. Methods Fluids*, **25**, 21–49.

LeVeque, R. J. (2002) *Finite Volume Methods for Hyperbolic Problems*. Cambridge: Cambridge University Press.

Li, Q. B. & Fu, S. (2003) Numerical simulation of high-speed planar mixing layer. *Comput. Fluids*, **32**, 1357–1377.

LI, Q. B., FU, S. & XU, K. (2004) A compressible Navier-Stokes flow solver with scalar transport. *J. Comput. Phys.*, **204**, 692–714.

MAY, G., SRINIVASAN, B. & JAMESON, A. (2005) Calculating three-dimensional transonic flow using a gas-kinetic BGK finite-volume method. *43rd AIAA Aerospace Sciences Meeting, AIAA Paper 2005-1397*.

MORTON, K. W. (1996) Numerical solution of convection-diffusion problems. *Applied Mathematics and Mathematical Computation*, vol. 12. London: Chapman & Hall.

MORTON, K. W. & SOBEY, I. J. (1993) Discretization of a convection-diffusion equation. *IMA J. Numer. Anal.*, **13**, 141–160.

OHWADA, T. (2002) On the construction of kinetic shemes. *J. Comput. Phys.*, **177**, 156–175.

OHWADA, T. & KOBAYASHI, S. (2004) Management of discontinuous reconstruction in kinetic schemes. *J. Comput. Phys.*, **197**, 116–138.

PERTHAME, B. (1992) Second-order Boltzmann schemes for compressible Euler equation in one and two space dimensions. *SIAM J. Numer. Anal.*, **29**, 1–19.

PULLIN, D. I. (1980) Direct simulation methods for compressible inviscid ideal gas flow. *J. Comput. Phys.*, **34**, 231.

PRENDERGAST, K. H. & XU, K. (1993) Numerical hydrodynamics from gas-kinetic theory. *J. Comput. Phys.*, **109**, 53.

SONG, J. A. & NI, G. X. (2004) A gamma-model BGK scheme for compressible multifluids. *Int. J. Numer. Methods Fluids*, **46**, 163–182.

VINCENTI, W. G. & KRUGER, JR., C. H. (1965) *Introduction to Physical Gas Dynamics*. New York: Wiley.

XU, K. (2001) A gas-kinetic BGK scheme for the Navier-Stokes equations and its connection with artificial dissipation and Godunov method. *J. Comput. Phys.*, **171**, 289–335.

XU, K., MAO, M. & TANG, L. (2005) A multidimensional gas-kinetic BGK scheme for hypersonic viscous flow. *J. Comput. Phys.*, **203**, 405–421.

Appendix A. BGK scheme

Due to the discontinuous evaluation the integration of (3.8) requires the calculation of half-space integrals of a Gaussian function. We define a generalized Heavyside function by

$$H_c(L,R) := \begin{cases} L & c < 0 \\ R & c > 0 \end{cases} \tag{A.1}$$

and provide the results for the first moments of this function with a Gaussian kernel. They are given by

$$\begin{aligned} \mathfrak{E}_a^{(0)}(L,R) &:= \int_{\mathbb{R}} H_c(L,R)\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}}\,\mathrm{d}c \\ &= L\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{a}{\sqrt{\varepsilon}}\right)\right) + R\tfrac{1}{2}\left(1-\operatorname{erf}\left(\frac{a}{\sqrt{\varepsilon}}\right)\right) \end{aligned} \tag{A.2}$$

$$\mathfrak{E}_a^{(1)}(L,R) := \int_{\mathbb{R}} cH_c(L,R)\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}}\,\mathrm{d}c = a\mathfrak{E}_a^{(0)}(L,R) + (L-R)\frac{1}{2}\sqrt{\frac{\varepsilon}{\pi}}\,\mathrm{e}^{-\frac{a^2}{\varepsilon}} \tag{A.3}$$

$$\mathfrak{E}_a^{(2)}(L,R) := \int_{\mathbb{R}} c^2H_c(L,R)\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}}\,\mathrm{d}c = a\mathfrak{E}_a^{(1)}(L,R) + \frac{\varepsilon}{2}\mathfrak{E}_a^{(0)}(L,R). \tag{A.4}$$

These formulas are essentially averaging expressions which weight the values L and R according to the value of a. For $\frac{a}{\sqrt{\varepsilon}} \gg 0$ we have

$$\mathfrak{E}_a^{(0)}(L, R) = L, \quad \mathfrak{E}_a^{(1)}(L, R) = aL, \quad \mathfrak{E}_a^{(0)}(L, R) = \left(a^2 + \frac{\varepsilon}{2}\right) L \tag{A.5}$$

and analogous expressions with R for $\frac{a}{\sqrt{\varepsilon}} \ll 0$. Since a is the advection velocity in the evolution equation (1.1) the equations (A.2)–(A.4) must be viewed as the core expressions for ‘kinetic upwinding’.

The functions $\mathfrak{E}_a^{(0,1,2)}$ can now be used for easy notation of the integration results for the BGK scheme. The interface value (3.16) is given by

$$u^n_{i+\frac{1}{2}} = \mathfrak{E}_a^{(0)}\left(u^n_{i+\frac{1}{2},L}, u^n_{i+\frac{1}{2},R}\right) - \tfrac{\tau}{2}((\delta_x u)_i - (\delta_x u)_{i+1})\sqrt{\frac{\varepsilon}{\pi}}e^{-\frac{a^2}{\varepsilon}}. \tag{A.6}$$

The time derivative $A^n_{i+\frac{1}{2}}$ follows after using the time-averaged interface distribution (3.8) in (3.18), the definition of $u^n_{i+\frac{1}{2}}$ and the definition of the weight function W_5. We obtain with $\omega = \varDelta t/\tau$

$$\int_{\mathbb{R}} (\partial_t g)^n_{i+\frac{1}{2}} \mathrm{d}c = -W_5(\omega) \int_{\mathbb{R}} c(\partial_x f)^n_{i+\frac{1}{2}} \mathrm{d}c - (1 - W_5(\omega)) \int_{\mathbb{R}} c(\partial_x g)^n_{i+\frac{1}{2}} \mathrm{d}c \tag{A.7}$$

from which $A^n_{i+\frac{1}{2}}$ can be calculated to give

$$A^n_{i+\frac{1}{2}} = -\left[\mathfrak{E}_a^{(1)}((\delta_x u)^n_i, (\delta_x u)^n_{i+1})W_5(\omega) + \mathfrak{E}_a^{(1)}\left(\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L}, \widetilde{(\delta_x u)}^n_{i+\frac{1}{2},R}\right)(1 - W_5(\omega))\right]. \tag{A.8}$$

The integration of the time-averaged interface distribution (3.8) yields

$$\begin{aligned}
\tilde{F}^n_{i+\frac{1}{2}} &= \int_{\mathbb{R}} c\tilde{f}\left(c, x_{i+\frac{1}{2}}\right)\mathrm{d}c \\
&= au^n_{i+\frac{1}{2}}(1 - W_1(\omega)) + \Bigg(\mathfrak{E}_a^{(1)}\left(u^n_{i+\frac{1}{2},L}, u^n_{i+\frac{1}{2},R}\right) \\
&\quad - \tfrac{\varepsilon\tau}{2}\mathfrak{E}_a^{(0)}((\delta_x u)^n_i, (\delta_x u)^n_{i+1})\Bigg) W_1(\omega) - \tau\,\mathfrak{E}_a^{(2)}((\delta_x u)^n_i, (\delta_x u)^n_{i+1})W_3(\omega) \\
&\quad - \tau\,\mathfrak{E}_a^{(2)}\left(\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L}, \widetilde{(\delta_x u)}^n_{i+\frac{1}{2},R}\right)W_2(\omega) + a\,\varDelta t\,A^n_{i+\frac{1}{2}}W_4(\omega).
\end{aligned} \tag{A.9}$$

In this formula, we introduce the result for $A^n_{i+\frac{1}{2}}$ and the representation of $W_4(\omega)$ and $W_5(\omega)$ according to Lemma 2 as well as the recursion for $\mathfrak{E}_a^{(2)}(L, R)$ given in (A.4). After rearrangement, usage of $W_1 + W_3 = 1 - W_2$ and re-introduction of W_5 we obtain the final result

$$\begin{aligned}
\tilde{F}^{(\mathrm{BGK})}_{i+\frac{1}{2}} &= au^n_{i+\frac{1}{2}}(1 - W_1(\omega)) + \mathfrak{E}_a^{(1)}\left(u^n_{i+\frac{1}{2},L}, u^n_{i+\frac{1}{2},R}\right)W_1(\omega) \\
&\quad - \frac{\varepsilon\tau}{2}\left[\mathfrak{E}_a^{(0)}((\delta_x u)^n_i, (\delta_x u)^n_{i+1})(1 - W_2(\omega)) + \mathfrak{E}_a^{(0)}\left(\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L}, \widetilde{(\delta_x u)}^n_{i+\frac{1}{2},R}\right)W_2(\omega)\right] \\
&\quad - \frac{a\,\varDelta t}{2}\left[\mathfrak{E}_a^{(1)}((\delta_x u)^n_i, (\delta_x u)^n_{i+1})W_5(\omega) + \mathfrak{E}_a^{(1)}\left(\widetilde{(\delta_x u)}^n_{i+\frac{1}{2},L}, \widetilde{(\delta_x u)}^n_{i+\frac{1}{2},R}\right)(1 - W_5(\omega))\right]
\end{aligned} \tag{A.10}$$

which gives (3.19).

Appendix B. Proofs for stability

B.1 *Kinetic upwinding for pure advection*

The assertion is given in Section 4.3.

Proof. Since the structure of the fluxes KIN1 and KIN2 in (3.23) and (3.24) is very similar these two methods can be investigated simultaneously. We define the functions

$$h_1(\alpha) = \operatorname{erf}\alpha \tag{B.1}$$

present in KIN1 and

$$h_2(\alpha) = \operatorname{erf}\alpha + \frac{1}{\alpha\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2} \tag{B.2}$$

for KIN2 with the parameter $\alpha = \frac{a}{\sqrt{\varepsilon}}$. We skip the details of the calculation of the amplification function and present the result in the form

$$\begin{aligned} g(\xi) &:= |G(\xi)| - 1 \\ &= 4\sin^2\frac{\xi}{2}\left(\lambda(\lambda - h(\alpha)) - \lambda^2(1 - h(\alpha)^2)\sin^2\frac{\xi}{2}\right), \end{aligned} \tag{B.3}$$

where h has to be replaced with h_1 for KIN1 and with h_2 for KIN2. The parameter $\lambda = \frac{a\varDelta t}{\varDelta x}$ represents the Courant number.

The stability condition transforms into

$$\max_{\xi\in[0,\pi]} g(\xi) \leqslant 0 \tag{B.4}$$

for the function g which gives

$$\lambda(\lambda - h(\alpha)) - \lambda^2(1 - h(\alpha)^2)\sin^2\frac{\xi}{2} \leqslant 0. \tag{B.5}$$

Since this expression is linear in $\sin^2\frac{\xi}{2}$, we evaluate it at two positions $\xi = 0$ and $\xi = \pi$ to obtain two necessary and sufficient conditions. These read

$$\lambda(\lambda - h(\alpha)) \leqslant 0 \quad \wedge \quad \lambda h(\alpha)(-1 + \lambda h(\alpha)) \leqslant 0. \tag{B.6}$$

Since we have $h(-\alpha) = -h(\alpha)$ for both function $h_{1,2}$ and $\operatorname{sign}\lambda = \operatorname{sign}\alpha = \operatorname{sign}a$, it follows that $\lambda h(\alpha) \geqslant 0$ and the conditions reduce to

$$|\lambda| \leqslant \frac{1}{|h(\alpha)|} \quad \wedge \quad |\lambda| \leqslant |h(\alpha)|. \tag{B.7}$$

At this stage we have to look at KIN1 and KIN2 separately. For KIN1 we have $|h_1(\alpha)| = |\operatorname{erf}\alpha| \leqslant 1$, hence the second condition is more restrictive and gives the assertion in the theorem for KIN1. In the case of KIN2 we will show that

$$|h_2(\alpha)| = \left|\operatorname{erf}\alpha + \frac{1}{\alpha\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}\right| \geqslant 1 \tag{B.8}$$

which gives a more restrictive first condition and after rearrangement the assertion of the theorem for KIN2.

The demonstration of the last inequality uses the continued fraction representation of the error function

$$\operatorname{erf}\alpha = 1 - \frac{1}{\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}\frac{1}{\alpha + \frac{1}{2\alpha + \frac{2}{\alpha + \frac{3}{2\alpha + \frac{4}{\cdots}}}}}, \tag{B.9}$$

which can be found in analysis text books. It follows directly

$$\alpha(1-\operatorname{erf}\alpha) = \frac{1}{\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}\frac{1}{1 + \frac{1}{2\alpha^2 + \frac{2}{1 + \frac{3}{2\alpha^2 + \frac{4}{\cdots}}}}} \leqslant \frac{1}{\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}, \tag{B.10}$$

which gives the above inequality.

The conditions in the second column of (4.16) consider the $\lambda_\varepsilon = \frac{\sqrt{\varepsilon}\,\Delta t}{\Delta x}$ which represents the Courant number with respect to the kinetic velocity $\sqrt{\varepsilon}$, i.e. a kinetic Courant number. With the relation $\alpha = \lambda/\lambda_\varepsilon$ the above necessary and sufficient conditions can be written as conditions on λ_ε. For KIN1 we find

$$\lambda \leqslant \operatorname{erf}\left(\frac{\lambda}{\lambda_\varepsilon}\right) \quad \Leftrightarrow \quad \lambda_\varepsilon \leqslant \frac{\lambda}{\operatorname{erf}^{-1}(\lambda)} \xrightarrow{\lambda\to 0} \lim_{\alpha\to 0}\operatorname{erf}'(\alpha) = \frac{2}{\sqrt{\pi}} \tag{B.11}$$

while for KIN2

$$\lambda \leqslant \frac{1}{h_2(\frac{\lambda}{\lambda_\varepsilon})} \quad \Leftrightarrow \quad \lambda_\varepsilon \leqslant \frac{\lambda}{h_2^{-1}(\lambda^{-1})} \xrightarrow{\lambda\to 0} \lim_{\alpha\to 0}\left(\frac{1}{h_2(\alpha)}\right)' = \sqrt{\pi}. \tag{B.12}$$

Due to the limit the final conditions are only necessary.

The amplification function of the method KIN3 is more involved than in the case of KIN1/KIN2 and we will give here only a sketch of the proof of the stability conditions. Similar to the upper case we can write the amplification function with a function $\hat{g}(x)$ defined by

$$4\sin^4\frac{\xi}{2}\hat{g}\left(\sin^2\frac{\xi}{2}\right) = |G(\xi)| - 1, \tag{B.13}$$

which has the form

$$\begin{aligned}\hat{g}(x) = {}& -\lambda^2(1-h_1^2)(1-\lambda h_2)^2x^2 - \lambda^2(1-\lambda h_2)(1-2\lambda h_1 + \lambda h_2)x \\ & + \lambda(\lambda^3 + \lambda + \lambda\, h_1h_2 - h_1 - 2\lambda^2 h_2).\end{aligned} \tag{B.14}$$

Here, the functions $h_{1,2}$ as defined above occur and we suppress the argument α for clarity. The function $\hat{g}$ is a upside down parabola. The stability condition reduces to $\max_{x\in[0,1]}\hat{g}(x) \leqslant 0$.

Numerical evaluations suggest that for large α the condition $\hat{g}(1) \leqslant 0$ is decisive. The relation $\hat{g}(1,\lambda,\alpha) = 0$ will provide the marginal value of λ depending on α for large α. This relation seen as function of λ has four roots, but only one root, $\lambda^{(0)}(\alpha)$ say, has the relevant behavior $\lambda \to 1$ for $\alpha \to \infty$. Hence, for large α we have the condition

$$\lambda \leqslant \lambda^{(0)}(\alpha) = \frac{h_1(\alpha) - \sqrt{4 + h_1(\alpha)^2 - 4h_1(\alpha)h_2(\alpha)}}{2(h_1(\alpha)h_2(\alpha) - 1)}. \tag{B.15}$$

For smaller values of α the maximum of the parabola $\hat{g}$ enters the interval [0, 1] from the right. Its value increases and hits the x-axis subsequently. The value of the maximum is given by

$$\begin{aligned} m\,(\lambda,\alpha) &= \lambda^3(4+h_2^2-4h_1h_2)+\lambda^2(8h_1^2h_2-4h_1-6h_2) \\ &\quad +\lambda(5-4h_1(h_1+h_2(h_1^2-1)))+4h_1(h_1^2-1) \end{aligned} \tag{B.16}$$

depending on λ and α. If the maximum is within the interval [0, 1] stability requires $m(\lambda,\alpha)\leqslant 0$ which yields a relation $\lambda^{(1)}(\alpha)$ for the marginal value of λ defined by

$$m(\lambda^{(1)}(\alpha),\alpha)=0, \tag{B.17}$$

which remains decisive for $\alpha\to 0$. The handing over between the two conditions $\lambda^{(0,1)}$ happens at the value $\alpha=\alpha_0$ when the maximum hits the x-axis at $x=1$, i.e. $m(\lambda^{(0)}(\alpha_0),\alpha_0)=0$ or $\lambda^{(0)}(\alpha_0)=\lambda^{(1)}(\alpha_0)$. The numerical solution of this equation gives $\alpha_0\approx 0.330726\ldots$ and we can state the final condition

$$\lambda\leqslant\lambda^{(\star)}(\alpha)=\begin{cases}\lambda^{(1)}(\alpha) & \alpha\leqslant\alpha_0\\[2ex] \lambda^{(0)}(\alpha) & \alpha>\alpha_0\end{cases} \tag{B.18}$$

for the KIN3 method. In terms of the kinetic Courant number λ_ε we obtain for small α

$$\lambda_\varepsilon\leqslant\frac{\lambda}{(\lambda^{(1)})^{-1}(\lambda)}\xrightarrow{\lambda\to 0}\lim_{\alpha\to 0}\lambda^{(1)\prime}(\alpha), \tag{B.19}$$

where the limit can be calculated with help of the algebra software Mathematica to give

$$\lambda_\varepsilon\leqslant 2\left(\sqrt{\pi}+\sqrt{\frac{7\pi-8}{3}}\cos\left(\frac{1}{3}\arctan\left(\frac{2}{9}\sqrt{75-\frac{96}{\pi}\frac{\pi-2}{3\pi-4}}\right)\right)\right) \tag{B.20}$$

which corresponds to the number $7.84\ldots$. □

B.2 *Full upwinding for advection–diffusion*

See Section 4.4 for the statement of the theorem.

Proof. We will use the notations $\lambda=\frac{a\,\Delta t}{\Delta x}$ and $\kappa=\frac{\varepsilon\tau}{a\,\Delta x}$. The amplification function can be written in the form

$$\begin{aligned} g(\xi) &= |G(\xi)|-1 \\ &= 4\sin^2\frac{\xi}{2}\left(-\lambda^2(\lambda+\kappa-1)^2\sin^4\frac{\xi}{2}+\lambda(\lambda+\kappa-1)(1+\lambda(\lambda+\kappa-1))\sin^2\frac{\xi}{2}-\lambda\kappa\right), \end{aligned} \tag{B.21}$$

where we substitute $x=\sin^2\frac{\xi}{2}$ and write the essential part

$$\tilde{g}(x)=-\lambda^2(\lambda+\kappa-1)^2x^2+\lambda(\lambda+\kappa-1)(1+\lambda(\lambda+\kappa-1))x-\lambda\kappa. \tag{B.22}$$

According to its definition we have $x\in[0,1]$ and the stability condition takes the form

$$\max_{x\in[0,1]}\tilde{g}(x)\leqslant 0 \tag{B.23}$$

for the function $\tilde{g}$. It is an upside down parabola with the two zeros

$$x_1 = \frac{1 + \lambda(\lambda + \kappa - 1) - \sqrt{(1 - \lambda(\lambda + \kappa - 1))^2 - 4(1 - \lambda)\lambda}}{2\lambda(\lambda + \kappa - 1)}, \tag{B.24}$$

$$x_2 = \frac{1 + \lambda(\lambda + \kappa - 1) + \sqrt{(1 - \lambda(\lambda + \kappa - 1))^2 - 4(1 - \lambda)\lambda}}{2\lambda(\lambda + \kappa - 1)}, \tag{B.25}$$

which can be complex eventually. It becomes clear that if the roots are not real we have stability, which, however, is not always the case.

At $x = 0$ we have $\tilde{g}(0) = -\lambda\kappa$ and thus $\lambda\kappa \geqslant 0$ is necessary for stability. Consider the case $\lambda < 0$ and $\kappa < 0$ in which the roots are always real. We have $x_1 x_2 = \frac{\kappa}{\lambda(\lambda+\kappa-1)^2} > 0$, hence, both zeros are on the same side of the ordinate. We find $x_2 > 0$ and, since

$$1 - \lambda(\lambda + \kappa - 1) - \sqrt{(1 - \lambda(\lambda + \kappa - 1))^2 - 4(1 - \lambda)\lambda} < 0, \tag{B.26}$$

it follows $x_1 < 1$. This shows that one zero, x_1, will always lie in the crucial interval $[0, 1]$ and prevents the method from being stable for $\lambda < 0$ and $\kappa < 0$. Hence, we consider $\lambda \geqslant 0$ and $\kappa \geqslant 0$, which gives the necessary conditions $a > 0$ and $\tau > 0$. In this setting we will always have $x_1 x_2 \geqslant 0$, so again both zeros lie on the same side of the ordinate, if they are real.

For $\lambda > 1$ and $\kappa \geqslant 0$ we again find that both zeros are real, $x_2 > 0$ and $x_1 < 1$. Hence, it follows $0 \leqslant \lambda \leqslant 1$ necessarily. The range $0 \leqslant \kappa < \infty$ is now split into the parts $0 \leqslant \kappa < 1$ and $\kappa > 1$ and it remains to show what marginal value λ may take depending on κ.

For the case $0 \leqslant \kappa < 1$ consider the value of the term $\lambda(\lambda + \kappa - 1)$ with $(\lambda, \kappa) \in [0, 1]^2$ which gives $\lambda(\lambda + \kappa - 1) \in [-\frac{1}{4}, 1]$. If the term is negative, we either have complex roots or both zeros are negative, since,

$$x_1 + x_2 = 1 + \frac{1}{\lambda(\lambda + \kappa - 1)} < 0. \tag{B.27}$$

On the other hand, if the term is positive, the roots are always real, but we have $x_1 > 1$ and $x_2 > 1$. It follows that there is no additional restriction for λ in the interval $0 \leqslant \kappa < 1$ which proves the first part of the assertion (4.17).

For the part $\kappa > 1$ we have $\lambda(\lambda + \kappa - 1) > 0$ and the roots can be complex or real. However, if the root is real, again $x_2 > 0$ and $x_1 < 1$ hold, which spoils stability. We conclude that the marginal value for λ is given at the point where the roots turn complex. The discriminant of the roots can be solved for κ to give

$$\kappa^{(\mathrm{crit})}(\lambda) = 1 + \frac{1}{\lambda} - \lambda - 2\sqrt{\frac{1 - \lambda}{\lambda}}, \tag{B.28}$$

which, at least formally, can be inverted to obtain $\lambda^{(\mathrm{crit})}(\kappa)$. □

B.3 *BGK scheme*

It follows the proof of the statement in Section 4.5.

Proof. We will write the amplification function in the form $g(\xi) = |G(\xi)| - 1$ and define the function $\tilde{g}(x)$ by

$$4 \sin^2 \frac{\xi}{2} \tilde{g}\left(\sin^2 \frac{\xi}{2}\right) = g(\xi). \tag{B.29}$$

The function $\tilde{g}(x)$ is a polynomial of third degree in x depending on the following parameter

$$\lambda = \frac{a\,\Delta t}{\Delta x}, \quad \kappa = \frac{\varepsilon\tau}{a\,\Delta x}, \quad \alpha = \frac{a}{\sqrt{\varepsilon}}, \quad \omega = \frac{\Delta t}{\tau}, \tag{B.30}$$

where ω can be related with the other parameters by $\omega = \frac{\lambda}{\kappa\alpha^2}$. A Taylor expansion around $\lambda = 0$ for small values of λ yields the form

$$\tilde{g}(x) = \left(\left(\kappa(1-(1-\mathrm{erf}^2\,\alpha)W_2(0)) - \mathrm{erf}\,\alpha - \frac{1}{\alpha\sqrt{\pi}}\,\mathrm{e}^{-\alpha^2}W_1(0)\right)x - \kappa\right)\lambda + \mathrm{O}(\lambda^2), \tag{B.31}$$

where the weight functions $W_{1,2}$ are given by (3.9). We have $W_1(0) = 1$, $W_2(0) = 0$ and obtain the essential function

$$\tilde{g}(x) = ((\kappa - h_2(\alpha))x - \kappa)\lambda \tag{B.32}$$

where $h_2(\alpha)$ is given by (B.2). For $\tilde{g}$ the stability condition takes the form $\max_{x\in[0,1]}\tilde{g}(x) \leqslant 0$. Since $\tilde{g}$ is linear in x we only need to consider the two evaluations

$$\tilde{g}(0) = -\kappa\lambda, \quad \tilde{g}(1) = -h_2(\alpha)\lambda \tag{B.33}$$

in order to check stability. The condition $\tilde{g}(0) \leqslant 0$ yields $\tau \geqslant 0$, while $\tilde{g}(1) \leqslant 0$ is always satisfied since $\mathrm{sign}\,\lambda = \mathrm{sign}\,\alpha$, $h_2(|\alpha|) \geqslant 0$ and $h_2(-\alpha) = -h_2(\alpha)$. □

Appendix C. Proofs for consistency

C.1 *General limit*

The exact solution of the BGK equation deviates from the solution of the advection–diffusion equation since the correspondence is only present asymptotically. The deviation can be estimated by evaluating the Chapman–Enskog expansion. We will present here some formal results for the error of the asymptotic expansion which will be used in the proof below.

If the Chapman–Enskog expansion (2.8) enters the BGK equation (2.3) balancing the powers of τ yields

$$f_n = (-1)^n D^n g[u] \tag{C.1}$$

for the coefficients in the expansion. Here, we use $D := \partial_t + c\,\partial_x$ for the microscopic total derivative. Hence, we write for the general Chapman–Enskog expanded distribution function of the BGK equation

$$f^{(N)} = \sum_{n=0}^{N}(-\tau)^n D^n g[u] \tag{C.2}$$

if we consider the distribution function up to Nth-order in τ.

For the time $T > 0$ we consider the domain $\tilde{\Omega} = \mathbb{R}\times\Omega\times(0,T)$ and assume f, g and its derivatives to be integrable on $\tilde{\Omega}$. We then have for small τ

$$\|D(f - f^{(N)})\|_{L^1(\tilde{\Omega})} = \mathrm{O}(\tau^{N+1}) \tag{C.3}$$

which follows from

$$\begin{aligned} Df - Df^{(N)} &= \frac{1}{\tau}\left(g - \sum_{n=0}^{\infty}(-\tau)^n D^n g[u]\right) - \sum_{n=0}^{N}(-\tau)^n D^{n+1} g[u] \\ &= \frac{1}{\tau}\left(-\sum_{n=1}^{\infty}(-\tau)^n D^n g[u] + \sum_{n=1}^{N+1}(-\tau)^n D^n g[u]\right) \\ &= \frac{1}{\tau}\sum_{n=N+2}^{\infty}(-\tau)^n D^n g[u] = \mathrm{O}(\tau^{N+1}), \end{aligned}$$

where we used the BGK equation, the result of the asymptotic expansion and the definition of $f^{(N)}$. Time-integration along the paths $\dot{x} = c$ gives the result

$$\|(f - f^{(N)})|_{t=T}\|_{L^1(\mathbb{R}\times\Omega)} = \|(f - f^{(N)})|_{t=0}\|_{L^1(\mathbb{R}\times\Omega)} + \mathrm{O}(T\tau^{N+1}). \tag{C.4}$$

We observe that even though the difference is smaller for higher N, starting from identical initial conditions the difference will increase linearly in time. This fact is used in the proof of the assertion given in Section 5.2.

Proof. The distribution function corresponding to the exact solution $u^{(\mathrm{ex})}$ is the Chapman–Enskog distribution $f^{(\mathrm{ex})} := f^{(1)}$. This distribution function would give the exact flux. We write the accuracy result (C.4) for $f^{(1)} = f^{(\mathrm{ex})}$ with $T = \Delta t$, i.e. after one time step. This gives

$$\|(f^{(\mathrm{BGK})} - f^{(\mathrm{ex})})|_{t=\Delta t}\|_{L^1(\mathbb{R}\times\mathbb{R})} = \|(f^{(\mathrm{BGK})} - f^{(\mathrm{ex})})|_{t=0}\|_{L^1(\mathbb{R}\times\mathbb{R})} + \mathrm{O}(\Delta t\tau^2), \tag{C.5}$$

where $f^{(\mathrm{BGK})}$ is the distribution function which follows from the full exact solution of the kinetic equation. At the initial time $t = 0$ we have

$$f^{(\mathrm{ex})}|_{t=0} = (u^{(\mathrm{ex})}|_{t=0} - \tau(c-a)\partial_x u^{(\mathrm{ex})}|_{t=0})\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}} \tag{C.6}$$

for the exact distribution function, while the exact ‘kinetic’ distribution function is given by the infinite expansion

$$f^{(\mathrm{BGK})}|_{t=0} = \sum_{n=0}^{\infty}(-\tau)^n D^n g[u^{(\mathrm{BGK})}|_{t=0}] \tag{C.7}$$

based on the initial numerical function $u^{(\mathrm{BGK})}|_{t=0}$. For the difference of these distribution functions we find

$$\begin{aligned} \|(f^{(\mathrm{BGK})} - f^{(ex)})|_{t=0}\|_{L^1(\mathbb{R}\times\mathbb{R})} \leqslant{}& \left\|(u^{(\mathrm{BGK})} - u^{(ex)})|_{t=0}\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}}\right\|_{L^1(\mathbb{R}\times\mathbb{R})} \\ &+ \tau\left\|(c-a)(\partial_x u^{(\mathrm{BGK})} - \partial_x u^{(ex)})|_{t=0}\frac{1}{\sqrt{\varepsilon\pi}}\,\mathrm{e}^{-\frac{(c-a)^2}{\varepsilon}}\right\|_{L^1(\mathbb{R}\times\mathbb{R})} \\ &+ \mathrm{O}(\tau^2). \end{aligned} \tag{C.8}$$

The initial numerical function $u^{(\mathrm{BGK})}$ is subject to a reconstruction procedure in which the function and its gradient is approximated with spatial order p and q. Hence, it follows

$$\|(f^{(\mathrm{BGK})} - f^{(\mathrm{ex})})|_{t=0}\|_{L^1(\mathbb{R}\times\mathbb{R})} \leqslant \tilde{C}_1 \varDelta x^p + \tilde{C}_2 \tau \varDelta x^q, \tag{C.9}$$

for small τ. The flux of the numerical method is defined in (3.3). Using the above findings we have

$$\|(F^{(\mathrm{BGK})} - F^{(\mathrm{ex})})|_{t=\varDelta t}\|_{L^1(\mathbb{R})} \leqslant \hat{C}_1 \varDelta x^p + \hat{C}_2 \tau \varDelta x^q + \hat{C}_3 \tau^2 \varDelta t \tag{C.10}$$

which can be transformed into

$$\|(u^{(\mathrm{BGK})} - u^{(\mathrm{ex})})|_{t=\varDelta t}\|_{L^1(\mathbb{R})} \leqslant C_1 \varDelta t\, \varDelta x^p + C_2 \tau\, \varDelta t\, \varDelta x^q + \mathcal{C}_3 \tau^2 \varDelta t^2 \tag{C.11}$$

by use of the finite-volume update (3.2). □

C.2 *Error grooves*

Here, we proove the theorem from Section 5.4.

Proof. The consistency error of the BGK numerical method for (1.1) has the shape

$$\frac{1}{\varDelta t}(u^{(\mathrm{BGK})} - u)\Big|_{t=\varDelta t} = \sum_{k=3}^{\infty} A_k(\varDelta t, \varDelta x, \varepsilon, \tau, a)\, u^{(k)}\Big|_{t=0} \tag{C.12}$$

found by Taylor expansion of the numerical method and the solution. The kth spatial derivative of the solution u is denoted by $u^{(k)}$. For scaling reasons the constants A_k have the representation

$$A_k = \frac{\varDelta x^k}{\varDelta t} \tilde{A}_k(\lambda, \kappa, W_1(\omega), W_2(\omega), W_5(\omega), h_0(\alpha), h_1(\alpha), h_2(\alpha)), \tag{C.13}$$

where $\tilde{A}_k$ is a dimensionless quantity depending polynomially on κ and λ as well as on the weight functions W_i from (3.9) and the auxiliary functions h_j defined in (B.1)/(B.2) and by

$$h_0(\alpha) = \frac{\alpha}{\sqrt{\pi}}\, \mathrm{e}^{-\alpha^2}. \tag{C.14}$$

The functional dependence of A_k on the parameters is strongly non-linear and non-monotone, hence the vanishing of some of the constants appears frequently in the parameter space. The vanishing of the leading constant A_3 is responsible for the super-convergence. In order to write A_3 in a suitable form we use the dimensionless parameters

$$\gamma = \frac{a^2 \varDelta t}{\varepsilon\tau}, \quad \kappa = \frac{\varepsilon\tau}{a\,\varDelta x}, \tag{C.15}$$

which represent the time step and the grid size. The constant A_3 can then be written as $A_3 = \frac{12a}{(\alpha\tau)^2}\hat{A}_3$ with

$$\begin{aligned}\hat{A}_3(\gamma, \kappa, \alpha) = (&\kappa^{-2} - 3\kappa^{-1}(h_1 + \gamma h_2 - 2(1 - W_1)h_0) \\ &+ 2(\gamma(\gamma + 3) + 3h_0(W_2 h_1 + (1 - W_5)\gamma h_2))),\end{aligned} \tag{C.16}$$

where the arguments of W_i and h_j are suppressed. Note that we have $\omega = \gamma/\alpha^2$ for the argument of the weights.

The relation

$$G(\varDelta x, \varDelta t) := \hat{A}_3\left(\frac{a^2\varDelta t}{\varepsilon\tau}, \frac{\varepsilon\tau}{a\varDelta x}, \frac{a}{\sqrt{\varepsilon}}\right) = 0 \tag{C.17}$$

corresponds to a subset of the $\varDelta t$-$\varDelta x$-plane where the leading constant A_3 vanishes. This subset has the shape of two lines $\kappa_{1,2}^{(\mathrm{opt})}(\gamma, \alpha)$ defined by

$$\hat{A}_3(\gamma, \kappa^{(\mathrm{opt})}(\gamma, \alpha), \alpha) = 0, \tag{C.18}$$

which has the form of a quadratic equation in κ^{-1} as can be seen in (C.16). We skip the explicit formula of these lines which contain the weights W_i and the auxiliary function h_j in a strongly non-linear way. The lines correspond to relations $\varDelta x_{1,2}^{(\mathrm{opt})}(\varDelta t, a, \varepsilon, \tau)$ for an optimal grid size depending on the other parameters of the scheme but not on the solution itself. Due to the high non-linearity introduced by the weights the relation cannot be inverted to find $\varDelta t^{(\mathrm{opt})}$ explicitly.

If A_3 vanishes the local consistency error turns out to be $\mathrm{O}(\varDelta x^4/\varDelta t)$. The remaining leading constant $\tilde{A}_4$ is of order unity for finite values of λ, κ, α and ω. □

C.3 *Error grooves for the LW-scheme*

An error representation like in (C.12) is also valid for the Lax–Wendroff scheme (4.11). A detailed evaluation of the Taylor expansions gives

$$\hat{A}_3^{(\mathrm{LW})}(\gamma, \kappa) = -\kappa^{-2} + \gamma^2 + 3\gamma, \tag{C.19}$$

which is the expression analogous to (C.16). We conclude that the leading error constant A_3 for the LW-scheme is vanishing along the line

$$\kappa^{(\mathrm{opt,LW})}(\gamma) = \frac{1}{\sqrt{3\gamma + \gamma^2}}, \tag{C.20}$$

which can be written

$$\varDelta x^{(\mathrm{opt,LW})}(\varDelta t, a, \varepsilon, \tau) = \sqrt{\left(3\frac{\varepsilon\tau}{a} + a\varDelta t\right) a\varDelta t} \tag{C.21}$$

for the grid size. In comparison to the BGK method the LW error groove is not twofold and also leaves the stability domain of the LW method earlier. Hence, the super-convergence of the BGK method will be more pronounced as observed in the empirical investigations.