

## ASYMPTOTIC ANALYSIS OF HIGH ORDER IMEX-RK METHODS FOR ES-BGK MODEL AT NAVIER–STOKES LEVEL

SEBASTIANO BOSCARINO<sup>1</sup>  AND SEUNG YEON CHO<sup>2,\*</sup> 

**Abstract.** Implicit–explicit Runge–Kutta (IMEX-RK) time discretization methods are very popular when solving stiff kinetic equations. In the work of Hu and Zhang in [*J. Sci. Comput.* **73** (2017) 797–818], an asymptotic analysis shows that a specific class of high-order IMEX-RK schemes can accurately capture the Navier–Stokes limit without needing to resolve the small scales dictated by the Knudsen number. In this work, we extend the asymptotic analysis to general IMEX-RK schemes, known in literature as Type I and Type II. We further introduce some IMEX-RK methods developed by Boscarino and Pareschi [*J. Comput. Appl. Math.* **316** (2017) 60–73] to attain uniform accuracy in the wide range of Knudsen numbers. Several numerical examples are presented to verify the validity of the obtained theoretical results and the effectiveness of the methods.

**Mathematics Subject Classification.** 82C40, 97N40, 65M12.

Received August 22, 2025. Accepted January 6, 2026.

### 1. INTRODUCTION

One of the most well-known kinetic models for rarefied gas dynamics is the Boltzmann transport equation (BTE). Its dimensionless form is written as

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f, f), \quad (1.1)$$

where  $f(t, x, v)$  is the distribution function which depends on time  $t > 0$ , on the position of particles  $x = (x_1, \dots, x_{d_x}) \in \mathbb{R}^{d_x}$  and on their velocity  $v = (v_1, \dots, v_{d_v}) \in \mathbb{R}^{d_v}$ . Here  $\varepsilon$  is the so-called Knudsen number, defined as the ratio of the mean free path of molecules and the characteristic length scale of the physical problem. The dimension of space and velocity domains are denoted by  $d_x$  and  $d_v$ , respectively.

The Boltzmann collision operator  $Q(f, f)$  is a non-linear operator that describes the binary collisions between molecules. It acts only on the velocity dependence of the distribution function  $f$ , and has the following fundamental properties

- of conserving mass momentum and energy:

$$\langle Q(f, f) \phi(v) \rangle = \mathbf{0} \in \mathbb{R}^{d_v+2}, \quad (1.2)$$

---

*Keywords and phrases.* Stiff kinetic equations, BGK/ES-BGK models, IMEX Runge–Kutta methods, compressible Euler equations, compressible Navier–Stokes equations.

<sup>1</sup> Department of Mathematics and Computer Science, University of Catania, 95125 Catania, Italy.

<sup>2</sup> Department of Mathematics, Gyeongsang National University, 52828 Jinju, Republic of Korea.

\*Corresponding author: [chosy89@gnu.ac.kr](mailto:chosy89@gnu.ac.kr)

for  $\phi(v) = (1, v, \frac{1}{2}|v|^2)^\top$ , and  $\langle g \rangle := \int_{\mathbb{R}^{d_v}} g(v) dv$ ,

– to satisfy  $H$ -theorem:

$$\int_{\mathbb{R}^{d_v}} Q(f, f) \ln f dv \leq 0, \quad (1.3)$$

– to vanish, *i.e.*,  $Q(f, f) = 0$  when  $f$  is the local Maxwellian:

$$\mathcal{M}[f](t, x, v) = \frac{\rho(t, x)}{(2\pi T(t, x))^{d_v/2}} \exp\left(-\frac{|v - u(t, x)|^2}{2T(t, x)}\right), \quad (1.4)$$

where  $\rho$ ,  $u = (u_1, \dots, u_{d_v})$ , and  $E$  are density, mean velocity and energy associated to  $f$ :

$$\langle f\phi(v) \rangle = (\rho, \rho u, E)^\top = U \in \mathbb{R}^{d_v+2}, \quad (1.5)$$

and temperature  $T$  is given by  $\frac{d_v \rho T}{2} = E - \frac{1}{2}\rho|u|^2$ . We introduce the vector  $U$  for later use.

When the Knudsen number is small, it is well known that BTE is closely related to fluid models such as compressible Euler or compressible Navier–Stokes (CNS) equations. The form of these fluid models associated to BTE are traditionally derived using the perturbation techniques like the Hilbert or Chapman–Enskog expansions [13, 14, 19]. Thus, BTE can be used for various Knudsen number, *i.e.*, from rarefied to continuum gas dynamics.

In spite of its good predictability and close relationship with fluid models, computation of the Boltzmann collision operator is very expensive due to its high dimensionality. Furthermore, the problem becomes more severe as the Knudsen number gets closer to zero (fluid regime). In this case, solving BTE by a standard explicit numerical scheme requires the use of a time step of the order of  $\varepsilon$ , which leads to very expensive numerical computations. Even if one adopts an implicit or semi-implicit time discretization for the collision part, it is still numerically challenging to construct an efficient implicit solver due to the complicate structure of the Boltzmann collision operator.

To circumvent the issue on computational cost of the Boltzmann collision operator  $Q(f, f)$  in (1.1), simpler kinetic models have been proposed to mimic the main properties of the full integral operator  $Q(f, f)$ . One such model is the BGK model [4]:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{\tau}{\varepsilon} (\mathcal{M}[f] - f), \quad (1.6)$$

where  $\tau$  is the collision frequency that depends on  $\rho$  and  $T$ . The BGK model replaces the Boltzmann collision operator with a simple relaxation toward the local Maxwellian. Note that the BGK model still maintains the conservation of mass, momentum, and energy, as well as the entropy inequality [13]. In addition, the BGK model describes the correct fluid limit as  $\varepsilon \rightarrow 0$ , *i.e.*, at the *leading order term*  $\varepsilon = 0$ , it yields the compressible Euler equations (see [13, 19]). Unfortunately, at the first-order correction in  $\varepsilon$ , the transport coefficients obtained at the Navier–Stokes level are not satisfactory. In particular, the Prandtl number defined by  $\text{Pr} = \frac{\gamma}{\gamma-1} \frac{\mu}{\kappa}$ , which relates the viscosity  $\mu$  to the heat conductivity  $\kappa$  of gases is fixed by 1. Here, the polytropic constant  $\gamma$  for monatomic molecule with translational motions is given by  $\gamma = \frac{d_v+2}{d_v}$ . Note that for most realistic gases, we have  $\text{Pr} < 1$ . In addition, in the hard-sphere model for monoatomic gases ( $\gamma = 5/3$ ) in Boltzmann equation, its Prandtl number is very close to  $2/3$ .

Many variants of the BGK model have been proposed in order to give the correct transport coefficients at the Navier–Stokes level. In [20], Holway proposed a model that possesses several desirable properties. It not only satisfies the correct conservation laws and the entropy condition, but also yields the Navier–Stokes equations with a Prandtl number less than one through the Chapman–Enskog expansion. This model is known as the ellipsoidal statistical model (ES-BGK) and reads:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{\tau}{\varepsilon} (\mathcal{G}[f] - f), \quad (1.7)$$

where the collision frequency  $\tau = \frac{\rho T}{\mu(1-\nu)}$  depends on the free parameter  $-\frac{1}{2} \leq \nu < 1$ . The anisotropic Gaussian distribution  $\mathcal{G}[f]$  is defined by

$$\mathcal{G}[f](t, x, v) = \frac{\rho(t, x)}{\sqrt{\det(2\pi\mathcal{T}(t, x))}} \exp\left(-\frac{(v - u(t, x))^\top \mathcal{T}(t, x)^{-1} (v - u(t, x))}{2}\right). \quad (1.8)$$

The temperature tensor  $\mathcal{T}(t, x)$  and stress tensor  $\Theta(t, x)$  are defined as

$$\mathcal{T}(t, x) = (1 - \nu)T(t, x)Id + \nu\Theta(t, x), \quad (1.9)$$

$$\Theta(t, x) = \frac{1}{\rho} \int_{\mathbb{R}^{d_v}} (v - u) \otimes (v - u) f(t, x, v) dv = \frac{1}{\rho} \langle (v - u) \otimes (v - u) f(v) \rangle, \quad (1.10)$$

respectively, where  $Id$  is the  $d_v \times d_v$  identity matrix and  $(v - u) \otimes (v - u)$  denotes the tensor product<sup>1</sup> of two vectors  $(v - u)$  and  $(v - u)$ .

As emphasized before, ES-BGK model also has a close relationship with fluid models. The Chapman–Enskog expansion applied to ES-BGK model gives the compressible Euler equations [16] at the *leading order term*:

$$\partial_t \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix} + \nabla_x \cdot \begin{pmatrix} \rho u \\ \rho u \otimes u + pId \\ (E + p)u \end{pmatrix} = 0, \quad (1.11)$$

where  $p = \rho T$ . While at the first-order correction in  $\varepsilon$ , we obtain the CNS equations [13] (for details see Appendix A):

$$\partial_t \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix} + \nabla_x \cdot \begin{pmatrix} \rho u \\ \rho u \otimes u + pId \\ (E + p)u \end{pmatrix} = \begin{pmatrix} 0 \\ \varepsilon \nabla_x \cdot (\mu \sigma(u)) \\ \varepsilon \nabla_x \cdot (\mu \sigma(u)u + q) \end{pmatrix} \quad (1.12)$$

where the stress tensor  $\sigma(u)$  and heat flux  $q$  are given by

$$\sigma(u) = \nabla_x u + (\nabla_x u)^\top - \frac{2}{d_v} \nabla_x \cdot u Id, \quad q = \kappa \nabla_x T,$$

with viscosity  $\mu = \frac{p}{(1-\nu)\tau}$  and thermal conductivity  $\kappa = \frac{d_v + 2}{2} \frac{p}{\tau}$ . Thus, the Prandtl number is given by

$$\frac{2}{3} \leq Pr = \frac{d_v + 2}{2} \frac{\mu}{\kappa} = \frac{1}{1 - \nu} < \infty,$$

and this implies that desired such number can be recovered by choosing  $\nu$  appropriately. Note that the equation for the total energy can be replaced by the equation for the temperature  $T$  [2] as follows:

$$\frac{d_v}{2} \rho (\partial_t T + u \cdot \nabla_x T) + \rho T \nabla_x \cdot u = \mathcal{O}(\varepsilon). \quad (1.13)$$

Considering the relationship with fluid models, when developing numerical methods for ES-BGK model, such methods should have a correct asymptotic behavior, *i.e.*, for small parameter  $\varepsilon$ , the schemes should degenerate into a good approximation of the fluid asymptotic (compressible Euler or CNS equations).

In the fluid dynamic limit, *i.e.*, in the case of the limiting compressible Euler equations, numerical methods that work effective while keeping the mesh size and time step fixed as the Knudsen number approaches zero, are referred to as asymptotic preserving (AP) [22]. Additionally, a scheme that not only preserves the correct

<sup>1</sup>The tensor product of two vectors  $a = (a_1, \dots, a_d), b = (b_1, \dots, b_d) \in \mathbb{R}^d$  means the outer product of  $u$  and  $v$ , *i.e.*,  $a \otimes b \in \mathbb{R}^{d \times d}$  with  $(a \otimes b)_{ij} = a_i b_j$ .

asymptotic behavior but also maintains high accuracy in the fluid dynamic regime is called asymptotically accurate (AA) [12]. In other words, a scheme that satisfies both the AP and AA properties exhibits robustness and high accuracy in the fluid dynamic regime, without resolving the small Knudsen number.

On the construction of AP and AA methods for kinetic models, IMEX time discretization [1, 5, 6, 8, 23, 24] have been successfully applied and studied. We focus our literature reviews to the works on the NS limit based on IMEX time discretization applied to BGK or ES-BGK model. There is, of course, a considerable amount of literature on the use of IMEX-RK methods for BGK-type equations, (see, *e.g.*, [12, 16, 17]). For instance, in [3, 26], the authors considered a micro-macro decomposition of the BGK equation and then applied IMEX-RK schemes to the resulting coupled system. In [18], authors introduced a first order IMEX method for ES-BGK model and showed that it is consistent to first order time discretization of CNS equations. High order IMEX-RK methods of type CK (or type II) [21] and IMEX multistep methods [17] are also similarly applied to ES-BGK model, and the methods are shown to be able to capture the NS limit under suitable conditions. However, reference [21] focuses solely on a specific subclass of IMEX schemes (namely, type CK or type II methods [12]).

This paper provides a broader and more unified analysis of the various IMEX-RK schemes, specifically type I and type II, extending and improving upon existing results in the literature for the ES-BGK model. In particular, we study their asymptotic behavior and generalize the results presented in [21], showing that both types of schemes are capable of capturing the compressible Navier–Stokes (CNS) limit without resolving the small parameter  $\varepsilon$ .

Another novelty of the present work is that we show numerically the uniform accuracy of specific type of IMEX-RK methods, originally introduced in [7] when applied to ES-BGK model for a wide range of Knudsen numbers. These IMEX-RK schemes were originally developed and analyzed for a specific class of hyperbolic relaxation systems. In the context of the ES-BGK model, we show that, under suitable assumptions on the coefficients of such schemes introduced in [7], consistency with CNS equations is ensured (see Thm. 3.8). Moreover, additionally order conditions allow the schemes to maintain the order of accuracy throughout the entire range of Knudsen numbers, thereby guaranteeing uniform accuracy. These conditions are sufficient but not necessary. However, the price to pay is that the IMEX-RK schemes proposed in [7] require more stages than those commonly used.

The outline of this paper is the following. In Section 2, we explain IMEX-RK methods for ES-BGK model. In Section 3, we perform asymptotic analysis at the Navier–Stokes level. Then in Section 4, we give several numerical tests in order to validate our theoretical findings. Finally, conclusion will be provided.

## 2. IMEX-RK SCHEMES FOR ES-BGK EQUATION

For the time discretization of (1.7) we consider an implicit-explicit (IMEX) Runge–Kutta (RK) scheme because the convection term in (1.7) is not stiff and the collision term is stiff when Knudsen number is small.

An IMEX-RK scheme usually can be represented by the double Butcher tables:

$$\begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline & \tilde{b}^\top \end{array}, \quad \begin{array}{c|c} c & A \\ \hline & b^\top \end{array}. \quad (2.1)$$

Here  $\tilde{A} = (\tilde{a}_{ij})$  with  $\tilde{a}_{ij} = 0$  for  $j \geq i$  and  $A = (a_{ij})$  with  $a_{ij} = 0$  for  $j > i$  are  $s \times s$  matrices, which are associated to the explicit and implicit time discretizations, respectively. The coefficients vectors  $\tilde{b} = (\tilde{b}_1, \dots, \tilde{b}_s)^\top$  and  $b = (b_1, \dots, b_s)^\top$  represent the weights, and the vectors  $\tilde{c} = (\tilde{c}_1, \dots, \tilde{c}_s)^\top$  and  $c = (c_1, \dots, c_s)^\top$  are the nodes defined as

$$\tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij}, \quad c_i = \sum_{j=1}^i a_{ij}, \quad i = 1, \dots, s. \quad (2.2)$$

Now we give some preliminary definitions. Based on the structure of matrix  $A$  in the implicit table, the IMEX schemes can be classified into the following types [5, 8, 12]:

**Definition 2.1.** An IMEX-RK method is of **type I** if the matrix  $A$  is invertible.

**Definition 2.2.** An IMEX-RK method is of **type II** if the matrix  $A$  can be written as

$$A = \begin{pmatrix} 0 & 0 \\ a & \hat{A} \end{pmatrix}, \quad (2.3)$$

where  $a = (a_{21}, \dots, a_{s1})^\top \in \mathbb{R}^{s-1}$  and the submatrix  $\hat{A} \in \mathbb{R}^{(s-1) \times (s-1)}$  is invertible [23]; in particular if  $a = \mathbf{0} \in \mathbb{R}^{s-1}$ ,  $b_1 = 0$ , the scheme is said of type ARS [1].

We note that IMEX-RK schemes of type II are very attractive because they allow some simplifying assumptions that make order conditions easier to treat, therefore permitting the construction of higher order IMEX-RK schemes [6, 8]. On the other hand, schemes of type I are more amenable to a theoretical analysis since the matrix  $A$  of the implicit scheme is invertible. Later, we start our analysis with the latter scheme, type I.

In the following, we will also make use of the following representation of the matrix  $\tilde{A}$  in the explicit part

$$\tilde{A} = \begin{pmatrix} 0 & 0 \\ \tilde{a} & \hat{A} \end{pmatrix}, \quad (2.4)$$

where  $\tilde{a} = (\tilde{a}_{21}, \dots, \tilde{a}_{s1})^\top \in \mathbb{R}^{s-1}$  and  $\hat{A} \in \mathbb{R}^{(s-1) \times (s-1)}$ . This representation of matrix  $\tilde{A}$  is useful for the analysis of IMEX-RK methods of type II.

Finally, we give the following definition [9, 10]:

**Definition 2.3.** If  $b_i = a_{si}$  for  $i = 1, \dots, s$ , the scheme is said to be **stiffly accurate** (SA) in the implicit tableau. Moreover, if  $\tilde{b}_i = \tilde{a}_{si}$  for  $i = 1, \dots, s$ , the scheme is said to be **globally stiffly accurate** (GSA).

The first condition in Definition 2.3 guarantees that an  $A$ -stable implicit tableau is  $L$ -stable<sup>2</sup>. The AP property of IMEX-RK schemes are strongly related to the  $L$ -stability of the implicit part of the scheme. Finally, if the IMEX-RK scheme is GSA we have  $f^{n+1} = f^{(s)}$ , *i.e.*, the numerical solution coincides with the last stage of the method.

Below, we review IMEX-RK methods of type I and type II applied to ES-BGK model (1.7).

– **IMEX-RK method of type I.** Applying an IMEX-RK method of type I to (1.7), we get in vector form

$$\begin{aligned} \mathbf{F} &= f^n \mathbf{e} - \Delta t \tilde{A} L(\mathbf{F}) + \frac{\Delta t}{\varepsilon} A \bar{\tau} (\mathcal{G}[\mathbf{F}] - \mathbf{F}), \\ f^{n+1} &= f^n - \Delta t \tilde{b}^\top L(\mathbf{F}) + \frac{\Delta t}{\varepsilon} b^\top \bar{\tau} (\mathcal{G}[\mathbf{F}] - \mathbf{F}), \end{aligned} \quad (2.5)$$

where  $\mathbf{F} = (f^{(1)}, f^{(2)}, \dots, f^{(s)})^\top \in \mathbb{R}^s$ ,  $L(\mathbf{F}) = (L(f^{(1)}), \dots, L(f^{(s)}))^\top \in \mathbb{R}^s$ , being  $L(f^{(k)}) = v \cdot \nabla_x f^{(k)}$  for all  $k = 1, \dots, s$ ,  $\mathcal{G}[\mathbf{F}] := (\mathcal{G}(f^{(1)}), \dots, \mathcal{G}(f^{(s)}))^\top \in \mathbb{R}^s$ . A diagonal matrix  $\bar{\tau} := \text{diag}(\tau^{(1)}, \dots, \tau^{(s)})$  is defined with the relaxation time  $\tau^{(i)}$  associated to  $f^{(i)}$  and  $\mathbf{e} = (1, 1, \dots, 1)^\top$  is a vector of length  $s$ .

From the first equation in (2.5), we get

$$\Delta t \bar{\tau} (\mathcal{G}[\mathbf{F}] - \mathbf{F}) = \varepsilon A^{-1} (\mathbf{F} - f^n \mathbf{e} + \Delta t \tilde{A} L(\mathbf{F})) \quad (2.6)$$

and inserting in the numerical solution, we obtain

$$f^{n+1} = (1 - b^\top A^{-1} \mathbf{e}) f^n + b^\top A^{-1} \mathbf{F} - \Delta t (\tilde{b}^\top - b^\top A^{-1} \tilde{A}) L(\mathbf{F}). \quad (2.7)$$

Therefore, for IMEX-RK schemes of type I the numerical solution is independent on  $\varepsilon$  and we are able to pass to the limit  $\varepsilon \rightarrow 0$  in (2.5).

<sup>2</sup>A Runge-Kutta method is  $L$ -stable if  $\lim_{z \rightarrow \infty} R(z) = 0$ , where  $R(z)$  is the stability function of the method. If the  $A$ -stable RK-scheme is SA, *i.e.*,  $b^\top A^{-1} \mathbf{e} = 1$ , it follows that  $R(\infty) = \lim_{z \rightarrow \infty} R(z) = 1 - b^\top A^{-1} \mathbf{e} = 0$  [25]. Here  $\mathbf{e} = (1, 1, \dots, 1)^\top$  is a vector of length  $s$ .

– **IMEX-RK method of type II.** Regarding the IMEX-RK method of type II applied to (1.7), with the notations in (2.3) and (2.4), the scheme reads in vector form

$$\begin{aligned} f^{(1)} &= f^n, \\ \hat{\mathbf{F}} &= f^n \hat{\mathbf{e}} - \Delta t \tilde{a} L(f^n) - \Delta t \hat{A} L(\hat{\mathbf{F}}) + \Delta t a \frac{\tau^n}{\varepsilon} (\mathcal{G}[f^n] - f^n) + \Delta t \hat{A} \frac{\hat{\tau}}{\varepsilon} (\mathcal{G}[\hat{\mathbf{F}}] - \hat{\mathbf{F}}), \\ f^{n+1} &= f^n - \Delta t \tilde{b}_1 L(f^n) - \Delta t \hat{b}^\top L(\hat{\mathbf{F}}) + \Delta t b_1 \frac{\tau^n}{\varepsilon} (\mathcal{G}[f^n] - f^n) + \Delta t \hat{b}^\top \frac{\hat{\tau}}{\varepsilon} (\mathcal{G}[\hat{\mathbf{F}}] - \hat{\mathbf{F}}), \end{aligned} \tag{2.8}$$

where  $\hat{\mathbf{F}} = (f^{(2)}, \dots, f^{(s)})^\top \in \mathbb{R}^{s-1}$ ,  $L(\hat{\mathbf{F}}) = (L(f^{(2)}), \dots, L(f^{(s)}))^\top \in \mathbb{R}^{s-1}$ ,  $\hat{\mathbf{e}} = (1, 1, \dots, 1)^\top \in \mathbb{R}^{s-1}$ ,  $\hat{\mathbf{b}} = (\tilde{b}_2, \tilde{b}_3, \dots, \tilde{b}_s)^\top$ ,  $\tau^n = \tau^{(1)}$ ,  $\hat{\tau} := \text{diag}(\tau^{(2)}, \dots, \tau^{(s)})$  and  $\mathcal{G}[\hat{\mathbf{F}}] := (\mathcal{G}(f^{(2)}), \dots, \mathcal{G}(f^{(s)}))^\top \in \mathbb{R}^{s-1}$ . From the second equation in (2.8) we get

$$\Delta t \hat{\tau} (\mathcal{G}[\hat{\mathbf{F}}] - \hat{\mathbf{F}}) = \varepsilon \hat{A}^{-1} [\hat{\mathbf{F}} - f^n \hat{\mathbf{e}} + \Delta t \tilde{a} L(f^n) + \Delta t \hat{A} L(\hat{\mathbf{F}})] - \Delta t \hat{A}^{-1} a \tau^n (\mathcal{G}[f^n] - f^n), \tag{2.9}$$

and substituting into the numerical solution in (2.8), we obtain

$$\begin{aligned} f^{n+1} &= \left(1 - \hat{b}^\top \hat{A}^{-1} \hat{\mathbf{e}}\right) f^n - \Delta t \left(\tilde{b}_1 - \hat{b}^\top \hat{A}^{-1} \tilde{a}\right) L(f^n) - \Delta t \left(\hat{b}^\top - \hat{b}^\top \hat{A}^{-1} \hat{A}\right) L(\hat{\mathbf{F}}) \\ &\quad + \Delta t \left(b_1 - \hat{b}^\top \hat{A}^{-1} a\right) \frac{\tau^n}{\varepsilon} (\mathcal{G}[f^n] - f^n) + \hat{b}^\top \hat{A}^{-1} \hat{\mathbf{F}}. \end{aligned} \tag{2.10}$$

Unfortunately, here the numerical solution depends on  $\varepsilon$ . Now, in order to be able to pass to the limit  $\varepsilon \rightarrow 0$ , we can either require that the initial conditions are well-prepared<sup>3</sup> [12], or impose that the additional condition  $b_1 - \hat{b}^\top \hat{A}^{-1} a = 0$  has to be satisfied, which allows us to take the limit  $\varepsilon \rightarrow 0$  in (2.8). Note that the condition  $b_1 - \hat{b}^\top \hat{A}^{-1} a = 0$ , is automatically satisfied if the scheme is SA, *i.e.*,  $\hat{b}^\top \hat{A}^{-1} = \hat{\mathbf{e}}_{s-1}^\top$ , with  $\hat{\mathbf{e}}_{s-1}^\top = (0, \dots, 0, 1)$  vector of length  $s - 1$ . Alternatively, this condition also holds for IMEX schemes of type ARS, where having  $a = 0$  and  $b_1 = 0$  condition  $b_1 - \hat{b}^\top \hat{A}^{-1} a = 0$  is also satisfied.

In both types, the methods require implicit solver for computing relaxation terms. However, the use of collision invariants allows us to treat the implicit terms explicitly. The details will be provided in Section 3.

### 3. ASYMPTOTIC PROPERTIES OF THE IMEX RK SCHEMES

In this section, we discuss in detail the asymptotic properties of the IMEX-RK schemes of type I and II with respect to the Euler and Navier–Stokes limits. We begin with a brief overview and proof of the results related to the Euler limit, highlighting the preservation of leading-order asymptotic. More details can be found in [12, 16, 18, 21]. Next, we will analyze the asymptotic properties of IMEX-RK schemes of type I and II in the Navier–Stokes limit.

#### 3.1. Preserving the Euler limit

We start with an arbitrary IMEX-RK scheme written for each stage  $k$ :

$$f^{(k)} = f^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} v \cdot \nabla_x f^{(\ell)} + \frac{\Delta t}{\varepsilon} \sum_{\ell=1}^k a_{k\ell} \tau^{(\ell)} (\mathcal{G}[f^{(\ell)}] - f^{(\ell)}), \quad k = 1, \dots, s \tag{3.1}$$

with numerical solution

$$f^{n+1} = f^n - \Delta t \sum_{k=1}^s \tilde{b}_k v \cdot \nabla_x f^{(k)} + \frac{\Delta t}{\varepsilon} \sum_{k=1}^s b_k \tau^{(k)} (\mathcal{G}[f^{(k)}] - f^{(k)}). \tag{3.2}$$

<sup>3</sup>The initial data for equation (1.1) are said well-prepared if  $f_0(x, v) = \mathcal{M}[f_0](x, y) + g_\varepsilon(x, v)$ ,  $\lim_{\varepsilon \rightarrow 0} g_\varepsilon(x, v) = 0$ .

At every  $k$ th stage of (3.1), the computation of  $f^{(k)}$  requires implicit treatment of  $\tau^{(k)}$  and  $\mathcal{G}[f^{(k)}]$ . This difficulty can be circumvented by approximating  $\tau^{(k)}$  and  $\mathcal{G}[f^{(k)}]$  explicitly. For this, we rewrite (3.1) as

$$f^{(k)} = \frac{\varepsilon}{\varepsilon + \tau^{(k)}\Delta t a_{kk}} f_*^{(k)} + \frac{\Delta t \tau^{(k)} a_{kk}}{\varepsilon + \tau^{(k)}\Delta t a_{kk}} \mathcal{G}[f^{(k)}] \quad (3.3)$$

with

$$f_*^{(k)} = f^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} v \cdot \nabla_x f^{(\ell)} + \frac{\Delta t}{\varepsilon} \sum_{\ell=1}^{k-1} a_{k\ell} \tau^{(\ell)} \left( \mathcal{G}[f^{(\ell)}] - f^{(\ell)} \right).$$

Taking the moments  $\langle \cdot, \phi \rangle := \int_{\mathbb{R}^{d_v}} \cdot \phi(v) dv$  with  $\phi(v) := (1, v, |v|^2/2)^\top$  on both sides of (3.3), and using the conservation properties (A.1), the implicit part is canceled, *i.e.*,

$$\left\langle \left( \mathcal{G}[f^{(k)}] - f^{(k)} \right) \phi \right\rangle = 0.$$

Therefore, one obtains the macroscopic quantities  $U := (\rho, \rho u, E)$  at every stage  $k$ :

$$U^{(k)} = \int_{\mathbb{R}^{d_v}} \left( f^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} v \cdot \nabla_x f^{(\ell)} \right) \phi(v) dv = \left\langle f^n \phi \right\rangle - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} \left\langle v \cdot \nabla_x f^{(\ell)} \phi \right\rangle, \quad (3.4)$$

and numerical solution

$$U^{n+1} = \left\langle f^n \phi \right\rangle - \Delta t \sum_{k=1}^s \tilde{b}_k \left\langle v \cdot \nabla_x f^{(k)} \phi \right\rangle. \quad (3.5)$$

Note that using (3.4) we can obtain  $\rho^{(k)}$ ,  $u^{(k)}$  and  $T^{(k)}$  at every stage  $k$ , which enables us to compute explicitly  $\tau^{(k)}$ . To evaluate  $\mathcal{G}[f^{(k)}]$ , however, we need to determine the stress tensor  $\Theta^{(k)}$  in (1.10). To achieve this, we first define the tensor  $\Sigma$  as follows

$$\Sigma = \int_{\mathbb{R}^{d_v}} v \otimes v f dv = \rho(\Theta + u \otimes u), \quad (3.6)$$

and consider the stage values:

$$\Sigma^{(k)} = \int_{\mathbb{R}^{d_v}} v \otimes v f^{(k)} dv = \rho^{(k)} \left( \Theta^{(k)} + u^{(k)} \otimes u^{(k)} \right). \quad (3.7)$$

Note that  $\Theta^{(k)}$  can be obtained by computing  $\Sigma^{(k)}$ . Next, we use (1.9) and (3.6) to get

$$\rho \mathcal{T} = \rho[(1 - \nu)TId + \nu\Theta] = \rho(1 - \nu)TId + \nu\Sigma - \nu\rho u \otimes u,$$

and combine this with

$$\int_{\mathbb{R}^{d_v}} v \otimes v \mathcal{G}[f](v) dv = \rho(\mathcal{T} + u \otimes u),$$

to obtain

$$\int_{\mathbb{R}^{d_v}} v \otimes v (\mathcal{G}[f] - f) dv = (1 - \nu)(\rho(TId + u \otimes u) - \Sigma). \quad (3.8)$$

Now, we multiply the scheme (3.1) by  $v \otimes v$  and integrate it over  $v$ , and use the relation (3.8) to derive the equation for  $\Sigma^{(k)}$ :

$$\Sigma^{(k)} = \frac{\varepsilon}{\varepsilon + (1 - \nu)\tau^{(k)}\Delta t a_{kk}} \Sigma^* + \frac{\Delta t(1 - \nu)\tau^{(k)} a_{kk}}{\varepsilon + (1 - \nu)\tau^{(k)}\Delta t a_{kk}} \rho^{(k)} \left( T^{(k)} Id + u^{(k)} \otimes u^{(k)} \right), \quad (3.9)$$

where

$$\Sigma^* = \Sigma^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k,\ell} \nabla_x \cdot \left\langle v \otimes v v f^{(\ell)} \right\rangle + \frac{\Delta t}{\varepsilon} \sum_{\ell=1}^{k-1} a_{k\ell} (1 - \nu) \tau^{(\ell)} \left[ \rho^{(\ell)} \left( T^{(\ell)} Id + u^{(\ell)} \otimes u^{(\ell)} \right) - \Sigma^{(\ell)} \right].$$

Then, the IMEX-RK scheme reads

$$\begin{aligned} \Sigma^{(k)} &= \frac{\varepsilon}{\varepsilon + (1 - \nu) \tau^{(k)} \Delta t a_{kk}} \Sigma^* + \frac{\Delta t (1 - \nu) \tau^{(k)} a_{kk}}{\varepsilon + (1 - \nu) \tau^{(k)} \Delta t a_{kk}} \rho^{(k)} \left( T^{(k)} Id + u^{(k)} \otimes u^{(k)} \right), \\ f^{(k)} &= \frac{\varepsilon}{\varepsilon + \tau^{(k)} \Delta t a_{kk}} f_*^{(k)} + \frac{\Delta t \tau^{(k)} a_{kk}}{\varepsilon + \tau^{(k)} \Delta t a_{kk}} \mathcal{G} \left[ f^{(k)} \right], \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} \Sigma^{n+1} &= \Sigma^n - \Delta t \sum_{k=1}^s \tilde{b}_k \nabla_x \cdot \left\langle v \otimes v v f^{(k)} \right\rangle + \frac{\Delta t}{\varepsilon} \sum_{k=1}^s b_k \left\langle v \otimes v \left( \tau^{(k)} \left( \mathcal{G} \left[ f^{(k)} \right] - f^{(k)} \right) \right) \right\rangle, \\ f^{n+1} &= f^n - \Delta t \sum_{k=1}^s \tilde{b}_k v \cdot \nabla_x f^{(k)} + \frac{\Delta t}{\varepsilon} \sum_{k=1}^s b_k \tau^{(k)} \left( \mathcal{G} \left[ f^{(k)} \right] - f^{(k)} \right). \end{aligned} \tag{3.11}$$

Regarding the leading order limit, *i.e.*, Euler equation, the following results for IMEX-RK schemes of type I and II have been proved in [12, 21].

**Proposition 3.1.** *Consider the IMEX-RK method (3.1) and (3.2) of type I. Then in the limit  $\varepsilon \rightarrow 0$ , for a fixed  $\Delta t$ , the scheme becomes the explicit RK scheme characterized by the pair  $(\tilde{A}, \tilde{b})$  applied to the limit Euler equations (1.11).*

**Corollary 3.2.** *Furthermore, if the scheme is GSA, then*

$$\lim_{\varepsilon \rightarrow 0} f^{n+1} = \lim_{\varepsilon \rightarrow 0} \mathcal{M} \left[ f^{n+1} \right]. \tag{3.12}$$

**Remark 3.3.** According to Proposition 3.1, the limit scheme is both AP and AA. In other words, as  $\varepsilon \rightarrow 0$ , the scheme remains stable and accurate, precisely matching the explicit tableau of the IMEX-RK method.

Corollary 3.2 claims that an important property of the IMEX schemes of type I is obtained if in the limit  $\varepsilon \rightarrow 0$  the distribution function is projected over the equilibrium, *i.e.*,  $f^{n+1} = \mathcal{M}(f^{n+1})$ . From (2.7) it is clear that this property is achieved if the scheme is GSA *i.e.*,  $b^\top A^{-1} = \mathbf{e}_s^\top$ , with  $\mathbf{e}_s^\top = (0, \dots, 0, 1)$  vector of length  $s$ , and  $b^\top A^{-1} \tilde{A} = \tilde{b}^\top$  [12, 16].

Note that GSA property is not essential for type I schemes to result in AP and AA for limit Euler equations, (see [5, 6, 8, 12]).

**Proposition 3.4.** *Consider a GSA IMEX-RK method (3.1) and (3.2) of type II, then in the limit  $\varepsilon \rightarrow 0$ , for fixed  $\Delta t$  and well-prepared initial data, the scheme becomes the explicit RK scheme characterized by the pair  $(\tilde{A}, \tilde{b})$  applied to the limit Euler equations (1.11).*

**Remark 3.5.** Since the scheme is GSA, it ensures that the initial value remains consistent at the next time step, *i.e.*,  $f^{n+1} = \mathcal{M}[f^{n+1}]$ . Without the assumption of GSA and the consistency of the initial data in Proposition 3.4, as discussed for the type II case in Section 2, IMEX-RK schemes of type ARS ( $a = 0$  and  $b_1 = 0$ ), remain asymptotic-preserving (AP) without requiring additional conditions (in (2.23) the quantity  $(b_1 - \hat{b}^\top A^{-1} a) = 0$ ), although they are not necessarily asymptotically accurate (AA). However, imposing the additional condition  $\tilde{b}_1 = 0$  ensures asymptotic accuracy (see for more details [12]).

Usually, to guarantee that an IMEX-RK scheme of type I or II remains robust and stable for any value of  $\varepsilon$ , that is, performs correctly across both the rarefied and fluid regimes, assuming GSA condition may achieve this task, particularly in the fluid regime, without additional order conditions [7, 12]. However, as noted in Remark 3.5, we can consider Type ARS schemes that are not GSA, but characterized by some assumptions on the coefficients of the scheme in order to guarantee both AP and AA properties.

We conclude this section by considering an explicit RK scheme characterized by  $(\tilde{A}, \tilde{b})$  applied to the limit compressible Euler equations (1.11) with  $T = p/\rho$ , which takes the form:

$$\begin{aligned}\rho^{(k)} &= \rho^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} \nabla_x \cdot \left( \rho^{(\ell)} u^{(\ell)} \right), \\ (\rho u)^{(k)} &= (\rho u)^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} \nabla_x \cdot \left( (\rho u)^{(\ell)} \otimes u^{(\ell)} + \rho^{(\ell)} T^{(\ell)} Id \right), \\ E^{(k)} &= E^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} \nabla_x \cdot \left( \left( E^{(\ell)} + \rho^{(\ell)} T^{(\ell)} \right) u^{(\ell)} \right),\end{aligned}\tag{3.13}$$

for  $k = 1, \dots, s$ , with numerical solution

$$\begin{aligned}\rho^{n+1} &= \rho^n - \Delta t \sum_{k=1}^s \tilde{b}_k \nabla_x \cdot \left( \rho^{(k)} u^{(k)} \right), \\ (\rho u)^{n+1} &= (\rho u)^n - \Delta t \sum_{k=1}^s \tilde{b}_k \nabla_x \cdot \left( (\rho u)^{(k)} \otimes u^{(k)} + \rho^{(k)} T^{(k)} Id \right), \\ E^{n+1} &= E^n - \Delta t \sum_{k=1}^s \tilde{b}_k \nabla_x \cdot \left( \left( E^{(k)} + \rho^{(k)} T^{(k)} \right) u^{(k)} \right).\end{aligned}\tag{3.14}$$

Note that in (3.13) and (3.14), to the leading-order, we can exchange the energy equation with the temperature one (1.13), so we get

$$T^{(k)} = T^n - \Delta t \sum_{\ell=1}^{k-1} \tilde{a}_{k\ell} \left( u^{(\ell)} \cdot \nabla_x T^{(\ell)} + \frac{2}{d_v} T^{(\ell)} \nabla_x \cdot u^{(\ell)} \right),\tag{3.15}$$

$$T^{n+1} = T^n - \Delta t \sum_{k=1}^s \tilde{b}_k \left( u^{(k)} \cdot \nabla_x T^{(k)} + \frac{2}{d_v} T^{(k)} \nabla_x \cdot u^{(k)} \right).\tag{3.16}$$

### 3.2. Preserving the Navier–Stokes limit

In this section, we analyze the asymptotic behavior of IMEX-RK schemes of type I and II for ES-BGK equations (1.7), and prove that for small value of  $\varepsilon$ , these schemes asymptotically capture the NS limit without resolving  $\varepsilon$ , providing a numerical scheme for the corresponding CNS equations (1.12). The two main Theorems below not only extend the asymptotic analysis in [21] by considering general IMEX RK methods, but also result in explicit-type RK schemes for CNS equations which are consistent with the results of Boscarino and Pareschi [7].

At the Navier–Stokes level, (for small but non-negligible Knudsen numbers,  $\varepsilon \ll 1$ ), the GSA condition is not crucial to ensure consistency with CNS equations for both types of IMEX-RK schemes. Therefore, for the subsequent analysis, we do not impose the GSA condition on the scheme.

- **IMEX-RK scheme of type I.** We first analyze the IMEX-RK method of type I. The following shows that the result in [7] can be generalized to ES-BGK model.

**Theorem 3.6.** *For small values of  $\varepsilon$  and with  $\Delta t^p + \varepsilon \Delta t + \frac{\varepsilon^2}{\Delta t} = o(\varepsilon)$ , the IMEX-RK of type I (2.5) asymptotically becomes a consistent macroscopic explicit-type RK scheme of order  $p$  characterized by the pair  $(\tilde{A}, \tilde{b})$  and  $(B, \omega)$  for the CNS equations (1.12) with*

$$B = \tilde{A}A^{-1}\tilde{A}, \quad \omega^\top = \tilde{b}^\top A^{-1}\tilde{A}.$$

Note that the pair  $(\tilde{A}, \tilde{b})$  is the explicit table of the IMEX-RK method of type I in (2.1), and  $A$  corresponds to the implicit part. Moreover, in the theorem we assume that the explicit-type RK method is of order  $p$ . We remark that the order  $p$  could be ensured by enforcing additional order conditions derived from the associated two pair of Butcher tableaux. Such conditions are derived in [7] up to order  $p = 3$ .

*Proof.* We start by considering the vector notation of the IMEX-RK scheme (2.5). By the first order discrete Chapman–Enskog expansion in  $\varepsilon$  of  $f^n$  and  $\mathbf{F}$ , we get

$$f^n = \mathcal{M}[f^n] + \varepsilon f_1^n, \quad \mathbf{F} = \mathcal{M}[\mathbf{F}] + \varepsilon \mathbf{f}_1, \tag{3.17}$$

where the vector function  $\mathbf{f}_1$  satisfies the so-called compatibility conditions  $\langle \phi \mathbf{f}_1 \rangle = 0$  in (A.5). Inserting this expansions (3.17) into (2.5), then multiplying by  $\phi(v)$  function and integrating on  $v$ , we get

$$\langle \phi \mathcal{M}[\mathbf{F}] \rangle = \langle \phi \mathcal{M}[f^n] \mathbf{e} \rangle - \Delta t \tilde{\mathbf{A}} \langle \phi L(\mathcal{M}[\mathbf{F}]) \rangle - \varepsilon \Delta t \tilde{\mathbf{A}} \nabla_x \cdot \langle v \phi \mathbf{f}_1 \rangle,$$

where  $\tilde{\mathbf{A}} := \tilde{A} \otimes_K I_{d_v+2} \in \mathbb{R}^{(2+d_v)s \times (2+d_v)s}$ ,  $I_{d_v+2}$  is the  $(d_v + 2) \times (d_v + 2)$  identity matrix, and the symbol  $\otimes_K$  denotes the Kronecker product<sup>4</sup>. Defining the flux vector  $\nabla_x \cdot F(U^{(i)}) = \langle \phi L(\mathcal{M}[F^{(i)}]) \rangle$  with  $U^{(i)} = \langle \phi \mathcal{M}[F^{(i)}] \rangle$ , it follows

$$\mathbf{U} = \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}) - \varepsilon \Delta t \tilde{\mathbf{A}} \nabla_x \cdot \langle v \phi \mathbf{f}_1 \rangle, \tag{3.18}$$

where  $\mathbf{e} = (1, 1, \dots, 1)^\top \in \mathbb{R}^s$ ,

$$\begin{aligned} \mathbf{U} &= (U^{(1)}, U^{(2)}, \dots, U^{(s)})^\top \in \mathbb{R}^{(2+d_v)s}, \\ U^{(i)} &= (\rho^{(i)}, \rho^{(i)} u^{(i)}, E^{(i)})^\top \in \mathbb{R}^{2+d_v} \\ \nabla_x \cdot F(\mathbf{U}) &= (\nabla_x \cdot F(U^{(1)}), \nabla_x \cdot F(U^{(2)}), \dots, \nabla_x \cdot F(U^{(s)}))^\top \in \mathbb{R}^{(2+d_v)s}, \\ \nabla_x \cdot F(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot (\rho^{(i)} u^{(i)}) \\ \nabla_x \cdot (\rho^{(i)} u^{(i)} \otimes u^{(i)} + \rho^{(i)} T^{(i)} Id) \\ \nabla_x \cdot ((E^{(i)} + \rho^{(i)} T^{(i)}) u^{(i)}) \end{pmatrix} \in \mathbb{R}^{2+d_v} \quad (i = 1, \dots, s). \end{aligned} \tag{3.19}$$

By Lemma A.3 in Appendix A, we obtain

$$\mathbf{U} = \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}) - \varepsilon \Delta t \tilde{\mathbf{A}} \nabla_x \cdot \mathbf{H}(\mathbf{U}), \tag{3.20}$$

and, similarly, for the numerical solution we have

$$U^{n+1} = U^n - \Delta t \tilde{\mathbf{b}} \nabla_x \cdot F(\mathbf{U}) - \varepsilon \Delta t \tilde{\mathbf{b}} \nabla_x \cdot \mathbf{H}(\mathbf{U}), \tag{3.21}$$

---

<sup>4</sup>We use the symbol  $\otimes_K$  to denote the Kronecker product: If  $A$  is a  $m \times n$  matrix and  $B$  is a  $p \times q$  matrix, then the Kronecker product of  $A, B$  is a  $mp \times nq$  matrix that can be written in block form as  $A \otimes_K B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}$ .

where  $\tilde{\mathbf{b}} := \tilde{b}^\top \otimes_K I_{d_v+2}$  and

$$\begin{aligned} \nabla_x \cdot \mathbf{H}(\mathbf{U}) &= \left( \nabla_x \cdot H(U^{(1)}), \nabla_x \cdot H(U^{(2)}), \dots, \nabla_x \cdot H(U^{(s)}) \right)^\top \in \mathbb{R}^{(2+d_v)s}, \\ \nabla_x \cdot H(U^{(i)}) &= \begin{pmatrix} 0 \\ \nabla_x \cdot (\rho^{(i)} \Theta_1^{(i)}) \\ \nabla_x \cdot (\mathbb{Q}_1^{(i)} + \rho^{(i)} \Theta_1^{(i)} u^{(i)}) \end{pmatrix} \in \mathbb{R}^{2+d_v} \quad (i = 1, \dots, s). \end{aligned}$$

Now we evaluate the quantities  $\Theta_1^{(i)}$  and  $\mathbb{Q}_1^{(i)}$ , for  $i = 1, \dots, s$ . To achieve this, we expand the anisotropic Gaussian  $\mathcal{G}$  with respect to  $\varepsilon$ , *i.e.*,

$$\mathcal{G}[\mathbf{F}] = \mathcal{M}[\mathbf{F}] + \varepsilon \mathbf{g}, \quad (3.22)$$

with vector  $\mathbf{g}$  defined as (A.13). Now inserting (3.17), (3.22) into (2.5), we get

$$\mathcal{M}[\mathbf{F}] = \mathcal{M}[f^n] \mathbf{e} - \Delta t \tilde{A} L(\mathcal{M}[\mathbf{F}]) - \varepsilon \left( \mathbf{f}_1 - f_1^n \mathbf{e} + \Delta t \tilde{A} L(\mathbf{f}_1) \right) + \Delta t A \bar{\tau} (\mathbf{g} - \mathbf{f}_1) \quad (3.23)$$

which implies

$$\bar{\tau} (\mathbf{f}_1 - \mathbf{g}) = -A^{-1} \left( \frac{\mathcal{M}[\mathbf{F}] - \mathcal{M}[f^n] \mathbf{e}}{\Delta t} + \tilde{A} L(\mathcal{M}[\mathbf{F}]) \right) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right). \quad (3.24)$$

Now, applying only the explicit part of IMEX-RK scheme to (A.20), we get

$$\frac{\mathcal{M}[\mathbf{G}] - \mathcal{M}[f^n] \mathbf{e}}{\Delta t} + \tilde{A} (L(\mathcal{M}[\mathbf{G}])) = \tilde{A} \text{diag}(\mathcal{M}[\mathbf{G}]) \left( A(\mathbf{V}) : \frac{\sigma(\mathbf{u})}{2} + 2B(\mathbf{V}) \cdot \nabla_x \sqrt{\mathbf{T}} \right) + \mathcal{O}(\varepsilon), \quad (3.25)$$

where  $\mathbf{G}$  denotes the vector of internal stage for the equation (A.20) and

$$\begin{aligned} \mathbf{u} &= \left( u^{(1)}, \dots, u^{(s)} \right)^\top \in \mathbb{R}^{d_v s}, \quad \mathbf{T} = \left( T^{(1)}, \dots, T^{(s)} \right)^\top \in \mathbb{R}^s, \\ \mathbf{V} &= \left( V^{(1)}, V^{(2)}, \dots, V^{(s)} \right)^\top \in \mathbb{R}^{d_v s} \quad V^{(i)} = \frac{v - u^{(i)}}{\sqrt{T^{(i)}}} \in \mathbb{R}^{d_v}, \\ A(\mathbf{V}) : \frac{\sigma(\mathbf{u})}{2} &\in \mathbb{R}^s, \quad \left( A(\mathbf{V}) : \frac{\sigma(\mathbf{u})}{2} \right)_i = \left( V^{(i)} \otimes V^{(i)} - \frac{1}{d_v} |V^{(i)}|^2 Id \right) : \sigma(u^{(i)}) \\ B(\mathbf{V}) \cdot \nabla_x \sqrt{\mathbf{T}} &\in \mathbb{R}^s, \quad \left( B(\mathbf{V}) \cdot \nabla_x \sqrt{\mathbf{T}} \right)_i = \frac{1}{2} V^{(i)} \left( |V^{(i)}|^2 - (d_v + 2) \right) \cdot \nabla_x \sqrt{T^{(i)}}, \quad i = 1, \dots, s. \end{aligned}$$

Integrating both sides of (3.25) macroscopic variables are associated to  $\mathbf{G}$  are obtained as follows:

$$\mathbf{U}_{\mathbf{G}} = \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}_{\mathbf{G}}) + \mathcal{O}(\varepsilon \Delta t), \quad (3.26)$$

where  $\mathbf{U}_{\mathbf{G}}$  is the macroscopic quantities associated to the stage values. Since the first stage values of  $\mathbf{F}$  and  $\mathbf{G}$  are the same  $f^n$ ,  $\mathbf{U}$  and  $\mathbf{U}_{\mathbf{G}}$  are obtained from the same  $\mathbf{U}^n$  and this implies that  $\|\mathbf{U} - \mathbf{U}_{\mathbf{G}}\|_\infty = \mathcal{O}(\varepsilon \Delta t)$ . Then, by Lemma A.1 we have  $\mathcal{M}[\mathbf{G}] = \mathcal{M}[\mathbf{F}] + \mathcal{O}(\varepsilon \Delta t)$ . Inserting this into (3.24), we get

$$\begin{aligned} \bar{\tau} (\mathbf{f}_1 - \mathbf{g}) &= -A^{-1} \left( \frac{\mathcal{M}[\mathbf{G}] - \mathcal{M}[f^n] \mathbf{e}}{\Delta t} + \tilde{A} L(\mathcal{M}[\mathbf{G}]) + \mathcal{O}(\varepsilon) + \mathcal{O}(\varepsilon \Delta t) \right) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ &= -A^{-1} \left( \tilde{A} \text{diag}(\mathcal{M}[\mathbf{G}]) \left( A(\mathbf{V}) : \frac{\sigma(\mathbf{u})}{2} + 2B(\mathbf{V}) \cdot \nabla_x \sqrt{\mathbf{T}} \right) + \mathcal{O}(\varepsilon) \right) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) + \mathcal{O}(\varepsilon \Delta t) \end{aligned} \quad (3.27)$$

where the second line comes from (3.25). Now, we again use  $\mathcal{M}[\mathbf{G}] = \mathcal{M}[\mathbf{F}] + \mathcal{O}(\varepsilon\Delta t)$  to derive

$$\bar{\tau}(\mathbf{f}_1 - \mathbf{g}) = -A^{-1} \left( \tilde{A} \operatorname{diag}(\mathcal{M}[\mathbf{F}]) \left( A(\mathbf{V}) : \frac{\sigma(\mathbf{u})}{2} + 2B(\mathbf{V}) \cdot \nabla_x \sqrt{\mathbf{T}} \right) \right) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(\varepsilon\Delta t).$$

Then, we have

$$\mathbf{f}_1 = \mathbf{g} - \bar{\tau}^{-1} A^{-1} \left( \tilde{A} \operatorname{diag}(\mathcal{M}[\mathbf{F}]) \left( A(\mathbf{V}) : \frac{\sigma(\mathbf{u})}{2} + 2B(\mathbf{V}) \cdot \nabla_x \sqrt{\mathbf{T}} \right) \right) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) + \mathcal{O}(\varepsilon). \tag{3.28}$$

Multiplying by  $v \otimes v$  or  $v|v|^2/2$  both side in (3.28) and taking the integration over  $v$ , we get from (A.21), (A.22) and (A.23) that

$$\begin{aligned} \rho^{(i)} \Theta_1^{(i)} &= \nu \rho^{(i)} \Theta_1^{(i)} - \frac{1}{\tau^{(i)}} \sum_{j=1}^s \left( A^{-1} \tilde{A} \right)_{ij} \rho^{(j)} T^{(j)} \sigma(u^{(j)}) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right), \quad i = 1, \dots, s \\ \mathbb{Q}_1^{(i)} + \rho^{(i)} \Theta_1^{(i)} u^{(i)} &= \nu \rho^{(i)} \Theta_1^{(i)} u^{(i)} - \frac{1}{\tau^{(i)}} \sum_{j=1}^s \left( A^{-1} \tilde{A} \right)_{ij} \left( \rho^{(j)} T^{(j)} \sigma(u^{(j)}) u^{(j)} + \frac{d_v + 2}{2} \rho^{(j)} T^{(j)} \nabla_x T^{(j)} \right) \\ &\quad + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right). \end{aligned}$$

Thus, the approximations of stress tensor and heat flux are given by

$$\begin{aligned} \rho \Theta_1 &= \left( \rho^{(1)} \Theta_1^{(1)}, \rho^{(2)} \Theta_1^{(2)}, \dots, \rho^{(s)} \Theta_1^{(s)} \right)^\top \\ &= -\frac{1}{1 - \nu} \left( \bar{\tau}^{-1} A^{-1} \tilde{A} \operatorname{diag}(\bar{p}) \otimes_K Id \right) \times \sigma(\mathbf{u}) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ &= -(\bar{\mu} \otimes_K Id) \times \sigma(\mathbf{u}) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ \mathbf{q} = \mathbb{Q}_1 &= \left( \mathbb{Q}_1^{(1)}, \mathbb{Q}_1^{(2)}, \dots, \mathbb{Q}_1^{(s)} \right)^\top \\ &= -\frac{d_v + 2}{2} \left( \bar{\tau}^{-1} A^{-1} \tilde{A} \operatorname{diag}(\bar{p}) \otimes_K Id \right) \times \nabla_x \mathbf{T} + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ &= -(\bar{\kappa} \otimes_K Id) \times \nabla_x \mathbf{T} + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \end{aligned}$$

where  $\sigma(u^{(i)}) = \nabla_x u^{(i)} + (\nabla_x u^{(i)})^\top - \frac{2}{d_v} \nabla_x \cdot u^{(i)} Id$ , and  $\bar{p} = (p^{(1)}, p^{(2)}, \dots, p^{(s)})^\top$  with  $p^{(i)} = \rho^{(i)} T^{(i)}$ . Here we defined the viscosity and thermal conductivity  $s \times s$  matrices  $\bar{\mu} = (\mu_{ij})$  and  $\bar{\kappa} = (\kappa_{ij})$  as

$$\bar{\mu} = \frac{1}{(1 - \nu)} \bar{\tau}^{-1} \left( A^{-1} \tilde{A} \right) \operatorname{diag}(\bar{p}), \quad \bar{\kappa} = \frac{d_v + 2}{2} \bar{\tau}^{-1} \left( A^{-1} \tilde{A} \right) \operatorname{diag}(\bar{p}). \tag{3.29}$$

This implies that

$$\bar{\tau}^{-1} \left( A^{-1} \tilde{A} \right) \operatorname{diag}(\bar{p}) = (1 - \nu) \bar{\mu} = \frac{2}{d_v + 2} \bar{\kappa}$$

and hence

$$\frac{d_v + 2}{2} \frac{\mu_{ij}}{\kappa_{ij}} = \frac{1}{1 - \nu}, \quad 1 \leq i, j \leq s.$$

This form is consistent to the Prandtl number of the ES-BGK model. To sum up, the scheme reads

$$\mathbf{U} = \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \tilde{\mathbf{A}} \nabla_x \cdot \underline{\mathbf{S}}(\mathbf{U}) + \mathcal{O}(\varepsilon^2), \tag{3.30}$$

and

$$U^{n+1} = U^n - \Delta t \tilde{\mathbf{b}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \tilde{\mathbf{b}} \nabla_x \cdot \underline{S}(\mathbf{U}) + \mathcal{O}(\varepsilon^2), \quad (3.31)$$

for the numerical solution, where

$$\begin{aligned} \nabla_x \cdot \underline{S}(\mathbf{U}) &= \left( \nabla_x \cdot \underline{S}(U^{(1)}), \nabla_x \cdot \underline{S}(U^{(2)}), \dots, \nabla_x \cdot \underline{S}(U^{(s)}) \right)^\top \in \mathbb{R}^{(d_v+2)s}, \\ \nabla_x \cdot \underline{S}(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot \left( \sum_{j=1}^{i-1} \bar{\mu}_{ij} \sigma(u^{(j)}) \right), \\ \nabla_x \cdot \left( \sum_{j=1}^{i-1} \bar{\mu}_{ij} \sigma(u^{(j)}) u^{(j)} + \bar{\kappa}_{ij} \nabla_x T^{(j)} \right) \end{pmatrix} \\ &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot \left( \frac{1}{\tau^{(i)}} \sum_{j=1}^{i-1} \frac{1}{(1-\nu)} \left( A^{-1} \tilde{A} \right)_{ij} p^{(j)} \sigma(u^{(j)}) \right), \\ \nabla_x \cdot \left( \frac{1}{\tau^{(i)}} \sum_{j=1}^{i-1} \frac{1}{(1-\nu)} \left( A^{-1} \tilde{A} \right)_{ij} p^{(j)} \sigma(u^{(j)}) u^{(j)} + \frac{d_v+2}{2} \left( A^{-1} \tilde{A} \right)_{ij} p^{(j)} \nabla_x T^{(j)} \right) \end{pmatrix} \in \mathbb{R}^{d_v+2}. \end{aligned}$$

For each  $i, j = 1, \dots, s$ , we now consider the approximation  $\tau^{(i)} = \tau^{(j)} + \mathcal{O}(\Delta t) + \mathcal{O}(\varepsilon^2)$ . Inserting this approximation into the  $\underline{S}(\mathbf{U})$  in (3.30) and (3.31), we obtain

$$\begin{aligned} \mathbf{U} &= \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \tilde{\mathbf{A}} \nabla_x \cdot \underline{S}(\mathbf{U}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2), \\ U^{n+1} &= U^n - \Delta t \tilde{\mathbf{b}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \tilde{\mathbf{b}} \nabla_x \cdot \underline{S}(\mathbf{U}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2), \end{aligned}$$

where

$$\begin{aligned} \nabla_x \cdot \underline{S}(\mathbf{U}) &= \left( \nabla_x \cdot \underline{S}(U^{(1)}), \nabla_x \cdot \underline{S}(U^{(2)}), \dots, \nabla_x \cdot \underline{S}(U^{(s)}) \right)^\top \in \mathbb{R}^{(d_v+2)s}, \\ \nabla_x \cdot \underline{S}(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot \left( \sum_{j=1}^s \frac{1}{(1-\nu)\tau^{(j)}} \left( A^{-1} \tilde{A} \right)_{ij} p^{(j)} \sigma(u^{(j)}) \right), \\ \nabla_x \cdot \left( \sum_{j=1}^s \frac{1}{(1-\nu)\tau^{(j)}} \left( A^{-1} \tilde{A} \right)_{ij} p^{(j)} \sigma(u^{(j)}) u^{(j)} + \frac{d_v+2}{2\tau^{(j)}} \left( A^{-1} \tilde{A} \right)_{ij} p^{(j)} \nabla_x T^{(j)} \right) \end{pmatrix} \\ &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \sum_{j=1}^s \left( A^{-1} \tilde{A} \right)_{ij} \nabla_x \cdot (\mu^{(j)} \sigma(u^{(j)})), \\ \sum_{j=1}^s \left( A^{-1} \tilde{A} \right)_{ij} \nabla_x \cdot (\mu^{(j)} \sigma(u^{(j)}) u^{(j)} + \kappa^{(j)} \nabla_x T^{(j)}) \end{pmatrix} \in \mathbb{R}^{d_v+2} \end{aligned}$$

with  $\mu^{(j)} = \frac{1}{(1-\nu)\tau^{(j)}} p^{(j)}$ ,  $\kappa^{(j)} = \frac{d_v+2}{2\tau^{(j)}} p^{(j)}$ . Defining

$$\begin{aligned} \nabla_x \cdot S(\mathbf{U}) &= \left( \nabla_x \cdot S(U^{(1)}), \nabla_x \cdot S(U^{(2)}), \dots, \nabla_x \cdot S(U^{(s)}) \right)^\top \in \mathbb{R}^{(d_v+2)s}, \\ \nabla_x \cdot S(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot (\mu^{(i)} \sigma(u^{(i)})), \\ \nabla_x \cdot (\mu^{(i)} \sigma(u^{(i)}) u^{(i)} + \kappa^{(i)} \nabla_x T^{(i)}) \end{pmatrix} \in \mathbb{R}^{d_v+2}, \end{aligned}$$

we finally derive

$$\begin{aligned} \mathbf{U} &= \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \tilde{\mathbf{B}} \nabla_x \cdot S(\mathbf{U}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2), \\ U^{n+1} &= U^n - \Delta t \tilde{\mathbf{b}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \tilde{\omega} \nabla_x \cdot S(\mathbf{U}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2), \end{aligned} \quad (3.32)$$

where  $\mathbf{B} = (\tilde{A}A^{-1}\tilde{A}) \otimes_K I_{d_v+2}$  and  $\boldsymbol{\omega} = (\tilde{b}^\top A^{-1}\tilde{A}) \otimes_K I_{d_v+2}$ . Next, we consider the following explicit-type RK method of order  $p$  based on the coefficients matrices  $\tilde{A}, B$  and the weights  $\tilde{b}, w$  with  $B = \tilde{A}A^{-1}\tilde{A}$ ,  $\boldsymbol{\omega}^\top = \tilde{b}^\top A^{-1}\tilde{A}$  applied to the CNS equations (1.12):

$$\begin{aligned}\mathbf{U} &= \mathbf{e} \otimes_K U^n - \Delta t \tilde{\mathbf{A}} \nabla_x \cdot F(\mathbf{U}) + \varepsilon \Delta t \mathbf{B} \nabla_x \cdot S(\mathbf{U}), \\ U^{n+1} &= U^n - \Delta t \tilde{\mathbf{b}} \nabla_x \cdot \mathbf{F}(\mathbf{U}) + \varepsilon \Delta t \boldsymbol{\omega} \nabla_x \cdot S(\mathbf{U}).\end{aligned}$$

Assuming that  $u(t)$  is the true solution of the CNS, then the local truncation error is  $\mathcal{O}(\Delta t^p)$ , i.e.,

$$\mathcal{O}(\Delta t^p) = \frac{u(t_{n+1}) - u(t_n)}{\Delta t} + \sum_{i=1}^s \tilde{b}_i \nabla_x \cdot F(u^{(i)}) - \sum_{i=1}^s \omega_i \varepsilon \nabla_x \cdot S(u^{(i)}) \quad (3.33)$$

where  $p$  is the order of the explicit scheme with

$$u^{(i)} = u(t_n) - \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \nabla_x \cdot F(u^{(j)}) - \Delta t \sum_{j=1}^{i-1} B_{ij} \varepsilon \nabla_x \cdot S(u^{(j)}),$$

where  $B = (B_{ij})$ . Now, we consider the local truncation of our method (3.32):

$$\text{L.T.E.} = \frac{u(t_{n+1}) - u(t_n)}{\Delta t} + \sum_{i=1}^s \tilde{b}_i \left( \nabla_x F(u^{(i)}) \right) - \sum_{i=1}^s \omega_i \left( \varepsilon \nabla_x S(u^{(i)}) \right) + \mathcal{O}\left(\frac{\varepsilon^2}{\Delta t}\right) + \mathcal{O}(\varepsilon \Delta t) \quad (3.34)$$

where  $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_s)^\top$ . This combined with (3.33), gives

$$\text{L.T.E.} = \mathcal{O}(\Delta t^p) + \mathcal{O}\left(\frac{\varepsilon^2}{\Delta t}\right) + \mathcal{O}(\varepsilon \Delta t). \quad (3.35)$$

□

Theorem 3.6 implies that, to accurately capture the Navier–Stokes equation, it is necessary that the local truncation error satisfies  $L.T.E. = o(\varepsilon)$ , which holds true if  $\Delta t^p + \varepsilon \Delta t + \frac{\varepsilon^2}{\Delta t} = o(\varepsilon)$ .

**Corollary 3.7.** *Setting  $\tau = 1$  in (1.7) the macroscopic explicit-type RK scheme has a local truncation error:  $L.T.E. = \mathcal{O}(\Delta t^p) + \mathcal{O}(\varepsilon^2/\Delta t)$ .*

Similarly, analogous considerations can be made for type II by introducing a new definition of the matrix  $B$  and weights  $w$  (for details, see paper [7]). However, in practice,  $\tau$  is closely related to the viscosity and heat conductivity of gases, and hence the form of  $\tau$  should be set carefully (see for example [3, 16, 18, 19]).

- **IMEX-RK scheme of type II.** In this section we analyze the asymptotic behavior of type II IMEX-RK schemes. As type I scheme, at this stage, we do not need to assume that the scheme satisfies the GSA condition.

**Theorem 3.8.** *For small values of  $\varepsilon$  and with  $\Delta t^p + \varepsilon \Delta t + \frac{\varepsilon^2}{\Delta t} = o(\varepsilon)$ , the IMEX-RK of type II (2.8) satisfying*

$$\hat{A}\hat{A}^{-1}a = \tilde{a}, \quad \hat{b}^\top \hat{A}^{-1}a = \tilde{b}_1, \quad (3.36)$$

with initial data  $f^n = \mathcal{M}[f^n] + \varepsilon f_1^n$ , where

$$f_1^n = g^n - \frac{1}{\tau^n} \mathcal{M}[f^n] \left( A(V^n) : \sigma(u^n) + 2B(V^n) \cdot \nabla_x \sqrt{T^n} \right) + \mathcal{O}(\varepsilon), \quad (3.37)$$

$V^n = \frac{v-u^n}{\sqrt{T^n}}$ , asymptotically becomes a consistent macroscopic explicit-type RK scheme of order  $p$  charterized by the pair  $(\tilde{A}, \tilde{b})$  and  $(B, \omega)$  for the CNS equations (1.12) with

$$\begin{aligned} B &= \begin{pmatrix} 0 & 0 \\ \mathbf{b}_1 & \hat{B} \end{pmatrix}, \quad \omega^\top = (\omega_1, \hat{\omega}^\top), \\ \mathbf{b}_1 &= \hat{A}\hat{A}^{-1}\tilde{a}, \quad \hat{B} = \hat{A}\hat{A}^{-1}\hat{A}, \quad \omega_1 = \hat{b}^\top \hat{A}^{-1}\tilde{a}, \quad \hat{\omega}^\top = \hat{b}^\top \hat{A}^{-1}\hat{A}. \end{aligned} \quad (3.38)$$

As in the case of Theorem 3.6, we remark that the order  $p$  could be ensured by enforcing additional order conditions derived up to order 3 for type II from the associated two pair of Butcher tableaux in [7].

*Proof.* The IMEX RK scheme of type II applied to (1.7) can be written in compact form (2.8) as

$$\begin{aligned} F^{(1)} &= f^n, \quad \hat{\mathbf{F}} = f^n \hat{\mathbf{e}} - \Delta t \tilde{a} L(f^n) - \Delta t \hat{\mathbf{A}} L(\hat{\mathbf{F}}) + \frac{\Delta t}{\varepsilon} a \tau^n (\mathcal{G}[f^n] - f^n) + \frac{\Delta t}{\varepsilon} \hat{A} \hat{\tau} (\mathcal{G}[\hat{\mathbf{F}}] - \hat{\mathbf{F}}), \\ f^{n+1} &= f^n - \Delta t \tilde{b}_1 L(f^n) - \Delta t \hat{\mathbf{b}}^\top L(\hat{\mathbf{F}}) + \frac{\Delta t}{\varepsilon} b_1 (\mathcal{G}[f^n] - f^n) + \frac{\Delta t}{\varepsilon} \hat{b}^\top (\mathcal{G}[\hat{\mathbf{F}}] - \hat{\mathbf{F}}). \end{aligned} \quad (3.39)$$

Now inserting expansions

$$f^n = \mathcal{M}[f^n] + \varepsilon f_1^n, \quad \hat{\mathbf{F}} = \mathcal{M}[\hat{\mathbf{F}}] + \varepsilon \hat{\mathbf{f}}_1, \quad \mathcal{G}[f^n] = \mathcal{M}[f^n] + \varepsilon g^n, \quad \mathcal{G}[\hat{\mathbf{F}}] = \mathcal{M}[\hat{\mathbf{F}}] + \varepsilon \hat{g}, \quad (3.40)$$

into (3.39), multiplying by  $\phi(v)$  function and integrating on  $v$ , we get

$$\begin{aligned} U^{(1)} &= U^n, \\ \hat{\mathbf{U}} &= \hat{\mathbf{e}} \otimes_K U^n - \Delta t \left( \tilde{a} \otimes_K \nabla_x \cdot F(U^n) + \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) \right) - \varepsilon \Delta t \left( \tilde{a} \otimes_K \langle \phi L(f_1^n) \rangle + \hat{\mathbf{A}} \langle \phi L(\hat{\mathbf{f}}_1) \rangle \right), \\ U^{n+1} &= U^n - \Delta t \left( \tilde{b}_1 \nabla_x \cdot F(U^n) + \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) \right) - \varepsilon \Delta t \left( \tilde{b}_1 \langle \phi L(f_1^n) \rangle + \hat{\mathbf{b}} \langle \phi L(\hat{\mathbf{f}}_1) \rangle \right). \end{aligned} \quad (3.41)$$

Here  $\hat{\mathbf{A}} := \hat{A} \otimes_K I_{d_v+2}$ ,  $\hat{\mathbf{b}} := \hat{b}^\top \otimes_K I_{d_v+2}$ , and we define  $\nabla_x \cdot F(U^n) = \langle \phi L(\mathcal{M}[f^n]) \rangle$  and  $\nabla_x \cdot F(\hat{\mathbf{U}}) = \langle \phi L(\mathcal{M}[\hat{\mathbf{F}}]) \rangle$ , i.e.,

$$\begin{aligned} \hat{\mathbf{U}} &= (U^{(2)}, U^{(3)}, \dots, U^{(s)})^\top \in \mathbb{R}^{(2+d_v)(s-1)}, \\ U^{(i)} &= (\rho^{(i)}, \rho^{(i)} u^{(i)}, E^{(i)})^\top \in \mathbb{R}^{2+d_v} \\ \nabla_x \cdot F(\hat{\mathbf{U}}) &= \left( \nabla_x \cdot F(U^{(2)}), \nabla_x \cdot F(U^{(3)}), \dots, \nabla_x \cdot F(U^{(s)}) \right)^\top \in \mathbb{R}^{(2+d_v)(s-1)}, \\ \nabla_x \cdot F(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot (\rho^{(i)} u^{(i)}) \\ \nabla_x \cdot (\rho^{(i)} u^{(i)} \otimes u^{(i)} + \rho^{(i)} T^{(i)} Id) \\ \nabla_x \cdot ((E^{(i)} + \rho^{(i)} T^{(i)}) u^{(i)}) \end{pmatrix} \in \mathbb{R}^{2+d_v} \quad (i = 1, \dots, s). \end{aligned} \quad (3.42)$$

Then, by Lemma A.3 we obtain

$$\begin{aligned} U^{(1)} &= U^n, \\ \hat{\mathbf{U}} &= \hat{\mathbf{e}} \otimes_K U^n - \Delta t \left( \tilde{a} \otimes_K \nabla_x \cdot F(U^n) + \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) \right) - \varepsilon \Delta t \left( \tilde{a} \otimes_K \nabla_x \cdot H(U^n) + \hat{\mathbf{A}} \nabla_x \cdot H(\hat{\mathbf{U}}) \right), \\ U^{n+1} &= U^n - \Delta t \left( \tilde{b}_1 \nabla_x \cdot F(U^n) + \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) \right) - \varepsilon \Delta t \left( \tilde{b}_1 \nabla_x \cdot H(U^n) + \hat{\mathbf{b}} \nabla_x \cdot H(\hat{\mathbf{U}}) \right), \end{aligned} \quad (3.43)$$

where

$$\begin{aligned} \nabla_x \cdot \mathbf{H}(\hat{\mathbf{U}}) &= \left( \nabla_x \cdot H(U^{(2)}), \nabla_x \cdot H(U^{(3)}), \dots, \nabla_x \cdot H(U^{(s)}) \right)^\top \in \mathbb{R}^{(2+d_v)(s-1)}, \\ \nabla_x \cdot H(U^{(i)}) &= \begin{pmatrix} 0 \\ \nabla_x \cdot (\rho^{(i)} \Theta_1^{(i)}) \\ \nabla_x \cdot (\mathbb{Q}_1^{(i)} + \rho^{(i)} \Theta_1^{(i)} u^{(i)}) \end{pmatrix} \in \mathbb{R}^{2+d_v}, \quad i = 1, \dots, s. \end{aligned}$$

Now we evaluate  $\hat{\mathbf{f}}_1$  in (3.41). Substituting (3.40) into (3.39), we obtain

$$\begin{aligned} \mathcal{M}[\hat{\mathbf{F}}] &= \mathcal{M}[f^n] \hat{\mathbf{e}} - \Delta t \left( \tilde{a} L(\mathcal{M}[f^n]) + \hat{A} L(\mathcal{M}[\hat{\mathbf{F}}]) \right) \\ &\quad - \varepsilon \left( \hat{\mathbf{f}}_1 - f_1^n \mathbf{e} + \Delta t \left( \tilde{a} L(f_1^n) + \hat{A} L(\hat{\mathbf{f}}_1) \right) \right) \\ &\quad + \Delta t \left( a \tau^n (g^n - f_1^n) + \hat{A} \hat{\tau} (\hat{\mathbf{g}} - \hat{\mathbf{f}}_1) \right). \end{aligned} \tag{3.44}$$

Following the same argument for type I in (3.25)–(3.27), we have

$$\begin{aligned} \hat{A} \hat{\tau} (\hat{\mathbf{f}}_1 - \hat{\mathbf{g}}) &= -\tilde{a} \left( \mathcal{M}[f^n] \left( A(V^n) : \frac{\sigma(u^n)}{2} + 2B(V^n) \cdot \nabla_x \sqrt{T^n} \right) \right) \\ &\quad - \hat{A} \left( \mathcal{M}[\hat{\mathbf{F}}] \left( A(\hat{V}) : \frac{\sigma(\hat{\mathbf{u}})}{2} + 2B(\hat{V}) \cdot \nabla_x \sqrt{\hat{\mathbf{T}}} \right) \right) \\ &\quad - a \tau^n (f_1^n - g^n) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) + \mathcal{O}(\varepsilon), \end{aligned} \tag{3.45}$$

where  $\mathbf{u}^\top = (u^{(1)}, \hat{\mathbf{u}}^\top)$ ,  $\mathbf{T}^\top = (T^{(1)}, \hat{\mathbf{T}}^\top)$  with  $\hat{\mathbf{u}}^\top \in \mathbb{R}^{d_v(s-1)}$ ,  $\hat{\mathbf{T}}^\top \in \mathbb{R}^{s-1}$  and

$$\mathbf{V}^\top = (V^{(1)}, \hat{V}^\top), \quad \hat{\mathbf{V}} = (V^{(2)}, V^{(2)}, \dots, V^{(s)})^\top \in \mathbb{R}^{d_v(s-1)} \quad V^{(i)} = \frac{v - u^{(i)}}{\sqrt{T^{(i)}}} \in \mathbb{R}^{d_v},$$

with  $u^{(1)} = u^n$ ,  $T^{(1)} = T^n$  and  $V^{(1)} = V^n$ . Now, at the initial step  $n$ , by the assumption (3.37) we have

$$\tau^n (f_1^n - g^n) = -\mathcal{M}[f^n] \left( A(V^n) : \frac{\sigma(u^n)}{2} + 2B(V^n) \cdot \nabla_x \sqrt{T^n} \right) + \mathcal{O}(\varepsilon),$$

and substituting this into (3.45), we get

$$\begin{aligned} \hat{\mathbf{f}}_1 &= \hat{\mathbf{g}} - \hat{\tau}^{-1} \hat{A}^{-1} (\tilde{a} - a) \left( \mathcal{M}[f^n] \left( A(V^n) : \frac{\sigma(u^n)}{2} + 2B(V^n) \cdot \nabla_x \sqrt{T^n} \right) \right) \\ &\quad - \hat{\tau}^{-1} \hat{A}^{-1} \hat{A} \left( \mathcal{M}[\hat{\mathbf{F}}] \left( A(\hat{V}) : \frac{\sigma(\hat{\mathbf{u}})}{2} + 2B(\hat{V}) \cdot \nabla_x \sqrt{\hat{\mathbf{T}}} \right) \right) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) + \mathcal{O}(\varepsilon). \end{aligned} \tag{3.46}$$

Recall that  $F^{(1)} = f^n$ ,  $U^{(1)} = U^n$ , we set  $g^{(1)} = g^n$ ,  $f_1^{(1)} = f_1^n$ . Now, we multiply by  $v \otimes v$  or  $v|v|^2/2$  both side in (3.46), and take the integration over  $v$ , which together with (A.21)–(A.23) gives

$$\begin{aligned} \rho^{(1)} \Theta_1^{(1)} &= \nu \rho^{(1)} \Theta_1^{(1)} - \frac{1}{\tau^n} \rho^n T^n \sigma(u^n) + \mathcal{O}(\varepsilon), \\ \mathbb{Q}_1^{(1)} + \rho^{(1)} \Theta_1^{(1)} u^{(1)} &= \nu \rho^{(1)} \Theta_1^{(1)} u^{(1)} - \frac{1}{\tau^n} \left( \rho^n T^n \sigma(u^n) u^n + \frac{d_v + 2}{2} \rho^n T^n \nabla_x T^n \right) + \mathcal{O}(\varepsilon), \end{aligned} \tag{3.47}$$

and

$$\begin{aligned}\rho^{(i)}\Theta_1^{(i)} &= \nu\rho^{(i)}\Theta_1^{(i)} - \frac{1}{\tau^{(i)}}\left(\hat{A}^{-1}(\tilde{a} - a)\right)_{i-1}\rho^n T^n \sigma(u^n) \\ &\quad - \frac{1}{\tau^{(i)}}\sum_{j=1}^{i-2}\left(\hat{A}^{-1}\hat{A}\right)_{ij}\rho^{(j+1)}T^{(j+1)}\sigma(u^{(j+1)}) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right), \\ \mathbb{Q}_1^{(i)} + \rho^{(i)}\Theta_1^{(i)}u^{(i)} &= \nu\rho^{(i)}\Theta_1^{(i)}u^{(i)} - \frac{1}{\tau^{(i)}}\left(\hat{A}^{-1}(\tilde{a} - a)\right)_{i-1}\left(\rho^n T^n \sigma(u^n)u^n + \frac{d_v + 2}{2}\rho^n T^n \nabla_x T^n\right) \\ &\quad - \frac{1}{\tau^{(i)}}\sum_{j=1}^{i-2}\left(\hat{A}^{-1}\hat{A}\right)_{ij}\left(\rho^{(j+1)}T^{(j+1)}\sigma(u^{(j+1)})u^{(j+1)} + \frac{d_v + 2}{2}\rho^{(j+1)}T^{(j+1)}\nabla_x T^{(j+1)}\right) \\ &\quad + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right), \quad i = 2, \dots, s\end{aligned}$$

with  $\hat{A}^{-1}(\tilde{a} - a) \in \mathbb{R}^{s-1}$ . Thus, the approximations of stress tensor and heat flux are given by

$$\begin{aligned}\rho^{(1)}\Theta_1^{(1)} &= \frac{1}{(1-\nu)\tau^n}\rho^n T^n \sigma(u^n) + \mathcal{O}(\varepsilon), \quad \mathbb{Q}_1^{(1)} = \frac{d_v + 2}{2\tau^n}\rho^n T^n \nabla_x T^n + \mathcal{O}(\varepsilon), \\ \rho\hat{\Theta}_1 &= \left(\rho^{(2)}\Theta_1^{(2)}, \rho^{(3)}\Theta_1^{(3)}, \dots, \rho^{(s)}\Theta_1^{(s)}\right)^\top, \\ &= -\frac{1}{1-\nu}\left(\hat{\tau}^{-1}\hat{A}^{-1}(\tilde{a} - a)p^n\right) \otimes_K \sigma(U^n) - \frac{1}{1-\nu}\left(\hat{\tau}^{-1}\hat{A}^{-1}\hat{A}\text{diag}(\hat{p}) \otimes_K Id\right) \times \sigma(\hat{\mathbf{u}}) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ &= -\bar{\mu}^{(1)} \otimes_K \sigma(U^n) - (\hat{\mu} \otimes_K Id) \times \sigma(\hat{\mathbf{u}}) + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ \mathbf{q} = \mathbb{Q}_1 &= \left(\mathbb{Q}_1^{(2)}, \mathbb{Q}_1^{(3)}, \dots, \mathbb{Q}_1^{(s)}\right)^\top, \\ &= -\frac{d_v + 2}{2}\left(\hat{\tau}^{-1}\hat{A}^{-1}(\tilde{a} - a)p^n\right) \otimes_K \nabla_x T^n - \frac{d_v + 2}{2}\left(\hat{\tau}^{-1}\hat{A}^{-1}\hat{A}\text{diag}(\hat{p}) \otimes_K Id\right) \times \nabla_x \hat{\mathbf{T}} + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right) \\ &= -\bar{\kappa}^{(1)} \otimes_K \nabla_x T^n - (\hat{\kappa} \otimes_K Id) \times \nabla_x \hat{\mathbf{T}} + \mathcal{O}\left(\frac{\varepsilon}{\Delta t}\right)\end{aligned}$$

where  $\sigma(\hat{\mathbf{u}}) = (\sigma(u^{(2)}), \dots, \sigma(u^{(s)}))^\top$ ,  $\nabla_x \hat{\mathbf{T}} = (\nabla_x T^{(2)}, \dots, \nabla_x T^{(s)})^\top$ ,  $\sigma(u^{(i)}) = \nabla_x u^{(i)} + (\nabla_x u^{(i)})^\top - \frac{2}{d_v}\nabla_x \cdot u^{(i)} Id$  ( $i = 2, \dots, s$ ) and  $\hat{p} = (p^{(2)}, p^{(3)}, \dots, p^{(s)})^\top$  with  $p^{(i)} = \rho^{(i)}T^{(i)}$  ( $i = 2, \dots, s$ ). Now we introduce the viscosity and thermal conductivity matrices given by

$$\bar{\mu} = \begin{pmatrix} 0 & 0 \\ \bar{\mu}^{(1)} & \hat{\mu} \end{pmatrix}, \quad \bar{\kappa} = \begin{pmatrix} 0 & 0 \\ \bar{\kappa}^{(1)} & \hat{\kappa} \end{pmatrix}, \quad (3.48)$$

where

$$\bar{\mu}^{(1)} = \frac{1}{1-\nu}\left(\hat{\tau}^{-1}\hat{A}^{-1}(\tilde{a} - a)p^n\right), \quad \bar{\kappa}^{(1)} = \frac{d_v + 2}{2}\left(\hat{\tau}^{-1}\hat{A}^{-1}(\tilde{a} - a)p^n\right), \quad (3.49)$$

and  $\hat{\mu} = (\mu_{ij})$  and  $\hat{\kappa} = (\kappa_{ij})$  are  $s-1 \times s-1$  matrices such that

$$\hat{\mu} = \frac{1}{(1-\nu)}\hat{\tau}^{-1}\left(\hat{A}^{-1}\hat{A}\right)\text{diag}(\hat{p}), \quad \hat{\kappa} = \frac{d_v + 2}{2}\hat{\tau}^{-1}\left(\hat{A}^{-1}\hat{A}\right)\text{diag}(\hat{p}). \quad (3.50)$$

Therefore, from (3.43) we obtain

$$U^{(1)} = U^n,$$

$$\hat{\mathbf{U}} = \hat{\mathbf{e}} \otimes_K U^n - \Delta t \tilde{a} \otimes_K \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) + \varepsilon \Delta t \tilde{a} \otimes_K \nabla \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{A}} \nabla_x \cdot \bar{S}(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2)$$

$$U^{n+1} = U^n - \Delta t \tilde{b}_1 \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) + \varepsilon \Delta t \tilde{b}_1 \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{b}} \nabla_x \cdot \bar{S}(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) \quad (3.51)$$

with

$$\begin{aligned} \nabla_x \cdot S(U^n) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot (\mu^n \sigma(u^n)), \\ \nabla_x \cdot (\mu^n \sigma(u^n) u^n + \kappa^n \nabla_x T^n) \end{pmatrix}, \quad \mu^n = \frac{1}{(1-\nu)\tau^n} p^n, \quad \kappa^n = \frac{d_v + 2}{2\tau^n} p^n, \\ \nabla_x \cdot \bar{S}(\hat{\mathbf{U}}) &= \left( \nabla_x \cdot \bar{S}(U^{(2)}), \nabla_x \cdot \bar{S}(U^{(3)}), \dots, \nabla_x \cdot \bar{S}(U^{(s)}) \right)^\top \in \mathbb{R}^{(d_v+2)(s-1)}, \\ \nabla_x \cdot \bar{S}(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot (\bar{\mu}_{i-1}^{(1)} \sigma(u^n)) + \nabla_x \cdot \left( \sum_{j=1}^{i-2} \hat{\mu}_{ij} \sigma(u^{(j+1)}) \right) \\ \nabla_x \cdot (\bar{\mu}_{i-1}^{(1)} \sigma(u^n) u^n + \bar{\kappa}_{i-1}^{(1)} \nabla_x T^n) + \nabla_x \cdot \left( \sum_{j=1}^{i-2} \hat{\mu}_{ij} \sigma(u^{(j+1)}) u^{(j)} + \hat{\kappa}_{ij} \nabla_x T^{(j+1)} \right) \end{pmatrix} \in \mathbb{R}^{d_v+2}. \end{aligned}$$

From  $\tau^{(i)} = \tau^{(j)} + \mathcal{O}(\Delta t) + \mathcal{O}(\varepsilon^2)$  for  $i, j = 1, \dots, s$ , we further get

$$\begin{aligned} U^{(1)} &= U^n, \\ \hat{\mathbf{U}} &= \hat{\mathbf{e}} \otimes_K U^n - \Delta t \tilde{a} \otimes_K \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\ &\quad + \varepsilon \Delta t \tilde{a} \otimes_K \nabla \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{A}} \nabla_x \cdot \bar{S}(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2) \\ U^{n+1} &= U^n - \Delta t \tilde{b}_1 \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\ &\quad + \varepsilon \Delta t \tilde{b}_1 \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{b}} \nabla_x \cdot \bar{S}(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2) \end{aligned}$$

with

$$\begin{aligned} \nabla_x \cdot \bar{S}(\mathbf{U}) &= \left( \nabla_x \cdot \bar{S}(U^{(2)}), \nabla_x \cdot \bar{S}(U^{(3)}), \dots, \nabla_x \cdot \bar{S}(U^{(s)}) \right)^\top \in \mathbb{R}^{(d_v+2)(s-1)}, \\ \nabla_x \cdot \bar{S}(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0} \\ \left( \hat{A}^{-1}(\tilde{a} - a) \right)_{i-1} \nabla_x \cdot (\mu^n \sigma(u^n)) \\ \left( \hat{A}^{-1}(\tilde{a} - a) \right)_{i-1} \nabla_x \cdot ((\mu^n \sigma(u^n) u^n) + \kappa^n \nabla_x T^n) \end{pmatrix} \\ &\quad + \begin{pmatrix} \nabla_x \cdot \mathbf{0} \\ \sum_{j=1}^{i-2} \left( \hat{A}^{-1} \hat{A} \right)_{ij} \nabla_x \cdot (\mu^{(j+1)} \sigma(u^{(j+1)})) \\ \sum_{j=1}^{i-2} \left( \hat{A}^{-1} \hat{A} \right)_{ij} \nabla_x \cdot (\mu^{(j+1)} \sigma(u^{(j+1)}) u^{(j+1)} + \kappa^{(j+1)} \nabla_x T^{(j+1)}) \end{pmatrix} \end{aligned}$$

where  $\mu^{(j)} = \frac{1}{(1-\nu)\tau^{(j)}} p^{(j)}$ ,  $\kappa^{(j)} = \frac{d_v+2}{2\tau^{(j)}} p^{(j)}$ . Defining

$$\begin{aligned} \nabla_x \cdot S(\hat{\mathbf{U}}) &= \left( \nabla_x \cdot S(U^{(2)}), \nabla_x \cdot S(U^{(3)}), \dots, \nabla_x \cdot S(U^{(s)}) \right)^\top \in \mathbb{R}^{(d_v+2)(s-1)}, \\ \nabla_x \cdot S(U^{(i)}) &= \begin{pmatrix} \nabla_x \cdot \mathbf{0}, \\ \nabla_x \cdot (\mu^{(i)} \sigma(u^{(i)})), \\ \nabla_x \cdot (\mu^{(i)} \sigma(u^{(i)}) u^{(i)} + \kappa^{(i)} \nabla_x T^{(i)}) \end{pmatrix} \in \mathbb{R}^{d_v+2}, \quad i = 2, \dots, s, \end{aligned}$$

we finally obtain

$$\begin{aligned}
U^{(1)} &= U^n, \\
\hat{\mathbf{U}} &= \hat{\mathbf{e}} \otimes_K U^n - \Delta t \tilde{a} \otimes_K \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\
&\quad + \varepsilon \Delta t \tilde{a} \otimes_K \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{A} \hat{A}^{-1} (\tilde{a} - a) \otimes_K \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{B}} \nabla_x \cdot S(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2) \\
U^{n+1} &= U^n - \Delta t \tilde{b}_1 \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\
&\quad + \varepsilon \Delta t \tilde{b}_1 \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{b}}^\top \hat{A}^{-1} (\tilde{a} - a) \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\omega} \nabla_x \cdot S(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2) \quad (3.52)
\end{aligned}$$

where  $\hat{\mathbf{B}} = \hat{B} \otimes_K I_{d_v+2}$  with  $\hat{B} = \hat{A} \hat{A}^{-1} \hat{A}$  and  $\hat{\omega} = \hat{\omega} \otimes_K I_{d_v+2}$  with  $\hat{\omega} = \hat{b}^\top \hat{A}^{-1} \hat{A}$ . By the extra condition (3.36), we get

$$\begin{aligned}
U^{(1)} &= U^n, \\
\hat{\mathbf{U}} &= \hat{\mathbf{e}} \otimes_K U^n - \Delta t \tilde{a} \otimes_K \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\
&\quad + \varepsilon \Delta t \mathbf{b}_1 \otimes_K \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{B}} \nabla_x \cdot S(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2) \\
U^{n+1} &= U^n - \Delta t \tilde{b}_1 \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\
&\quad + \varepsilon \Delta t \omega_1 \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\omega} \nabla_x \cdot S(\hat{\mathbf{U}}) + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \Delta t^2)
\end{aligned}$$

where and  $\mathbf{b}_1 = \hat{A} \hat{A}^{-1} \tilde{a}$  and  $\omega_1 = \hat{b}^\top \hat{A}^{-1} \tilde{a}$ . Next, we consider the following explicit-type RK method of order  $p$  applied to the CNS equations (1.12):

$$\begin{aligned}
U^{(1)} &= U^n, \\
\hat{\mathbf{U}} &= \hat{\mathbf{e}} \otimes_K U^n - \Delta t \tilde{a} \otimes_K \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{A}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\
&\quad + \varepsilon \Delta t \mathbf{b}_1 \otimes_K \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\mathbf{B}} \nabla_x \cdot S(\hat{\mathbf{U}}) \\
U^{n+1} &= U^n - \Delta t \tilde{b}_1 \nabla_x \cdot F(U^n) - \Delta t \hat{\mathbf{b}} \nabla_x \cdot F(\hat{\mathbf{U}}) \\
&\quad + \varepsilon \Delta t \omega_1 \nabla_x \cdot S(U^n) + \varepsilon \Delta t \hat{\omega} \nabla_x \cdot S(\hat{\mathbf{U}}) \quad (3.53)
\end{aligned}$$

characterized by the pair  $(\tilde{A}, \tilde{b})$  and  $(B, \omega)$  given by (3.38) for type II schemes.

The same conclusions regarding the local truncation error presented in Theorem 3.6 can be applied here for the case of type II.  $\square$

### 3.3. Uniform accuracy

In the stiff case, *i.e.*,  $\Delta t \gg \varepsilon$ , the phenomenon of order reduction may occur, particularly in the worst case when  $\Delta t \approx \varepsilon$  (intermediate regime of  $\varepsilon$ ). In this scenario, the classical order of the method can drop off, leading to a loss of accuracy in IMEX-RK schemes within this regime. This phenomenon was investigated in detail in [5, 7, 12]. In particular in [7] the authors considered a prototype hyperbolic relaxation system and performed an asymptotic expansion up to  $\mathcal{O}(\varepsilon)$  for IMEX-RK methods of type I and II. They showed that, under additional order conditions, these schemes effectively reduce to explicit-type RK methods at the  $\mathcal{O}(\varepsilon)$  level. These methods were carefully designed to accurately match the  $\mathcal{O}(\varepsilon)$  terms up to a desired order  $p$ , ensuring that effectively behave as explicit RK schemes of order  $p$  for the convection-diffusion equation when  $\varepsilon$  is small but not negligible.

However, constructing high-order IMEX-RK schemes of type I requires additional stages to satisfy extra-order conditions, their construction is particularly challenging [7, 12]. In this work, we therefore focus on IMEX-RK schemes of type II [7], since their structures make them easier to construct.

We note that if conditions (3.36) are not satisfied, order reduction is likely to occur in these intermediate regimes of  $\varepsilon$ , for example, in the numerical tests (Test 1 and Test 2), a classical ARS scheme of type II in (4.1) show this phenomenon of order reduction. Alternatively, we adopt IMEX-RK schemes of type II introduced in [7] that satisfy conditions (3.36) ensuring the consistency of the schemes (Thm. 3.8). Furthermore, these schemes satisfy additional order conditions, preventing order reduction in the intermediate regime and ensuring uniform accuracy over the full range of Knudsen numbers. In particular, we examine two IMEX-RK schemes of type II called IMEX-II-GSA3 (4.2) and IMEX-II-ISA3 (4.3). These two schemes are specifically designed in [7] to improve the resolution of hyperbolic relaxation systems, especially in the intermediate regime, *i.e.*,  $\varepsilon \approx \Delta t$ . In Section 4, we compare these methods with other IMEX-RK schemes of type I or II existing in literature.

Finally, we note that IMEX-II-ISA3 scheme fulfills  $\tilde{b} = b$  and the quantity in (3.52) becomes

$$\hat{b}^\top \hat{A}^{-1}(a - \tilde{a}) = \hat{b}^\top \hat{A}^{-1}(a - \tilde{a}) = \hat{e}_s^\top (a - \tilde{a}) \tag{3.54}$$

which equals zero if  $\tilde{a}_{s1} = a_{s1}$ . It is worth noting that the assumption  $\tilde{b} = b$ , plays an important rule in mitigating the lost of accuracy, especially in regimes where  $\varepsilon \approx \Delta t$ , as demonstrated in [5]. In Section 4, we will show that IMEX-II-GSA3 scheme, which is GSA and  $\tilde{b} \neq b$  leads to a mild order reduction for small values of  $\varepsilon$ . On the contrary, IMEX-II-ISA3 with  $\tilde{b} = b$ , does not exhibit order reduction, (as confirmed by the numerical results in the next section).

### 4. NUMERICAL RESULTS

In this section, we test the performance of different types of IMEX RK schemes using two distinct models: the 1 + 1 BGK model and the 1 + 2 ES-BGK model. We use the fifth-order finite difference WENO method [15] to approximate spatial derivatives. The discretization of the velocity domain is achieved through uniform grid points within a sufficiently large interval or domain. The choice of free parameter  $\nu$  and collision frequency  $\tau$  will be explained in each problem.

Throughout this section, we consider IMEX-RK schemes of type I and II. In particular, we take into account the GSA IMEX-II-GSA3 scheme and IMEX-II-ISA3 with  $\tilde{b} = b$ . For comparison, we consider the SI-IMEX(4, 4, 3) scheme of type I not GSA in [12], and the scheme GSA ARS(4, 4, 3) of type II in [1].

In the following, these schemes are represented, as usual, by the double Butcher tableau.

- Third-order SI-IMEX(4, 4, 3) scheme of type I in [12]. (Top: explicit method. Bottom: implicit method).

0		0	0	0	0
$\gamma$	$\gamma$	0	0	0	0
$c_3$	1.243893189483362	-0.525959928729133	0	0	0
1	0.630412558152867	0.786580740199155	-0.416993298352022	0	0
	0	1.208496649176010	-0.644363170684468	$\gamma$	
$\gamma$	$\gamma$	0	0	0	0
$\gamma$	0	$\gamma$	0	0	0
$c_3$	0	0.282066739245771	$\gamma$	0	0
1	0	1.208496649176010	-0.644363170684468	$\gamma$	$\gamma$
	0	1.208496649176010	-0.644363170684468	$\gamma$	$\gamma$

$$\gamma = 0.435866521508459, c_3 = 0.7179332607542294.$$

– Third-order GSA ARS(4, 4, 3) in [1]:

0	0	0	0	0	0	0	0	0	0	0	0	(4.1)
1/2	1/2	0	0	0	0	1/2	0	1/2	0	0	0	
2/3	11/18	1/18	0	0	0	2/3	0	1/6	1/2	0	0	
1/2	5/6	-5/6	1/2	0	0	1/2	0	-1/2	1/2	1/2	0	
1	1/4	7/4	3/4	-7/4	0	1	0	3/2	-3/2	1/2	1/2	
	1/4	7/4	3/4	-7/4	0		0	3/2	-3/2	1/2	1/2	

– Third-order GSA IMEX-II-GSA3 of type ARS [7]. (Top: explicit method. Bottom: implicit method):

0	0	0	0	0	0	0	0	0	(4.2)
43/100	43/100	0	0	0	0	0	0	0	
336/929	0	336/929	0	0	0	0	0	0	
-29/42	0	-29/42	0	0	0	0	0	0	
581/527	0	-1213/770	2491/956	267/3701	0	0	0	0	
2/3	0	-197/1238	499/743	0	581/3768	0	0	0	
1	0	263/620	134/16589	1040/22119	0	4777/9174	0	0	
	0	263/620	134/16589	1040/22119	0	4777/9174	0	0	
0	0	0	0	0	0	0	0	0	
43/100	0	43/100	0	0	0	0	0	0	
336/929	0	-168/2459	43/100	0	0	0	0	0	
-29/42	0	-2353/2100	0	43/100	0	0	0	0	
581/527	0	889/1322	0	0	43/100	0	0	0	
2/3	0	247/2416	0	408/3035	0	43/100	0	0	
1	0	872/1201	0	139/4081	-50/237	434/20817	43/100	0	
	0	872/1201	0	139/4081	-50/237	434/20817	43/100	0	

– Third-order IMEX-II-ISA3 [7]. (Top: explicit method. Bottom: implicit method):

0	0	0	0	0	0	0	0	(4.3)
1/5	1/5	0	0	0	0	0	0	
1/3	0	1/3	0	0	0	0	0	
2/3	0	557/867	7/289	0	0	0	0	
3/4	0	16/289	803/1156	0	0	0	0	
1	0	13348/3993	-9355/3993	0	0	0	0	
1	0	75/154	0	-3/14	8/11	0	0	
	0	-155/112	251/80	-547/280	2/3	1/3	1/5	
0	0	0	0	0	0	0	0	
1/5	0	1/5	0	0	0	0	0	
1/3	0	2/15	1/5	0	0	0	0	
2/3	0	7/15	0	1/5	0	0	0	
3/4	0	1137/1004	-731/1255	0	1/5	0	0	
1	0	447/565	0	-636/613	519/496	1/5	0	
1	0	-155/112	251/80	-547/280	2/3	1/3	1/5	
	0	-155/112	251/80	-547/280	2/3	1/3	1/5	

#### 4.1. Accuracy test

*Test 1.* In this test we investigate numerically the convergence rate solving the BGK model for a smooth solution. The initial data is taken as

$$\begin{aligned} \rho_0(x) &= 1 + 0.2 \sin(\pi x), \quad u_0(x) = 1, \quad T_0(x) = \frac{1}{\rho_0(x)}, \\ f_0(x, v) &= \mathcal{M}(x, v) - \frac{\varepsilon}{\tau} (I - \Pi_{\mathcal{M}})(v_1 \partial_x \mathcal{M}) \end{aligned} \quad (4.4)$$

where

$$\mathcal{M}(x, v) = \frac{\rho_0(x)}{2\pi T_0(x)} \exp\left(-\frac{|v - u_0(x)|^2}{2T_0(x)}\right).$$

For an explicit definition of the term  $(I - \Pi_{\mathcal{M}})(v_1 \partial_x \mathcal{M})$ , please refer to Remark A.2 in the appendix.

We use 32 uniform points for velocity in the interval  $v \in [-10, 10]$ , and  $N_x$  uniform points for space in the interval  $x \in [0, 2]$  with periodic boundary conditions. We set  $\Delta t = \text{CFL} \Delta x / 10$  with  $\text{CFL} = 0.9$  and  $\tau = 1$ .

Since the exact solution is not available for this test, as well as for the subsequent test, the  $L^1$  errors are computed as the difference of the numerical solutions on two consecutive meshes in space, *i.e.*, we use the numerical solution on a finer mesh with mesh size  $\Delta x/2$  as the reference solution to compute the error for solutions on the mesh size of  $\Delta x$ .

Table 1 shows results for third order accurate IMEX-RK schemes at  $t = 0.25$ . First we can see order reduction in the intermediate regime for ARS(4, 4, 3) and SI-IMEX(4, 4, 3) scheme. Although we observe order reductions for IMEX-RK at small Knudsen numbers, such as  $10^{-4}$ , it is a typical issue of IMEX-RK methods. In general, it is highly nontrivial to avoid order reduction of IMEX-RK schemes in the intermediate regime (for example, see papers [5, 6, 8]).

For comparison, we test the two third-order schemes introduced in [7] IMEX-II-GSA3 (4.2) and IMEX-II-ISA3 (4.3) for this example. The third order accuracy is achieved for IMEX-II-ISA3 for all values of  $\varepsilon$  even in the intermediate regime, *i.e.*, the scheme maintains uniform accuracy, on the other hand IMEX-II-GSA3 shows a slight order reduction for the value  $\varepsilon = 10^{-4}$ . Note that, as discussed in Section 2, since the IMEX-II-ISA3 scheme has same weights, it performs better than the other methods for intermediate values of  $\varepsilon$ .

*Test 2.* In this test we investigate numerically the convergence rate solving the ES-BGK model for a smooth solution considering  $d_v = 2$  with  $v = (v_1, v_2)$ , and the initial well-prepared initial data (4.4), where

$$\mathcal{M}(x, v) = \frac{\rho_0(x)}{2\pi T_0(x)} \exp\left(-\frac{(v_1 - u_0(x))^2 + v_2^2}{2T_0(x)}\right). \quad (4.5)$$

For the velocity discretization, we use  $32 \times 32$  uniform points in the domain  $v \in [-10, 10] \times [-10, 10]$ . Note that, for the SI-IMEX(4, 4, 3) scheme in order to achieve the expected order of convergence, we required more points in velocity. In this test we use  $N_v = 80$  uniform points for the velocity. We use  $N_x$  uniform points for space in the interval  $x \in [0, 2]$  with periodic boundary conditions. The time step is taken as  $\Delta t = \text{CFL} \Delta x / 10$  with  $\text{CFL} = 0.9$ ,  $\tau = 1$  and  $\nu = -1/2$ .

Table 2 shows the result for the third order accurate IMEX-RK schemes at  $t = 0.25$ . As the one dimensional case we can see order reduction in the intermediate regime ( $\varepsilon = 10^{-4}$ ) for the ARS(4, 4, 3) scheme. The third order scheme IMEX-II-ISA3 achieves the third order accuracy for all the values of  $\varepsilon$  even in the intermediate regime, whereas IMEX-II-GSA3 and SI-IMEX(4, 4, 3) schemes show a slight order reduction for the value  $\varepsilon = 10^{-4}$ .

At the theoretical level, the ES-BGK model approximates the NSE equations with an error of order  $\mathcal{O}(\varepsilon^2)$ , as is well known from the Chapman–Enskog expansion. However, verifying this convergence numerically is extremely challenging. Indeed, the numerical method approximates the ES-BGK model, not the NSE system directly. Hence, the observed error will always be a combination of the modeling error, between ES-BGK and

TABLE 1. Accuracy test for the BGK equation. Initial data is given in (4.4),  $L^1$  error on the density  $\rho$  at  $T = 0.25$  and  $N_v = 32$ .

Relative $L^1$ error and order of density								
SI-IMEX(4, 4, 3)-WENO35								
	$\varepsilon = 1$		$\varepsilon = 10^{-2}$		$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-6}$	
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	5.5435e-06		2.4034e-07		2.4841e-07		2.5178e-07	
(80, 160)	1.9598e-07	4.82	1.5473e-08	3.96	8.0955e-09	4.94	8.1595e-09	4.95
(160, 320)	8.6077e-09	4.51	1.9379e-09	3.00	2.9470e-10	4.78	2.5731e-10	4.99
(320, 640)	6.0318e-10	3.84	2.5092e-10	2.95	3.4689e-11	3.09	8.1584e-12	4.98
(640, 1280)	6.2088e-11	3.28	3.2067e-11	2.97	9.1516e-12	1.92	3.3436e-13	4.61
GSA ARS(4, 4, 3)-WENO35								
	$\varepsilon = 1$		$\varepsilon = 10^{-2}$		$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-6}$	
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	1.6273e-05		1.2738e-06		1.1499e-06		1.1482e-06	
(80, 160)	8.5850e-07	4.24	4.2936e-08	4.89	3.5346e-08	5.02	3.5193e-08	5.02
(160, 320)	5.9885e-08	3.84	3.5424e-09	3.60	1.3376e-09	4.72	1.2315e-09	4.83
(320, 640)	6.2342e-09	3.26	4.7790e-10	2.88	8.0292e-10	0.73	5.6966e-11	4.43
(640, 1280)	7.4782e-10	3.06	6.2449e-11	2.93	4.7953e-10	0.74	6.3566e-12	3.16
IMEX-II-GSA3-WENO35								
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	1.3047e-05		1.2245e-06		1.1307e-06		1.1301e-06	
(80, 160)	4.2711e-07	4.93	3.5404e-08	5.11	3.2615e-08	5.11	3.2616e-08	5.11
(160, 320)	1.4238e-08	4.90	1.0951e-09	5.01	8.8667e-10	5.20	8.8936e-10	5.20
(320, 640)	2.4517e-09	2.53	1.1533e-10	3.24	2.3656e-11	5.22	2.1240e-11	5.38
(640, 1280)	3.2531e-10	2.92	1.5141e-11	2.93	5.8846e-12	2.00	3.5767e-12	2.57
IMEX-II-ISA3-WENO35								
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	1.5114e-05		1.2600e-06		1.1486e-06		1.1426e-06	
(80, 160)	6.8701e-07	4.45	3.9869e-08	4.98	3.4483e-08	5.05	3.4369e-08	5.05
(160, 320)	3.6504e-08	4.23	1.4366e-09	4.79	1.1093e-09	4.95	1.1169e-09	4.94
(320, 640)	3.2821e-09	3.47	7.6254e-11	4.23	4.9755e-11	4.47	4.4882e-11	4.63
(640, 1280)	3.7817e-10	3.11	7.0441e-12	3.43	4.4539e-12	3.48	4.7061e-12	3.25

NSE, and the numerical discretization error in approximating both models. With this view point, to show the error between ES-BGK and NSE is  $\mathcal{O}(\varepsilon^2)$ , we consider the same initial data used in Test 2, compute very accurate solutions for each model for a range of Knudsen numbers, and check whether the discrepancy of the numerical solutions show expected order  $\mathcal{O}(\varepsilon^2)$ . In view of this, we compute numerical solutions with sufficiently small  $\Delta t$  and  $\Delta x$  (for both ES-BGK and NSE) and large number of velocity grid points (for ES-BGK), and measure the discrepancy in relative  $L^1$ -norm. In the following Figure 1, it appears that the convergence rates are close to 2 when  $10^{-4} \leq \varepsilon \leq 10^{-2}$ . On the other cases, the second order is not obtained. Note that when  $\varepsilon$  is relatively small ( $\varepsilon < 10^{-4}$ ), the result is improved as we increase the accuracy of numerical solutions. However, when  $\varepsilon$  is relatively large ( $\varepsilon > 10^{-2}$ ), the errors are not improved even for high accurate solutions. This can be interpreted as modelling errors between ES-BGK and NSE.

TABLE 2. Accuracy test for the ES-BGK equation. Initial data is given in (4.4) with (4.5),  $L^1$  error on the density  $\rho$  at  $T = 0.25$ .

Relative $L^1$ error and order of density								
SI-IMEX (4, 4, 3)-WENO35 $N_v = 80$								
	$\varepsilon = 1$		$\varepsilon = 10^{-2}$		$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-6}$	
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	3.1641e-05		1.4962e-06		1.1980e-06		1.1999e-06	
(80, 160)	1.8285e-06	4.11	4.4260e-08	5.07	3.5725e-08	5.06	3.6019e-08	5.05
(160, 320)	8.6363e-08	4.40	1.7243e-09	4.68	1.0440e-09	5.09	1.1081e-09	5.02
(320, 640)	6.2789e-09	3.78	1.9614e-10	3.13	5.0039e-11	4.38	3.4509e-11	5.00
(640, 1280)	4.1337e-10	3.92	2.5833e-11	2.92	1.0886e-11	2.20	3.4304e-12	3.33
GSA ARS(4, 4, 3)-WENO35 $N_v = 32$								
	$\varepsilon = 1$		$\varepsilon = 10^{-2}$		$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-6}$	
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	5.8625e-06		8.7440e-08		6.1776e-08		6.2235e-08	
(80, 160)	4.0040e-07	3.87	5.0511e-09	4.11	2.3576e-09	4.71	2.4420e-09	4.67
(160, 320)	4.0934e-08	3.29	5.0181e-10	3.33	2.5187e-10	3.22	1.4842e-10	4.04
(320, 640)	4.8923e-09	3.06	6.0480e-11	3.05	1.1787e-10	1.09	1.4216e-11	3.38
(640, 1280)	6.0623e-10	3.01	7.5418e-12	3.00	4.6731e-11	1.33	1.8633e-12	2.93
IMEX-II-GSA3-WENO35 $N_v = 32$								
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	3.3509e-06		6.5315e-08		1.1307e-06		4.5811e-07	
(80, 160)	1.1158e-07	4.90	1.4444e-09	5.49	3.2615e-08	5.11	1.2718e-08	5.17
(160, 320)	1.5075e-08	2.88	1.4088e-10	3.35	8.8667e-10	5.20	3.8152e-10	5.05
(320, 640)	3.7304e-09	2.01	3.6735e-11	1.93	2.3656e-11	5.22	1.1960e-11	5.00
(640, 1280)	2.6209e-10	3.83	3.2301e-12	3.50	5.8846e-12	2.00	1.7831e-12	2.75
IMEX-II-ISA3-WENO35 $N_v = 32$								
$(N_x, 2N_x)$	$L_1$	Order	$L_1$	Order	$L_1$	Order	$L_1$	Order
(40, 80)	4.9602e-06		7.9449e-08		5.9187e-08		4.6149e-07	
(80, 160)	2.6934e-07	4.20	3.4997e-09	4.50	2.0718e-09	4.83	1.3395e-08	5.10
(160, 320)	2.3668e-08	3.50	2.5659e-10	3.76	1.0526e-10	4.29	4.7348e-10	4.82
(320, 640)	2.7055e-09	3.12	2.7401e-11	3.22	9.1594e-12	3.52	2.0956e-11	4.50
(640, 1280)	3.3206e-10	3.02	3.2314e-12	3.08	1.1581e-12	2.98	2.6896e-12	2.96

*Test 3. Riemann Problem.* In this problem, we consider a Riemann problem in 1D space and 2D velocity domain. The same test has been adopted in [18] to show the consistency between the ES-BGK model and BTE or NSE. As initial macroscopic states, we use

$$(\rho_0, u_{x0}, u_{y0}, T_0) = \begin{cases} (1, M\sqrt{2}, 0, 1), & -1 \leq x \leq 0.5, \\ (\frac{1}{8}, 0, 0, \frac{1}{4}), & \text{otherwise,} \end{cases}$$

where  $M = 2.5$  is the Mach number. Here we impose the free-flow boundary condition in space  $x \in [-1, 2]$ . We truncate the velocity domain with  $v_{\max} = 15$ . We compare numerical solutions at time  $t = 0.15, 0.2, 0.4$  with the time step that corresponds to CFL = 0.5. Here we set  $\nu = -1$  and  $\tau = 0.9\pi\rho/2$  following [18], which gives

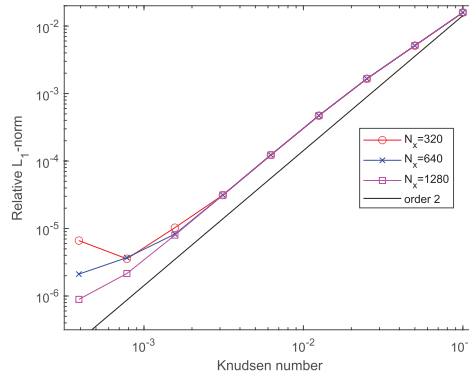


FIGURE 1. Knudsen number ( $\varepsilon$ ) vs. relative  $L_1$ -norm. For the comparison, we used the same initial data for test 2, and compute numerical solutions up to  $t = 0.5$ .

the following viscosity and heat conductivity:

$$\mu = \frac{1}{0.9\pi}T, \quad \kappa = \frac{1}{0.9\pi}T.$$

We consider these transport coefficients when computing the reference solutions to NSE. We remark that this choice matches the viscosity of ES-BGK model and the one derived from BTE for Maxwellian molecules [18].

In Figures 2 and 3, we present numerical results for BTE, NSE and ES-BGK model for different Knudsen numbers  $\varepsilon = 0.5, 0.1$  respectively. For BTE, we use an explicit fourth-order semi-Lagrangian scheme [11], while for NSE, we use the classical RK4 and WENO23. For ES-BGK, we use the combination of IMEX-II-ISA3 for the time discretization and WENO23 for the space one. For  $\varepsilon = 0.5$ , the numerical solution of the ES-BGK model is very close to that of BTE, while solutions to NSE are deviated from the other solutions. On the other hand, for  $\varepsilon = 0.1$ , the discrepancy between three solutions become smaller, which confirms the consistency of the three models.

*Test 4. The Lax Shock Tube Problem.* Finally we test IMEX-RK schemes solving the ES-BGK model for the Lax shock tube problem [21]. We set the collision frequency  $\tau$  for ES-BGK model as

$$\tau = \frac{2}{3}\rho\sqrt{T},$$

which yields the viscosity and heat conductivity:

$$\mu = \sqrt{T}, \quad \kappa = \frac{15}{4}\sqrt{T},$$

where  $\nu = -1/2$ , and this implies that the Prandtl number is  $2/3$ . We consider the initial macroscopic variables

$$\begin{pmatrix} \rho \\ u \\ p \end{pmatrix} = \begin{cases} (0.445, 0.698, 3.528)^\top, & -0.5 \leq x \leq 0, \\ (0.5, 0, 0.571)^\top, & 0 < x \leq 0.5, \end{cases}$$

on the free-flow condition  $x \in [-5, 5]$ . We take well-prepared initial conditions

$$f_0(x, v) = \mathcal{M}(x, v) - \frac{\varepsilon}{\tau}(I - \Pi_{\mathcal{M}})(v_1 \partial_x \mathcal{M}),$$

where  $I$  is the identity operator and  $\Pi_{\mathcal{M}}$  is the projection operator defined as (A.16). We first use  $80 \times 80$  uniform points for the velocity domain  $[-20, 20] \times [-20, 20]$ , and  $N_x = 200$  uniform points for the spatial discretization.

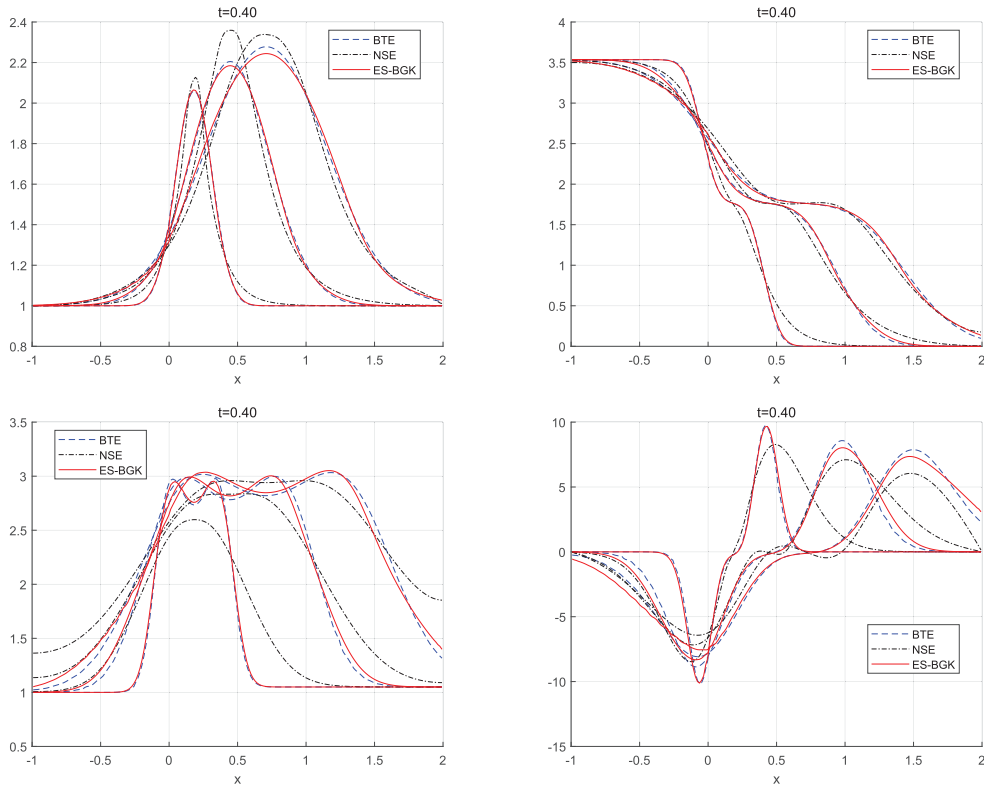


FIGURE 2. Comparison between reference solutions of BTE and NSEs with the numerical solutions for ES-BGK model. We report the case of density, velocity, temperature and heat flux at time  $t = 0.1, 0.25, 0.4$ . The Knudsen number is  $\varepsilon = 0.5$ .

The CFL number is taken as 0.2 and final time 1.3. We consider IMEX-II-ISA3 scheme coupled with WENO35 for solving the ES-BGK model. In Figure 4, we compare the numerical solution of the ES-BGK model with those of the compressible Navier–Stokes equations and the compressible Euler equations. The reference solution of the Euler equations and of the Navier–Stokes equations was generated by using spectral method for the spatial derivatives and fourth-order RK method for the time on a grid of 3200 for  $\varepsilon = 0, 10^{-2}, 10^{-4}$  respectively. As we can see from the Figure 4 the numerical solution of the ES-BGK model is very close to that of the NS equations. Furthermore for a better visualization, we zoomed the two solutions on a portion of the domain. In Figure 5 we reported the moments

$$\begin{aligned}
 -\mu\sigma(u) &= \int_{\mathbb{R}^{d_v}} f_1(v-u) \otimes (v-u) \, dv = \langle f_1(v-u) \otimes (v-u) \rangle, \\
 -\kappa\nabla T &= \int_{\mathbb{R}^{d_v}} f_1 \frac{1}{2}(v-u)|v-u|^2 \, dv = \langle f_1(v-u)|v-u|^2 \rangle,
 \end{aligned}
 \tag{4.6}$$

*i.e.*, the shear stress and heat flux, computed (1) from the function  $f_1 = \frac{g-f}{\varepsilon}$  (2) from the macroscopic quantities and (3) from the NS equation for the case of  $\varepsilon = 10^{-2}$  and  $\varepsilon = 10^{-4}$ .

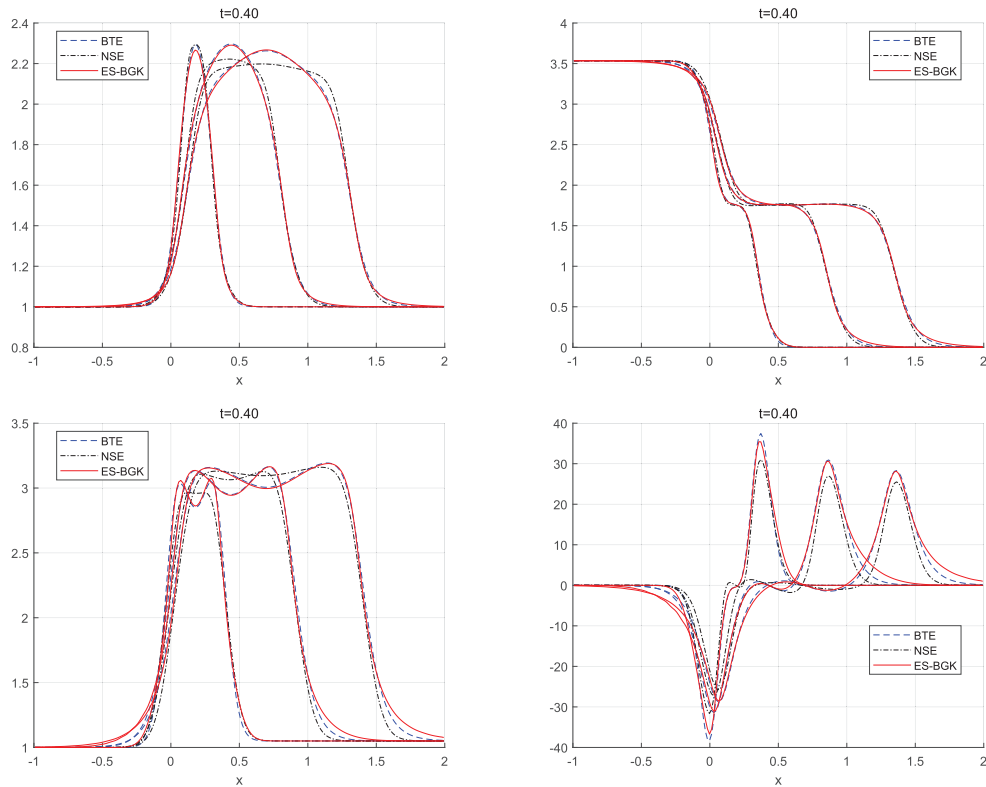


FIGURE 3. Comparison between reference solutions of BTE and NSEs with the numerical solutions for ES-BGK model. We report the case of density, velocity, temperature and heat flux at time  $t = 0.1, 0.25, 0.4$ . The Knudsen number is  $\varepsilon = 0.1$ .

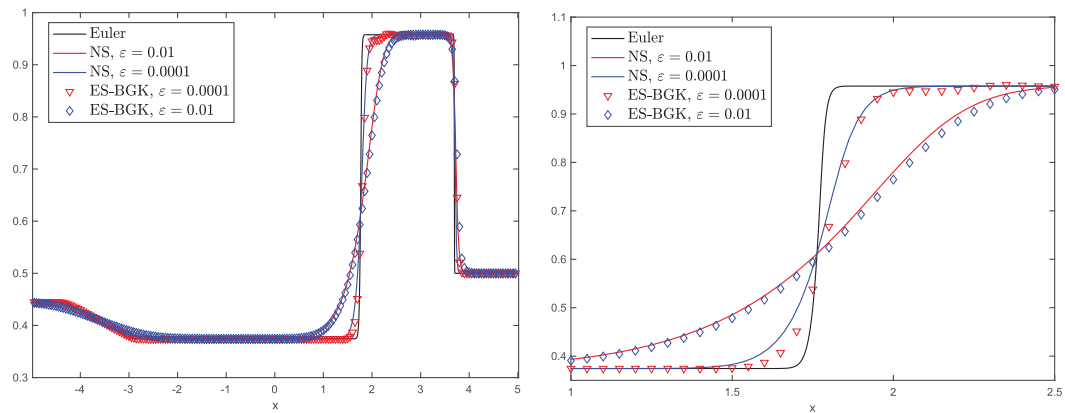


FIGURE 4. Comparison between reference solutions of Navier–Stokes equations and the numerical solutions of IMEX-II-GSA3 (left) and a zoom (right). Uniform  $40 \times 40$  points for the velocity domain  $[-20, 20] \times [-20, 20]$ .

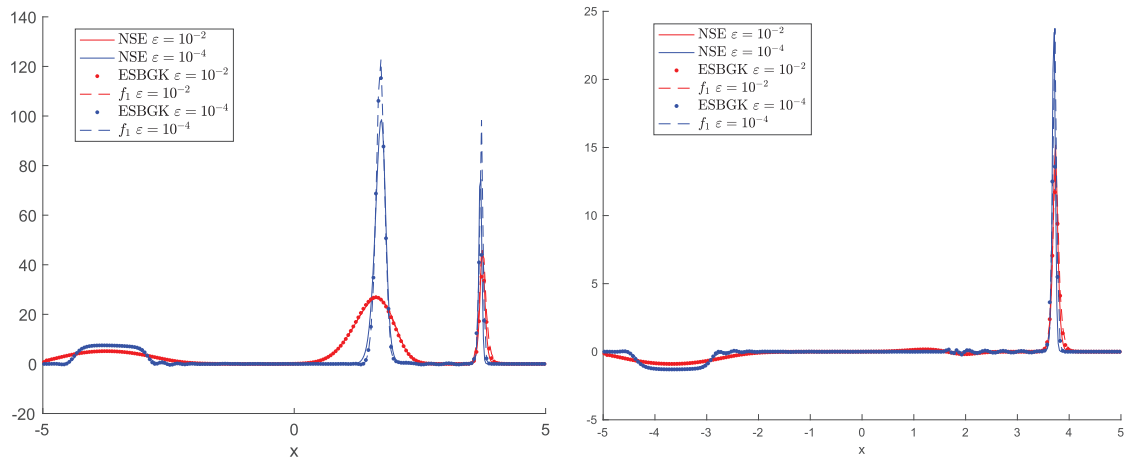


FIGURE 5. The shear stress and heat flux for  $\varepsilon = 10^{-2}, 10^{-4}$ . Uniform  $40 \times 40$  points for the velocity domain  $[-20, 20] \times [-20, 20]$ .

## 5. CONCLUSIONS

In this work, we investigated the asymptotic behavior at the Navier–Stokes level, reproducing theoretical results, Theorems 3.6 and 3.8. We showed that existing IMEX–RK schemes including type I and II can accurately capture the NS limit without resolving the small scales determined by the Knudsen number. To address the issues on order reduction of these schemes at the Navier–Stokes level, we propose IMEX–RK schemes of type II developed in [7]. In particular, the IMEX–II–ISA3 scheme of type ARS satisfying additional conditions (3.36) guarantees uniform accuracy avoiding order reduction. Finally, we provided numerical examples that validate the theoretical results.

### FUNDING

Sebastiano Boscarino is supported for this work by (1) the Spoke 1 “FutureHPC & BigData” of the Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC) funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di “campioni nazionali di R&S (M4C2-19)”, (CN00000013); by (2) the Italian Ministry of Instruction, University and Research (MIUR) to support this research with funds coming from PRIN Project 2022 (2022KA3JBA), entitled “Advanced numerical methods for time dependent parametric partial differential equations and applications”; (3) from Italian Ministerial grant PRIN 2022 PNRR “FIN4GEO: Forward and Inverse Numerical Modeling of hydrothermal systems in volcanic regions with application to geothermal energy exploitation”, (No. P2022BNB97). S. Boscarino is a member of the INdAM Research group GNCS. S. Y. Cho was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00166144), and Learning & Academic research institution for Master’s Ph.D students, and Post-docs (LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301974).

### DATA AVAILABILITY STATEMENT

The research data associated with this article are included in the article.

### REFERENCES

- [1] U.M. Ascher, S.J. Ruuth and R.J. Spiteri, Implicit–explicit Runge–Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.* **25** (1997) 151–167.

- [2] C. Bardos, F. Golse and D. Levermore, Fluid dynamic limits of kinetic equations. I. Formal derivations. *J. Stat. Phys.* **63** (1991) 323–344.
- [3] M. Bennoune, M. Lemou and L. Mieussens, Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier–Stokes asymptotics. *J. Comput. Phys.* **227** (2008) 3781–3803.
- [4] 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. *Phys. Rev.* **94** (1954) 511.
- [5] S. Boscarino, Error analysis of IMEX Runge–Kutta methods derived from differential-algebraic systems. *SIAM J. Numer. Anal.* **45** (2007) 1600–1621.
- [6] S. Boscarino, On an accurate third order implicit–explicit Runge–Kutta method for stiff problems. *Appl. Numer. Math.* **59** (2009) 1515–1528.
- [7] S. Boscarino and L. Pareschi, On the asymptotic properties of IMEX Runge–Kutta schemes for hyperbolic balance laws. *J. Comput. Appl. Math.* **316** (2017) 60–73.
- [8] S. Boscarino and G. Russo, On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *SIAM J. Sci. Comput.* **31** (2009) 1926–1945.
- [9] S. Boscarino and G. Russo, Flux-explicit IMEX Runge–Kutta schemes for hyperbolic to parabolic relaxation problems. *SIAM J. on Numer. Anal.* **51** (2013) 163–190.
- [10] S. Boscarino, L. Pareschi and G. Russo, Implicit–explicit, Runge–Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit. *SIAM J. Sci. Comput.* **35** (2013) A22–A51.
- [11] S. Boscarino, S.Y. Cho and G. Russo, A conservative semi-Lagrangian method for inhomogeneous Boltzmann equation. *J. Comput. Phys.* **496** (2024) 112633.
- [12] S. Boscarino, L. Pareschi and G. Russo, Implicit–explicit methods for evolutionary partial differential equations (2024).
- [13] C. Cercignani, Rarefied Gas Dynamics. Cambridge Texts in Applied Mathematics (2000).
- [14] S. Chapman and T.G. Cowling, The Mathematical Theory of Non-Uniform Gases: An Account of the Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases. Cambridge University Press (1990).
- [15] B. Cockburn, C.-W. Shu, C. Johnson, E. Tadmor and C.-W. Shu, Essentially Non-Oscillatory and Weighted Essentially Non-Oscillatory Schemes for Hyperbolic Conservation Laws. Springer (1998).
- [16] G. Dimarco and L. Pareschi, Asymptotic preserving implicit–explicit, Runge–Kutta methods for nonlinear kinetic equations. *SIAM J. Numer. Anal.* **51** (2013) 1064–1087.
- [17] G. Dimarco and L. Pareschi, Implicit–explicit linear multistep methods for stiff kinetic equations. *SIAM J. Numer. Anal.* **55** (2017) 664–690.
- [18] F. Filbet and S. Jin, An asymptotic preserving scheme for the ES-BGK model of the Boltzmann equation. *J. Sci. Comput.* **46** (2011) 204–224.
- [19] F. Golse, The Boltzmann equation and its hydrodynamic limits, in Handbook of Differential Equations: Evolutionary Equations. Vol. 2. Elsevier (2005) 159–301.
- [20] L.H. Holway, Jr., Kinetic theory of shock structure using an ellipsoidal distribution function. *Rare. Gas. Dyn.* **1** (1965) 193.
- [21] J. Hu and X. Zhang, On a class of implicit–explicit, Runge–Kutta schemes for stiff kinetic equations preserving the Navier–Stokes limit. *J. Sci. Comput.* **73** (2017) 797–818.
- [22] S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review, in Lecture Notes for Summer School on Methods and Models of Kinetic Theory (M&MKT), Porto Ercole (Grosseto, Italy) (2010) 177–216.
- [23] C.A. Kennedy and M.H. Carpenter, Additive Runge–Kutta schemes for convection–diffusion–reaction equations. *Appl. Numer. Math.* **44** (2003) 139–181.
- [24] L. Pareschi and G. Russo, Implicit–explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.* **25** (2005) 129–155.
- [25] G. Wanner and E. Hairer, Solving Ordinary Differential Equations II. Vol. 375. Springer Berlin Heidelberg New York (1996).
- [26] T. Xiong, J. Jang, F. Li and J.-M. Qiu, High order asymptotic preserving nodal discontinuous Galerkin IMEX schemes for the BGK equation. *J. Comput. Phys.* **284** (2015) 70–94.



**Please help to maintain this journal in open access!**

This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting [subscribers@edpsciences.org](mailto:subscribers@edpsciences.org).

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.

APPENDIX A. APPENDIX

**A.1. Basic properties of Gaussian distribution function**

By definition,  $\mathcal{G}[f]$  satisfies

$$\begin{aligned} \langle f \rangle &= \int_{\mathbb{R}^{d_v}} f(v) \, dv = \int_{\mathbb{R}^{d_v}} \mathcal{G}[f](v) \, dv = \langle \mathcal{G}[f] \rangle = \rho, \\ \langle vf \rangle &= \int_{\mathbb{R}^{d_v}} vf(v) \, dv = \int_{\mathbb{R}^{d_v}} v\mathcal{G}[f](v) \, dv = \langle v\mathcal{G}[f](v) \rangle = \rho u, \\ \langle \frac{|v|^2}{2} f \rangle &= \int_{\mathbb{R}^{d_v}} \frac{|v|^2}{2} f(v) \, dv = \int_{\mathbb{R}^{d_v}} \frac{|v|^2}{2} \mathcal{G}[f](v) \, dv = \langle \frac{|v|^2}{2} \mathcal{G}[f](v) \rangle = E. \end{aligned} \tag{A.1}$$

Note that from

$$\begin{aligned} \int_{\mathbb{R}^{d_v}} (v-u) \otimes (v-u) f(v) \, dv &= \rho \Theta, \\ \int_{\mathbb{R}^{d_v}} (v-u) \otimes (v-u) \mathcal{G}[f](v) \, dv &= \rho \mathcal{T}, \end{aligned} \tag{A.2}$$

we further have

$$\begin{aligned} f = \mathcal{G}[f] &\iff f = \mathcal{M}[f]; \\ f = \mathcal{G}[f] + \mathcal{O}(\varepsilon) &\text{ implies } \mathcal{G}[f] = \mathcal{M}[f] + \mathcal{O}(\varepsilon). \end{aligned} \tag{A.3}$$

The proof is straightforward. In fact, on both side of  $f = \mathcal{G}[f]$  using (A.2) we get  $\Theta = \mathcal{T}$  and from (1.9) we get  $\mathcal{T} = TId$  hence  $\mathcal{G}[f]$  is just the isotropic Maxwellian  $\mathcal{M}[f]$ . Similarly if on both side of  $f = \mathcal{M}[f]$  takes (A.2). For the second point in (A.3) the proof is similar and we omit the details.

**Lemma A.1.** *Let  $U_f = (\rho_f, \rho_f u_f, E_f)$  and  $U_g = (\rho_g, \rho_f u_g, E_g)$  be macroscopic variables associated to  $f$  and  $g$ , respectively. Then, for a positive constant  $\delta$  the assumption*

$$\|U_f - U_g\|_\infty = \mathcal{O}(\delta)$$

*implies that*

$$|\mathcal{M}[f] - \mathcal{M}[g]| = \mathcal{O}(\delta).$$

*Proof.* The differential form of  $\mathcal{M}$  satisfies

$$d\mathcal{M} = \mathcal{M} \left[ \frac{1}{\rho} d\rho + \frac{(v-u)}{T} \cdot du + \left( \frac{|v-u|^2}{2T^2} - \frac{d_v}{2T} \right) dT \right].$$

This further implies that

$$\mathcal{M}[f] - \mathcal{M}[g] = \mathcal{M}[g] \left[ \frac{1}{\rho_g} (\rho_f - \rho_g) + \frac{(v-u_g)}{T_g} \cdot (u_f - u_g) + \left( \frac{|v-u_g|^2}{2T_g^2} - \frac{d_v}{2T_g} \right) (T_f - T_g) \right] + \mathcal{O}(\delta^2).$$

This gives the desired estimate. □

### A.2. Chapman–Enskog expansion of ES-BGK model

Here we perform the first order Chapman–Enskog expansion of ES-BGK model with respect to  $\varepsilon$ :

$$f = \mathcal{M}[f] + \varepsilon f_1, \quad (\text{A.4})$$

which implies that  $f_1$  satisfies the so-called compatibility relations:

$$\langle \phi f_1 \rangle = 0, \quad \phi = (1, v, |v|^2). \quad (\text{A.5})$$

Now we take the expansion in terms of  $\varepsilon$  for the stress tensor

$$\Theta = TId + \varepsilon \Theta_1, \quad (\text{A.6})$$

and the heat flux

$$\mathbb{Q} := \left\langle \frac{|v-u|^2}{2} (v-u) f \right\rangle = 0 + \varepsilon \mathbb{Q}_1 \quad (\text{A.7})$$

with

$$\Theta_1 = \frac{1}{\rho} \int_{\mathbb{R}^{d_v}} f_1 (v-u) \otimes (v-u) dv = \frac{1}{\rho} \langle f_1 (v-u) \otimes (v-u) \rangle, \quad (\text{A.8})$$

and

$$\mathbb{Q}_1 = \int_{\mathbb{R}^{d_v}} f_1 \frac{1}{2} (v-u) |v-u|^2 dv = \left\langle f_1 \frac{1}{2} (v-u) |v-u|^2 \right\rangle. \quad (\text{A.9})$$

Inserting (A.4) and these latter expansions for  $\Theta$  and  $\mathbb{Q}$  into the conservation laws

$$\langle \phi f \rangle + \langle \phi v \cdot \nabla_x f \rangle = 0,$$

or

$$\langle \phi \mathcal{M}[f] \rangle + \langle \phi v \cdot \nabla_x \mathcal{M}[f] \rangle = -\varepsilon \nabla_x \cdot \langle \phi v f_1 \rangle,$$

it gives the equations:

$$\partial_t U + \nabla_x F(U) = -\varepsilon \nabla_x \cdot H(\mathbf{U}) \quad (\text{A.10})$$

with

$$U = \langle \phi f \rangle = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}, \quad F(U) = \begin{pmatrix} \rho u \\ \rho u \otimes u + \rho T Id \\ (E + \rho T)u \end{pmatrix}, \quad H(\mathbf{U}) = \begin{pmatrix} 0 \\ \rho \Theta_1 \\ \mathbb{Q}_1 + \rho \Theta_1 u \end{pmatrix}. \quad (\text{A.11})$$

To evaluate the quantities  $\rho \Theta_1$  and  $\mathbb{Q}_1$ , we seek the form of  $f_1$ . For this aim, we consider the expansion of the anisotropic Gaussian  $\mathcal{G}(f)$  with respect to  $\varepsilon$ , *i.e.*,

$$\mathcal{G}[f] = \mathcal{M}[f] + \varepsilon g. \quad (\text{A.12})$$

By (1.9) and considering the expansion (A.6), we get  $\mathcal{T} = TId + \nu \varepsilon \Theta_1$ , and using the fact that  $tr(\Theta_1) = 0$ , this gives  $\det(\mathcal{T}) = T^{d_v} + \mathcal{O}(\varepsilon^2)$  and

$$\mathcal{T}^{-1} = \frac{1}{T} \left( Id - \frac{\nu \varepsilon}{T} \Theta_1 \right) + \mathcal{O}(\varepsilon^2),$$

and therefore by (1.8) we obtain

$$g := \frac{\mathcal{G}[f] - \mathcal{M}[f]}{\varepsilon} = \nu \frac{\mathcal{M}[f]}{2T^2} (v-u)^\top \Theta_1 (v-u). \quad (\text{A.13})$$

where we neglect the term  $\mathcal{O}(\varepsilon^2)$ . Now the next step is to insert expansions (A.4) and (A.12) into the equation

$$\partial_t f + v \cdot \nabla_x f = \frac{\tau}{\varepsilon} (\mathcal{G}[f] - f).$$

Then, we get

$$\tau(f_1 - g) = -(\partial_t \mathcal{M} + v \cdot \nabla_x \mathcal{M} + \varepsilon(\partial_t f_1 + v \cdot \nabla_x f_1)). \quad (\text{A.14})$$

Before proceeding, we review a Hilbert space and an associated operator considered in [3] and some lemmas.

**Remark A.2.** Given a Maxwellian function  $\mathcal{M}[f]$ ,  $\Pi_{\mathcal{M}}(f)$  is the orthogonal projection in the Hilbert space  $L^2_{\mathcal{M}} = \{\phi \text{ such that } \phi \mathcal{M}^{-1/2} \in L_2(\mathbb{R}^{d_v})\}$  onto

$$\mathcal{N} = \text{Span}\{\mathcal{M}, v\mathcal{M}, |v|^2\mathcal{M}\},$$

equipped with the following weighted inner product

$$\langle \phi \psi \rangle_{\mathcal{M}} = \langle \phi \psi \mathcal{M}^{-1} \rangle = \int_{\mathbb{R}^{d_v}} \phi \psi \mathcal{M}^{-1} dv.$$

Then, by using the orthogonal basis of  $\mathcal{N}$

$$\mathcal{B} = \left\{ \frac{1}{\rho} \mathcal{M}, \frac{(v-u)}{\sqrt{T}} \frac{1}{\rho} \mathcal{M}, \left( \frac{|v-u|^2}{2T} - \frac{d_v}{2} \right) \frac{1}{\rho} \mathcal{M} \right\}, \quad (\text{A.15})$$

its explicit form is written as

$$\Pi_{\mathcal{M}}(f) = \frac{1}{\rho} \left[ \langle f \rangle + \frac{(v-u) \cdot \langle (v-u)f \rangle}{T} + \left( \frac{|v-u|^2}{2T} - \frac{d_v}{2} \right) \frac{2}{d_v} \left\langle \left( \frac{|v-u|^2}{2T} - \frac{d}{2} \right) f \right\rangle \right] \mathcal{M}. \quad (\text{A.16})$$

Note that a direct calculation gives  $\Pi_{\mathcal{M}}(\partial_t \mathcal{M}) = 0$  and  $\Pi_{\mathcal{M}}(f_1 - g) = f_1 - g$ . Then, from (A.14) we have

$$\begin{aligned} \tau(f_1 - g) &= (I - \Pi_{\mathcal{M}})(\tau(f_1 - g)) \\ &= -(I - \Pi_{\mathcal{M}})(\partial_t \mathcal{M} + v \cdot \nabla_x \mathcal{M}) + \mathcal{O}(\varepsilon) \\ &= -(I - \Pi_{\mathcal{M}})(v \cdot \nabla_x \mathcal{M}) + \mathcal{O}(\varepsilon). \end{aligned} \quad (\text{A.17})$$

Moreover, by Golse [19] and Filbet and Jin [18] it follows that

$$(I - \Pi_{\mathcal{M}})(v \cdot \nabla_x \mathcal{M}) = \mathcal{M} \left( A(V) : \frac{\sigma(u)}{2} + 2B(V) \cdot \nabla_x \sqrt{T} \right) + \mathcal{O}(\varepsilon), \quad (\text{A.18})$$

where

$$V = \frac{v-u}{\sqrt{T}}, \quad A(V) = V \otimes V - \frac{1}{d_v} |V|^2 Id, \quad B(V) = \frac{1}{2} V (|V|^2 - (d_v + 2)), \quad (\text{A.19})$$

and

$$\sigma(u) = \nabla_x u + (\nabla_x u)^\top - \frac{2}{d_v} \nabla_x \cdot u Id.$$

To sum up, the relations (A.14), (A.18) and (A.17) imply that

$$\partial_t \mathcal{M} + v \cdot \nabla_x \mathcal{M} = \mathcal{M} \left( A(V) : \frac{\sigma(u)}{2} + 2B(V) \cdot \nabla_x \sqrt{T} \right) + \mathcal{O}(\varepsilon). \quad (\text{A.20})$$

**Lemma A.3.** For  $\phi = (1, v, |v|^2/2)^\top$ , it follows

$$\langle v \phi f_1 \rangle = H(U), \quad \text{where } H(U) = (0, \rho \Theta_1, \mathbb{Q}_1 + \rho \Theta_1 u)^\top, \quad (\text{A.21})$$

with  $U = (\rho, \rho u, E)^\top$ .

**Lemma A.4.** For  $\phi = (1, v, |v|^2/2)^\top$ , it follows

$$\langle v \phi g \rangle = G(U), \quad \text{where } G(U) = (0, \nu \rho \Theta_1, \nu \rho \Theta_1 u)^\top, \quad (\text{A.22})$$

with  $U = (\rho, \rho u, E)^\top$ .

**Lemma A.5.** For  $\phi = (1, v, |v|^2/2)^\top$ , it follows

$$\langle v\phi(I - \Pi_{\mathcal{M}})(v \cdot \nabla \mathcal{M}) \rangle = \left\langle v\phi \mathcal{M} \left( A(V) : \frac{\sigma(u)}{2} + 2B(V) \cdot \nabla_x \sqrt{T} \right) \right\rangle = \mathcal{F}(U) \quad (\text{A.23})$$

where

$$\mathcal{F}(U) = \left( 0, \rho T \sigma(u), \rho T \sigma(u)u + \frac{d_v + 2}{2} \rho T \nabla T \right)^\top,$$

with  $U = (\rho, \rho u, E)^\top$ .

By Remark A.2, we can rewrite (A.14) as

$$f_1 = g - \frac{1}{\tau} \left( \mathcal{M} \left( A(V) : \frac{\sigma(u)}{2} + 2B(V) \cdot \nabla_x \sqrt{T} \right) \right) + \mathcal{O}(\varepsilon). \quad (\text{A.24})$$

Next, we multiply both sides of (A.24) by  $v\phi$  and take integration and use Lemmas A.3–A.5 to obtain

$$\begin{aligned} \langle v\phi f_1 \rangle &= \langle v\phi g \rangle - \frac{1}{\tau} \left\langle v\phi \mathcal{M} \left( A(V) : \sigma(u) + 2B(V) \cdot \nabla_x \sqrt{T} \right) \right\rangle + \mathcal{O}(\varepsilon), \\ H(U) &= G(U) - \frac{1}{\tau} \mathcal{F}(U) + \mathcal{O}(\varepsilon), \end{aligned} \quad (\text{A.25})$$

that is,

$$\begin{pmatrix} 0 \\ \rho \Theta_1 \\ \mathbb{Q}_1 + \rho \Theta_1 u \end{pmatrix} = \begin{pmatrix} 0 \\ \nu \rho \Theta_1 \\ \nu \rho \Theta_1 u \end{pmatrix} - \frac{1}{\tau} \begin{pmatrix} 0 \\ \rho T \sigma(u) \\ \rho T \sigma(u)u + \frac{d_v + 2}{2} \rho T \nabla_x T \end{pmatrix} + \mathcal{O}(\varepsilon). \quad (\text{A.26})$$

Then, we get

$$\rho \Theta_1 = -\mu \sigma(u) + \mathcal{O}(\varepsilon),$$

with viscosity  $\mu = \frac{\rho T}{(1-\nu)\tau}$  and

$$\mathbb{Q}_1 = -\kappa \nabla_x T + \mathcal{O}(\varepsilon),$$

with the thermal conductivity

$$\kappa = \frac{d_v + 2}{2} \frac{p}{\tau},$$

where  $p = \rho T$ . Finally, from (A.10) we derive the CNS equations

$$\partial_t U + \nabla_x \cdot F(U) = \varepsilon \nabla_x \cdot S(U), \quad (\text{A.27})$$

with

$$\begin{aligned} S(U) &= \begin{pmatrix} 0 \\ \mu \sigma(u) \\ \mu \sigma(u)u + \kappa \nabla_x T \end{pmatrix}, \\ \sigma(u) &= \nabla_x u + (\nabla_x u)^\top - \frac{2}{d_v} \nabla_x \cdot u I. \end{aligned} \quad (\text{A.28})$$