Adaptation of the Unified Gas-Kinetic Scheme to ES-BGK models

C. Baranger[1], A. Coëpeau*[1,2], and L. Mieussens[2]

[1]CEA-CESTA 15 avenue des Sablières CS 60001 33114 Le Barp Cedex France
[2]Univ. Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, F-33400, Talence, France

January 9, 2026

Abstract

The Unified Gas-Kinetic Scheme (UGKS) stands out from traditional deterministic numerical approaches to gas dynamics by enabling the resolution of flow regimes ranging from rarefied to continuum, within a simulation time independent of the regime. Its effectiveness in solving complex flows has primarily been demonstrated with Shakhov or Rykov models. Furthermore, it has also been applied to ES-BGK type models, even if its application has so far been limited to purely monoatomic effects.

In this paper, we aim to explore the application of the unified gas-kinetic scheme to ES-BGK models in greater depth, extending its use from monoatomic to polyatomic gases by employing recent formulations. Particular attention will be given to validating this adaptation through numerical comparisons with simulation codes from the literature, such as SPARTA for non-vibrating gas flows and PIClas for other cases, as it can handle polyatomic ES-BGK models, including vibrational phenomena.

Keywords: Ellipsoidal Statistical BGK models, Unified Gas-Kinetic Scheme, diatomic gas

1 Introduction

The kinetic theory of gases relies primarily on the Boltzmann equation to describe the behavior of a gas at the microscopic scale. This fundamental equation governs the time evolution of the mass distribution of gas in phase space, accounting for variations in position, particle velocity, and, when necessary, internal energy such as rotation or vibration. Despite its complexity, the Boltzmann equation provides a framework for modeling gases across the entire range of rarefaction, from the free molecular to the continuum flow regimes. One of its most remarkable features lies in its natural asymptotic behavior in low Knudsen number where it naturally derives to hydrodynamic models such as the Euler and Navier-Stokes equations. This property not only bridges microscopic and macroscopic descriptions of gas dynamics but also emphasizes the singular and essential importance of this fundamental equation in both the multi-scale and multi-regime modeling.

Nevertheless, the Boltzmann equation presents significant challenges, both in terms of theoretical analysis [9] and numerical resolution. These difficulties have naturally motivated the development of simplified models that preserve the fundamental properties of the original equation such as its asymptotic behavior, conservation of mass, momentum and total energy, and the dissipation of entropy. One of the earliest simplified models of the Boltzmann equation is the Bhatnagar-Gross-Krook (BGK) model [5, 41]. It was developed to enable simpler and faster numerical simulations of transport phenomena within a monoatomic gas in a transitional state between rarefied and continuous regimes. Although the BGK model successfully reproduces many of the fundamental properties of the Boltzmann equation, it is limited to simulating gas flows with a unitary Prandtl number. As a result, it cannot accurately reproduce the experimentally observed propagation of thermal effects relative to dynamic effects in the hydrodynamic limit. To address this limitation, more advanced models such as the Ellipsoidal-Statistical BGK (ES-BGK) model [17] and the Shakhov-BGK model [36] were introduced. These models are capable of correctly recovering a Prandtl number of 2/3 for monoatomic gases, thereby providing a more physically

*Corresponding author: alexis.coepeau@u-bordeaux.fr

accurate representation of thermal dynamics and a more realistic continuous asymptotic limit behavior. Unlike the Shakhov model, the monoatomic ES-BGK model ensures the positivity of the mass distribution, and satisfies the H-theorem [2]. Consequently, it retains the same fundamental properties as the original BGK model and the Boltzmann equation. This model has subsequently been extended to diatomic gases [2] and has undergone various modifications, such as modeling the rotation and vibration energy distribution and their induced relaxation processes toward equilibrium [13, 26, 33].

To resolve rarefied gas flows, several methodologies have been considered in the literature. One of the most widely used is certainly the Direct Simulation Monte-Carlo (DSMC) [6], which emulates the physics described by the Boltzmann equation by alternatively simulating the transport and collisions of numerical particles, each representing an agglomeration of real gas molecules. While DSMC is centered on detailed collision models, other particle stochastic approaches focus on simpler treatments of particles interactions induced by simplified considerations as ES-BGK models [1, 8, 16, 34, 39]. Although such both particle-based methods mimic physical processes and yield excellent results in the free molecular regime, they face significant limitations in simulating gas flow in the continuum regime. In particular, they are constrained by strict time-step and mesh-size constraints and subject to statistical noise inherently associated with stochastic sampling.

In contrast to particle-based methods, deterministic approaches rely on the numerical discretization of kinetic models and dedicated numerical schemes. Most of theses methods employ a Discrete Velocity Model (DVM) or Discrete Ordinate Method (DOM) [18], in which the continuous velocity space is discretized, yielding to a system of coupled partial differential equations. Due to the conservation properties inherent in kinetic models, finite volume schemes are generally preferred. Several deterministic solvers which use for most of them the DVM technique can be found in the review [27]. When combined with high-order spatial reconstructions, these schemes can achieve high accuracy free of statistical noise. Moreover, an implicit or semi-implicit time discretization is commonly adopted to handle the particle interaction term of the model equation which can become stiff in the continuum regime. However, for a long time, the numerical fluxes of these finite volume solvers were derived from a purely kinetic viewpoint, treating transport and collision processes separately. This methodology significantly impairs the accuracy of the schemes in the hydrodynamic limit where the Navier-Stokes equations apply. Indeed, for this type of scheme in this regime, extra numerical diffusion is commonly observed in addition to the physical diffusion induced by the Navier-Stokes asymptotics. With these schemes, accurate results free from this diffusion can only be obtained when the mesh size and time step resolve the kinetic scales, making simulations prohibitively expensive as the flow approaches the continuum regime. In some sense, this deterministic splitting and its consequences can be compared with the particle-based splitting methods.

Since 2010, the research group led by K. Xu has been developing a unified method known as the "Unified Gas-Kinetic Scheme" [43, 46]. This method introduces a coupled treatment of transport and collision processes within the flux evaluation, enabling accurate simulations across all flow regimes with a computational cost that is independent of the gas rarefaction and free from extra-diffusion phenomena. Crucially, UGKS naturally recovers both the correct free molecular and hydrodynamic behavior described by the Navier-Stokes equations without requiring kinetic-scale resolution making it an Asymptotic Preserving (AP) scheme. Furthermore, it also naturally leads to a truly multi-scale scheme making it particularly well-suited for configurations involving strong variations in the degree of rarefaction such as those encountered downstream of a nozzle in atmospheric reentry or space propulsion applications, where continuum and rarefied regimes coexist.

The UGKS has only been applied so far to monoatomic ES-BGK models [11, 24]. Here, we propose to extend it to phenomena typical of diatomic models, such as rotation and vibration energy exchanges, as described by [2, 13, 26, 33]. This extension raises two main difficulties: first, the incorporation of the pseudo-equilibrium of ES-BGK models into the construction of UGKS fluxes; and second, the development of a robust temporal approximation for the stiff relaxation operator. Indeed, a naive application of the UGKS approach fails here, in the near-continuum regime, where we observe some instabilities. In this work, we analyze the sources of this lack of robustness and we propose various approximations that make our scheme much more robust.

The outline of the paper is as follows. Section 2 is dedicated to the presentation of various ellipsoidal statistical BGK models. Section 3 presents and adapts the unified gas kinetic scheme to ES-BGK models. The results of numerical simulations conducted on several test cases are provided in Section 4. Finally, Section 5 presents the conclusions of this paper.

2 The model

2.1 Mass distribution and relationship with macroscopic quantities

To describe the dynamics of a polyatomic gas, a microscopic distribution of mass F is used. It is defined over the phase space $(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i}) \in \mathbb{R}^+ \times \mathbb{R}^{D_{\boldsymbol{x}}} \times \mathbb{R}^{D_{\boldsymbol{v}}} \times \mathbb{R}^+ \times \mathbb{N}$ such that, at any time $t \in \mathbb{R}^+$, $F(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i})\, d\boldsymbol{x}\, d\boldsymbol{v} d\epsilon$ represents the mass of gas with vibration energy associated with the ith discrete excitation level, in the volume $d\boldsymbol{x} d\boldsymbol{v} d\epsilon$ centered at the spatial point $\boldsymbol{x}$, the particle velocity $\boldsymbol{v}$ and the rotational energy ϵ. Formally, $D_{\boldsymbol{x}}$ and $D_{\boldsymbol{v}}$ represent the number of spatial and kinetic dimensions, respectively, and δ will denote the number of continuous rotational degrees of freedom.

In practice, $D_{\boldsymbol{x}}$ and $D_{\boldsymbol{v}}$ are typically set to 3, as particles can exist and move in three-dimensional space, and δ is set to 2 for diatomic molecules. For such molecules, the harmonic oscillator model is commonly used to determine the vibration energy distribution, relying on a gas-characteristic vibration temperature T_0. In this model, the vibration energy of the ith excitation level is $\mathrm{i}R_s T_0$, where R_s is the constant of the gas. In more general cases ($\delta \geqslant 2$), the vibration energy could be modeled by a summation over multiple harmonic oscillators corresponding to multiple vibration modes. For the purposes of this paper, only diatomic molecules are considered.

The density, momentum, and total energy, which are macroscopic quantities depending on space and time only, are recovered as velocity and internal energy moments of the microscopic distribution:

$$\rho = \langle F \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}, \qquad \rho \boldsymbol{u} = \langle \boldsymbol{v} F \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}, \qquad E = E_c + E_{tr} + E_{rot} + E_{vib}, \tag{1}$$

$$E_c = \frac{1}{2}\rho |\boldsymbol{u}|^2, \qquad E_{tr} = \left\langle \frac{1}{2}|\boldsymbol{v} - \boldsymbol{u}|^2 F \right\rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}, \qquad E_{rot} = \langle \epsilon F \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}\,, \qquad E_{vib} = \langle \mathrm{i} R_s T_0 F \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}\,, \tag{2}$$

with $\langle \chi \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}} = \sum_{\mathrm{i}=0}^{+\infty} \int_{\mathbb{R}^+} \int_{\mathbb{R}^{D_{\boldsymbol{v}}}} \chi d\boldsymbol{v} d\epsilon$ for any distribution $\chi(\boldsymbol{v}, \epsilon, \mathrm{i})$. Moreover, additional quantities, such as the temperatures associated with the different energy modes, can be derived as follows:

$$E_{tr} = \frac{D_{\boldsymbol{v}}}{2}\rho R_s T_{tr}, \quad E_{rot} = \frac{\delta}{2}\rho R_s T_{rot}, \quad E_{vib} = \rho \frac{R_s T_0}{\exp\left(T_0/T_{vib}\right) - 1}. \tag{3}$$

In these expressions, E_c is the density of kinetic energy and the subscripts tr, rot and vib refer to the translational, the rotational and the vibrational modes of energy. The equilibrium temperature T_{eq} is associated with all these internal modes simultaneously:

$$E_{tr} + E_{rot} + E_{vib} = \frac{D_{\boldsymbol{v}}}{2}\rho R_s T_{eq} + \frac{\delta}{2}\rho R_s T_{eq} + \rho \frac{R_s T_0}{\exp\left(T_0/T_{eq}\right) - 1}. \tag{4}$$

Finally, for further needs, we introduce the following invertible energy functions with respect to any positive temperature T:

$$e_{tr}(T) = \frac{D_{\boldsymbol{v}}}{2} R_s T, \qquad e_{rot}(T) = \frac{\delta}{2} R_s T, \qquad e_{vib}(T) = \frac{R_s T_0}{\exp(T_0/T) - 1}, \tag{5}$$

so that:

$$E_{tr} = \rho e_{tr}(T_{tr}), \qquad E_{rot} = \rho e_{rot}(T_{rot}), \qquad E_{vib} = \rho e_{vib}(T_{vib}). \tag{6}$$

2.2 The Ellipsoidal-Statistical BGK models for diatomic gases

The distribution F is governed by a Boltzmann type of equation for dilute polyatomic gas in the absence of any external force field:

$$(\partial_t F + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} F)(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i}) = Q(F(t, \boldsymbol{x}, \cdot, \cdot, \cdot))(\boldsymbol{v}, \epsilon, \mathrm{i}). \tag{7}$$

The right-hand term of this equation is the *collision operator*, which is the term modeled to simplify and speed up simulations of rarefied flows. In the ES-BGK models framework [2, 13, 17, 26, 33], it is proposed to model the collision operator as a relaxation toward a local anisotropic equilibrium:

$$Q(F(t, \boldsymbol{x}, \cdot, \cdot, \cdot))(\boldsymbol{v}, \epsilon, \mathrm{i}) = \frac{1}{\tau}(G[F] - F)(\boldsymbol{v}, \epsilon, \mathrm{i}), \tag{8}$$

where $G[F](\boldsymbol{v},\epsilon,\mathrm{i})$ is a combination of multiple pseudo-equilibrium distributions, each corresponding to a specific energy mode. In the most general case, accounting for translational, rotational, and vibrational degrees of freedom of particles, the near-equilibrium state $G[F]$ is expressed as the product of G_{tr}, G_{rot}, and G_{vib} defined as follows:

$$G_{tr}(\boldsymbol{v}) = \frac{\rho}{\sqrt{\det(2\pi\mathcal{T})}}\exp\left(-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{u})^{\top}\mathcal{T}^{-1}(\boldsymbol{v}-\boldsymbol{u})\right), \tag{9}$$

$$G_{rot}(\epsilon) = \frac{\Lambda_\delta}{(R_s T_{rot}^{rel})^{\delta/2}}\epsilon^{\frac{\delta-2}{2}}\exp\left(-\frac{\epsilon}{R_s T_{rot}^{rel}}\right), \tag{10}$$

$$G_{vib}(\mathrm{i}) = (1-\exp(-T_0/T_{vib}^{rel}))\exp\left(-\mathrm{i}\frac{T_0}{T_{vib}^{rel}}\right). \tag{11}$$

Here, the constant Λ_δ is defined in terms of the standard gamma function as $\Lambda_\delta = 1/\Gamma(\delta/2)$. The terms T_{rot}^{rel} and T_{vib}^{rel} represent the rotational and vibrational relaxation temperatures, respectively, while $\mathcal{T}/R_s$ denotes a relaxation temperature tensor. These three last quantities describe the exchange of energy between translational, rotational and vibrational modes. The tensor $\mathcal{T}/R_s$ is related to the anisotropic tensor of temperature Θ/R_s, the Prandtl number Pr, and the relaxation translational temperature T_{tr}^{rel} as follows:

$$\mathcal{T} = R_s T_{tr}^{rel} I + \left(1-\frac{1}{\mathrm{Pr}}\right)[\Theta - R_s T_{tr} I], \qquad \Theta = \frac{1}{\rho}\langle(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})F\rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}, \tag{12}$$

where I is the unit tensor of $\mathbb{R}^{D_{\boldsymbol{v}}}$, and the modal relaxation temperatures are defined by:

$$T_{tr}^{rel} = e_{tr}^{-1}(e_{tr}^{rel}), \qquad T_{rot}^{rel} = e_{rot}^{-1}(e_{rot}^{rel}), \qquad T_{vib}^{rel} = e_{vib}^{-1}(e_{vib}^{rel}). \tag{13}$$

While the previous expressions are common to all ES-BGK models, relaxation modal energies $e_{tr}^{rel}, e_{rot}^{rel}$ and e_{vib}^{rel} need to be defined. In fact, they differ from one ES-BGK model to another, depending on the energy relaxation process modeled.

The first ES-BGK model was proposed by Holway [17]. Since it focuses solely on monoatomic gases, the rotational and vibrational modes of particles are not considered. Thus, the phases ϵ and i of distributions F and G in (7–8) are omitted, and G_{rot}, G_{vib}, e_{rot}^{rel} and e_{vib}^{rel} are not taken into account. As a result, only the relaxation translation energy e_{tr}^{rel} needs to be defined:

$$\text{Holway [17]}: \; e_{tr}^{rel} = e_{tr}(T_{tr}). \tag{14}$$

Notably, the monoatomic BGK model can be also formulated within this formalism simply by considering a unitary Prandtl number in the Holway model. Thus, the BGK model is considered here as belonging to the "ES-BGK class of models", and all the concepts developed in this paper naturally apply to it.

In the Andriès model [2], polyatomic molecules are described using only continuous internal degrees of freedom, thereby generally neglecting vibrational excitation. Consequently, phase i of distributions F and G in (7–8) is omitted, and G_{vib} and e_{vib}^{rel} are not taken into account. As a result, only translational and rotational energies remain and are modeled to relax toward their equilibrium state with a characteristic time $\tau_{rot} = Z_{rot}\tau$, where Z_{rot} is interpreted as the average number of collisions required to involve energy exchange with the rotational mode. Thus, the Andriès model [2] is constructed to satisfy the following local relaxation laws:

$$\text{Andriès [2]}: \begin{cases} \dfrac{d}{dt}e_{tr}(T_{tr}) = \dfrac{1}{Z_{rot}\tau}\left[e_{tr}(T_{eq}) - e_{tr}(T_{tr})\right], \\ \dfrac{d}{dt}e_{rot}(T_{rot}) = \dfrac{1}{Z_{rot}\tau}\left[e_{rot}(T_{eq}) - e_{rot}(T_{rot})\right], \end{cases} \tag{15}$$

and thus, the corresponding energy relaxation terms must be defined as:

$$\text{Andriès [2]}: \begin{cases} e_{tr}^{rel} = e_{tr}(T_{tr}) + \dfrac{1}{Z_{rot}}\left[e_{tr}(T_{eq}) - e_{tr}(T_{tr})\right], \\ e_{rot}^{rel} = e_{rot}(T_{rot}) + \dfrac{1}{Z_{rot}}\left[e_{rot}(T_{eq}) - e_{rot}(T_{rot})\right]. \end{cases} \tag{16}$$

The Dauvois model [13] is based on the same concepts as the Andriès model. Moreover, it additionally incorporates the vibrational mode of energy, which relaxes toward its equilibrium state with a characteristic time

$\tau_{vib} = Z_{vib}\tau$. The parameter Z_{vib} also serves a similar role of that of Z_{rot}. However, the relaxation dynamic of translational and rotational modes differs from the Andriès model. It is generally accepted that the vibrational mode relaxes more slowly than the two others. Consequently, the translational and rotational modes are supposed to reach a pseudo-equilibrium state of temperature $T_{tr,rot}$ accounting for these modes only, before relaxing toward the final equilibrium state. Thus $T_{tr,rot}$ is defined such as:

$$e_{tr}(T_{tr,rot}) + e_{rot}(T_{tr,rot}) = e_{tr}(T_{tr}) + e_{rot}(T_{rot}), \tag{17}$$

and this model is constructed to satisfy the following local relaxation laws:

$$\text{Dauvois [13]} : \begin{cases} \dfrac{d}{dt}e_{tr}(T_{tr}) = \dfrac{1}{Z_{rot}\tau}\left[e_{tr}(T_{tr,rot}) - e_{tr}(T_{tr})\right] + \dfrac{1}{Z_{vib}\tau}\left[e_{tr}(T_{eq}) - e_{tr}(T_{tr,rot})\right], \\ \dfrac{d}{dt}e_{rot}(T_{rot}) = \dfrac{1}{Z_{rot}\tau}\left[e_{rot}(T_{tr,rot}) - e_{rot}(T_{rot})\right] + \dfrac{1}{Z_{vib}\tau}\left[e_{rot}(T_{eq}) - e_{rot}(T_{tr,rot})\right], \\ \dfrac{d}{dt}e_{vib}(T_{vib}) = \dfrac{1}{Z_{vib}\tau}\left[e_{vib}(T_{eq}) - e_{vib}(T_{vib})\right]. \end{cases} \tag{18}$$

Hence, the corresponding energy relaxation terms are:

$$\text{Dauvois [13]} : \begin{cases} e_{tr}^{rel} = e_{tr}(T_{tr}) + \dfrac{1}{Z_{rot}}\left[e_{tr}(T_{tr,rot}) - e_{tr}(T_{tr})\right] + \dfrac{1}{Z_{vib}}\left[e_{tr}(T_{eq}) - e_{tr}(T_{tr,rot})\right], \\ e_{rot}^{rel} = e_{rot}(T_{rot}) + \dfrac{1}{Z_{rot}}\left[e_{rot}(T_{tr,rot}) - e_{rot}(T_{rot})\right] + \dfrac{1}{Z_{vib}}\left[e_{rot}(T_{eq}) - e_{rot}(T_{tr,rot})\right], \\ e_{vib}^{rel} = e_{vib}(T_{vib}) + \dfrac{1}{Z_{vib}}\left[e_{vib}(T_{eq}) - e_{vib}(T_{vib})\right] \end{cases} \tag{19}$$

Finally, the last and certainly most recent ES-BGK model for diatomic gases is the Pfeiffer model [26, 33]. As the previous kinetic model, it also involves vibrational modes of energy based on the simple harmonic oscillator model. However, relaxation processes differ. As realized in the description of relaxations [32] and in DSMC methods [6], relaxation times τ_{rot} and τ_{vib} are defined relative to a collision time τ_C rather than the ES-BGK model relaxation time τ. Moreover, the rotation and vibration energy modes do not relax directly toward their equilibrium state, but toward their state associated with the translational temperature. This results in reproducing the following local Launder-Teller and Jeans relaxation laws:

$$\text{Pfeiffer [26, 33]} : \begin{cases} \dfrac{d}{dt}e_{tr}(T_{tr}) = -\dfrac{d}{dt}e_{rot}(T_{rot}) - \dfrac{d}{dt}e_{vib}(T_{vib}), \\ \dfrac{d}{dt}e_{rot}(T_{rot}) = \dfrac{1}{Z_{rot}\tau_C}\left[e_{rot}(T_{tr}) - e_{rot}(T_{rot})\right], \\ \dfrac{d}{dt}e_{vib}(T_{vib}) = \dfrac{1}{Z_{vib}\tau_C}\left[e_{vib}(T_{tr}) - e_{vib}(T_{vib})\right], \end{cases} \tag{20}$$

and the corresponding energy relaxation terms:

$$\text{Pfeiffer [26, 33]} : \begin{cases} e_{tr}^{rel} = e_{tr}(T_{tr}) - \dfrac{\tau}{Z_{rot}\tau_C}\left[e_{rot}(T_{tr}) - e_{rot}(T_{rot})\right] - \dfrac{\tau}{Z_{vib}\tau_C}\left[e_{vib}(T_{tr}) - e_{vib}(T_{vib})\right], \\ e_{rot}^{rel} = e_{rot}(T_{rot}) + \dfrac{\tau}{Z_{rot}\tau_C}\left[e_{rot}(T_{tr}) - e_{rot}(T_{rot})\right], \\ e_{vib}^{rel} = e_{vib}(T_{vib}) + \dfrac{\tau}{Z_{vib}\tau_C}\left[e_{vib}(T_{tr}) - e_{vib}(T_{vib})\right]. \end{cases} \tag{21}$$

The mean collision time τ_C is defined through a collision model, such as the VSS model [21] in which it is related to the pressure p, the fluid viscosity μ, the viscosity index ω, and the diffusion factor α by the following expression:

$$\tau_C^{VSS} = \frac{\alpha(5-2\omega)(7-2\omega)}{5(\alpha+1)(\alpha+2)}\frac{\mu}{p}. \tag{22}$$

In practice, Z_{rot} and Z_{vib} typically range from 3 to 20 and from 50 to 100,000, respectively. These values can either be treated as constants or expressed as functions of the translational temperature, as described in [25, 31,

32]. In the same way, the Prandtl number Pr can be taken as constant, tabulated, or computed using the heat capacity ratio γ and the Eucken formula:

$$\mathrm{Pr} = \frac{4\gamma}{9\gamma - 5}. \tag{23}$$

Furthermore, the relaxation time τ of ES-BGK models is related to the fluid viscosity and the pressure as follows:

$$\tau = \frac{\mu}{\mathrm{Pr}p}, \qquad p = \rho R_s T_{tr}. \tag{24}$$

Finally, all these models have been proven [2, 13, 26], under reasonable conditions, to be well defined, conserve mass, momentum, and total energy, and admit the "Maxwellian state" M below as full equilibrium:

$$M[F](\boldsymbol{v}, \epsilon, \mathrm{i}) = M_{tr}(\boldsymbol{v}) M_{rot}(\epsilon) M_{vib}(\mathrm{i}), \tag{25}$$

with

$$M_{tr}(\boldsymbol{v}) = \frac{\rho}{(2\pi R_s T_{eq})^{D_{\boldsymbol{v}}/2}} \exp\left(-\frac{|\boldsymbol{v}-\boldsymbol{u}|^2}{2R_s T_{eq}}\right), \tag{26}$$

$$M_{rot}(\epsilon) = \frac{\Lambda_\delta}{(R_s T_{eq})^{\delta/2}} \epsilon^{\frac{\delta-2}{2}} \exp\left(-\frac{\epsilon}{R_s T_{eq}}\right), \tag{27}$$

$$M_{vib}(\mathrm{i}) = (1 - \exp(-T_0/T_{eq})) \exp\left(-\mathrm{i}\frac{T_0}{T_{eq}}\right). \tag{28}$$

They also yield correct transport coefficients and Prandtl number in the hydrodynamic limit, and satisfy the H-theorem.

2.3 The reduced models

As presented previously, the microscopic distribution F is defined through variables $(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i}) \in \mathbb{R}^+ \times \mathbb{R}^{D_{\boldsymbol{x}}} \times \mathbb{R}^{D_{\boldsymbol{v}}} \times \mathbb{R}^+ \times \mathbb{N}$. Directly solving an ES-BGK model (8) can be computationally expensive due to unnecessary memory storage and computation of all phases. Indeed, for most aerodynamic problems, only macroscopic quantities such as mass density, velocity, temperature of different energy modes, pressure, stress, or heat flux are generally sufficient. Thus, the internal energy phases can be ignored in computations by using a reduction technique, that reduces the number of phases without any approximation. This kind of technique was primary used in [12] to reduce velocity space from $\mathbb{R}^3$ to $\mathbb{R}$ in the study of 1D shock problems with the monoatomic BGK model, and also used in [19] for a polyatomic model. Here, following the work realized by [13, 26], we introduce three reduced distributions:

$$\begin{pmatrix} f \\ g \\ h \end{pmatrix}(t, \boldsymbol{x}, \boldsymbol{v}) = \sum_{\mathrm{i}=0}^{+\infty} \int_{\mathbb{R}^+} \begin{pmatrix} 1 \\ \epsilon \\ \mathrm{i} R_s T_0 \end{pmatrix} F(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i}) d\epsilon. \tag{29}$$

The macroscopic quantities can be exactly recovered from f, g and h as follows:

$$\rho = \langle f \rangle, \quad \rho \boldsymbol{u} = \langle \boldsymbol{v} f \rangle, \quad E = \underbrace{\langle \tfrac{1}{2}|\boldsymbol{v}|^2 f \rangle}_{E_c + E_{tr}} + \underbrace{\langle g \rangle}_{E_{rot}} + \underbrace{\langle h \rangle}_{E_{vib}}, \quad \Theta = \tfrac{1}{\rho}\langle (\boldsymbol{v}-\boldsymbol{u}) \otimes (\boldsymbol{v}-\boldsymbol{u}) f \rangle, \tag{30}$$

where $\langle \psi \rangle$ is defined as $\int_{\mathbb{R}^{D_{\boldsymbol{v}}}} \psi d\boldsymbol{v}$ for any distribution $\psi(\boldsymbol{v})$. Complementary macroscopic quantities, such as temperatures, are still computed using the relations (3) and finally, the reduced equations derived for f, g, and h are:

$$\partial_t f + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f = \frac{1}{\tau}(G_{tr} - f), \tag{31}$$

$$\partial_t g + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} g = \frac{1}{\tau}(e_{rot}^{rel} G_{tr} - g), \tag{32}$$

$$\partial_t h + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} h = \frac{1}{\tau}(e_{vib}^{rel} G_{tr} - h). \tag{33}$$

All mathematical properties of the original model are maintained in the reduced one. Furthermore, additional reduced distributions can be introduced in a similar way for 2D or 1D flows to reduce the memory footprint and speed up their numerical resolution.

3 The Unified-Gas Kinetic Scheme

The numerical method will be presented in a 1D spatial framework for simplicity, though it can be extended to 2D or 3D either by a directional splitting approach, as used in the numerical experiments section of this paper and in [20], or through a truly multi-dimensional construction as in [45] for the Gas-Kinetic Scheme (GKS). We begin by outlining the framework before constructing the UGKS for an unreduced ES-BGK model that includes vibration energy [13, 26]. This approach naturally lends itself to the development of a practical scheme compatible with any chosen reduction technique, including, for instance, the one exposed previously. Note that the same methodology could be applied to ES-BGK models accounting solely for translational or translational-rotational energy [2, 17].

3.1 A Discrete Velocity Model and Finite Volume framework

The ES-BGK model is an integro-differential equation expressed in an advection-relaxation form, which makes the finite volume framework intrinsically well suited. In that sense, the time space $\mathbb{R}^+$ and the physical space $\mathbb{R}^{D_{\boldsymbol{x}}}$ ($D_{\boldsymbol{x}} = 1$) are divided into intervals $[t^n, t^{n+1}]$ and $[x_{i-1/2}, x_{i+1/2}]$, respectively. For simplicity, the spatial interval length will be constant, denoted as $\Delta x = x_{i+1/2} - x_{i-1/2}$. Finally, following the methodology of Discrete Velocity Models (DVM), we consider a finite velocity set $\mathcal{V} \subset \mathbb{R}^{D_{\boldsymbol{v}}}$ and projection of the model equation (7–8) on this set:

$$\forall \boldsymbol{v}_k \in \mathcal{V}, \quad (\partial_t F + \boldsymbol{v}_k \cdot \nabla_{\boldsymbol{x}} F)(t, \boldsymbol{x}, \boldsymbol{v}_k, \epsilon, \mathrm{i}) = \frac{1}{\tau}(G[F] - F)(t, \boldsymbol{x}, \boldsymbol{v}_k, \epsilon, \mathrm{i}). \tag{34}$$

Note that we use the italic font for the index i of a space cell, while the roman font is used for the ith excitation level of vibrational energy.

In association with the velocity set $\mathcal{V}$, we choose a quadrature rule on the velocity phase that enables the computation of the moments of the microscopic distributions set. The choice of the quadrature is not the primary focus here; it can be, for example, a midpoint rule or a Newton-Cotes quadrature. Nevertheless, care must be taken in defining the discrete equilibrium distribution to ensure the conservation of moments and the decrease of entropy [28, 29]. In the following, $\langle\cdot\rangle_{\mathcal{V}}$ will denote the velocity quadrature rule integration of a distribution $\psi(\boldsymbol{v})$, while $\langle\cdot\rangle_{\mathcal{V},\epsilon,\mathrm{i}}$ will refer to the combination of this velocity quadrature rule integration, continuous integration over the phase ϵ, and summation over the phase i of a distribution $\Psi(\boldsymbol{v}, \epsilon, \mathrm{i})$.

As is common in finite volume methods, we introduce the distributions $F^n_{i,k}(\cdot,\cdot)$ and $G^n_{i,k}(\cdot,\cdot)$ defined as the mean values of F and G on a spatial cell $[x_{i-1/2}, x_{i+1/2}]$ at time t^n, and velocity $\boldsymbol{v}_k$:

$$F^n_{i,k}(\epsilon, \mathrm{i}) = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} F(t^n, x, \boldsymbol{v}_k, \epsilon, \mathrm{i})dx, \qquad G^n_{i,k}(\epsilon, \mathrm{i}) = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} G(t^n, x, \boldsymbol{v}_k, \epsilon, \mathrm{i})dx. \tag{35}$$

Integrating (34) over the spatial volume $[x_{i-1/2}, x_{i+1/2}]$ for the time interval $[t^n, t^{n+1}]$ leads to the classical finite volume formulation:

$$F^{n+1}_{i,k}(\epsilon, \mathrm{i}) - F^n_{i,k}(\epsilon, \mathrm{i}) + \frac{\Delta t}{\Delta x}\left[\phi^n_{i+\frac{1}{2},k} - \phi^n_{i-\frac{1}{2},k}\right](\epsilon, \mathrm{i}) = \int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{G - F}{\tau}(t, x, \boldsymbol{v}_k, \epsilon, \mathrm{i})dxdt, \tag{36}$$

$$\phi^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \boldsymbol{v}_k F(t, x_{i+\frac{1}{2}}, \boldsymbol{v}_k, \epsilon, \mathrm{i})dt. \tag{37}$$

Finally, for further need, we also introduce the discrete moments $\boldsymbol{W}^n_i$ of $F^n_{i,k}(\cdot,\cdot)$ with respect to the operator $\langle\cdot\rangle_{\mathcal{V},\epsilon,\mathrm{i}}$:

$$\boldsymbol{W}^n_i = \left\langle \boldsymbol{m}_k(\epsilon, \mathrm{i}) F^n_{i,k}(\epsilon, \mathrm{i})\right\rangle_{\mathcal{V},\epsilon,\mathrm{i}}, \qquad \boldsymbol{m}_k(\epsilon, \mathrm{i}) = \begin{pmatrix} 1 \\ \boldsymbol{v}_k \\ \boldsymbol{v}_k \otimes \boldsymbol{v}_k \\ \epsilon \\ \mathrm{i}R_s T_0 \end{pmatrix}. \tag{38}$$

Integrating (36–37) with $\langle\cdot\rangle_{\mathcal{V},\epsilon,\mathrm{i}}$ naturally leads to a finite volume scheme on moments $\boldsymbol{W}$.

The finite-volume formulation (36–37) is an exact expression. However, the relaxation and flux terms must be approximated. Generally, they are considered separately using a splitting method. The relaxation term is commonly approximated by a quadrature method in time (e.g., forward Euler, backward Euler, trapezoidal formula), while the flux term is estimated either by a high-order reconstruction or by characteristic techniques, applied to the collisionless transport equation of F.

3.2 The UGKS fluxes

3.2.1 A multi-scale formulation of the flux part of the numerical scheme

The key idea of the UGKS [43] is to use the entire model equation (7–8) to express the evolution of the distribution F during the time interval $[t^n, t]$ at cell interface position $(x_{i\pm1/2}, \boldsymbol{v}_k, \epsilon, \mathrm{i})$, which is required to compute the numerical fluxes (37). This approach differs from conventional methods by accounting not only for free transport of F but also for its relaxation toward equilibrium during its transport itself. Specifically, by using the characteristic method on the model equation (7–8), for a time-independent τ over the interval $[t^n, t]$ and for any $(\boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i})$, we get:

$$F(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i}) = \exp\left(-\frac{t-t^n}{\tau}\right) F(t^n, \boldsymbol{x} - \boldsymbol{v}(t-t^n), \epsilon, \mathrm{i}) + \int_{t^n}^{t} \exp\left(-\frac{t-s}{\tau}\right) \frac{1}{\tau} G(s, \boldsymbol{x} - \boldsymbol{v}(t-s), \boldsymbol{v}, \epsilon, \mathrm{i}) ds. \quad (39)$$

The above equation, which expresses the distribution F at $(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i})$, is a balance between the diminishing collisionless transport of the initial distribution and the transport of the emerging equilibrium distribution. This balance is governed by the time difference $t - t^n$ and the relaxation time τ. The larger the time difference and the smaller the relaxation time, the greater the influence of the equilibrium distribution on the instantaneous microscopic distribution.

3.2.2 Second order reconstruction of both microscopic and macroscopic parts of the flux

An exact numerical flux based on equation (39) would require knowledge of the distributions F and G at any $(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i})$. For practical computations, approximations must be made. To develop a second-order scheme, these distributions are approximated by linear reconstructions based on discretized distributions (35). A common assumption in the finite volume framework is to consider these mean values at the center of the spatial cells. In the UGKS framework [43], τ is considered constant near the interface, reconstructions for F are performed for each cell, while reconstructions for G are realized for each cell interface as illustrated in Figure 1 and formalized below:

$$\begin{aligned} &F(t^n, x, \boldsymbol{v}_k, \epsilon, \mathrm{i}) \approx \mathbb{F}^n_{i,k}(x, \epsilon, \mathrm{i}) = F^n_{i,k}(\epsilon, \mathrm{i}) + \delta_x F^n_{i,k}(\epsilon, \mathrm{i})(x - x_i), \\ &G(t, x, \boldsymbol{v}_k, \epsilon, \mathrm{i}) \approx \mathbb{G}_{i+\frac{1}{2},k}(t, x, \epsilon, \mathrm{i}) = G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) + \delta_x G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i})(x - x_{i+\frac{1}{2}}) + \delta_t G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i})(t - t^n), \end{aligned} \quad (40)$$

where $\delta_x F^n_{i,k}$, and $(G, \delta_x G, \delta_t G)^n_{i+1/2,k}$ should be discrete approximations of microscopic and equilibrium distributions, and their partial derivatives.

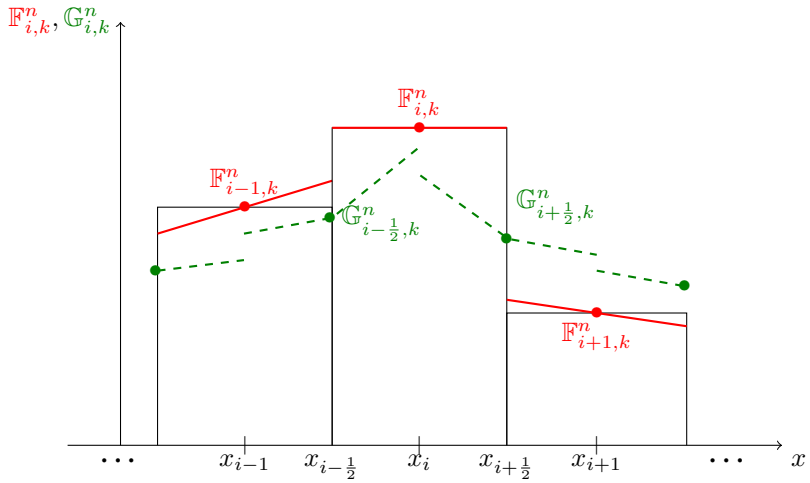


Figure 1: Spatial reconstructions at time t^n and fixed ϵ, i of microscopic and equilibrium distributions in the UGKS framework.

For stability purposes, $\delta_x F$ is defined in each cell as a limited slope based on forward and backward slope and a Total Variation Diminishing (TVD) limiter ξ:

$$\delta_x F^n_{i,k}(\epsilon, \mathrm{i}) = \xi\left(\frac{F^n_{i,k}(\epsilon, \mathrm{i}) - F^n_{i-1,k}(\epsilon, \mathrm{i})}{\Delta x}, \frac{F^n_{i+1,k}(\epsilon, \mathrm{i}) - F^n_{i,k}(\epsilon, \mathrm{i})}{\Delta x}\right). \quad (41)$$

The choice of the limiter is arbitrary. A multitude of possibility exists in the literature. Commonly, the van Leer limiter is used by the authors of UGKS [10, 40, 43]. Then, the microscopic part of (39) is replaced by the appropriate linear approximation of the microscopic distribution, depending on the cell location of $\boldsymbol{x} - \boldsymbol{v}(t - t^n)$. By the way and for further need, the slope of F at the cell interface is defined by the following upwind approximation:

$$\delta_x F^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \begin{cases} \delta_x F^n_{i,k}(\epsilon, \mathrm{i}) & \text{if } v_{k,x} \geqslant 0, \\ \\ \delta_x F^n_{i+1,k}(\epsilon, \mathrm{i}) & \text{if } v_{k,x} < 0. \end{cases} \tag{42}$$

For the macroscopic part of (39), we first define $G^n_{i+1/2,k}(\epsilon, \mathrm{i})$. This term is entirely determined by the moments $\boldsymbol{W}^n_{i+1/2}$ of F at the cell interface $i + 1/2$ and time t^n, see (8–13). Consequently, the linear reconstructions $\mathbb{F}^n_{i,k}$ and $\mathbb{F}^n_{i+1,k}$ are employed to approximate this distribution and its moments:

$$F(t^n, x_{i+\frac{1}{2}}, \boldsymbol{v}_k, \epsilon, \mathrm{i}) \approx F^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \begin{cases} \mathbb{F}^n_{i,k}(x_{i+\frac{1}{2}}, \epsilon, \mathrm{i}) & \text{if } v_{k,x} \geqslant 0, \\ \\ \mathbb{F}^n_{i+1,k}(x_{i+\frac{1}{2}}, \epsilon, \mathrm{i}) & \text{if } v_{k,x} < 0, \end{cases} \tag{43}$$

$$\begin{pmatrix} \rho \\ \rho \boldsymbol{u} \\ \rho \boldsymbol{u} \otimes \boldsymbol{u} + \rho \Theta \\ E_{rot} \\ E_{vib} \end{pmatrix} (t^n, x_{i+\frac{1}{2}}) \approx \boldsymbol{W}^n_{i+\frac{1}{2}} = \left\langle \boldsymbol{m}_k(\epsilon, \mathrm{i}) F^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) \right\rangle_{\mathcal{V},\epsilon,\mathrm{i}}. \tag{44}$$

Finally, $(\mathcal{T}, e^{rel}_{rot}, e^{rel}_{vib})^n_{i+1/2}$ are determined when necessary, using $\boldsymbol{W}^n_{i+1/2}$, and the appropriate formulation of the relaxation energies and tensor (12, 14, 16, 19) or (21), after which $G^n_{i+1/2,k}(\epsilon, \mathrm{i})$ can be deduced.

3.2.3 Construction of discrete macroscopic derivative terms $\delta_x G$ and $\delta_t G$ and numerical flux induced

The remaining terms in (40) to be approximated are the discrete derivative components of $\mathbb{G}$, denoted by $\delta_x G$ and $\delta_t G$. For the BGK model, spatial derivatives of the equilibrium distribution are computed on both sides of the interface as proposed in [43] and illustrated in Figure 1. This approach is inherited from a modification introduced in [42] of the original Gas-Kinetic Scheme (GKS). Applying this approach to ES-BGK models naturally yields:

$$\delta_x G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \begin{cases} \delta^-_x G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) & \text{if } v_{kx} \geqslant 0, \\ \\ \delta^+_x G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) & \text{if } v_{kx} < 0, \end{cases} \tag{45}$$

where $\delta^-_x G$ and $\delta^+_x G$ are defined as follows. Using an exponential form of the continuous state $G = \epsilon^{(\delta-2)/2} \exp(\boldsymbol{m} \cdot \boldsymbol{\beta})$ with $\boldsymbol{m} = (1, \boldsymbol{v}, \boldsymbol{v} \otimes \boldsymbol{v}, \epsilon, \mathrm{i} R_s T_0)^\top$ and $\boldsymbol{\beta}$ a vector related to macroscopic quantities, it can be shown that the derivative of the continuous pseudo-equilibrium state G with respect to x is the inner product of $\boldsymbol{m}$ and a macroscopic vector: $\partial_x G = (\boldsymbol{m} \cdot \partial_x \boldsymbol{\beta}) G$. This property results from the independence of the phases $\boldsymbol{v}, \epsilon, \mathrm{i}$ with respect to x. So, we are looking for discrete derivatives $\delta^\pm_x G$ in the form:

$$\delta^\pm_x G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \boldsymbol{m}_k(\epsilon, \mathrm{i}) \cdot \delta^\pm_x \boldsymbol{\beta}^n_{i+\frac{1}{2}} G^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}). \tag{46}$$

Using a chain rule, we get $\partial_x \boldsymbol{\beta} = \partial_{\boldsymbol{W}} \boldsymbol{\beta}\ \partial_x \boldsymbol{W}$, where $\boldsymbol{W}$ are the moments of F with respect to the microscopic vector of moments $\boldsymbol{m}$ and the quadrature rule $\langle \cdot \rangle_{\mathcal{V},\epsilon,\mathrm{i}}$ (38), and $\partial_{\boldsymbol{W}} \boldsymbol{\beta}$ is a Jacobian matrix that can be derived analytically. Thus, $\delta^\pm_x \boldsymbol{\beta}$ is computed as:

$$\delta^\pm_x \boldsymbol{\beta}^n_{i+\frac{1}{2}} = \left[\partial_{\boldsymbol{W}} \boldsymbol{\beta} \right]^n_{i+\frac{1}{2}} \delta^\pm_x \boldsymbol{W}^n_{i+\frac{1}{2}}, \tag{47}$$

where the left and right-sided discrete derivatives of $\boldsymbol{W}$ are:

$$\delta^-_x \boldsymbol{W}^n_{i+\frac{1}{2}} = \frac{\boldsymbol{W}^n_{i+\frac{1}{2}} - \boldsymbol{W}^n_i}{\Delta x/2}, \qquad \delta^+_x \boldsymbol{W}^n_{i+\frac{1}{2}} = \frac{\boldsymbol{W}^n_{i+1} - \boldsymbol{W}^n_{i+\frac{1}{2}}}{\Delta x/2}. \tag{48}$$

For the temporal derivative, the same methodology applies. We set:

$$\delta_t G^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}) = \boldsymbol{m}_k(\epsilon,\mathrm{i}) \cdot \delta_t \boldsymbol{\beta}^n_{i+\frac{1}{2}} G^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}), \tag{49}$$

with

$$\delta_t \boldsymbol{\beta}^n_{i+\frac{1}{2}} = \left[\partial_{\boldsymbol{W}} \boldsymbol{\beta}\right]^n_{i+\frac{1}{2}} \delta_t \boldsymbol{W}^n_{i+\frac{1}{2}}. \tag{50}$$

For the discrete temporal macroscopic derivative $\delta_t \boldsymbol{W}$, which is the discrete equivalent of $\partial_t \boldsymbol{W} = \partial_t \langle \boldsymbol{m} F \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}$, we proceed as follows. Taking the moments of (7–8), we have:

$$\partial_t \boldsymbol{W} = -\langle \boldsymbol{m}\ (\boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} F)\rangle_{\boldsymbol{v},\epsilon,\mathrm{i}} + \frac{1}{\tau}\left(\boldsymbol{V} - \boldsymbol{W}\right), \tag{51}$$

where $\boldsymbol{V}$ is the vector of moments of G with respect to $\boldsymbol{m}$. Consequently, we set at the discrete level:

$$\delta_t \boldsymbol{W}^n_{i+\frac{1}{2}} = -\left\langle \boldsymbol{m}_k(\epsilon,\mathrm{i}) v_{kx} \delta_x F^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}) \right\rangle_{\mathcal{V},\epsilon,\mathrm{i}} + \frac{1}{\tau^n_{i+\frac{1}{2}}} \left(\boldsymbol{W} - \boldsymbol{V}\right)^n_{i+\frac{1}{2}}. \tag{52}$$

This completes the construction of the discrete derivative terms of the pseudo-equilibrium G. Substituting these terms, together with the previously constructed quantities (42–44) entering the Taylor-similar expansion (40), into the numerical flux definition (37) leads to:

$$\phi^n_{i+\frac{1}{2},k} = v_{kx}\left[q_1 G + q_2 v_{kx}\delta_x G + q_3 \delta_t G + q_4 F + q_5 v_{kx}\delta_x F\right]^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}), \tag{53}$$

where the coefficients $(q_j)_{1\leqslant j\leqslant 5}$ are defined later.

3.2.4 Approximation of the discrete macroscopic derivative terms $\delta_x G$ and $\delta_t G$ by $\delta_x M$ and $\delta_t M$

Constructing the discrete spatial and temporal derivatives of (40) is a major step in assembling the flux of the UGKS. The previous section described the natural way to construct them, following the procedure applied in [43] for the BGK model. However, while this strategy performs well in the rarefied regime, numerical instabilities appear in the continuum regime when $\Delta t \gg \tau$, often leading to simulation failure.

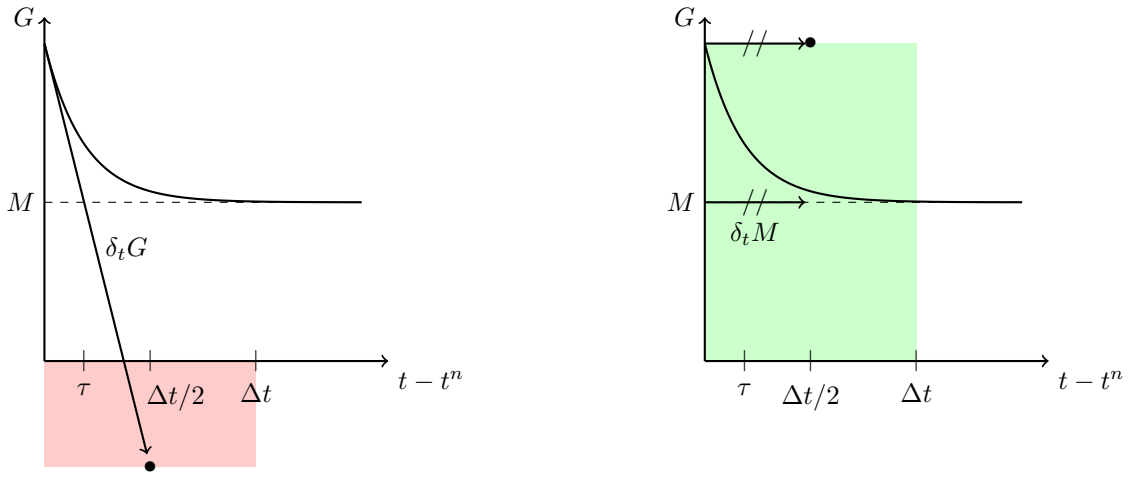
Our interpretation is that, in the continuum regime with $\Delta t \gg \tau$, due to the behavior of coefficients $(q_j)_{1\leqslant j\leqslant 5}$, the UGKS numerical flux (53) constructed with $\delta_x G$ and $\delta_t G$ becomes asymptotically equivalent to:

$$\phi^n_{i+\frac{1}{2},k} \approx v_{kx}\left[G + \frac{\Delta t}{2}\delta_t G - \tau(\delta_t G + v_{kx}\delta_x G)\right]^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}). \tag{54}$$

In this expression, the discrete derivative terms play two roles: they contribute to the viscous part associated with the Navier-Stokes asymptotic behavior, and the temporal derivative enables second-order accuracy by providing an indirect approximation of the pseudo-equilibrium G at time $t^n + \Delta t/2$. This approximation can be viewed as a forward Euler step applied to G^n using $\delta_t G^n$. However, the pseudo-equilibrium G permanently relaxes toward the Maxwellian M with relaxation time τ. Consequently, when $\Delta t \gg \tau$, a forward Euler prediction based on $\delta_t G$ may yield an excessively far state, unrepresentative of the true evolution of G between t^n and t^{n+1}. In strongly anisotropic flows, this may even result in a non-positive G at time $t^n + \Delta t/2$, as illustrated in Figure 2a. This is precisely the consequence of the stiff relaxation term that appears in the construction of $\delta_t \boldsymbol{W}$ in (52), which is then used to defined $\delta_t G$. As we can see in (52), this term is not negligible in the final numerical flux in case of strong anisotropy and large numerical step $\Delta t \gg \tau$.

Hence, the key idea employed here to prevent instabilities is to replace the pseudo-equilibrium derivatives $\delta_x G$ and $\delta_t G$ with those of the corresponding Maxwellian distribution, namely $\delta_x M$ and $\delta_t M$. As illustrated in Figure 2b, this strategy naturally avoids an "over-relaxation" of G toward M and effectively assumes that G evolves consistently with the trend of M. Since the characteristic evolution time of M is the same as that of the conservative quantities, it is in practice of the same order as Δt, which ensures a properly scaled time-derivative term.

Furthermore, this substitution remains sufficiently accurate to preserve the benefits brought by the UGKS. First, it is important to note that the continuous states G and M differ by $O(\tau)$, and therefore their derivatives also differ by $O(\tau)$. This follows from the classical Chapman-Enskog analysis (see, e.g., [2]). As a result, during a large time step $\Delta t \gg \tau$, from a continuous point of view, the average time derivative of M is very close to the average time derivative G, with an error depending on $\tau \ll \Delta t$. Second, the discrete derivative contribution in the



(a) Based on $\delta_t G$,

(b) Based on $\delta_t M$,

Figure 2: Approximation of G at time $t^n + \Delta t/2$ by using a forward Euler method starting from time t^n, illustrated in the case of an homogeneous relaxation ($\delta_t M = 0$). *The bullet points represent the approximated states at time* $t^n + \Delta t/2$ *and the shaded regions represent their corresponding time integrations over the time step* Δt.

numerical fluxes is mostly important in the continuum regime and express mainly the viscous phenomena of the Navier-Stokes equation. Thus, an approximation of $O(\tau)$ on the derivatives of equilibrium is fully consistent with the Chapman-Enskog expansion.

For these reasons, $\delta_x G$ and $\delta_t G$ are replaced respectively by $\delta_x M$ and $\delta_t M$ in (40) in the following. At the contrary, the leading term G in (40) is conserved and not replaced by its Maxwellian equivalent since this term is essential to capture the correct Prandtl number in the Navier-Stokes asymptotics. Section 4.2 provides an example that highlights the importance of this formulation for ensuring the asymptotic property and the correct behavior in the hydrodynamic limit of the scheme. Note that although the construction of the discrete derivative terms for the Shakhov and the monoatomic ES-BGK model is not explicitly detailed in [11, 44, 24], a similar substitution was applied to the Rykov model in [23] without any justification.

3.2.5 Construction of discrete macroscopic derivative terms $\delta_x M$ and $\delta_t M$

Here, the discrete spatial derivatives of the Maxwellian distribution are defined in the same manner as for the pseudo-equilibrium G in Section 3.2.3. We set:

$$\delta_x M^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \begin{cases} \delta_x^- M^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) & \text{if } v_{kx} \geqslant 0, \\ \\ \delta_x^+ M^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) & \text{if } v_{kx} < 0, \end{cases} \tag{55}$$

where $\delta_x^- M$ and $\delta_x^+ M$ are defined as follows. The equilibrium state M can be written as an inner product between the velocity-internal energy dependent vector $\boldsymbol{\eta} = (1, \boldsymbol{v}, \frac{1}{2}|\boldsymbol{v}|^2 + \epsilon + \mathrm{i}R_s T_0)^\top$ and a vector $\boldsymbol{\alpha}$ related to conservative quantities, namely: $M = \epsilon^{(\delta-2)/2} \exp(\boldsymbol{\eta} \cdot \boldsymbol{\alpha})$. Since $\boldsymbol{\eta}$ is independent of x, the continuum derivative of the equilibrium state M with respect to x takes the form: $\partial_x M = (\boldsymbol{\eta} \cdot \partial_x \boldsymbol{\alpha}) M$. Accordingly, we seek discrete derivatives $\delta_x^\pm M$ of the form:

$$\delta_x^\pm M^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}) = \boldsymbol{\eta}_k(\epsilon, \mathrm{i}) \cdot \delta_x^\pm \boldsymbol{\alpha}^n_{i+\frac{1}{2}} M^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}). \tag{56}$$

Using a chain rule, we get $\partial_x \boldsymbol{\alpha} = \partial_{\boldsymbol{U}} \boldsymbol{\alpha}\ \partial_x \boldsymbol{U}$, where $\boldsymbol{U} = (\rho, \rho \boldsymbol{u}, E)^\top$ are the conservative moments defining the Maxwellian state (25) and $\partial_{\boldsymbol{U}} \boldsymbol{\alpha}$ is a Jacobian matrix that can be derived analytically. The vector $\boldsymbol{U}$ can be deduced from $\boldsymbol{W}$ and is both used to compute $M^n_{i+1/2,k}$ using (25) and $\delta_x^\pm \boldsymbol{\alpha}$ as:

$$\delta_x^- \boldsymbol{\alpha}^n_{i+\frac{1}{2}} = \left[\partial_{\boldsymbol{U}} \boldsymbol{\alpha}\right]^n_{i+\frac{1}{2}} \frac{\boldsymbol{U}^n_{i+\frac{1}{2}} - \boldsymbol{U}^n_i}{\Delta x/2}, \qquad \delta_x^+ \boldsymbol{\alpha}^n_{i+\frac{1}{2}} = \left[\partial_{\boldsymbol{U}} \boldsymbol{\alpha}\right]^n_{i+\frac{1}{2}} \frac{\boldsymbol{U}^n_{i+1} - \boldsymbol{U}^n_{i+\frac{1}{2}}}{\Delta x/2}. \tag{57}$$

For the time derivative, the same technique is applied. However, we first have to construct a discrete temporal macroscopic derivative $\delta_t \boldsymbol{U}$, which is the discrete equivalent of $\partial_t \boldsymbol{U} = \partial_t \langle \boldsymbol{\eta} F \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}} = -\langle \boldsymbol{\eta}\ (\boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} F) \rangle_{\boldsymbol{v},\epsilon,\mathrm{i}}$. Consequently, we set at the discrete level:

$$\delta_t \boldsymbol{U}^n_{i+\frac{1}{2}} = -\left\langle \boldsymbol{\eta}_k(\epsilon,\mathrm{i}) v_{kx} \delta_x F^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}) \right\rangle_{\mathcal{V},\epsilon,\mathrm{i}}, \tag{58}$$

$$\delta_t \boldsymbol{\alpha}^n_{i+\frac{1}{2}} = \left[\partial_{\boldsymbol{U}} \boldsymbol{\alpha}\right]^n_{i+\frac{1}{2}} \delta_t \boldsymbol{U}^n_{i+\frac{1}{2}}, \tag{59}$$

$$\delta_t M^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}) = \boldsymbol{\eta}_k(\epsilon,\mathrm{i}) \cdot \delta_t \boldsymbol{\alpha}^n_{i+\frac{1}{2}} M^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}). \tag{60}$$

3.2.6 Conclusion on the flux part of the scheme

All terms of the Taylor-similar expansion (40) have been defined previously. Injecting the expansion in the numerical flux definition (37) leads to a linear combination of terms weighted by coefficients $(q_j)_{1\leqslant j\leqslant 5}$:

$$\phi^n_{i+\frac{1}{2},k} = v_{kx} \left[q_1 G + q_2 v_{kx} \delta_x M + q_3 \delta_t M + q_4 F + q_5 v_{kx} \delta_x F\right]^n_{i+\frac{1}{2},k}(\epsilon,\mathrm{i}). \tag{61}$$

The coefficients $(q_j)_{1\leqslant j\leqslant 5}$ are also defined at the cell interface $i+1/2$ and at time t^n. They depend locally on the time step Δt and the relaxation time $\tau^n_{i+1/2}$ as described below:

1. $q^n_{1,i+\frac{1}{2}} = 1 - \frac{\tau}{\Delta t}(1 - e^{-\Delta t/\tau})$,
2. $q^n_{2,i+\frac{1}{2}} = -\tau(1 + e^{-\Delta t/\tau}) + 2\frac{\tau^2}{\Delta t}(1 - e^{-\Delta t/\tau})$,
3. $q^n_{3,i+\frac{1}{2}} = \frac{\Delta t}{2} - \tau + \frac{\tau^2}{\Delta t}(1 - e^{-\Delta t/\tau})$,
4. $q^n_{4,i+\frac{1}{2}} = \frac{\tau}{\Delta t}(1 - e^{-\Delta t/\tau})$,
5. $q^n_{5,i+\frac{1}{2}} = \tau e^{-\Delta t/\tau} - \frac{\tau^2}{\Delta t}(1 - e^{-\Delta t/\tau})$,

where, for readability, τ stands for $\tau^n_{i+1/2}$. These non-constant coefficients determine the behavior of the scheme. The more the flow is in the continuum regime and $\Delta t \gg \tau$, the more the macroscopic part of the scheme dominates. Conversely, the more the flow is rarefied and $\Delta t \ll \tau$, the more the microscopic part of the scheme takes over. Thus, the numerical flux of the scheme is a sophisticated combination of equilibrium and non-equilibrium components, automatically adapting to the nature of the flow and the mesh resolution to achieve optimal accuracy.

3.3 The UGKS relaxation part

3.3.1 The trapezoidal formula

Following [24, 43], a trapezoidal formula is employed for the relaxation term. This approach can handle the stiffness of this term, particularly in the continuum regime while maintaining second-order accuracy. The numerical scheme for the microscopic distribution is therefore:

$$F^{n+1}_{i,k}(\epsilon,\mathrm{i}) = F^n_{i,k}(\epsilon,\mathrm{i}) - \frac{\Delta t}{\Delta x}\left[\phi^n_{i+\frac{1}{2},k} - \phi^n_{i-\frac{1}{2},k}\right](\epsilon,\mathrm{i}) + \frac{\Delta t}{2}\left[\left(\frac{G-F}{\tau}\right)^n_{i,k} + \left(\frac{G-F}{\tau}\right)^{n+1}_{i,k}\right](\epsilon,\mathrm{i}). \tag{62}$$

In the above microscopic scheme (62), the pseudo-equilibrium state G at time t^{n+1} is required. To overcome this issue, a finite volume scheme on macroscopic quantities $\boldsymbol{W}$ is first used to predict them, which then serve to compute G at time t^{n+1} [43]. This is obtained by applying the quadrature rule $\langle\cdot\rangle_{\mathcal{V},\epsilon,\mathrm{i}}$ on the microscopic scheme (62). This approach leads to the following macroscopic scheme:

$$\boldsymbol{W}^{n+1}_i = \boldsymbol{W}^n_i - \frac{\Delta t}{\Delta x}\left[\boldsymbol{\Phi}^n_{i+\frac{1}{2}} - \boldsymbol{\Phi}^n_{i-\frac{1}{2}}\right] + \frac{\Delta t}{2}\left[\left(\frac{\boldsymbol{V}-\boldsymbol{W}}{\tau}\right)^n_i + \left(\frac{\boldsymbol{V}-\boldsymbol{W}}{\tau}\right)^{n+1}_i\right], \tag{63}$$

where $\boldsymbol{\Phi}$ is the numerical flux of macroscopic quantities $\boldsymbol{W}$ and $\boldsymbol{V}$ is the vector of moments of G with respect to the microscopic vector of moments $\boldsymbol{m}$ and the quadrature rule $\langle\cdot\rangle_{\mathcal{V},\epsilon,\mathrm{i}}$ (38). Formally, these quantities are both defined as:

$$\boldsymbol{\Phi} = \langle \boldsymbol{m}\phi\rangle_{\mathcal{V},\epsilon,\mathrm{i}}, \quad \boldsymbol{V} = \langle \boldsymbol{m}G\rangle_{\mathcal{V},\epsilon,\mathrm{i}}. \tag{64}$$

From the macroscopic scheme (63), the mass density, momentum, and total energy at time t^{n+1} can be easily determined. Indeed, since these quantities are collision invariants, the relaxation term in (63) vanishes for them, rendering their computation fully explicit. This simplification does not apply to the calculation of Θ and the modal energies at time t^{n+1}, making their resolution implicit. To solve these, we use the relationship between $\boldsymbol{V}$ and $\boldsymbol{W}$ induced by the definition of $\mathcal{T}$ in (12) and T_{tr}^{rel} in (14, 16, 19) or (21). This resolution is locally iterative since the relaxation time τ is a non-linear function of T_{tr}, as can also be Z_{rot} [14, 25, 32], Z_{vib} [31, 32], or $\mathcal{T}$ in the Pfeiffer model [26, 33]. At this point, all necessary terms for determining the microscopic distribution F at time t^{n+1} are known and the microscopic scheme (62) can be used to compute it. Ultimately, both microscopic and macroscopic unknowns, F and $\boldsymbol{W}$, at time t^{n+1} are fully determined with second-order accuracy using this natural technique.

3.3.2 The deficiency of the trapezoidal formula

In a strong non-equilibrium situation, the methodology presented above could lead to a non-positive tensor Θ and could alter the calculation of a realistic equilibrium G^{n+1}. To illustrate this behavior, let us consider the energy relaxation processes of a monoatomic gas described by the Holway model (12–14). In an adiabatic bath, the macroscopic evolution of $\boldsymbol{W}$ given by (63) reduces to:

$$\boldsymbol{W}_i^{n+1} = \boldsymbol{W}_i^n + \frac{\Delta t}{2}\left[\left(\frac{\boldsymbol{V}-\boldsymbol{W}}{\tau}\right)_i^n + \left(\frac{\boldsymbol{V}-\boldsymbol{W}}{\tau}\right)_i^{n+1}\right]. \tag{65}$$

This ensures the conservation of mass density, momentum and total energy for the same reason as previously. Since the evolution of macroscopic and microscopic quantities are spatially located, the index i is omitted in the following. Regarding the anisotropic tensor of temperature, the induced scheme is:

$$\Theta^{n+1} = \Theta^n + \frac{\Delta t}{2}\left[\left(\frac{\mathcal{T}-\Theta}{\tau}\right)^n + \left(\frac{\mathcal{T}-\Theta}{\tau}\right)^{n+1}\right]. \tag{66}$$

Since we are focusing on the Holway model (12–14), the translational temperature is equal to the equilibrium temperature and is conserved during the relaxation process. Thus, the relaxation time τ is constant in time and its index n is omitted. Moreover, using the expression of $\mathcal{T}$ leads to the following scheme:

$$\mathcal{T}^{n+1} = R_s T_{tr} I + \left(1 - \frac{1}{\mathrm{Pr}}\right)\frac{1-\zeta}{1+\zeta}[\Theta^n - R_s T_{tr} I]. \tag{67}$$

where ζ is a time ratio parameter defined as:

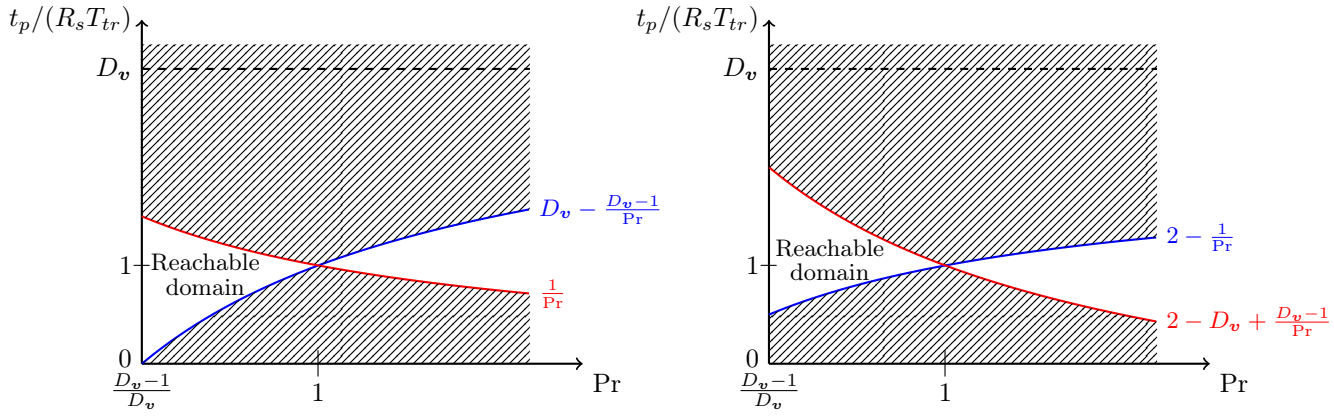
$$\zeta = \frac{\Delta t}{2\tau \mathrm{Pr}} > 0. \tag{68}$$

Similar manipulations to those in [2] carried out on (67), yields that $\mathcal{T}^{n+1}$ is unconditionally positive-definite for realistic Prandtl number ($\mathrm{Pr} \leqslant 1$), provided that Θ^n is as well. In details, if $(t_p)_{1\leqslant p\leqslant D_{\boldsymbol{v}}}$ denote the eigenvalues of $\mathcal{T}^{n+1}$, realizability of Θ^n ensures that they satisfy the following constraints:

- For $\zeta \leqslant 1$, $\left(t_p/(R_s T_{tr})\right)_p$ are bounded between $D_{\boldsymbol{v}} - \frac{D_{\boldsymbol{v}}-1}{\mathrm{Pr}}$ and $1/\mathrm{Pr}$;
- For $\zeta > 1$, $\left(t_p/(R_s T_{tr})\right)_p$ are bounded between $2 - 1/\mathrm{Pr}$ and $2 - D_{\boldsymbol{v}} + \frac{D_{\boldsymbol{v}}-1}{\mathrm{Pr}}$.

These bounds are represented in Figure 3 for both numerical resolutions $\zeta \leqslant 1$ and $\zeta > 1$. In both cases, and for a realistic Prandtl number ($\mathrm{Pr} \leqslant 1$), Figure 3 illustrates that the eigenvalues of $\mathcal{T}^{n+1}$ lie between 0 and $D_{\boldsymbol{v}}$, provided that the eigenvalues of Θ^n do as well. This ensures that $\mathcal{T}^{n+1}$ is positive-definite which is essential for ensuring the computation of a consistent equilibrium state G^{n+1} even for large numerical steps Δt.

However, in contrast to $\mathcal{T}^{n+1}$, Θ^{n+1} is not guaranteed to be positive-definite for arbitrary Δt, even if Θ^n is. This behavior is a consequence of the fact that the trapezoidal formula is not L-stable. The issue becomes particularly



(a) $\zeta \leqslant 1$, (b) $\zeta > 1$,

Figure 3: Reachable values of the eigenvalues of $\mathcal{T}^{n+1}$ computed using (67) and a realizable Θ^n. *The monoatomic ES-BGK model is well-defined only for* $\mathrm{Pr} \geqslant (D_{\boldsymbol{v}} - 1)/D_{\boldsymbol{v}}$ [2].

pronounced under strong anisotropic conditions, or whenever a specific direction is strongly favored in Θ, with its corresponding eigenvalue exceeding $2R_sT_{tr}$. For example, let us assume Θ^n to be the following anisotropic tensor:

$$\Theta^n = \begin{pmatrix} 5/2 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} R_sT_{tr}. \tag{69}$$

In this particular case, Θ^{n+1} would be:

$$\Theta^{n+1} = \frac{1}{4(1+\zeta)} \begin{pmatrix} 10 - 2\zeta & 0 & 0 \\ 0 & 1 + 7\zeta & 0 \\ 0 & 0 & 1 + 7\zeta \end{pmatrix} R_sT_{tr}. \tag{70}$$

For this case, it clearly appears that Θ^{n+1} is symmetric positive-definite if and only if $\zeta < 5$. That is to say Δt is less than $10\tau\mathrm{Pr}$, which is a too restrictive numerical constraint for the UGKS which aims to use a numerical time step $\Delta t \gg \tau$, especially in the continuum regime.

In conclusion, even if Θ is positive-definite at time t^n, the tensor Θ^{n+1} may become non-positive. This indicates that the mass distribution F^{n+1} is not guaranteed to remain positive everywhere, and that $\mathcal{T}^{n+2}$ may also lose positivity, either of which may lead to a failure of the simulation.

3.3.3 The explicit predictive-corrective approach

In [11], the relaxation part of the ES-BGK model is treated differently. As in [24, 43], the relaxation term in the microscopic scheme follows a trapezoidal formulation:

$$F_{i,k}^{n+1}(\epsilon, \mathrm{i}) = F_{i,k}^{n}(\epsilon, \mathrm{i}) - \frac{\Delta t}{\Delta x}\left[\phi^n_{i+\frac{1}{2},k} - \phi^n_{i-\frac{1}{2},k}\right](\epsilon, \mathrm{i}) + \frac{\Delta t}{2}\left[\left(\frac{G-F}{\tau}\right)^n_{i,k} + \frac{G^\star_{i,k} - F^{n+1}_{i,k}}{\tau^\star_{i,k}}\right](\epsilon, \mathrm{i}). \tag{71}$$

However, in contrast to (62), the equilibrium state G at time t^{n+1} is replaced in (71) by an alternative equilibrium state, denoted by $G^\star$. To ensure consistency with the ES-BGK model, $G^\star$ has to approximate G^{n+1}. Thus, a prediction step is realized to determine $G^\star$ via the estimation of macroscopic quantities $\boldsymbol{W}^\star$. In [11], this prediction is performed by replacing the trapezoidal term in the macroscopic scheme (63) with a forward Euler relaxation term, as expressed in the following formulation:

$$\boldsymbol{W}_i^\star = \boldsymbol{W}_i^n - \frac{\Delta t}{\Delta x}\left[\boldsymbol{\Phi}^n_{i+\frac{1}{2}} - \boldsymbol{\Phi}^n_{i-\frac{1}{2}}\right] + \Delta t\left[\frac{\boldsymbol{V} - \boldsymbol{W}}{\tau}\right]^n_i. \tag{72}$$

Once $G^\star$ is determined, equation (71) is used to compute the microscopic distribution function F at time t^{n+1}. Finally, the macroscopic variables $\boldsymbol{W}^{n+1}$ are obtained from F^{n+1} using (38) to ensure consistency between the microscopic and macroscopic levels of description. Since ρ, $\rho\boldsymbol{u}$, and E are collision invariants, they remain unaffected by the numerical relaxation term. Thus, they are already determined and equal to their predicted values $\rho^\star$, $(\rho\boldsymbol{u})^\star$, and $E^\star$; still, corrections are required for Θ and the modal energies.

Unfortunately, this scheme has major deficiencies, as it is neither A-stable nor L-stable. Indeed, the necessary condition to compute the equilibrium state requires $\mathcal{T}^\star$ to be positive-definite. However, by following the same reasoning as before, one can see that ζ must be less than $1/(2-2\mathrm{Pr})$ for realistic Prandtl numbers ($\mathrm{Pr} \leqslant 1$). This is once again too restrictive a condition for the UGKS.

3.3.4 The implicit predictive-corrective approach

We propose to modify the previous predictive-corrective approach by using a backward Euler scheme for the macroscopic relaxation term. The corresponding formulation reads:

$$\boldsymbol{W}_i^\star = \boldsymbol{W}_i^n - \frac{\Delta t}{\Delta x}\left[\boldsymbol{\Phi}^n_{i+\frac{1}{2}} - \boldsymbol{\Phi}^n_{i-\frac{1}{2}}\right] + \Delta t\left[\frac{\boldsymbol{V}-\boldsymbol{W}}{\tau}\right]_i^\star. \tag{73}$$

As before, the relaxation term affects only the evolution of non-conservative quantities. Compared to the previous approach, the predicted starred states $\Theta^\star$ and $\mathcal{T}^\star$, as computed from (73), exhibit favorable properties. In the case of an adiabatic bath, they remain positive-definite and less anisotropic than Θ^n and $\mathcal{T}^n$, respectively, for any $\zeta > 0$. This behavior is consistent with the relaxation process described by ES-BGK models. Moreover, this enable the evaluation of a realizable equilibrium state $G^\star$, which is required in (71) for the computation of F^{n+1}.

Once again, consistency between the microscopic and macroscopic levels requires correcting the evolution of the anisotropic tensor of temperature using the updated distribution F^{n+1}. In the case of an adiabatic bath, this is equivalent to:

$$\Theta^{n+1} = \Theta^n + \frac{\Delta t}{2}\left(\frac{\mathcal{T}^n - \Theta^n}{\tau^n} + \frac{\mathcal{T}^\star - \Theta^{n+1}}{\tau^\star}\right). \tag{74}$$

However, this correction does not ensure the positivity of Θ^{n+1}, even if Θ^n is positive-definite. Indeed, substituting the relation between $\mathcal{T}^\star$ and $\Theta^\star$, the prediction of $\Theta^\star$ given by (73), and the definition of ζ in (68), we obtain:

$$\left[1+\mathrm{Pr}\zeta\right]\Xi^{n+1} = \frac{1}{1+2\zeta}\left[1+\mathrm{Pr}\zeta-2\zeta^2\right]\Xi^n. \tag{75}$$

where Ξ denotes the deviator tensor of Θ, that is:

$$\Xi = \Theta - R_s T_{tr} I. \tag{76}$$

Relation (75) clearly shows that Ξ converges monotonically to zero for small values of ζ. Indeed, whenever:

$$\zeta < \frac{1+\sqrt{1+8/\mathrm{Pr}^2}}{4}\mathrm{Pr}, \tag{77}$$

the scaling factor of Ξ^n in (75) remains strictly positive, and in this case, relation (75) yields:

$$|\Xi^{n+1}| < \frac{1}{(1+\mathrm{Pr}\zeta)(1+2\zeta)}\left[1+\mathrm{Pr}\zeta\right]|\Xi^n| = \frac{1}{1+2\zeta}|\Xi^n|. \tag{78}$$

Thus, under condition (77), Θ^{n+1} is a relaxation of Θ^n toward $R_s T_{tr}$, and therefore preserves symmetric positive definiteness whenever Θ^n is as well, which enables (but not guaranties) the realizability of F^{n+1}. However, the monotone decay of Ξ is not reproduced for larger value of ζ, since an asymptotic analysis of (75) shows that:

$$\Xi^{n+1} \underset{\zeta\to+\infty}{\sim} -\frac{1}{\mathrm{Pr}}\Xi^n. \tag{79}$$

Consequently, because the Prandtl number is typically set to 2/3 or 0.7, this approach may induce unbounded oscillations of Θ for sufficiently large values of ζ, which may in turn produce a non positive prediction of Θ and thereby causing the simulation to fail.

In conclusion, this scheme cannot be used in its current form within the UGKS framework, whose key feature is the use of a time step $\Delta t \gg \tau$ in the continuum regime.

3.3.5 A more robust method

As mentioned before, an implicit treatment is mandatory to handle the stiffness of the relaxation term, particularly in the continuum regime. This involves computing an approximation of G^{n+1}, which requires a predicted tensor $\mathcal{T}^\star \approx \mathcal{T}^{n+1}$ to be positive-definite. The trapezoidal formula cannot guarantee this condition for the entire simulation, as Θ^{n+1} is not ensured to be positive even if Θ^n is. The implicit predictive-corrective approach has the advantage of producing predicted approximations $\Theta^\star$ and $\mathcal{T}^\star$ that are positive-definite whenever Θ^n is, even with large Δt. However, ensuring consistency between the microscopic and macroscopic levels necessitates correcting the prediction $\Theta^\star$ to obtain Θ^{n+1}, which is not guaranteed to remain positive definite.

Thus, we adopt an implicit predictive **non-corrective** approach, using (71) and (73), while enforcing $\Theta^{n+1} = \Theta^\star$ at the macroscopic level. In the adiabatic bath case at least, this ensures that Θ and $\mathcal{T}$ remain positive definite for any positive-definite initial tensor Θ^0. In general cases, this approach remains consistent with the ES-BGK model, has only a minor impact on the evolution of conservative quantities, and does not alter the results in steady simulations. Although this is achieved at the cost of a macro-micro inconsistency in the non-conservative quantities, in practice only small discrepancies are observed between the imposed Θ^{n+1} and the one obtained as the moment of F^{n+1}.

Because of the complex structure of the numerical fluxes, establishing a rigorous theoretical proof of the robustness and stability of this approach is challenging. In practice we observe improved robustness in very strongly anisotropic flows, such as those occurring in unsteady unresolved strong shocks, compared with the original trapezoidal formula. However, for common aerodynamic flows, adequate results can still be obtained with the trapezoidal formula.

3.4 Summary of the UGKS for ES-BGK models

The UGKS for ES-BGK models can be summarized through the following relations. First, as described in Section 3.2, the microscopic numerical fluxes ϕ are defined through a sophisticated formulation involving the distributions F, G and M:

$$\phi^n_{i+\frac{1}{2},k} = v_{kx}\left[q_1 G + q_2 v_{kx}\delta_x M + q_3 \delta_t M + q_4 F + q_5 v_{kx}\delta_x F\right]^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}). \tag{80}$$

The required discrete terms are constructed according to relations (42–44) and (55–60), and the coefficients $(q_j)_{1\leqslant j\leqslant 5}$ are defined in Section 3.2.6. Second, as detailed in Section 3.3, the macroscopic quantities $\boldsymbol{W}$ are advanced from time t^n to t^{n+1} using:

$$\boldsymbol{W}^{n+1}_i = \boldsymbol{W}^n_i - \frac{\Delta t}{\Delta x}\left[\boldsymbol{\Phi}^n_{i+\frac{1}{2}} - \boldsymbol{\Phi}^n_{i-\frac{1}{2}}\right] + \Delta t\left[\frac{\boldsymbol{V}-\boldsymbol{W}}{\tau}\right]^{n+1}_i, \tag{81}$$

where $\boldsymbol{\Phi}$ denotes the macroscopic fluxes obtained as the moments the microscopic fluxes ϕ. Finally, the discrete microscopic distribution F is updated according to the following numerical scheme:

$$F^{n+1}_{i,k}(\epsilon, \mathrm{i}) = F^n_{i,k}(\epsilon, \mathrm{i}) - \frac{\Delta t}{\Delta x}\left[\phi^n_{i+\frac{1}{2},k} - \phi^n_{i-\frac{1}{2},k}\right](\epsilon, \mathrm{i}) + \frac{\Delta t}{2}\left[\left(\frac{G-F}{\tau}\right)^n_{i,k} + \left(\frac{G-F}{\tau}\right)^{n+1}_{i,k}\right](\epsilon, \mathrm{i}). \tag{82}$$

3.5 Discrete treatments of boundary conditions

Boundary conditions are fundamental components in numerical simulations. First, they need to accurately represent the flow behavior at the boundaries of the simulation domain (inflow, outflow, body interactions, etc.). Second, the implementation of these conditions is often challenging, sensitive, and requires careful attention, as it significantly impacts the reliability and precision of the simulation results.

From a practical point of view, ghost cells are commonly employed in finite volume methods to implement boundary conditions. In this framework, inflow, outflow, and symmetric flow conditions are relatively easy to simulate. For inflow conditions, microscopic distributions and macroscopic quantities in ghost cells are set to inflow equilibrium conditions. For outflow conditions, the states in the ghost cells are directly related to those of the cells in the last real layers. Finally, symmetric conditions are reproduced by duplicating the real state through a symmetric transformation in the corresponding ghost cell.

Undoubtedly, the most challenging boundary condition to implement is the gas-surface interaction. The first reason is the complexity of modeling the gas-body interaction, which depends on factors such as the structure of the solid surface layer, its roughness, the interaction between impacting gas molecules and solid molecules [9], and physical processes we aim to model. The second reason concerns the numerical scheme behavior that must be

preserved. Specific reconstruction techniques are required at the boundaries to maintain the accuracy of the scheme [4] or to ensure its correct asymptotic behavior in the continuum regime [10].

Here, we focus solely on the diffuse reflection model. In this model, a gas particle colliding with a wall is reflected according to the Maxwellian distribution (25) associated with a unitary mass density, and the wall temperature and velocity. Formally, let T_w and $\boldsymbol{u}_w$ represent the wall temperature and velocity, respectively, and let $v_n = (\boldsymbol{v}-\boldsymbol{u}_w)\cdot\boldsymbol{n}$ be the inner product of the relative particle velocity to the wall velocity and the normal direction of the wall directed into the gas. Moreover, the wall is assumed to do not move in the normal direction $\boldsymbol{n}$, while it may still shift in tangential direction ($\boldsymbol{n}\cdot\boldsymbol{u}_w = 0$). Then, at the interface $\boldsymbol{x}_w$, we have for the microscopic distribution representing re-emitting particles:

$$F(t,\boldsymbol{x}_w,\boldsymbol{v}_{|v_n>0},\epsilon,\mathrm{i}) = \rho_w(t,\boldsymbol{x})M(\boldsymbol{v},\epsilon,\mathrm{i}), \tag{83}$$

where the quantity ρ_w ensure the vanishing of the mass flux across the wall boundary:

$$\rho_w(t,\boldsymbol{x}_w) = -\left(\sum_{i=0}^{+\infty}\int_{\mathbb{R}^+}\int_{v_n<0} v_n F(t,\boldsymbol{x}_w,\boldsymbol{v},\epsilon,\mathrm{i})d\boldsymbol{v}d\epsilon\right)\left(\sum_{i=0}^{+\infty}\int_{\mathbb{R}^+}\int_{v_n>0} v_n M(\boldsymbol{v},\epsilon,\mathrm{i})d\boldsymbol{v}d\epsilon\right)^{-1}. \tag{84}$$

This model assumes that the re-emitted gas is in perfect equilibrium, with temperature T_w and velocity $\boldsymbol{u}_w$. In this framework, all modal temperatures of the re-emitted gas are equal to T_w. In [40], only the translational temperature is modeled to thermalized to T_w due to the assumed short interaction time between the gas and the wall. This consideration is interesting but will not be the main focus here, as it concerns the fine modeling of the interaction between the gas and the surface. The following treatment can, however, certainly be adapted to account for this assumption.

At the discrete level, the difficulty lies in defining the numerical flux at the wall interface in such a way that it corresponds both to the desired numerical method and modeling of the boundary. Commonly, the numerical flux at the wall interface is defined in two parts, each corresponding either to the boundary modeling or the numerical scheme, depending on the considered particle velocity:

$$\phi^n_{w,k}(\epsilon,\mathrm{i}) = \begin{cases} \phi^{n,out}_{w,k}(\epsilon,\mathrm{i}) & \text{if } v_{kn} < 0, \\ v_{kn}\rho^n_w M_k(\epsilon,\mathrm{i}) & \text{if } v_{kn} \geqslant 0. \end{cases} \tag{85}$$

In the previous expression, the subscript w denotes the wall location, v_{kn} is defined as the inner product $(\boldsymbol{v}_k-\boldsymbol{u}_w)\cdot\boldsymbol{n}$, and ρ^n_w represents to the average mass re-emitted by the wall over $[t^n, t^{n+1}]$. This quantity is defined so as to enforce a vanishing net mass flux across the wall during the time step:

$$\rho^n_w = -\left\langle \phi^{n,out}_{w,k}(\epsilon,\mathrm{i})\right\rangle_{\mathcal{V}_{w^+},\epsilon,\mathrm{i}}\left[\langle v_{kx}M_k(\epsilon,\mathrm{i})\rangle_{\mathcal{V}_{w^-},\epsilon,\mathrm{i}}\right]^{-1}. \tag{86}$$

In the above expression, $\langle\cdot\rangle_{\mathcal{V}_{w^-},\epsilon,\mathrm{i}}$ and $\langle\cdot\rangle_{\mathcal{V}_{w^+},\epsilon,\mathrm{i}}$ refer to the continuous integration over the phase ϵ and i, and the velocity quadrature rule integration respectively on negative and positive relative velocity $v_{kx} - u_{xw} = v_{kx}$ set. With (85–86), the discrete treatment of the boundary condition is reduced to specifying the outgoing flux $\phi^{n,\mathrm{out}}_{w,k}$.

In order to both conserve the second-order accuracy of UGKS and its asymptotic behavior in the continuum regime, the idea is to build $\phi^{n,out}_{w,k}$ of (85) in the same way as in (80). For the sake of simplicity, let us suppose the wall at left of an interface located at $x_w = x_{1/2}$. Furthermore, although the mesh is assumed to be uniform here, this boundary condition can be easily adapted to non-uniform meshes. [4]. Figure 4 provides with a visual summary of our methodology explained here after. Since $\phi^{n,out}_{1/2,k}$ would be used only with outgoing velocity, the only microscopic terms required are $F^n_{1,k}$ and $\delta_x F^n_{1/2,k}$. To define $\delta_x F^n_{1/2,k}$ as a limited slope (42) while maintaining second-order accuracy at the boundary, we define $F^n_{0,k}$ for outgoing velocities such that the forward and backward slope of F in cell 1 are equal [4]:

$$\forall k \text{ s.t. } v_{kx} < 0, \qquad \frac{F^n_{1,k}(\epsilon,\mathrm{i}) - F^n_{0,k}(\epsilon,\mathrm{i})}{\Delta x} = \frac{F^n_{2,k}(\epsilon,\mathrm{i}) - F^n_{1,k}(\epsilon,\mathrm{i})}{\Delta x}. \tag{87}$$

To define the equilibrium state at the interface, we first need its corresponding moments. In [10], the interface velocity $\boldsymbol{u}^n_{1/2}$ and temperature $T^n_{1/2}$ are set to the wall values $\boldsymbol{u}_w$ and T_w, and the pressure $p^n_{1/2}$ at the interface is assumed equal to the pressure in the first adjacent cell. Surely, this choice neglects the temperature jump and velocity slip that exist even in the near-continuum regime. However, it only impacts the equilibrium terms of the flux, and non-equilibrium effects are still represented and taken into account through microscopic terms. In [10],

satisfying results from rarefied to continuum flows are obtained with this choice. Our methodology mainly differs in constructing the equilibrium state at this interface. Here, a temperature jump and a slip velocity are considered by constructing moments on the interface based on outgoing and reflected-incoming microscopic distributions at time t^n and located exactly at the wall interface. Firstly, the outgoing microscopic distribution (i.e. with $v_{kx} <0$ here) is constructed by an extrapolation based on the two first layers of real cells:

$$\forall k \text{ s.t. } v_{kx} < 0, \qquad F^n_{1/2,k}(\epsilon, \mathrm{i}) = F^n_{1,k} - \frac{\Delta x}{2} \frac{F^n_{2,k}(\epsilon, \mathrm{i}) - F^n_{1,k}(\epsilon, \mathrm{i})}{\Delta x}. \tag{88}$$

Then the related incoming distribution (i.e. with $v_{kx} > 0$ here) is deduced following a diffuse reflection:

$$\forall k \text{ s.t. } v_{kx} > 0, \qquad F^n_{1/2,k}(\epsilon, \mathrm{i}) = \rho^\star_{1/2} M_k(\epsilon, \mathrm{i}), \tag{89}$$

where $\rho^\star_{1/2}$ is defined so that the mass flux at time t^n across the wall interface vanishes:

$$\rho^\star_{1/2} = -\left\langle v_{kx} F^n_{1/2,k}(\epsilon, \mathrm{i}) \right\rangle_{\mathcal{V}_{w^+},\epsilon,\mathrm{i}} \left[\langle v_{kx} M_k(\epsilon, \mathrm{i}) \rangle_{\mathcal{V}_{w^-},\epsilon,\mathrm{i}} \right]^{-1}. \tag{90}$$

Thus, the moments on interface are defined as usual using (38) and one can define the equilibrium state at the wall interface.

Finally, since the macroscopic derivative term $\delta_x^- M$ is not used for outgoing velocities, the remaining terms to be constructed are $\delta_x^+ M$ and $\delta_t M$. The first term is constructed as usual and the second is built using equations (58–60) by assuming $\delta_x F^n_{0,k} = \delta_x F^n_{1,k}$ for positive velocities, which does not compromise the second order of the scheme.

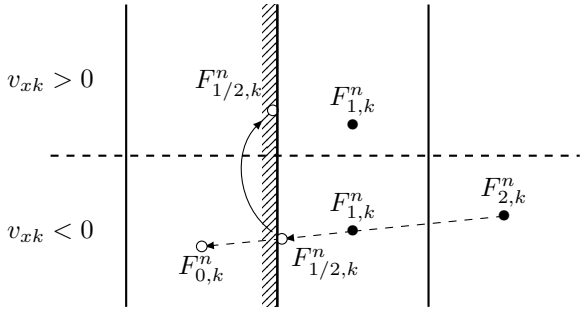


Figure 4: Definition of ghost cells quantities for the UGKS. Figure adapted from [4].

As a result, this boundary treatment is second-order accurate for any Knudsen number. It prevents a decrease in order at boundary [4] and its multi-scale approach is a key ingredient in ensuring the correct recovery of viscous boundary layer effects in the Navier-Stokes asymptotic limit, as illustrated in [10].

3.6 The UGKS on the reduced model

In practice, the ϵ and i energy phases are ignored in the discretization by using reduced distribution techniques as presented previously in Section 2.3. The same reducing operations can, in principle, be applied to the numerical scheme (62) for F with any rotational and vibrational energy quadrature error since these energy phases were not discretized. However, because of the nonlinearity of the TVD limiter, the slopes $\delta_x F$ cannot be rigorously reduced, unlike all other terms of the flux. Nevertheless, in practice, this term is replaced by the limited slope of reduced distributions. Indeed, in regions where the limiter does not really act, this reduction is rigorously valid. In other regions, the primary purpose of this limiter is to suppress spurious oscillations before ensuring second-order accuracy. The same mechanism is then reproduced with the reduced distributions f, g, and h which legitimizes the following reduced flux expressions that are used in practice:

$$\begin{aligned}
\mathcal{F}^n_{i+\frac{1}{2},k} &= v_{kx} \left[q_1 G_{tr} + q_2 v_{kx} \delta_x M_{tr} + q_3 \delta_t M_{tr} + q_4 f + q_5 v_{kx} \delta_x f \right]^n_{i+\frac{1}{2},k}, \\
\mathcal{G}^n_{i+\frac{1}{2},k} &= v_{kx} \left[q_1 [e^{rel}_{rot} G_{tr}] + q_2 v_{kx} \delta_x [e_{rot}(T_{eq}) M_{tr}] + q_3 \delta_t [e_{rot}(T_{eq}) M_{tr}] + q_4 g + q_5 v_{kx} \delta_x g \right]^n_{i+\frac{1}{2},k}, \\
\mathcal{H}^n_{i+\frac{1}{2},k} &= v_{kx} \left[q_1 [e^{rel}_{vib} G_{tr}] + q_2 v_{kx} \delta_x [e_{vib}(T_{eq}) M_{tr}] + q_3 \delta_t [e_{vib}(T_{eq}) M_{tr}] + q_4 h + q_5 v_{kx} \delta_x h \right]^n_{i+\frac{1}{2},k}.
\end{aligned} \tag{91}$$

As presented, the construction of the reduced flux terms can be achieved by calculating analytically the internal-energy integral of the terms of the original UGKS fluxes in (62). However, directly constructing the new terms is both simpler and equally accurate. This is why the theoretical UGKS framework is generally applied directly to reduced models rather than to complete distributions like $F(t, \boldsymbol{x}, \boldsymbol{v}, \epsilon, \mathrm{i})$. Indeed, the discrete microscopic terms $f, g, h, \delta_x f, \delta_x g, \delta_x h$ are constructed in the same manner as in (41) and (43). For the macroscopic part, the pseudo-equilibrium G and Maxwellian distributions associated with internal energy are not necessary. Consequently, the complete pseudo-equilibrium G and Maxwellian terms in the original numerical flux are replaced by the product of their translational counterparts with internal energies. The expression for the Maxwellian to consider is now $M_{tr} = \exp(\boldsymbol{\eta}_{tr} \cdot \boldsymbol{\alpha}_{tr})$, where $\boldsymbol{\eta}_{tr} = (1, \boldsymbol{v}, \frac{1}{2}|\boldsymbol{v}|^2)^\top$. In addition, the discrete spatial derivative terms $\delta_x M_{tr}$, $\delta_x[e_{rot}(T_{eq})]$, and $\delta_x[e_{vib}(T_{eq})]$, as well as their temporal counterparts, are constructed analogously to $\delta_x M$ and $\delta_t M$ in Section 3.2.5, based on discrete derivatives of the conservative moments ρ, $\rho\boldsymbol{u}$, and E.

4 Numerical results

For efficient numerical simulations, the UGKS is implemented for the reduced model described in Section 3.6. This approach dispenses with the discretization of the rotational and vibrational energy phases, ϵ and i, thereby significantly reducing both the memory footprint and the computational cost while preserving accuracy. For one and two-dimensional flows, the orthogonal velocity components are also omitted from the discretization through the introduction of an additional reduced distribution function. This distribution accounts for the translational energy associated with motion orthogonal to the computational domain. The resulting formulation is highly efficient, requiring discretization only of the (t, x, v_x) phases for 1D flows and (t, x, y, v_x, v_y) phases for 2D flows.

To ensure accurate simulations, the velocity set $\mathcal{V}$ has to be:

- large enough to encompass most of the reduced distribution functions in every spatial cell;
- fine enough to provide an accurate discretization of the reduced distribution functions in every spatial cell.

The UGKS can be implemented using locally adapted velocity grids, which improve the computational efficiency [3]. However, for simplicity, the results presented here are obtained using a Cartesian velocity grid defined as $\mathcal{V} = [v_{x,\min}, v_{x,\max}]$ for one-dimensional flows, and $\mathcal{V} = [v_{x,\min}, v_{x,\max}] \times [v_{y,\min}, v_{y,\max}]$ for two-dimensional flows. The velocity grid is constructed with a uniform spacing Δv_x or Δv_y in each direction, and the overall velocity domain $\mathcal{V}$ is chosen so as to satisfy, for every cell center $\boldsymbol{x_c}$:

- $v_{\alpha,min} \leqslant u_\alpha(\boldsymbol{x_c}) - 4\sqrt{R_s T(\boldsymbol{x_c})}$ and $u_\alpha(\boldsymbol{x_c}) + 4\sqrt{R_s T(\boldsymbol{x_c})} \leqslant v_{\alpha,max}$, for $\alpha = x, y$;
- $\Delta v_\alpha \leqslant \sqrt{R_s T(\boldsymbol{x_c})}$, for $\alpha = x, y$.

For each test case, the bounds of $\mathcal{V}$ and the number of velocity points in each direction are specified.

The time step Δt is constrained by a Courant-Friedrichs-Lewy (CFL) condition that depends on both the spatial discretization and the discrete velocity set $\mathcal{V}$. For a two-dimensional Cartesian spatial grid, this condition reads:

$$\Delta t \max_{\boldsymbol{v} \in \mathcal{V}} \left(\frac{|v_x|}{\Delta x} + \frac{|v_y|}{\Delta y} \right) < 1, \tag{92}$$

and can be generalized to multi-dimensional curvilinear meshes. Equation (92) highlights that Δt is not constrained by the flow rarefaction, which is a necessary requirement for the scheme to be asymptotic preserving (AP). In the following cases, the left-hand side of (92) is limited to less than a number named CFL, except for the first test case where another condition on Δt is applied.

For unsteady flows, the simulation is carried out up to a prescribed final time, specified for each case. For steady flows, the computation is advanced until convergence is reached. Convergence is evaluated using the L_2 norm for the reduced mass distribution f. Specifically, convergence is considered achieved when the L_2 norm of the difference between two successive iterations, f^{n+1} and f^n ,relative to that between f^1 and f^0, is less than 10^{-5}.

The complexity of the numerical cases presented here increases progressively. In each case, the results are compared with either analytical solutions, simulation results, or experimental data. First, the relaxation rate of energy modes is examined using a bath relaxation approach. Subsequently, the 1D Couette flow is simulated to assess the scheme ability to accurately recover the correct Prandtl number in a viscous flow. Following this, non-equilibrium effects are investigated through the stationary shock flow problem. The classic Sod shock tube problem

is then performed to attest both the robustness of handling different wave structures and the accurate simulation of unsteady non-equilibrium flows. Afterward, a two-dimensional supersonic flow simulation is compared with experimental data to evaluate the performance of the scheme in more realistic and complex scenarios. Finally, a hypersonic flow around an infinite cylinder is simulated with UGKS to compare our approach with another simulation code that resolves the ES-BGK model of Pfeiffer.

Finally, the van Leer limiter is employed in the simulation, and the viscosity μ is computed as: $\mu = \mu_{ref}(T_{tr}/T_{ref})^{\omega}$, with μ_{ref}, T_{ref} and ω given for each test case.

4.1 Homogeneous relaxation of energies

In this first case, we focus on energy transfer phenomena. The gas is supposed to be spatially homogeneous, with no interacting bounds, and initially away from equilibrium. The reduced ES-BGK model (31–33) is therefore simplified as:

$$\begin{cases} \partial_t f = \frac{1}{\tau}(G_{tr} - f), \\ \partial_t g = \frac{1}{\tau}(e_{rot}^{rel} G_{tr} - g), \\ \partial_t h = \frac{1}{\tau}(e_{vib}^{rel} G_{tr} - h). \end{cases} \tag{93}$$

Since density, momentum, and total energy are three collision invariants, they are conserved during the relaxation process toward equilibrium. Moreover, related macroscopic variables, such as flow velocity and equilibrium temperature, are also conserved. The only effects observed are the homogenization of the translation, rotation and vibration energy modes with characteristic times depending on the specific ES-BGK model considered: Andriès [2], Dauvois [13] or Pfeiffer [26] as presented in Section 2.2. In any case, these energy exchanges are described by simple relaxation differential equations, which can be analytically solved only for constant times τ and τ_C, vibrational degree of freedom δ', collision numbers Z_{rot} and Z_{vib}.

In case of non vibrating gas, using the ES-BGK model of Andriès et al. [2], we get:

$$\begin{aligned} \Theta(t) &= R_s T_{tr}(t) I + (\Theta^0 - R_s T_{tr}^0 I) e^{-t/(\mathrm{Pr}\tau)}, \\ T_{tr}(t) &= T_{eq} + (T_{tr}^0 - T_{eq}) e^{-t/(Z_{rot}\tau)}, \\ T_{rot}(t) &= T_{eq} + (T_{rot}^0 - T_{eq}) e^{-t/(Z_{rot}\tau)}. \end{aligned} \tag{94}$$

In case of vibrating gas, the translational, rotational and vibrational temperatures undergo two relaxation processes with two characteristics times λ_- and λ_+ such as:

$$\begin{pmatrix} T_{tr} \\ T_{rot} \\ T_{vib} \end{pmatrix}(t) = \begin{pmatrix} T_{eq} \\ T_{eq} \\ T_{eq} \end{pmatrix} + \begin{pmatrix} C_{t1} \\ C_{r1} \\ C_{v1} \end{pmatrix} e^{-t/\lambda_-} + \begin{pmatrix} C_{t2} \\ C_{r2} \\ C_{v2} \end{pmatrix} e^{-t/\lambda_+}, \tag{95}$$

where constants $C_{r1}, ...C_{v2}$ depend on the initial energy state. In the particular case of $D_{\boldsymbol{v}} = 3$, $\delta = \delta' = 2$, and when the ES-BGK model of Pfeiffer et al. [26] is used, the characteristic times λ_- and λ_+ are related to Z_{rot}, Z_{vib} and τ_C by:

$$[\lambda_\pm]^{-1} = \frac{5}{6}\left(\frac{1}{Z_{rot}\tau_C} + \frac{1}{Z_{vib}\tau_C}\right) \pm \frac{1}{6}\left[25\left(\frac{1}{Z_{rot}\tau_C} - \frac{1}{Z_{vib}\tau_C}\right)^2 + 16\frac{1}{Z_{rot}\tau_C}\frac{1}{Z_{vib}\tau_C}\right]^{1/2} \tag{96}$$

These relations are invertible and can be used to express the modal energy relaxation times $\tau_{rot} = Z_{rot}\tau_C$ and $\tau_{vib} = Z_{vib}\tau_C$, in terms of the apparent relaxation times λ_- and λ_+:

$$[\tau_{rot}]^{-1} = \frac{3}{5}\frac{\lambda_-^{-1} + \lambda_+^{-1}}{2} + \left[\left(\frac{3}{5}\frac{\lambda_-^{-1} + \lambda_+^{-1}}{2}\right)^2 - \frac{3}{7}\lambda_-^{-1}\lambda_+^{-1}\right]^{1/2} \tag{97}$$

$$[\tau_{vib}]^{-1} = \frac{3}{5}\frac{\lambda_-^{-1} + \lambda_+^{-1}}{2} - \left[\left(\frac{3}{5}\frac{\lambda_-^{-1} + \lambda_+^{-1}}{2}\right)^2 - \frac{3}{7}\lambda_-^{-1}\lambda_+^{-1}\right]^{1/2} \tag{98}$$

Such relaxation problems are solved using the UGKS. First, the vibrational mode of the gas is deactivated, and the ES-BGK model of Andriès et al. [2] is employed. This leads to the relaxation processes described in

(94), which are obtained by assuming a constant relaxation time τ rather than the standard definition (24). This relaxation time is therefore assigned an arbitrary value, as it is the characteristic time of the overall relaxation. The characteristic times of the modal relaxations are related to the Prandtl number and Z_{rot}, which are fixed to 0.71 and 5, respectively. The degrees of freedom are $D_{\boldsymbol{v}} = 3$ and $\delta = 2$, and the numerical time step Δt is set to one quarter of τ. Initially, the gas is assumed to have the following temperatures:

$$\Theta^0/R_s = \begin{pmatrix} 0.5 & & \\ & 1.1 & \\ & & 2.0 \end{pmatrix} T_{eq}, \qquad T^0_{tr} = 1.2T_{eq}, \qquad T^0_{rot} = 0.7T_{eq}. \tag{99}$$

Figure 5 illustrates the relaxation of energies induced by the initial temperatures presented above. First, the figure demonstrates a perfect match between the numerical data (point style) and the analytical expression (94) (dashed lines). Second, by performing linear regressions on a semi-log scale, the characteristic decay times of the deviations of the energy modes from equilibrium, as well as their associated regression errors, can be evaluated. This yields estimates of Z_{rot} for the translational and rotational curves, denoted $Z_{rot,tr}$ and $Z_{rot,rot}$ in Figure 5, and estimates of the Prandtl number, denoted $\mathrm{Pr}_{x,y,z}$, for each directional temperature curve. Here, the regression errors are very low, indicating that exponential decays are well reproduced. Additionally, the post-simulation estimated relaxation coefficients $\mathrm{Pr}^{num} \approx 0.70$ and $Z^{num}_{rot} \approx 5.0$ agree well with the values imposed at the beginning of the simulation. Consequently, the relaxation of energies is well reproduced by the relaxation term of the numerical scheme.

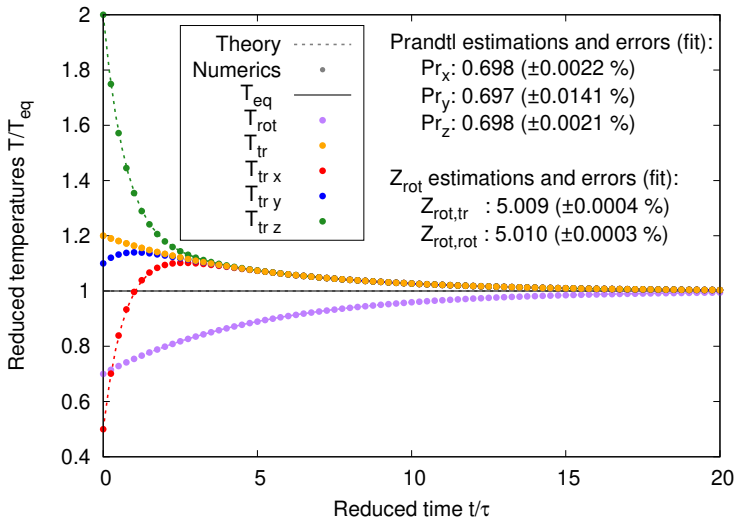


Figure 5: Relaxation of directional and internal temperatures toward the equilibrium temperature, described by the ES-BGK model of Andriès et al. [2] ($\mathrm{Pr} = 0.71$, $Z_{rot} = 5$).

Then, the vibrational mode is enabled using the ES-BGK model of Pfeiffer et al. [26]. This results in relaxation processes described in (95–98). The relaxation time τ is assigned a constant value for the same reason as before, and the mean collision time τ_C is arbitrarily set to the constant $\tau_C = 3\tau/2$. The Prandtl number is fixed to 0.70, Z_{rot} and Z_{vib} are set to 5 and 20 respectively and the degrees of freedom are $D_{\boldsymbol{v}} = 3$ and $\delta = \delta' = 2$. The numerical time step Δt is set to $\tau/10$ until time $t = 7\tau$ and then set to 5τ until the end of the simulation. Initially, the gas is assumed to have the following temperatures:

$$\Theta^0/R_s = \begin{pmatrix} 0.3 & & \\ & 0.7 & \\ & & 2.0 \end{pmatrix} T_{eq}, \qquad T^0_{tr} = T_{eq}, \qquad T^0_{rot} = 1.6T_{eq}, \qquad T^0_{vib} = 0.4T_{eq}. \tag{100}$$

Figure 6 illustrates the relaxation of energies induced by the initial temperatures presented above. As in the previous case, the exponential decays of energy are well reproduced since numerical data perfectly match the analytical solution. Regressions are again performed to estimate the Prandtl, Z_{rot}, and Z_{vib} numbers, with the naming conventions in Figure 6 identical to those previously used. These post-simulation estimated relaxation coefficients

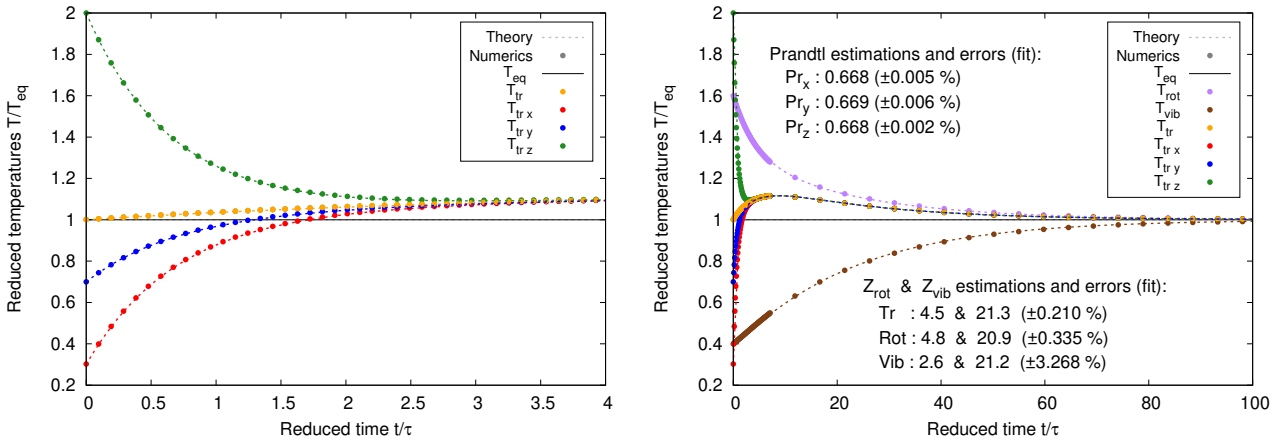


Figure 6: Relaxation of directional and internal temperatures toward the equilibrium temperature, described by the ES-BGK model of Pfeiffer et al. [26] (Pr = 0.70, $\tau_C = 1.5\tau$, $Z_{rot} = 5$, $Z_{vib} = 20$).

are in good agreement with the values imposed at the beginning of the simulation, except for the estimation of Z_{vib} based on the evolution of T_{vib} (estimated at 2.6). This discrepancy is likely due to the concentration of numerical points in the early time period ($t/\tau < 7$), where the effect of vibrational relaxation through the Z_{vib} parameter is negligible compared to translational and rotational relaxation. Using a smaller time step would likely reduce the discrepancy observed in the Z_{vib} estimation.

4.2 Couette flow

In the Couette configuration, the flow is situated between two parallel, isothermal, infinite plates. One plate is stationary, while the other moves with a finite velocity u_w, as illustrated in Figure 7. In this configuration, significant simplifications can be made. Indeed, owing to the invariance in the y- and z-directions, the steady nature, and the planarity of the problem, the Compressible Navier-Stokes (CNS) equations under the continuum assumption reduce to:

$$\begin{array}{cccc} \text{Continuity:} & \text{x-Momentum:} & \text{y-Momentum:} & \text{Energy:} \\ u_x = 0, & \partial_x p = 0, & \partial_x(\mu \partial_x u_y) = 0, & 0 = -\partial_x q + \partial_x(\mu u_y \partial_x u_y). \end{array}$$

Considering constant viscosity μ and constant thermal conductivity κ, Fourier's law for the heat flux q, and no slip boundary conditions, the flow can be theoretically solved within the continuum assumption as follows:

$$u_y(x) = \frac{x}{L} u_w, \qquad T(x) = T_w + \frac{1}{2}\frac{\mu}{\kappa}\frac{x(L-x)}{L^2} u_w^2, \qquad p(x) = cst. \tag{101}$$

The Prandtl number can be identified in the parabolic coefficient of the temperature expression, using the specific heat at constant pressure c_p:

$$T(x) = T_w + \frac{1}{2}\frac{\mathrm{Pr}}{c_p}\frac{x(L-x)}{L^2} u_w^2. \tag{102}$$

Finally, a Knudsen number can be associated with the flow, defined as:

$$\mathrm{Kn} = \frac{\mu_{ref}}{L\rho\sqrt{R_s T_w}}. \tag{103}$$

To assess the effectiveness of incorporating the collision process into the characteristic method (39) used for defining numerical fluxes, we introduce the KO2 scheme, which is formulated without these relaxation effects. Namely, the KO2 scheme employs standard second-order upwind fluxes, as derived from the REA approach (see [22]). This scheme naturally recovers the same behavior as UGKS in the free molecular regime. However, it behaves differently in the continuum or near-continuum regimes, where the two schemes diverge significantly. Formally, this scheme is expressed in the same way as the UGKS (80), simply by imposing $q_1 = q_2 = q_3 = 0$, $q_4 = 1$, and $q_5 = -\Delta t/2$.

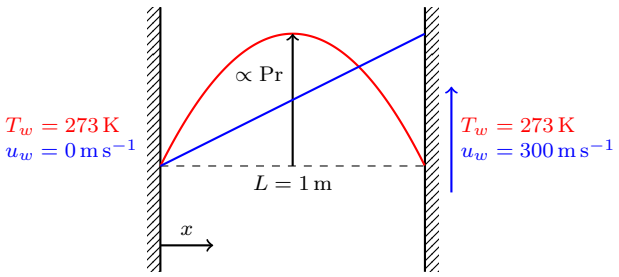


Figure 7: Macroscopic profiles in a Couette flow.

To further investigate the precise formulation of the UGKS fluxes given in (80), two alternative versions are considered in this specific test case. The first formulation corresponds to the UGKS flux of (80), denoted here as UGKS (G), in which the equilibrium G is used as leading term of the equilibrium part. The second formulation, referred to as UGKS (M), employs an alternative flux expression in which the leading equilibrium term G is replaced by its Maxwellian counterpart M. This flux is expressed by:

$$\phi^{n,M}_{i+\frac{1}{2},k} = v_{kx}\left[q_1 M + q_2 v_{kx}\delta_x M + q_3\delta_t M + q_4 F + q_5 v_{kx}\delta_x F\right]^n_{i+\frac{1}{2},k}(\epsilon, \mathrm{i}), \tag{104}$$

which allows assessing the impact of the equilibrium representation on the overall flux behavior.

Simulations are conducted at different Knudsen numbers by varying the initial density of the flow. All configurations are in the near-continuum regime, allowing for comparison with the analytical CNS solution (102) as presented in Figure 8. The temperature profiles are thus expected to be parabolic, with the Prandtl number determining the leading coefficient. Since the problem is planar, a reduced distribution technique allows us to avoid discretizing the z-component of the particle velocities. The parameters of flows are listed in Table 1, and the ES-BGK model of Andriès [2] is used since the vibrational mode of energy is not excited.

GAS				WALL			
gas	*non real*	R_s	$296.8\,\mathrm{J\,kg^{-1}\,K^{-1}}$	type	diffuse	L	$1.0\,\mathrm{m}$
μ	$1.656\times10^{-5}\,\mathrm{Pa\,s}$	Pr	0.71	u_w	$300.0\,\mathrm{m\,s^{-1}}$	T_w	$273.0\,\mathrm{K}$
δ	2.0	Z_{rot}	5.0				

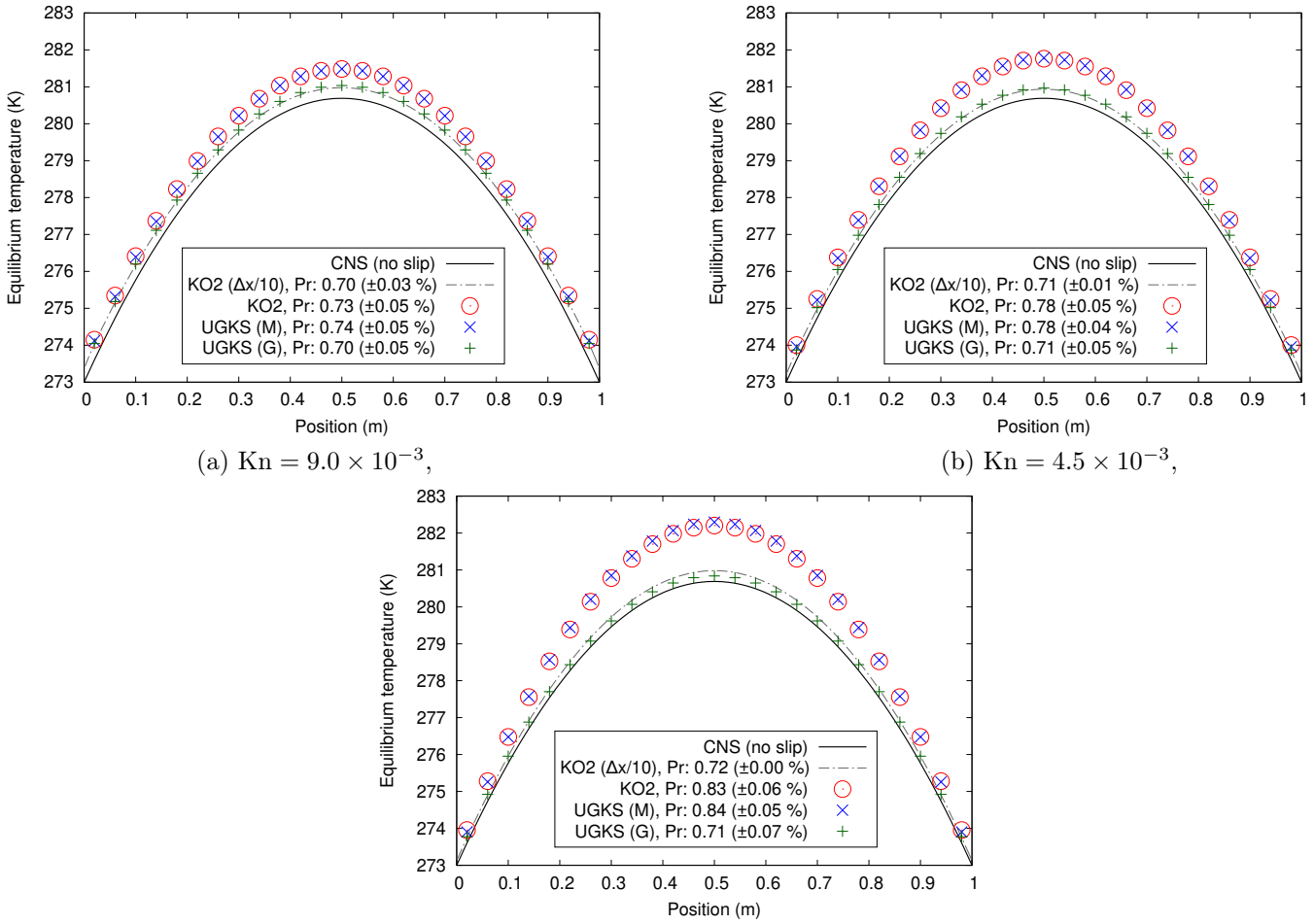
NUMERICAL PARAMETERS	
spatial cells	25
velocity bounds	$[\pm1200]\times[-1200,+1500]\,\mathrm{m\,s^{-1}}$
velocity numbers	50×50

Table 1: Parameters used in 1D Couette flows.

In Figure 8, comparisons of the UGKS (G), UGKS (M) and KO2 schemes with the CNS solutions at different Knudsen numbers are presented. A simulation with the kinetic KO2 scheme and a finer mesh has been conducted to illustrate the deviation of the CNS solution from the rarefied flow solutions. Indeed, the presented flows cannot be accurately modeled using no-slip boundary conditions, as assumed in the derivation of the CNS solution (101).

First, Figure 8 demonstrates the capability of both UGKS and KO2 schemes to resolve flows for Knudsen numbers associated with near-continuum conditions. However, it appears that UGKS (G) achieves better accuracy than UGKS (M) and KO2 schemes on a same mesh. Indeed, the temperature profiles from UGKS (G) simulations are closer to the CNS solutions and KO2 with a refined mesh ($\Delta x/10$). More precisely, as the Knudsen number decreases, the KO2 and UGKS (M) predictions increasingly deviate, while the UGKS (G) predictions remain accurate.

Second, using the temperature equation (102), it is possible to estimate the Prandtl number after each simulation through a parabolic regression using the least squares method. Although this methodology may be criticized for near-continuum flows where slip conditions affect the boundary values and the leading coefficient of the temperature parabola, acceptable results can still be obtained. Figure 8 shows an excellent capability of UGKS (G) to simulate a correct Prandtl number for a given mesh, whereas the KO2 scheme deviates as the Knudsen number decreases



(a) $\mathrm{Kn} = 9.0 \times 10^{-3}$, (b) $\mathrm{Kn} = 4.5 \times 10^{-3}$,

(c) $\mathrm{Kn} = 2.2 \times 10^{-3}$,

Figure 8: Comparison of UGKS (G), UGKS (M) and KO2 schemes with CNS solution in 1D Couette flow and their Prandtl estimations using regressions (Pr=0.71, $Z_{rot} = 5.0$).

unless the mesh is sufficiently refined. The same observation can be made with UGKS (M), which also requires a fine mesh to guarantee accurate results.

This illustrates the advantage of the Asymptotic Preservation (AP) property in enabling UGKS [10] to recover the Navier-Stokes solutions, unlike classical kinetic schemes such as KO2. While this property is obtained by incorporating the relaxation of the mass distribution F toward its equilibrium within the flux expression, it critically requires to correctly treat the discrete equilibrium counterpart in the numerical flux to maintain this AP property as illustrated with the comparisons between UGKS (G) and UGKS (M). The results obtained in this test case validate the choices made in Section 3.2.4, and justify the selection of the UGKS (G) formulation as the appropriate expression of the UGKS.

4.3 Normal shock wave

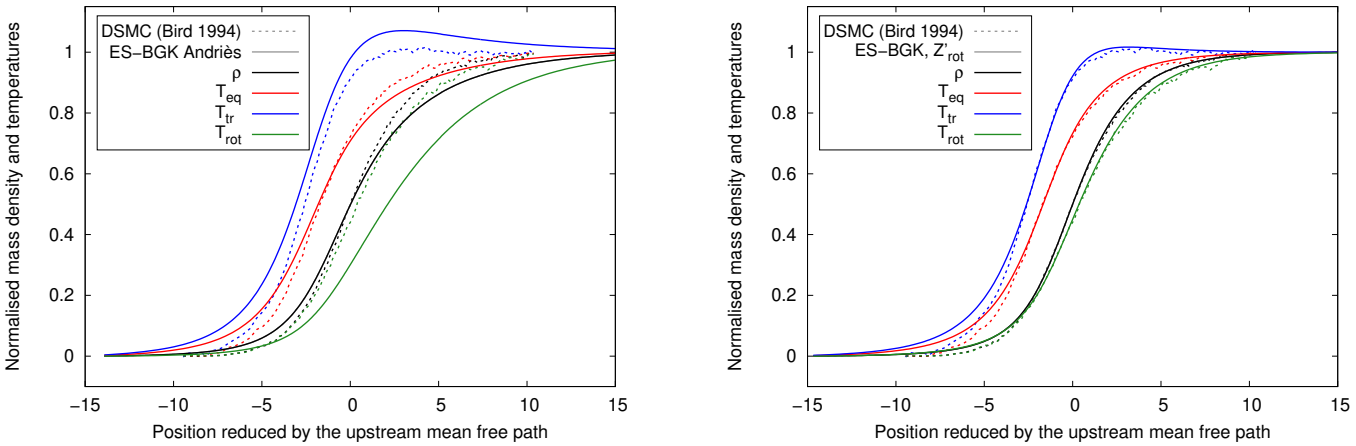
In this third case, we are interested in steady and non-equilibrium flows. The 1D normal shock wave serves as a standard reference in the study of internal relaxation processes. This type of flow is divided into three main regions: the supersonic inflow, the associated outflow linked by the Rankine-Hugoniot relation, and the shock between them. When the heat capacity ratio γ of the gas is constant across the shock, the Rankine-Hugoniot relation explicitly links the downstream state to the upstream conditions of the gas as follows:

$$\frac{\rho_{out}}{\rho_{in}} = \frac{(\gamma+1)\mathrm{Ma}^2}{2+(\gamma-1)\mathrm{Ma}^2}, \qquad \frac{u_{out}}{u_{in}} = \frac{\rho_{in}}{\rho_{out}}, \qquad \frac{p_{out}}{p_{in}} = 1 + \frac{2\gamma}{\gamma+1}(\mathrm{Ma}^2-1). \tag{105}$$

When γ is constant, the upstream Mach number can be considered here as the only macroscopic parameter influencing the structure of the shock. Indeed, the shock thickness is a few mean free paths long for all Knudsen numbers and is a similarity parameter in shock studies. The upstream mean free path is determined using the *Variable Hard Spheres* (VHS) model [6], which defines it as:

$$\lambda = \frac{2(5-2\omega)(7-2\omega)}{15}\sqrt{\frac{1}{2\pi R_s T_{ref}}}\left(\frac{T}{T_{ref}}\right)^{\omega-\frac{1}{2}}\frac{\mu_{ref}}{\rho}. \tag{106}$$

The UGKS is compared to DSMC results [6] for a stationary shock of dinitrogen at Mach 1.71 without any vibrational excitation. Due to Knudsen similarity, the only determining parameters are the index viscosity $\omega = 0.74$, the Prandtl number $\mathrm{Pr} = 0.71$, and the collision number $Z_{rot} = 5$. Finally, since the problem is fully 1D, a reduced distribution technique allows us to discretize the particle velocities phase with only 20 points, accounting solely for the x-component.



(a) Andriès ES-BGK model [2],

(b) Modified ES-BGK model to account for same energy relaxations as [6],

Figure 9: Comparison of density and temperatures within a stationary shock of dinitrogen at Mach 1.71, simulated using DSMC [6] and UGKS on different ES-BGK models.

The comparison of mass density and temperature modes is presented in Figure 9(a). Clearly, the mode separation predicted by UGKS and DSMC differs significantly. In details, the ES-BGK model of Andriès et al. [2] does not reproduce the same relaxation times of the energy modes as DSMC does. Indeed, according to (94), for this ES-BGK model, the relaxation time of the rotational energy is $\tau_{rot}^{ES,And} = Z_{rot}\tau$, whereas for DSMC results used here, it is $\tau_{rot}^{DSMC} = Z_{rot}\tau_C$, where τ_C is the mean collision time, defined in the VHS model as:

$$\tau_C^{VHS} = \frac{(5-2\omega)(7-2\omega)}{30}\frac{\mu}{p}. \tag{107}$$

A sufficient condition to recover the same relaxation time of internal energy as DSMC [6] is to modify Z_{rot} for this ES-BGK model to $Z'_{rot} = Z_{rot}\tau_C/\tau$. The result is presented in Figure 9(b) and clearly shows better agreements. However, one can notice the deviation of the ES-BGK results in the upstream region where the shock is a bit more extended. This behavior is commonly observed with the ES-BGK class of models and should be related to the independence of the time τ on higher moments of the microscopic distribution, the distribution itself [6] or particle velocities [30].

4.4 Sod shock tube

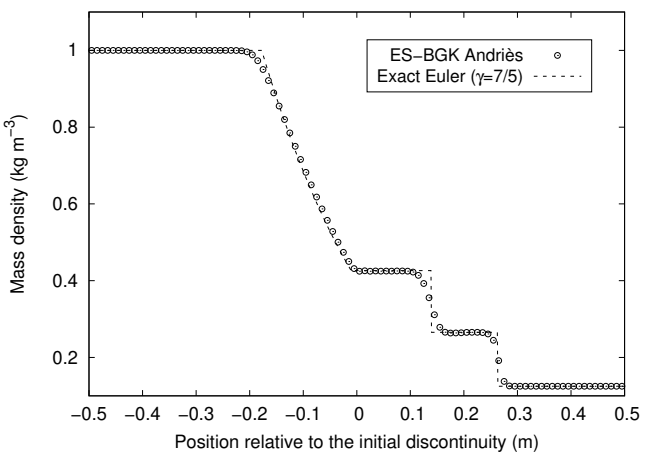
The Sod shock tube problem is a classic 1D unsteady case used to validate numerical methods from rarefied dynamics to fluid dynamics. It involves a tube divided, by a diaphragm, into two sections at different states of pressure and density. Upon removal of the diaphragm, a shock wave propagates into the low-pressure region, a rarefaction wave travels into the high-pressure region, and a contact discontinuity forms between them. This setup provides a robust

test for assessing the accuracy and stability of numerical schemes in capturing shock waves, contact discontinuities, and rarefaction waves.

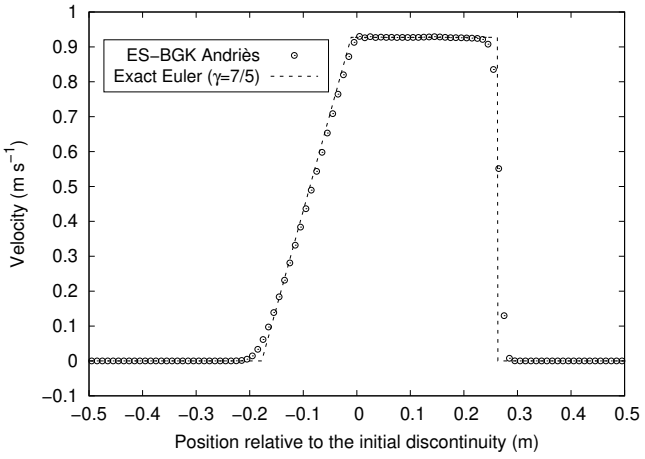
Here, the UGKS is tested on different ES-BGK models by simulating the classic Sod shock tube problem first at a low Knudsen number and then in a transitional regime. Since the problem is fully 1D, a reduced distribution technique allows us to take into account only the x-component of the particle velocities.

4.4.1 Low Knudsen number

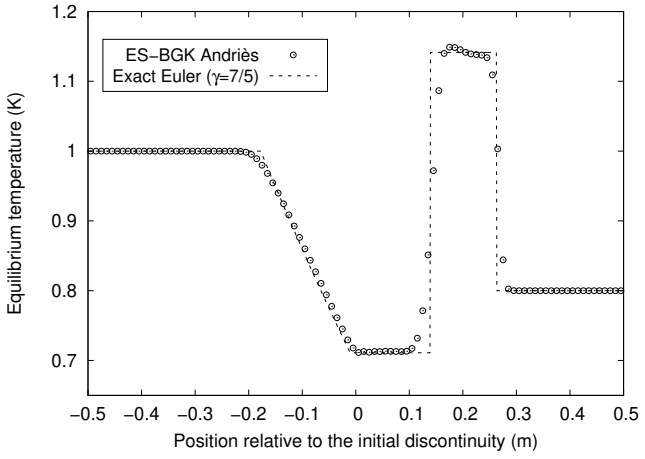
The parameters for the first simulation are listed in Table 2. The model of Andriès [2] is firstly used and results are compared with the exact solution of the Euler equations as shown in Figure 10. The exact solutions of the Euler equation are provided by a program developed by Toro [37] and integrated into the NUMERICA library.



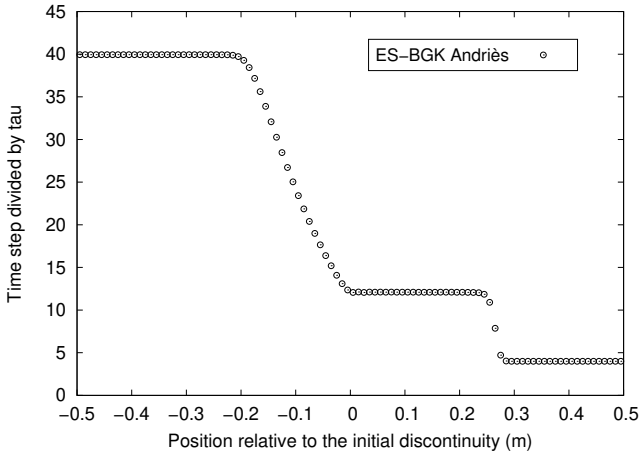
(a) Mass density,



(b) Velocity,



(c) Equilibrium temperature,



(d) Numerical resolution $\Delta t/\tau$,

Figure 10: Comparison of macroscopic quantities and numerical resolution in a Sod shock tube, computed in continuum configuration with UGKS applied on ES-BGK and with exact solution of Euler equations [37].

According to Figure 10, and in comparison to the Euler solution, the plateau values and wave positions computed by the UGKS are accurate for each flow region, consistent with the behavior of a diatomic molecule with a constant specific heat ratio γ=7/5. This case confirms that the UGKS can effectively handle unsteady flows in the continuum regime. Furthermore, Figure 10d shows the ratio $\Delta t/\tau$ which is clearly greater than 1 in every cells. This demonstrate the scheme ability to simulate continuum flow, even when the mesh is under-resolved ($\Delta t > \tau$). This capability is a key feature of the UGKS framework [43] to permit reasonable simulations of flow in continuum regime.

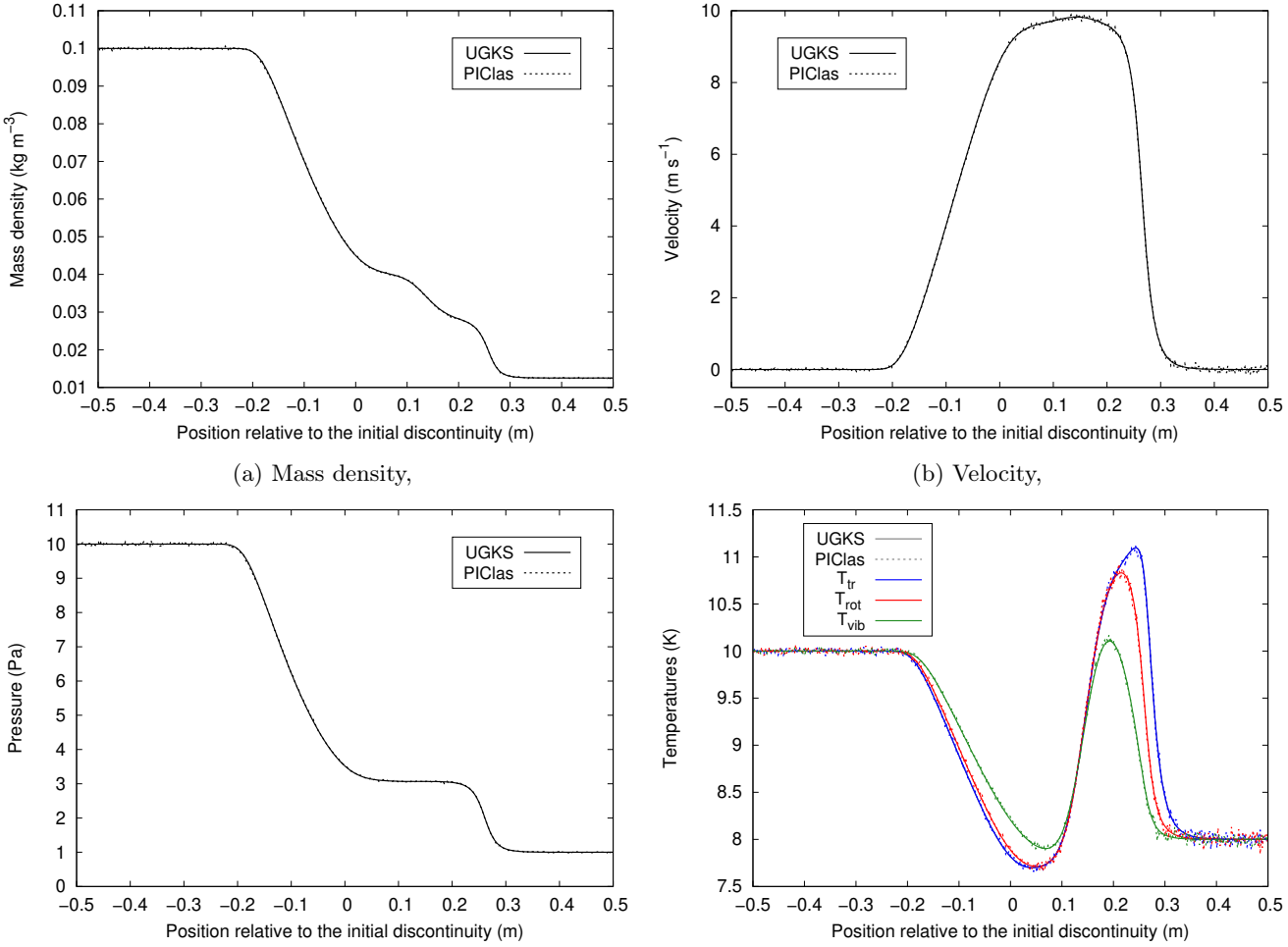
INITIAL CONDITIONS			
ρ_L	$1.0\,\mathrm{kg\,m^{-3}}$	ρ_R	$0.125\,\mathrm{kg\,m^{-3}}$
u_L	$0.0\,\mathrm{m\,s^{-1}}$	u_R	$0.0\,\mathrm{m\,s^{-1}}$
p_L	$1.0\,\mathrm{Pa}$	p_R	$0.1\,\mathrm{Pa}$

GAS	
gas	*non real*
δ	2.0
R_s	$1.0\,\mathrm{J\,kg^{-1}\,K^{-1}}$
μ	$5.0\times10^{-5}\,\mathrm{Pa\,s}$
Pr	0.71
Z_{rot}	5.0

NUMERICAL PARAMETERS	
length L	$1.0\,\mathrm{m}$
final time	$0.15\,\mathrm{s}$
spatial cells	100
velocity bounds	$[-8, 8]\,\mathrm{m\,s^{-1}}$
velocity numbers	30

Table 2: Parameters used for the low-Knudsen 1D Sod flow.

4.4.2 Transitional Knudsen number



(a) Mass density,

(b) Velocity,

(c) Pressure,

(d) Temperatures,

Figure 11: Comparison of macroscopic quantities in a transitional Sod shock tube, computed with UGKS and PIClas, both solving the ES-BGK model of Pfeiffer et al. [26].

The second Sod shock tube is performed with the ES-BGK model of Pfeiffer et al. [26], accounting for translational, rotational and vibrational modes of energy. The parameters for this second simulation are listed in Table 3. The results are compared with those obtained using the PIClas code [15, 34], which also solves this model through a particle stochastic approach, see Figure 11. In these simulations, the PIClas code (version 3.3.1) was used with a time step of $\Delta t = 1.0\times10^{-5}$ s, different from that used in the UGKS simulations, and 40 million numerical particles, each representing 1.0×10^{15} real physical particles.

Since the Knudsen number is larger than previously and due to the shock, the flow is not in an equilibrium

state. Figure 11d clearly shows that the modal temperatures differ. Furthermore, the solutions obtained with PIClas exhibit excellent agreement with those computed using our deterministic UGKS-ES-BGK solver. This confirms that UGKS is also well-suited for accurately capturing unsteady flows in non-equilibrium states.

INITIAL CONDITIONS			
ρ_L	$0.1\,\mathrm{kg\,m^{-3}}$	ρ_R	$0.0125\,\mathrm{kg\,m^{-3}}$
u_L	$0.0\,\mathrm{m\,s^{-1}}$	u_R	$0.0\,\mathrm{m\,s^{-1}}$
p_L	$10.0\,\mathrm{Pa}$	p_R	$1.0\,\mathrm{Pa}$

GAS (HS)			
gas	*non real*	Pr	Eucken (23)
R_s	$10.0\,\mathrm{J\,kg^{-1}\,K^{-1}}$	Z_{rot}	5.0
μ_{ref}	$1.0\times10^{-3}\,\mathrm{Pa\,s}$	Z_{vib}	20.0
T_{ref}	$10.0\,\mathrm{K}$	T_0	$12.0\,\mathrm{K}$
δ	2.0		

NUMERICAL PARAMETERS	
length L	$1.0\,\mathrm{m}$
final time	$0.015\,\mathrm{s}$
spatial cells	1000
velocity bounds	$[-80, 80]\,\mathrm{m\,s^{-1}}$
velocity numbers	30

Table 3: Parameters used for the 1D Sod flow including vibrational mode of energy.

4.5 Supersonic flow passing an infinite flat plate

This case is based on an experiment conducted by Tsuboi and Matsumoto [38] and depicted in Figure 12. In this experiment, a flat plate is immersed in a supersonic near-continuum dinitrogen cold flow. The experimental set up and the parameters for subsequent numerical comparisons are detailed in Table 4. A strong shock wave forms in front of and beneath the wedge. Above the plate near the leading edge, a merged layer occurs before the separation of the boundary layer and the weak shock. A lot of kinetic energy from the supersonic nature of the flow is converted into thermal and internal energy both in the shock and also within the boundary layer due to viscous effects. The internal energy is entirely associated with the rotational mode, since the reached temperature is too low compared to the characteristic temperature of vibration $T_0 = 3371\,\mathrm{K}$ to initiate significant excitation of the vibrational mode.

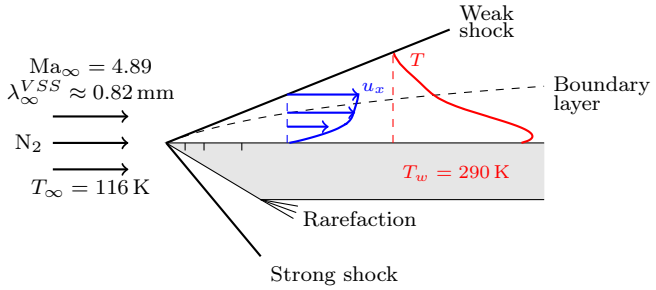


Figure 12: Illustration of the experiment No. 34 from Tsuboi and Matsumoto [38].

GAS (VSS)			
gas	$\mathrm{N_2}$	T_{ref}	$273.15\,\mathrm{K}$
δ	2.0	R_s	$296.8\,\mathrm{J\,kg^{-1}\,K^{-1}}$
α	1.36	Pr	0.71
ω	0.74	Z_{rot}	5.0
μ_{ref}	$1.656\times10^{-5}\,\mathrm{Pa\,s}$		

INFLOW		WALL	
λ_∞^{VSS}	$0.82\,\mathrm{mm}$	type	diffuse
ρ_∞	$6.16\times10^{-5}\,\mathrm{kg\,m^{-3}}$	T_w	$290.0\,\mathrm{K}$
Ma_∞	4.89		
T_∞	$116.0\,\mathrm{K}$		

Table 4: Numerical simulation conditions of the supersonic flow passing a flat plate.

For the computations, the velocity domain is reduced to 2D using a reduced distribution technique. The computational domain in space is meshed using a 2D structured mesh all around the plate, as illustrated in Figure 13. A total of 40×40 elements are used above the plate. The velocity grid is bounded within $[-1200, 2200]\times[-1700, 1500]$

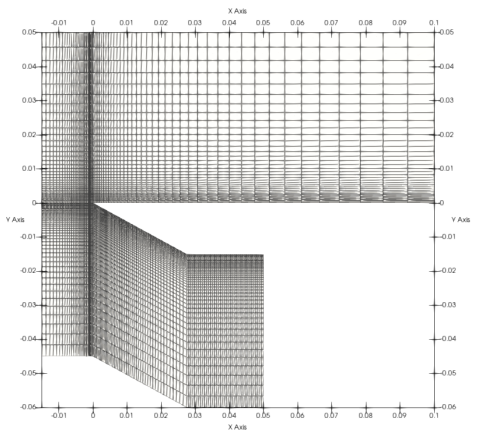


Figure 13: Mesh used for the computation of the supersonic flow passing the flat plate.

($\mathrm{m\,s^{-1}}$) and discretized with 25 points in both x and y-directions. Comparisons with experimental measurements of rotational temperature, taken at 5 and 20 mm from the stagnation point on vertical lines, are presented in Figure 14.

The results provided by the ES-BGK model of both Andriès al. [2] and Pfeiffer et al. [26] are in good agreement with the experimental data. The shape of the rotational temperature is well reproduced, although some discrepancies arise due to uncertainties in accurately modeling the real relaxation processes involved. As seen with the stationary shock problem, a simple replacement in Z_{rot} can align the relaxation behavior with that of other authors. To match the relaxation processes used by Bird [6] based on the ES-BGK model of Andriès [2], the required adjustments is $Z'_{rot} = \tau_C Z_{rot}/\tau$. This modifications significantly influence the profiles of rotational temperature, as illustrated in Figure 14.

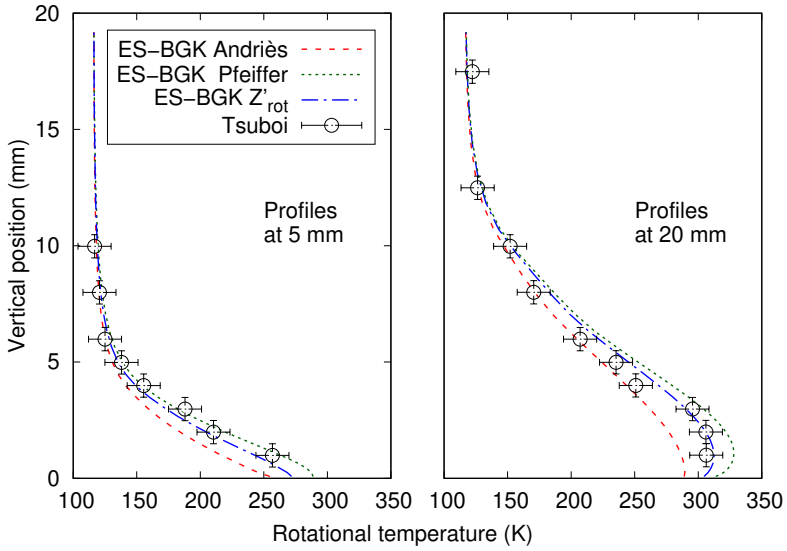


Figure 14: Comparison of rotational temperature predictions between ES-BGK models and experimental data [38] above the flat plate in the supersonic flow of dinitrogen.

4.6 Hypersonic flow passing an infinite cylinder

Finally, the last flow studied here is a two-dimensional hypersonic flow of dinitrogen around an infinite cylinder, representative of atmospheric conditions at an altitude of 70 km [7]. Given the hypersonic nature of the flow, a detached shock forms ahead of the cylinder, as illustrated in Figure 15. This shock is smooth, and the non-equilibrium effects occurring within it are particularly noticeable due to the rarefaction of the flow. Additionally,

the conversion of a major part of kinetic energy to thermal energy behind the shock results in a sufficient increase in temperature for the vibrational mode of the particles to be excited. Consequently, to model these non-equilibrium phenomena, we adopt the ES-BGK model developed by Pfeiffer et al. [26].

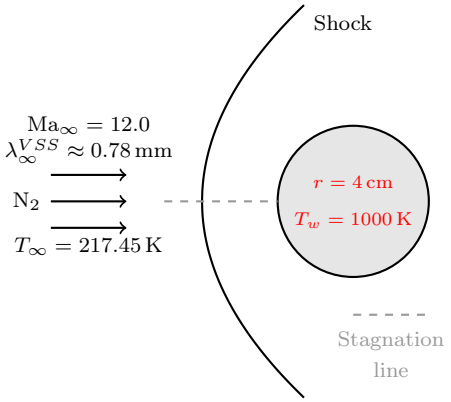


Figure 15: Illustration of the hypersonic flow of dinitrogen around an infinite cylinder.

The computational domain is reduced to two dimensions using a reduced distribution approach and discretized with a structured 2D mesh surrounding a cylinder of radius 4 cm. The velocity space is similarly bounded and discretized in both the x- and y-directions. Finally, the properties of the gas flow and numerical parameters are summarized in Table 5.

GAS (VSS)				INFLOW		WALL	
gas	N_2	R_s	$296.8\,\mathrm{J\,kg^{-1}\,K^{-1}}$	λ_∞^{VSS}	$0.78\,\mathrm{mm}$	type	diffuse
δ	2.0	Pr	Eucken (23)	ρ_∞	$7.48 \times 10^{-5}\,\mathrm{kg\,m^{-3}}$	T_w	$1000\,\mathrm{K}$
α	1.36	Z_{rot}	5.0	u_∞	$3608\,\mathrm{m\,s^{-1}}$	r	$4\,\mathrm{cm}$
ω	0.74	Z_{vib}	200.0	T_∞	$217.45\,\mathrm{K}$		
μ_{ref}	$1.656 \times 10^{-5}\,\mathrm{Pa\,s}$	T_0	$3371\,\mathrm{K}$				
T_{ref}	$273.15\,\mathrm{K}$						

NUMERICAL PARAMETERS	
spatial cells	350 normal $\times$ 100 azimutal
velocity bounds	$[-6000, 9500] \times [\pm 8000]\,\mathrm{m\,s^{-1}}$
velocity numbers	55×60

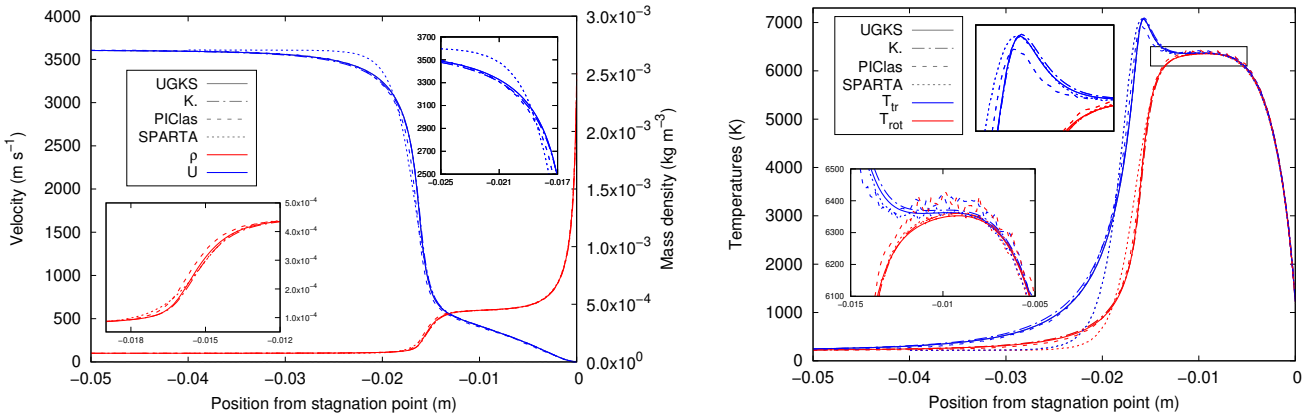
Table 5: Parameters for the UGKS simulation of the hypersonic flow of dinitrogen around the infinite cylinder.

To validate our implementation, we compare the macroscopic profiles computed along the stagnation line with those obtained from three reference solvers: the open-source DSMC code SPARTA [35], the particle stochastic ES-BGK solver PICLas (version 3.3.1) [15, 34], and the deterministic finite volume ES-BGK solver developed at CEA, referred to as code K. [3]. Since theses codes do not model the same relaxation processes for internal energies, we first consider a configuration excluding the vibrational energy mode. In this case, rotational and translational degrees of freedom are fully excited, and it is possible to ensure consistent energy relaxations across solvers by adjusting their respective rotational collision number Z_{rot}, as done in Sections 4.3 and 4.5. Subsequently, a second set of simulations includes vibrational energy and compares results between UGKS and PICLas since their energy relaxation processes are both consistent with the same ES-BGK model.

Lastly, since our objective here is to assess the consistency of our UGKS implementation with the model of Pfeiffer, we intentionally use a fine numerical resolution. The PICLas simulations are also run with a small time step, $\Delta t = 2 \times 10^{-8}\,\mathrm{s}$, and use 200 million numerical particles, each representing 10 billion real physical particles. Comparisons between simulation results are performed along the stagnation line and are presented in Figures 16 and 17, corresponding to the non-vibrational and vibrational cases, respectively.

4.6.1 Without the vibrational mode of energy

Good agreement can be observed between all compared codes for the non vibrating case as illustrated in Figure 16. The primary difference is the shock expansion predicted by ES-BGK based codes compared to SPARTA which emulates the Boltzmann equation. This well-known behavior of BGK and ES-BGK models is generally attributed to the independence of the relaxation time τ on higher moments of the microscopic distribution, the distribution itself [6] or particle velocities [30]. Another difference appears in the peak translational temperature: while UGKS, K., and SPARTA yield similar values (within 0.15 %), PIClas slightly underpredicts this maximum, with a deviation of -2.12 % as compared to UGKS, even if the shapes of temperature profiles are well reproduced.

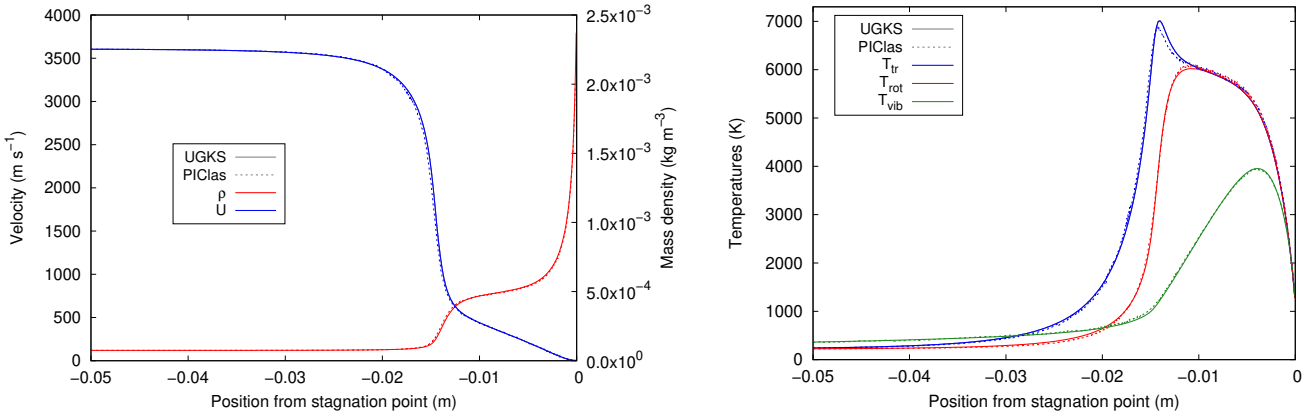


(a) Mass density and velocity,

(b) Temperatures,

Figure 16: Comparison of mass density, velocity and temperatures profiles along the stagnation line of a cylinder in a hypersonic flow where the vibrational mode of energy is ignored, between UGKS, K., PIClas and SPARTA simulations.

4.6.2 With the vibrational mode of energy



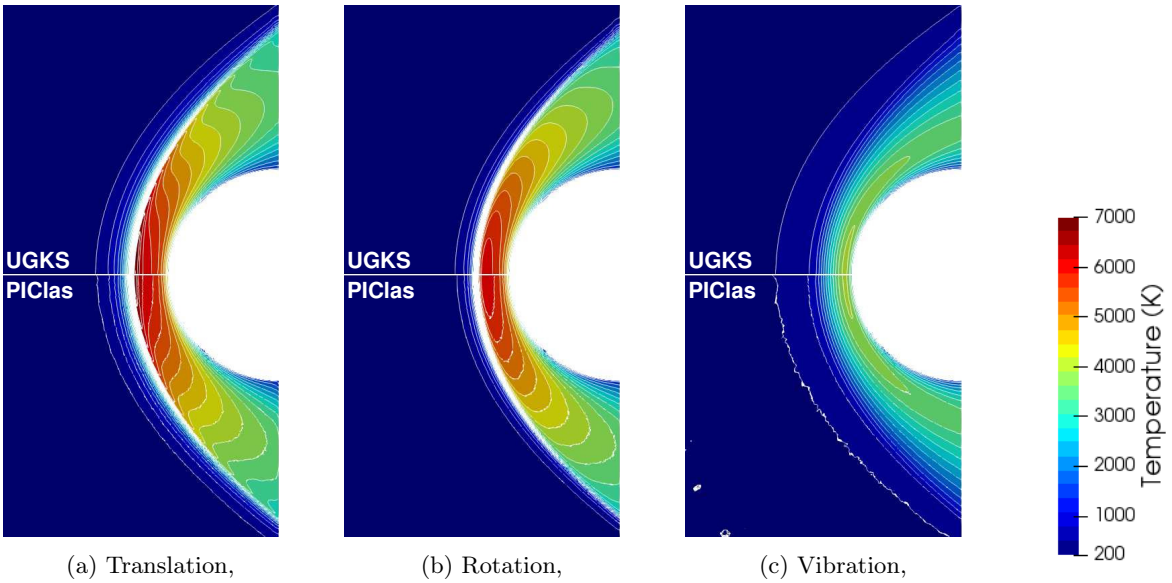
(a) Mass density and velocity,

(b) Temperatures,

Figure 17: Comparison of mass density, velocity and temperature profiles along the stagnation line of a cylinder in a hypersonic flow where the vibrational mode of energy is considered, between UGKS and PIClas simulations.

The vibrational mode is then taken into account, and the UGKS implementation is only compared with PIClas, since only both of them are consistent with the ES-BGK model of Pfeiffer. Unlike in the case without vibrational energy, the mesh for the PIClas simulation must be carefully designed. In particular, it must be sufficiently refined near the cylinder surface to accurately resolve the local mean free path. Simultaneously, it should be coarser in

the free-stream region to avoid numerical artifacts caused by the quantization of vibrational energy, as discussed in [7], which may occur when too few numerical particles are present per cell. Nevertheless, good agreement can be obtained with an appropriate mesh, as illustrated in Figures 17 and 18. The temperature profiles from both UGKS and PIClas simulations are in close agreement, with the main discrepancy observed at the shock location, where the peak translational temperature predicted by PIClas deviates from UGKS by 1.71 %. Such small deviation was primarily observed in the absence of vibration and seems to be related to PIClas. Except in this region, excellent agreement can be observed all around the cylinder, as illustrated in Figure 18.



(a) Translation, (b) Rotation, (c) Vibration,

Figure 18: Comparison of temperature fields around the cylinder simulated with UGKS (above) and PIClas (below).

5 Conclusion

We proposed an extension of the Unified Gas-Kinetic Scheme (UGKS) to ES-BGK models that incorporates specific diatomic energy modes like rotation and vibration. This adaptation extends the UGKS-ES-BGK framework from monoatomic to diatomic gases using techniques similar to those employed in extending UGKS-BGK to Shakhov and Rykov models. This work demonstrates that adapting UGKS to a reduced distribution equation that accounts for additional energy phenomena can be accomplished with minimal difficulty. Therefore, it is possible to sequentially introduce various energy effects without altering the existing scheme structure.

Moreover, the scheme has been validated across several test cases as energy relaxations or one-dimensional problems, including viscous-driven and shock flows. The Couette flow illustrated the Asymptotic Preserving property of the scheme and its advantage over classical kinetic schemes in capturing the correct Navier-Stokes viscous effects without introducing excessive numerical diffusion. The low Knudsen number Sod shock tube demonstrated the robustness of the scheme in unsteady flows, even when the mesh does not resolve the kinetic scale. Finally, comparisons with results from experiments or from simulation codes like SPARTA or PIClas confirm the excellent capability of the UGKS in solving the ES-BGK model for diatomic gases, including vibrational excitation, even in two-dimensional configurations.

Acknowledgments

This work has been undertaken under the auspice of LRC ANABASE, which is a joined research laboratory between Institut de Mathématiques de Bordeaux and CEA-CESTA devoted to the development of innovative numerical methods for the simulation of complex fluid flows. Computer time for this study was provided by the computing facilities of the MCIA (Mésocentre de Calcul Intensif Aquitain).

References

[1] P. Andriès, J.-F. Bourgat, P. le Tallec, and B. Perthame. "Numerical comparison between the Boltzmann and ES-BGK models for rarefied gases". In: *Computer Methods in Applied Mechanics and Engineering* 191.31 (2002), pp. 3369–3390. ISSN: 0045-7825. DOI: 10.1016/S0045-7825(02)00253-0.

[2] P. Andriès, P. Le Tallec, J.-P. Perlat, and B. Perthame. "The Gaussian-BGK model of Boltzmann equation with small Prandtl number". In: *European Journal of Mechanics - B/Fluids* 19.6 (2000), pp. 813–830. ISSN: 0997-7546. DOI: 10.1016/S0997-7546(00)01103-1.

[3] C. Baranger, J. Claudel, N. Hérouard, and L. Mieussens. "Locally refined discrete velocity grids for stationary rarefied flow simulations". In: *Journal of Computational Physics* 257 (2014), pp. 572–593. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2013.10.014.

[4] C. Baranger, N. Hérouard, J. Mathiaud, and L. Mieussens. "Numerical boundary conditions in Finite Volume and Discontinuous Galerkin schemes for the simulation of rarefied flows along solid boundaries". In: *Mathematics and Computers in Simulation* 159 (2019), pp. 136–153. ISSN: 0378-4754. DOI: 10.1016/j.matcom.2018.11.011.

[5] P. L. Bhatnagar, E. P. Gross, and M. Krook. "A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems". In: *Phys. Rev.* 94 (3 May 1954), pp. 511–525. DOI: 10.1103/PhysRev.94.511.

[6] G. A. Bird. *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Oxford Science Publications, 1994, pp. 1–476. ISBN: 9780198561958.

[7] I. D. Boyd and T. E. Schwartzentruber. "Models for Nonequilibrium Thermochemistry". In: *Nonequilibrium Gas Dynamics and Molecular Simulation*. Cambridge Aerospace Series. Cambridge University Press, 2017, pp. 252–310.

[8] J. Burt and I. Boyd. "Evaluation of a Particle Method for the Ellipsoidal Statistical Bhatnagar-Gross-Krook Equation". In: *44th AIAA Aerospace Sciences Meeting and Exhibit*. Jan. 2006. DOI: 10.2514/6.2006-989.

[9] C. Cercignani. *The Boltzmann Equation and Its Application*. Vol. 67. Applied Mathematical Sciences. Springer New York, NY, 1988, pp. 1–455. ISBN: 978-0-387-96637-3. DOI: 10.1007/978-1-4612-1039-9.

[10] S. Chen and K. Xu. "A comparative study of an asymptotic preserving scheme and unified gas-kinetic scheme in continuum flow limit". In: *Journal of Computational Physics* 288 (2015), pp. 52–65. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2015.02.014.

[11] S. Chen, K. Xu, and Q. Cai. "A Comparison and Unification of Ellipsoidal Statistical and Shakhov BGK Models". In: *Advances in Applied Mathematics and Mechanics* 7.2 (2015), pp. 245–266. DOI: 10.4208/aamm.2014.m559.

[12] C. K. Chu. "Kinetic-Theoretic Description of the Formation of a Shock Wave". In: *The Physics of Fluids* 8.1 (Jan. 1965), pp. 12–22. ISSN: 0031-9171. DOI: 10.1063/1.1761077.

[13] Y. Dauvois, J. Mathiaud, and L. Mieussens. "An ES-BGK model for polyatomic gases in rotational and vibrational nonequilibrium". In: *European Journal of Mechanics - B/Fluids* 88 (2021), pp. 1–16. ISSN: 0997-7546. DOI: 10.1016/j.euromechflu.2021.02.006.

[14] D. R. Chapman F. E. Lumpkin and C. Park. "A new rotational relaxation model for use in hypersonic computational fluid dynamics". In: *24th Thermophysics Conference*. June 1989. DOI: 10.2514/6.1989-1737.

[15] S. Fasoulas, C.-D. Munz, M. Pfeiffer, J. Beyer, T. Binder, S. Copplestone, A. Mirza, P. Nizenkov, P. Ortwein, and W. Reschke. "Combining particle-in-cell and direct simulation Monte Carlo for the simulation of reactive plasma flows". In: *Physics of Fluids* 31.7 (July 2019), p. 072006. ISSN: 1070-6631. DOI: 10.1063/1.5097638.

[16] M. A. Gallis and J. R. Torczynski. "Investigation of the ellipsoidal-statistical Bhatnagar–Gross–Krook kinetic model applied to gas-phase transport of heat and tangential momentum between parallel walls". In: *Physics of Fluids* 23.3 (Mar. 2011), p. 030601. ISSN: 1070-6631. DOI: 10.1063/1.3558869.

[17] L. H. Holway. "Kinetic theory of shock structure using an ellipsoidal distribution function". In: *Rarefied Gas Dynamics, Volume 1* 1 (1965), pp. 193–215.

[18] A. B. Huang, D. P. Giddens, and C. W. Bagnal. "Rarefied Gas Flow between Parallel Plates Based on the Discrete Ordinate Method". In: *The Physics of Fluids* 10.3 (Mar. 1967), pp. 498–502. ISSN: 0031-9171. DOI: 10.1063/1.1762143.

[19] A. B. Huang and P. F. Hwang. "Test of statistical models for gases with and without internal energy states". In: *The Physics of Fluids* 16.4 (Apr. 1973), pp. 466–475. ISSN: 0031-9171. DOI: 10.1063/1.1694368.

[20] J.-C. Huang, K. Xu, and P. Yu. "A Unified Gas-Kinetic Scheme for Continuum and Rarefied Flows II: Multi-Dimensional Cases". In: *Communications in Computational Physics* 12.3 (2012), pp. 662–690. DOI: 10.4208/cicp.030511.220911a.

[21] K. Koura and H. Matsumoto. "Variable soft sphere molecular model for inverse-power-law or Lennard-Jones potential". In: *Physics of Fluids A: Fluid Dynamics* 3.10 (Oct. 1991), pp. 2459–2465. ISSN: 0899-8213. DOI: 10.1063/1.858184.

[22] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[23] S. Liu, P. Yu, K. Xu, and C. Zhong. "Unified gas-kinetic scheme for diatomic molecular simulations in all flow regimes". In: *Journal of Computational Physics* 259 (2014), pp. 96–113. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2013.11.030.

[24] S. Liu and C. Zhong. "Investigation of the kinetic model equations". In: *Phys. Rev. E* 89 (3 Mar. 2014), p. 033306. DOI: 10.1103/PhysRevE.89.033306.

[25] J. A. Lordi and R. E. Mates. "Rotational Relaxation in Nonpolar Diatomic Gases". In: *The Physics of Fluids* 13.2 (Feb. 1970), pp. 291–308. ISSN: 0031-9171. DOI: 10.1063/1.1692920.

[26] J. Mathiaud, L. Mieussens, and M. Pfeiffer. "An ES-BGK model for diatomic gases with correct relaxation rates for internal energies". In: *European Journal of Mechanics - B/Fluids* 96 (2022), pp. 65–77. ISSN: 0997-7546. DOI: 10.1016/j.euromechflu.2022.07.003.

[27] L. Mieussens. "A survey of deterministic solvers for rarefied flows". In: *AIP Conference Proceedings* 1628.1 (Dec. 2014), pp. 943–951. ISSN: 0094-243X. DOI: 10.1063/1.4902695.

[28] L. Mieussens. "Discrete velocity model and implicit scheme for the BGK equation of rarefied gas dynamics". In: *Mathematical Models and Methods in Applied Sciences* 10.08 (2000), pp. 1121–1149. DOI: 10.1142/S0218202500000562.

[29] L. Mieussens. "Discrete-Velocity Models and Numerical Schemes for the Boltzmann-BGK Equation in Plane and Axisymmetric Geometries". In: *Journal of Computational Physics* 162.2 (2000), pp. 429–466. ISSN: 0021-9991. DOI: 10.1006/jcph.2000.6548.

[30] L. Mieussens and H. Struchtrup. "Numerical comparison of Bhatnagar–Gross–Krook models with proper Prandtl number". In: *Physics of Fluids* 16.8 (Aug. 2004), pp. 2797–2813. ISSN: 1070-6631. DOI: 10.1063/1.1758217.

[31] R. C. Millikan and D. R. White. "Systematics of Vibrational Relaxation". In: *The Journal of Chemical Physics* 39.12 (Dec. 1963), pp. 3209–3213. ISSN: 0021-9606. DOI: 10.1063/1.1734182.

[32] J. G. Parker. "Rotational and Vibrational Relaxation in Diatomic Gases". In: *The Physics of Fluids* 2.4 (July 1959), pp. 449–462. ISSN: 0031-9171. DOI: 10.1063/1.1724417.

[33] M. Pfeiffer. "Extending the particle ellipsoidal statistical Bhatnagar-Gross-Krook method to diatomic molecules including quantized vibrational energies". In: *Physics of Fluids* 30.11 (Nov. 2018), p. 116103. ISSN: 1070-6631. DOI: 10.1063/1.5054961.

[34] M. Pfeiffer. "Particle-based fluid dynamics: Comparison of different Bhatnagar-Gross-Krook models and the direct simulation Monte Carlo method for hypersonic flows". In: *Physics of Fluids* 30.10 (Oct. 2018), p. 106106. ISSN: 1070-6631. DOI: 10.1063/1.5042016.

[35] S. J. Plimpton, S. G. Moore, A. Borner, A. K. Stagg, T. P. Koehler, J. R. Torczynski, and M. A. Gallis. "Direct simulation Monte Carlo on petaflop supercomputers and beyond". In: *Physics of Fluids* 31.8 (Aug. 2019), p. 086101. ISSN: 1070-6631. DOI: 10.1063/1.5108534.

[36] E. M. Shakhov. "Generalization of the Krook kinetic relaxation equation". In: *Fluid Dynamics* 3.5 (1968), pp. 95–96. DOI: 10.1007/BF01029546.

[37] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer Berlin, Heidelberg, Oct. 2010. ISBN: 978-3-540-25202-3. DOI: 10.1007/b79761.

[38] N. Tsuboi and Y. Matsumoto. "Experimental and Numerical Study of Hypersonic Rarefied Gas Flow over Flat Plates". In: *AIAA Journal* 43.6 (2005), pp. 1243–1255. DOI: 10.2514/1.10950.

[39] O. Tumuklu, Z. Li, and D. A. Levin. "Particle Ellipsoidal Statistical Bhatnagar–Gross–Krook Approach for Simulation of Hypersonic Shocks". In: *AIAA Journal* 54.12 (2016), pp. 3701–3716. DOI: 10.2514/1.J054837.

[40] Z. Wang, H. Yan, Q. Li, and K. Xu. "Unified gas-kinetic scheme for diatomic molecular flow with translational, rotational, and vibrational modes". In: *Journal of Computational Physics* 350 (2017), pp. 237–259. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2017.08.045.

[41] P. Welander. "On the temperature jump in a rarefied gas". In: *Ark. Fys.* 7 (1954), pp. 507–553.

[42] K. Xu. "A Gas-Kinetic BGK Scheme for the Navier–Stokes Equations and Its Connection with Artificial Dissipation and Godunov Method". In: *Journal of Computational Physics* 171.1 (2001), pp. 289–335. ISSN: 0021-9991. DOI: 10.1006/jcph.2001.6790.

[43] K. Xu and J.-C. Huang. "A unified gas-kinetic scheme for continuum and rarefied flows". In: *Journal of Computational Physics* 229.20 (2010), pp. 7747–7764. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2010.06.032.

[44] K. Xu and J.-C. Huang. "An improved unified gas-kinetic scheme and the study of shock structures". In: *IMA Journal of Applied Mathematics* 76.5 (Mar. 2011), pp. 698–711. ISSN: 0272-4960. DOI: 10.1093/imamat/hxr002.

[45] K. Xu, M. Mao, and L. Tang. "A multidimensional gas-kinetic BGK scheme for hypersonic viscous flow". In: *Journal of Computational Physics* 203.2 (2005), pp. 405–421. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2004.09.001.

[46] Y. Zhu and K. Xu. *The first decade of unified gas kinetic scheme*. 2021. arXiv: 2102.01261 [physics.comp-ph]. URL: https://arxiv.org/abs/2102.01261.