Year-Five Progress Report on Numerical Methods for BGK-Type Kinetic Equations

Evan Habbershaw[‡], Jack Haight[†], and Steven M. Wise[†]

[‡]Department of Mathematics, The Pennsylvania State University, University Park, PA 16802

[†]Department of Mathematics, The University of Tennessee, Knoxville, TN, 37996

Reporting Period: August 1, 2025–July 31, 2026

Prepared for Oak Ridge National Laboratory
ORNL Subcontract UTB-CW24420

Version 1.2 — October 10, 2026

Corresponding author: Steven M. Wise, swise1@utk.edu

Abstract

This cumulative Year-Five progress report, which adds the most recent support period, August 1, 2025 through July 31, 2026, documents research supported by Oak Ridge National Laboratory on the analysis and numerical approximation of multispecies kinetic equations. Relative to the Year Four Report, the multispecies Boltzmann chapter is not new; it has been substantially expanded and reorganized to provide a more self-contained account of collision structure, conservation laws, entropy production (equivalently, information dissipation), moment equations, and equilibrium. Two new chapters develop the Chapman–Enskog derivation of the Euler and Navier–Stokes–Fourier equations and survey finite moment-closure methods. The chapter on unary and multispecies ellipsoidal statistical BGK (ES-BGK) models has also been substantially updated, including a positivity-preserving implicit collision update and a careful statement of the additional hypotheses required for mixture models. The principal new research direction is dynamical low-rank approximation (DLRA): a new chapter develops its mathematical and computational foundations and presents simulations for BGK and high-resolution finite-volume algorithms. Previously reported results on positivity, well-posedness, asymptotic relaxation, and an implicit nonlinear moment solver for a space-homogeneous multispecies BGK model are retained, together with earlier material on BGK models, finite-volume and discrete-velocity discretizations, and prototype computations. The cumulative format is itself an important project outcome: continued consolidation, correction, and polishing make the report a durable teaching resource for future graduate students who join the project and a common technical reference for the research team. These additions and updates provide a foundation for integrating implicit moment solvers, ES-BGK collision models, and low-rank representations in full transport simulations.

Keywords: Boltzmann equation, BGK and ES-BGK approximations, multispecies BGK and ES-BGK models, Chapman–Enskog expansion, moment closure, finite-volume schemes, implicit-explicit time stepping, discrete-velocity methods, dynamical low-rank approximation, asymptotic relaxation.

Contents

List of Figures **vii**

List of Tables **xi**

Abbreviations and Acronyms **xii**

Acknowledgements **xiii**

Executive Summary **1**

What Changed in Year Five 2

Selected Verifiable Indicators 3

Milestones 3

1 Introduction and Report Organization **5**

1.1 A Toy Model 5

1.2 Quick Introduction to the Unary Boltzmann Equation 8

1.3 The BGK Approximation Model 10

1.4 Organization 11

1.5 Terminology and Notation Conventions 12

1.6 Review Status and Limitations 13

2 Multispecies Boltzmann Equations **14**

2.1 The Binary Collision Transformation 15

2.2 The Boltzmann Collision Operator 21

2.3 Collision Invariants 28

2.4 Information, Entropy, and Boltzmann's H-Theorem 34

2.5 Moments and Moment Equations 42

2.6 The Maxwellian and Maximal Entropy States 50

2.7 Space-Homogeneous Model 54

3 Introduction to the Chapman–Enskog Analysis 58
3.1 Formal Status and Hydrodynamic Scaling 59
3.2 Unary Macroscopic Variables and the Closure Problem 62
3.3 Local Maxwellians and the Euler Closure 63
3.4 The Linearized Collision Operator 65
3.5 The Chapman–Enskog Ansatz 69
3.6 The First-Order Source 77
3.7 Viscosity and Heat Conduction 79
3.8 The Compressible Navier–Stokes–Fourier System 82
3.9 What Has and Has Not Been Proved 84

4 Moment Closure Methods 86
4.1 Finite Collections of Moments 87
4.2 Reconstruction and Closure 88
4.3 The Five-Moment Maxwellian Closure 90
4.4 Grad's Polynomial Moment Method 91
4.5 Maximum-Entropy Closure 93
4.6 Quadrature and Discrete-Velocity Closures 96
4.7 Local, Differential, and Extended Closures 96
4.8 What Should a Closure Preserve? 97
4.9 Extension to Multispecies Systems 99
4.10 Summary of the Methodology 99

5 Single-Species BGK Kinetic Models 101
5.1 Basic Properties of Solutions 101
5.1.1 Collision Invariants 102
5.1.2 Space-Homogeneous Problem 106
5.1.3 The H-Theorem 108
5.2 Discrete-Velocity BGK Models 112
5.2.1 Discrete-Velocity BGK Model in One Space Dimension 112
5.2.2 Space-Homogeneous Discrete-Velocity Model 120

6 Numerics for the Single-Species BGK Model 123
6.1 Finite Volume Space and Velocity Discretization 123

6.2 Implicit-Explicit Runge–Kutta Time Stepping . . . 127
6.3 Fully Discrete Scheme . . . 129
6.4 Poisson Solver for Vlasov–Poisson–BGK Equation . . . 130
6.5 Sample Computations and Accuracy Tests . . . 132
6.5.1 Relaxation Test 1 . . . 132
6.5.2 Relaxation Test 2 . . . 135
6.5.3 Sod Shock Tube (Euler Equation Limit of BGK) . . . 137
6.5.4 Square Pulse Rotation . . . 140
6.5.5 Two-Stream Instability (Vlasov–Poisson) . . . 142
6.5.6 Landau Damping (Vlasov–Poisson) . . . 143
6.5.7 Quantitative Evidence and Reproducibility Status . . . 144

7 Discrete-Velocity Numerical Methods for the Single-Species BGK Problem 145
7.1 A Fully Explicit Scheme . . . 145
7.2 An Explicit Scheme for the Space-Homogeneous Case . . . 148
7.3 An Implicit Scheme for the Space-Homogeneous Case . . . 150

8 Multispecies BGK Equations 153
8.1 Boltzmann Structure to Be Preserved . . . 153
8.2 Formal Ad Hoc Construction from the Boltzmann Operator . . . 154
8.3 Pair Parameters and Their Positivity . . . 157
8.4 Conservation Laws . . . 159
8.5 H-Theorem and Equilibria . . . 160
8.6 Fluid-Limit Connection and Scope . . . 161

9 Asymptotic Relaxation and Numerics of Multispecies Moment Equations 162
9.1 Model Equations . . . 163
9.1.1 Model Properties . . . 164
9.1.2 Moment Equations . . . 165
9.2 Properties of the ODE System . . . 167
9.2.1 Positivity of the Temperature . . . 167
9.2.2 Velocity Bounds . . . 169
9.2.3 Existence and Uniqueness . . . 170
9.3 Long-time Behavior . . . 172
9.3.1 Null Spaces of $D - A$, Z, $F - B$, and $\widehat{Z}$ . . . 174
9.3.2 Velocity Relaxation Proof . . . 175

9.3.3 Energy Relaxation Proof . . . 176
9.3.4 Temperature Relaxation . . . 179
9.4 Numerical Confirmation of Theory . . . 180
9.5 A Novel Nonlinear GST Iterative Solver . . . 184
9.5.1 Sod problem with two identical species . . . 185
9.5.2 Three Species Mixture: Ar–Kr–Xe . . . 187
9.6 Conclusions . . . 190

10 ES-BGK Models and Numerics 192
10.1 Purpose and Organization . . . 192
10.2 The Single-Species ES-BGK Model . . . 192
10.3 Chapman–Enskog Expansion . . . 200
10.4 Implicit Collision Update . . . 207
10.5 Numerical Demonstration of the IMEX Method . . . 209
10.5.1 Slab problem and initial data . . . 209
10.5.2 Discretization and parameters . . . 210
10.6 Multispecies ES-BGK Models . . . 212
10.6.1 Pair Gaussians and their moments . . . 212
10.6.2 The homogeneous moment system . . . 214
10.6.3 Restricted tensor model . . . 216
10.6.4 General tensor model . . . 217
10.6.5 Nonlinear moment-first GST solve . . . 219
10.6.6 Distribution update and admissibility . . . 221
10.7 Conclusions . . . 222

11 Dynamical Low-Rank Approximation Methods 223
11.1 Foundations of Dynamical Low-Rank Approximation . . . 223
11.1.1 Matrix Differential Equations and Low-Rank Factorizations . . . 223
11.1.2 The Fixed-Rank Matrix Manifold . . . 229
11.1.3 Tangent-Space Projection and Factor Evolution Equations . . . 239
11.1.4 Projector-Splitting and BUG Integrators . . . 241
11.1.5 Continuous Dynamical Low-Rank Approximation . . . 256
11.1.6 Computational Benefits and Limitations . . . 261
11.2 Projector-Splitting DLRA for the $1x1v$ BGK Equation . . . 262
11.2.1 The $1x1v$ BGK Model and Its Moments . . . 262

11.2.2 Finite-Volume Matrix Formulation and DLRA Update 270
11.2.3 Sod Shock-Tube Experiments 275
11.3 DLRA for Slope-Limited Collisionless Transport 279
11.3.1 Slope-Limited Finite-Volume Discretization 279
11.3.2 Projector-Splitting Time Integration 286
11.3.3 Rank-Growth and Accuracy Experiments 286
11.4 Conclusions 302

12 Progress Summary and Next Steps **304**
12.1 Year-Five Progress 304
12.2 Numerical Prototypes at a Glance 306
12.3 Next Steps 306

A An Important Integral **308**

B Picard–Lindelöf Theorem **310**

C Bounding the Eigenvalues of Z and $\widehat{Z}$ **311**

D Null Space Components of $\widetilde{W}$ and $\widetilde{\boldsymbol{\xi}}$ **313**

E Some Theory of Manifolds **314**
E.1 Definitions 314
E.2 Theorems 315

F 1x1v Prototype Code **318**
F.1 Code Availability, Run Connection, and Verification Checklist 318
F.2 Main Driver: `vlasovPoissonBGKMain.m` 319
F.3 `vlasovPoissonBGKSolver.m` 326
F.4 `BGKCollision.m` 330
F.5 `divFlux.m` 333
F.6 `slopeReconstruction.m` 335
F.7 `minMod.m` 337
F.8 `sodSoln.m` 338
F.9 `pStarSolve.m` 342

G Bulk-Quantity Notation Map **344**
G.1 Kinetic Inputs, Densities, and Velocity Frames 345
G.2 Energy Densities and Temperatures 346
G.3 Stress and Flux Quantities 348
G.4 Information and Entropy Quantities 349
G.5 Conserved Variables and Special Cases 350

References **351**

List of Figures

1.1 A fully discrete particle-count analogy. Each printed integer is the exact value of $f(i, v, t^n)$. Positive-velocity particles tend to move toward larger spatial indices, negative-velocity particles toward smaller indices, and collisions may reassign velocity labels without changing spatial cells. The displayed counts sum to 35 in both panels. 6

1.2 Velocity marginals of the two discrete phase-space states. Each bar is the sum of one velocity column in Figure 1.1. Transport redistributes particles among spatial cells, while the sum of the bar heights remains 35. 7

2.1 An equal-mass binary collision with $\boldsymbol{\sigma} = 2^{-1/2}(1, -1)^\top$. The components parallel to $\boldsymbol{\sigma}$ are exchanged; in this particular configuration, the two velocity vectors are exchanged completely. 17

6.1 Relaxation Test 1 for the space-homogeneous BGK model (5.24): initial data (6.47) (left) and the numerical solution at $t_\mathrm{f} = 1$ (right). Here $\lambda = 1$ and $N_x = N_v = 128$. 134

6.2 Relaxation Test 1: signed pointwise error $f_\mathrm{exact} - f_\mathrm{num}$ at $t_\mathrm{f} = 1$ for $\lambda = 1$ and $N_x = N_v = 128$. The displayed error has magnitude $\mathcal{O}(10^{-5})$. 134

6.3 Relaxation Test 2 for the space-homogeneous BGK model (5.24): four-Gaussian initial data (6.48) and numerical solutions at $t = 1$ and $t = 2$. Here $\lambda = 1$, and the initial moments are evaluated with the trapezoidal rule. 135

6.4 Relaxation Test 2: signed pointwise errors $f_\mathrm{exact} - f_\mathrm{num}$ at $t = 1$ and $t = 2$ for $\lambda = 1$ and the four-Gaussian initial data (6.48). 136

6.5 Sod shock tube: initial distribution and computed phase-space distribution at $t_\mathrm{f} = 0.2$ for the BGK model with $\tau = 10^{-4}$ on $\Omega \times V = [-0.75, 0.75] \times [-10, 10]$. 138

6.6 Sod shock tube at $t_f = 0.2$ with $\tau = 10^{-4}$: BGK moments (number density, bulk velocity, temperature, and pressure) compared with the exact Euler Riemann solution. The agreement tests the highly collisional Euler limit; a norm-based mesh study is reserved for the code-verification pass described in Table 6.1. 139
6.7 Rotation Problem: $\frac{\partial}{\partial t}f + v\frac{\partial}{\partial x}f - x\frac{\partial}{\partial v}f = 0$. Grid: $N_v = Nx = 400$. 141
6.8 Two-Stream instability. The numerical solution is shown at $t_f = 5$ (left) and $t_f = 45$ (right). The numerical grid has $N_v \times N_x = 800 \times 800$ cells on the grid $\Omega \times V = [-2\pi, 2\pi] \times [-2\pi, 2\pi]$. The fine filamentation of the particle streams is adequately resolved and captured. 142
6.9 Landau damping for $\alpha = 0.01$ and $k = 0.5$: the computed L^2 norm of the electric field follows the reference exponential decay rate -0.1533 over the displayed interval. A regression estimate and fit interval will be reported when the archived output data are regenerated. 143

9.1 Temperature Decay of Ar–Kr–Xe Mixture in Example 1. 181
9.2 Velocity and Temperature Decay of Ar–Kr–Xe Mixture in Example 2. 182
9.3 Velocity and Temperature Decay of He–Kr–Xe Mixture in Example 3. 183
9.4 The total density, total bulk velocity, and total temperature moments are compared to the solution of the Euler equations at time $t = 0.2$, with $\varepsilon = 10^{-4}$.186
9.5 The number of iterations required for convergence is decreased by the application of Anderson Acceleration (AA) to the GST algorithm. The maximum iteration number over the spatial grid is strictly greater for the GST algorithm, compared to the Anderson Acceleration approach. 186
9.6 Noble gas mixture of Ar, Kr, and Xe: Moment plots of species number density, bulk velocity, and temperature, are given at $t_f = 1$ μs, for a mixture with $\varepsilon \approx 0.00103$. M-BGK uses the GST algorithm for the moment solve. M-BGK_{EX} uses a fully explicit method. 188
9.7 The number of iterations required for convergence is decreased by the application of Anderson Acceleration (AA) to the GST algorithm. The maximum iteration number over the spatial grid is strictly greater for the GST algorithm, compared to the Anderson Acceleration approach. 189

10.1 Moment profiles for the $1x2v$ ES–BGK slab problem at $t_F = 0.25$ with $\varepsilon = 10^{-3}$.211

11.1 A finite-volume cell $C_{i,j}$ in truncated phase space $\Omega_x \times \Omega_v$. 264

11.2 Sod results at kinetic scale 10^{-3} 276
11.3 Sod numerical-rank histories at kinetic scale 10^{-3} 277
11.4 Macroscopic quantities and distribution error for the one-dimensional Sod experiment with $\varepsilon = 10^{-4}$. The DLRA results are compared with the slope-limited full finite-volume reference solution. 278
11.5 Numerical ranks of the DLRA approximation Y, the sampled Maxwellian $M[Y]$, and the slope-limiter matrix $\Lambda[Y]$ for the Sod experiment with $\varepsilon = 10^{-4}$. Numerical rank is measured using the relative threshold $\sigma_k/\sigma_1 > 10^{-10}$. . . 279
11.6 Smooth rank-one solution comparison 292
11.7 Smooth rank-one slope compressibility 293
11.8 Smooth rank-one numerical-rank histories 294
11.9 Final-time solutions for the smooth rank-two initial condition. The full-matrix Forward Euler solution is shown on the left, and the rank-16 projector-splitting approximation is shown on the right. Both calculations use the same slope-limited finite-volume spatial discretization and timestep. 296
11.10 Same-rank compression error of the slope-limiter matrix for the smooth rank-two experiment. Each curve shows the best rank-r relative Frobenius approximation error of $\Sigma(Y_r(t))$ for the corresponding prescribed DLRA rank. Smaller values indicate greater compressibility at the state rank. 297
11.11 Numerical-rank histories for the state matrix $Y_r(t)$ and the slope-limiter matrix $\Sigma(Y_r(t))$ in the smooth rank-two experiment. Numerical ranks are computed using the relative criterion $\sigma_k/\sigma_1 > 10^{-8}$. The state retains its prescribed rank, whereas the slope-limiter matrix rapidly reaches substantially higher numerical rank. 298
11.12 Final-time solutions for the discontinuous rank-one initial condition. The full-matrix Forward Euler solution is shown on the left, and the rank-16 projector-splitting approximation is shown on the right. The discontinuity produces considerably greater rank growth and a more demanding low-rank approximation problem than the smooth initial conditions. 300
11.13 Same-rank compression error of the slope-limiter matrix for the discontinuous rank-one experiment. Compared with the smooth experiments, the larger errors show that the slope-limiter matrix is less accurately representable at the prescribed state rank. 301

11.14Numerical-rank histories for the state matrix $Y_r(t)$ and the slope-limiter matrix $\Sigma(Y_r(t))$ in the discontinuous rank-one experiment. Numerical ranks are computed using the relative criterion $\sigma_k/\sigma_1 > 10^{-8}$. The numerical rank of the limiter matrix grows rapidly, particularly for the higher-rank state approximations. 302

List of Tables

6.1 Audit of the quantitative evidence retained with the unary BGK and Vlasov–Poisson test cases. Entries identified as pending require regenerated solver output and are not inferred from figures. 144

9.1 Archived nonlinear-solver and time-step counts. For the Ar–Kr–Xe test, the IMEX method uses approximately 8.30 times fewer time steps than the fully explicit method. AA denotes Anderson acceleration applied to the GST iteration. 189

11.1 Smooth rank-one accuracy and conservation 291
11.2 Smooth rank-one limiter ranks 291
11.3 Smooth rank-two accuracy and conservation 295
11.4 Smooth rank-two limiter ranks 295
11.5 Discontinuous rank-one accuracy and conservation 299
11.6 Discontinuous rank-one limiter ranks 299

12.1 Numerical-validation matrix 306

Abbreviations and Acronyms

AA	Anderson acceleration
BE	backward Euler
BGK	Bhatnagar–Gross–Krook
BUG	basis-update and Galerkin
CFL	Courant–Friedrichs–Lewy
DLRA	dynamical low-rank approximation
ES-BGK	ellipsoidal statistical BGK
GST	Gauss–Seidel-type
HHM	Haack–Hauck–Murillo
IMEX	implicit-explicit
ODE	ordinary differential equation
PDE	partial differential equation
PSI	projector-splitting integrator
QR	QR matrix factorization
RK	Runge–Kutta
SPD	symmetric positive definite
SPSD	symmetric positive semidefinite
SVD	singular value decomposition

Abbreviations local to a single derivation are defined where they first occur.

Acknowledgements

This work is supported by Oak Ridge National Laboratory under subcontract UTB-CW24420, "BGK Kinetic Equations," to the University of Tennessee, Knoxville. The authors wish to thank Jeff Haack (LANL), Cory Hauck (ORNL), Kyle Schwiebert (ORNL), Ryan Glasby (ORNL), and Stefan R. Schnake (ORNL) for useful discussions concerning numerical methods for BGK-type equations and related kinetic models. The authors also gratefully acknowledge permission to include results from joint work in this report.

Artificial intelligence tools assisted with language editing, report organization, notation checks, figure refinement, code commentary, and selected symbolic or algebraic checks. The authors independently reviewed all AI-assisted text, code, figures, and calculations against the governing mathematical definitions, source material, or numerical output, as appropriate. AI-generated output was not treated as independent verification. The human authors remain solely responsible for the report's content.

Executive Summary

This report records work over a five-year period and adds work completed during the most recent support period, August 1, 2025–July 31, 2026. It retains, clarifies, and expands selected material from earlier years as a cumulative technical and training resource. This cumulative design is deliberate: continued consolidation and polishing transform a sequence of annual results into a coherent teaching resource for future graduate students who join the project, reducing the time required to master the background and begin contributing to the research. The document is therefore a valuable project outcome in its own right, as well as a record of technical progress. Relative to the Year Four Report, the principal changes are an expanded multispecies Boltzmann chapter, two new chapters on Chapman–Enskog analysis and moment methods, a substantial update of the ES-BGK chapter, and a new chapter on dynamical low-rank approximation (DLRA) with numerical simulations. The multispecies Boltzmann material and the published analysis and implicit solver for multispecies BGK moment equations were already present in the Year Four Report and are therefore treated here as expanded or retained foundations, not as new Year-Five contributions.

The principal deliverables and their connection to the next stage of the project are summarized below.

Objective	Year-Five change	Evidence	Next action
Kinetic foundations and models	Expansion of the existing multispecies Boltzmann chapter; new stand-alone Chapman–Enskog and moment-method chapters; substantial revision of unary and multispecies ES-BGK analysis and numerics.	Chapters 2–4 and 10.	Calibrate BGK collision frequencies and extend the framework to multispecies transport coefficients.
Implicit moment methods	Previously reported analysis and the convergent Gauss–Seidel-type solver are retained and reorganized as foundations for subsequent transport work.	Two 2025 SIAM publications; Chapter 9.	Incorporate the solver into full IMEX transport schemes.
Reduced kinetic methods	A new DLRA chapter develops the mathematical framework, algorithms, BGK applications, and simulations using high-resolution finite-volume discretizations.	Chapter 11.	Integrate rank adaptivity with multispecies BGK and ES-BGK solvers.
Training and dissemination	Continued refinement of the cumulative report as a structured teaching and onboarding resource; a graduate course used material from the report, with recorded lectures made publicly available.	Self-contained foundational chapters; *Nonequilibrium Thermodynamics*; YouTube course playlist.	Continue polishing the report and use it to prepare future graduate students and postdoctoral researchers to contribute to the project.

What Changed in Year Five

The following roadmap distinguishes material that is genuinely new to the cumulative report from material that was substantially or modestly updated. The labels describe the status of the report material, not the publication date or originality of every cited result.

Update class	Location	Scope
Expanded	Chapter 2	The multispecies Boltzmann material already appeared in the Year Four Report; Year Five substantially expands and reorganizes it into a more self-contained treatment.
Genuinely new	Chapters 3 and 4	Stand-alone treatments of the Chapman–Enskog derivation and finite moment-closure methods.
Significantly updated	Chapter 10	Expanded unary and multispecies ES-BGK analysis and numerics, including a closed positivity-preserving implicit collision update.
Genuinely new	Chapter 11	Matrix-manifold foundations, DLRA integrators, BGK applications, and numerical simulations with high-resolution finite-volume methods.
Updated or retained	Chapters 5–9 and supporting appendices	Earlier unary and multispecies BGK material, including the published moment analysis and solver, is retained as a technical foundation and training resource, with reorganization, corrections, and improved cross-references.

Selected Verifiable Indicators

The cumulative project record includes two peer-reviewed SIAM journal articles, completion of one PhD, one graduate course with publicly available recordings, and preliminary MATLAB test codes for DLRA and high-resolution finite-volume algorithms. The two articles and the associated multispecies moment solver were already reported in the Year Four Report; the course and the DLRA work are among the activities emphasized in this Year-Five report. Chapter 9 reports archived iteration and time-step counts for the multispecies solver; other numerical claims are stated qualitatively when the underlying output data are not available for a new norm-based analysis.

Milestones

The following cumulative milestones provide context for the Year-Five changes. The two publications in Items 1 and 2 were already reported in the Year Four Report and are retained

here for continuity; they are not newly claimed Year-Five results.

1. We published *Asymptotic Relaxation of Moment Equations for a Multi-species, Homogeneous BGK Model*, *SIAM Journal on Applied Mathematics* 85 (2025), 294–313 [33]. This paper establishes long-time dynamics of the multispecies BGK moment equations. doi:10.1137/23M1596314

2. We published *Implicit Update of the Moment Equations for a Multi-Species, Homogeneous BGK Model*, *SIAM Journal on Numerical Analysis* 63 (2025), 881–907 [34]. This paper establishes stability, well-posedness, and convergence of an iterative moment solver. doi:10.1137/24M165421X

3. Evan Habbershaw completed his PhD in August 2025 after defending the thesis *Numerical Methods for Multi-Species Kinetic Models with Approximate Collision Operators*. Dr. Habbershaw is now a postdoctoral researcher at the Pennsylvania State University and a co-author of this report.

4. In 2025, the PI (Wise) recruited two UTK Mathematics graduate students to work on the project, Jack Haight and Daniel Bond. Bond later left the project. Haight continues project-related research without support from this subcontract.

5. In 2026, the PI recruited one new UTK Mathematics graduate student, Alli Brophy. She will begin her research on the project in Fall 2026 or Spring 2027.

6. The PI developed and taught a graduate course, *Nonequilibrium Thermodynamics*, using material from this report. Recordings of the course are publicly available in a YouTube playlist.

7. We conducted preliminary research on dynamical low-rank approximation and moment methods, establishing a foundation for reduced representations of kinetic and hyperbolic equations.

8. We developed MATLAB test codes for DLRA algorithms, including experiments for BGK and high-resolution finite-volume discretizations.

Chapter 1

Introduction and Report Organization

Continuum macroscopic models describe fluids using fields such as density, bulk velocity, and temperature. Their validity depends on scale separation and near-local thermodynamic equilibrium. The continuum velocity is a local mean of the microscopic particle velocities, which may remain broadly distributed.

To make the scale separation precise, let the molecular mean free path be λ_{mfp} and a characteristic macroscopic length be L. The **Knudsen number** is $K_n = \lambda_{\text{mfp}}/L$. Chapter 3 denotes the same ratio by ε when it is used as the Chapman–Enskog expansion parameter. When $K_n \ll 1$, frequent collisions keep the velocity distribution close to a local Maxwellian, permitting closure by continuum models such as the Euler or Navier–Stokes–Fourier equations [74].

When $K_n = \mathcal{O}(1)$, nonequilibrium features of the velocity distribution generally cannot be represented adequately by a continuum closure, and a kinetic description such as the Boltzmann transport equation is required [74]. In three spatial dimensions, the distribution depends on time, three position variables, and three velocity variables. The position and velocity variables together form six-dimensional phase space. The Boltzmann equation is challenging because of both this high dimensionality and its collision operator, which is nonlinear and nonlocal in velocity while remaining local in position and time.

1.1 A Toy Model

For those readers not familiar with kinetic equations, the following toy model will give some idea of how these models keep track of particles over time and space. We begin with a deliberately simplified $1x1v$ treatment in which position, velocity, and time are all discrete.

For example, let

$$\mathcal{X} = \{1, 2, \ldots, 8\}, \qquad \mathcal{V} = \{-4, -3, \ldots, 3, 4\}, \qquad \mathcal{T}_{\Delta t} = \{t^n = n\Delta t \mid n = 0, 1, 2, \ldots\}.$$

The state is an integer-valued counting function

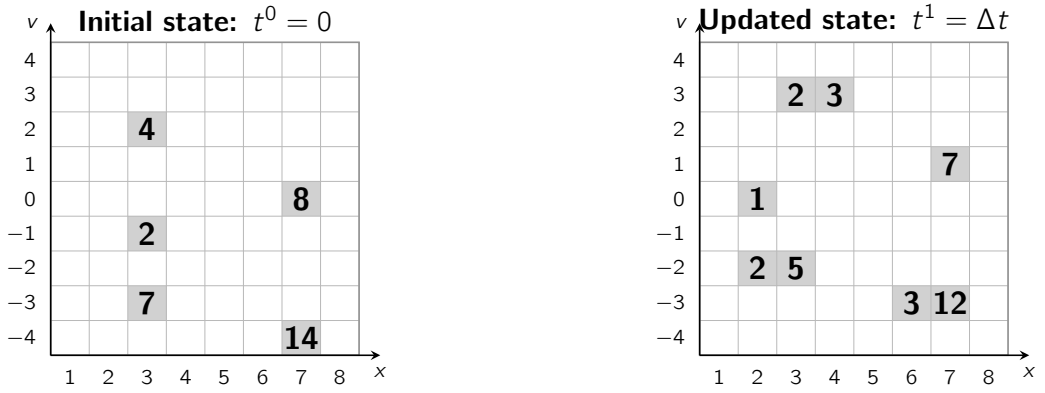
$$f : \mathcal{X} \times \mathcal{V} \times \mathcal{T}_{\Delta t} \longrightarrow \mathbb{N}_0. \tag{1.1}$$

The value $f(i, v, t^n)$ is the number of particles in spatial cell i with velocity v at time t^n. Here v can take only the nine listed values; there are no intermediate velocities. In particular,

$$f(3, -3, 0) = 7$$

means exactly that at time zero there are seven particles in spatial cell 3 whose velocity is -3. It is not a density, a cell average, or a quadrature approximation. Every unmarked entry in the grid below is understood to be zero.

In Figure 1.1, we give the state at two time instances, $t = 0$ and $t = t^1$. Particles can move by advection with their labeled velocities, moving from cell to cell, and collide with each other and, thereby, exchanging momentum.



(a) The initial integer-valued state. **(b)** One possible updated integer-valued state.

Figure 1.1: A fully discrete particle-count analogy. Each printed integer is the exact value of $f(i, v, t^n)$. Positive-velocity particles tend to move toward larger spatial indices, negative-velocity particles toward smaller indices, and collisions may reassign velocity labels without changing spatial cells. The displayed counts sum to 35 in both panels.

Summing the columns of the states over all velocity labels removes the velocity coordinate

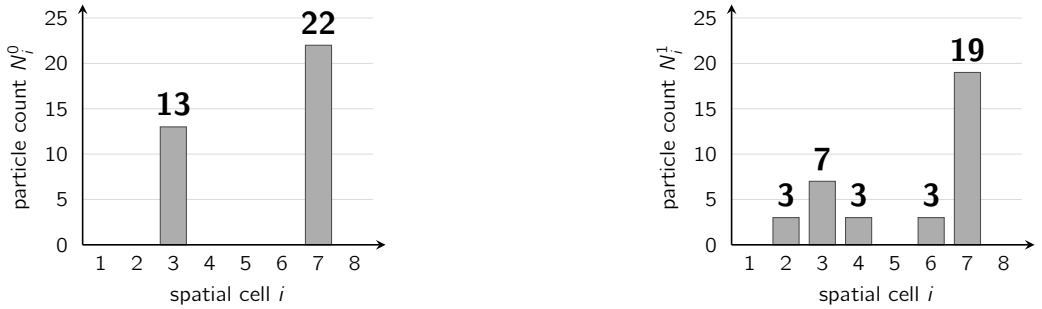
and gives the number of particles in a spatial cell. Define the discrete velocity marginal

$$N_i^n := \sum_{v \in \mathcal{V}} f(i, v, t^n), \qquad i \in \mathcal{X}. \tag{1.2}$$

For the two states in Figure 1.1, these column sums are

$$(N_i^0)_{i=1}^8 = (0, 0, 13, 0, 0, 0, 22, 0), \qquad (N_i^1)_{i=1}^8 = (0, 3, 7, 3, 0, 3, 19, 0).$$

They form one-dimensional histograms of the spatial particle counts.



(a) Initial spatial histogram. **(b)** Updated spatial histogram.

Figure 1.2: Velocity marginals of the two discrete phase-space states. Each bar is the sum of one velocity column in Figure 1.1. Transport redistributes particles among spatial cells, while the sum of the bar heights remains 35.

The quantities N_i^n are particle counts, not yet number densities. If spatial cell i represents an interval C_i of width Δx, its cell-averaged number density is

$$\bar{n}_i^n := \frac{N_i^n}{\Delta x}, \qquad N_i^n = \bar{n}_i^n \Delta x. \tag{1.3}$$

Thus the histogram of N_i^n and the histogram of $\bar{n}_i^n$ have the same shape when all cells have the same width; only the vertical scale differs. This velocity sum is the discrete analogue—not a discretization—of the continuous velocity marginal

$$n(\boldsymbol{x}, t) = \int_{\mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{v}, t)\, \mathrm{d}\boldsymbol{v},$$

which is the continuous number density used later in the chapter.

A single discrete update of the state can be imagined as two successive bookkeeping operations,

$$f^n \xrightarrow{\text{transport between spatial cells}} f^{n+\frac{1}{2}} \xrightarrow{\text{collisions within each spatial cell}} f^{n+1}. \tag{1.4}$$

During the transport stage, a particle retains its discrete velocity but may move to another spatial cell; the precise movement rule depends on the chosen analogy.

During the collision stage, particles stay in the same spatial cell but may be reassigned to different allowed velocities. For now, we do not specify how and why particles may collide or if those collisions are elastic or otherwise. It is enough to know that collisions lead to changes in the velocity labels of particles.

In either stage the entries remain nonnegative integers. Define the total particle count at time t^n by

$$\mathcal{N}^n := \sum_{i \in \mathcal{X}} N_i^n = \sum_{i \in \mathcal{X}} \sum_{v \in \mathcal{V}} f(i, v, t^n).$$

For a periodic or closed grid, this count does not change during an update. The two panels from Figure 1.1 therefore satisfy $\mathcal{N}^0 = \mathcal{N}^1 = 35$.

This construction is only an analogy. It is not obtained by discretizing the Boltzmann equation, and the displayed update is not proposed as a numerical method. Its purpose is to isolate two ideas: transport changes the spatial cells occupied by particles, whereas collisions change the distribution of their discrete velocities. From this point onward, f again denotes a continuous phase-space density rather than an integer particle count.

1.2 Quick Introduction to the Unary Boltzmann Equation

The adjective **unary** means that only one particle species is present. Instead of following individual particles, kinetic theory describes their statistical distribution. Standard mathematical introductions and surveys include Bressan [9], Cercignani [10], Cercignani, Illner, and Pulvirenti [12], and Villani [78]. In the continuous model, the quantity

$$f(\boldsymbol{x}, \boldsymbol{v}, t)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{v}$$

is interpreted as the expected number of particles whose positions and velocities lie in small volume elements about $\boldsymbol{x}$ and $\boldsymbol{v}$ at time t. Thus the continuous f is a density on phase space, not the trajectory of a single particle.

The **Vlasov–Boltzmann equation** for a unary dilute gas is

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f + \boldsymbol{a} \cdot \nabla_{\boldsymbol{v}} f = Q[f, f](\boldsymbol{x}, \boldsymbol{v}, t), \tag{1.5}$$

where $f(\boldsymbol{x}, \boldsymbol{v}, t)$ is the number-density distribution in phase space and $\boldsymbol{a} = \boldsymbol{a}(\boldsymbol{x}, \boldsymbol{v}, t)$ is the acceleration produced by an external field. For example, an electric field produces a velocity-

independent acceleration, whereas the magnetic part of the Lorentz force depends on velocity. We assume here that $\nabla_{\boldsymbol{v}} \cdot \boldsymbol{a} = 0$, as in those examples; otherwise, the conservative phase-space-density equation also contains $f \nabla_{\boldsymbol{v}} \cdot \boldsymbol{a}$ on the left-hand side. The **Boltzmann transport equation** results from setting $\boldsymbol{a} \equiv \boldsymbol{0}$. The transport operator follows from the chain rule along a particle trajectory:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}, \boldsymbol{v}, t) &= \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} \frac{\mathrm{d}x_i}{\mathrm{d}t} + \sum_{i=1}^{d} \frac{\partial f}{\partial v_i} \frac{\mathrm{d}v_i}{\mathrm{d}t} + \frac{\partial f}{\partial t} \\ &= \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f + \boldsymbol{a} \cdot \nabla_{\boldsymbol{v}} f + \frac{\partial f}{\partial t}, \end{aligned}$$

where $d = 3$ for the standard model. Thus the left-hand side is the rate of change of f observed while moving along a particle trajectory. The Vlasov–Boltzmann equation asserts that this rate is produced by collisions:

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}, \boldsymbol{v}, t) = Q[f, f](\boldsymbol{x}, \boldsymbol{v}, t).$$

The collision operator is not merely a collision frequency: it is the net rate at which collisions add particles to, or remove particles from, each velocity. Its definition requires additional physical input. Collisions are assumed binary, instantaneous, local in space, and elastic [74]. We state its generic three-dimensional form for the moment:

$$Q[f, f](\boldsymbol{x}, \boldsymbol{v}, t) = \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f' f'_\star - f f_\star \right] B(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star,$$

where

$$\begin{aligned} f' &\coloneqq f(\boldsymbol{x}, \boldsymbol{v}', t), \quad f'_\star \coloneqq f(\boldsymbol{x}, \boldsymbol{v}'_\star, t), \\ f &\coloneqq f(\boldsymbol{x}, \boldsymbol{v}, t), \quad f_\star \coloneqq f(\boldsymbol{x}, \boldsymbol{v}_\star, t); \end{aligned}$$

B is the collision kernel describing the frequency and angular distribution of binary collisions, $\boldsymbol{\sigma}$ is a unit vector called the apse direction, and the primed velocities are computed from

$$\boldsymbol{v}' = \boldsymbol{v} - \left[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) \right] \boldsymbol{\sigma}, \tag{1.6}$$

$$\boldsymbol{v}'_\star = \boldsymbol{v}_\star + \left[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) \right] \boldsymbol{\sigma}. \tag{1.7}$$

The transport terms move the distribution through position and velocity space. The collision term acts only in velocity space and is local in $(\boldsymbol{x}, t)$: particles collide only when they

are at the same position and time. The gain term $f'f'_\star$ represents collisions arriving at the pair $(\boldsymbol{v}, \boldsymbol{v}_\star)$, while the loss term $ff_\star$ represents collisions leaving that pair. The binary collision transformation and the kernel B will be specified carefully below.

The computation of the Boltzmann collision operator is expensive. In three dimensions, its five-dimensional integral must be evaluated at every phase-space point $(\boldsymbol{x}, \boldsymbol{v})$, and closed-form evaluation is available only in special cases. Deterministic spectral or discrete-velocity methods and stochastic Monte Carlo methods offer different accuracy–cost tradeoffs; neither is universally preferable. For an N-species mixture, one must additionally account for all N^2 ordered collision terms. A common reduced description replaces these integrals by a less costly BGK-type relaxation. The initials honor Bhatnagar, Gross, and Krook, who introduced the approximation in 1954 [6].

1.3 The BGK Approximation Model

The full collision integral is costly, but its most important qualitative effect is easy to describe: collisions preserve mass, momentum, and energy while driving the velocity distribution toward a Maxwellian. The **BGK collision model** builds these two facts directly into the simpler nonlinear relaxation operator

$$Q^{BGK}[f](\boldsymbol{x}, \boldsymbol{v}, t) = \lambda(M_f(\boldsymbol{x}, \boldsymbol{v}, t) - f(\boldsymbol{x}, \boldsymbol{v}, t)), \tag{1.8}$$

where $\lambda = \tau^{-1}$ is a relaxation frequency, $\tau > 0$ is a characteristic collision time, and M_f is the **Maxwellian** constructed from the hydrodynamic moments of f,

$$n(\boldsymbol{x}, t) \coloneqq \int_{\mathbb{R}^d} f \,\mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u}(\boldsymbol{x}, t) \coloneqq \frac{1}{n}\int_{\mathbb{R}^d} f\boldsymbol{v} \,\mathrm{d}\boldsymbol{v}, \qquad \theta(\boldsymbol{x}, t) \coloneqq \frac{m}{nd}\int_{\mathbb{R}^d} f|\boldsymbol{v} - \boldsymbol{u}|^2 \,\mathrm{d}\boldsymbol{v}, \tag{1.9}$$

and is defined as

$$M_f(\boldsymbol{x}, \boldsymbol{v}, t) \coloneqq M[f](\boldsymbol{x}, \boldsymbol{v}, t) \coloneqq n\left(\frac{m}{2\pi\theta}\right)^{d/2} \exp\left(-\frac{m|\boldsymbol{v} - \boldsymbol{u}|^2}{2\theta}\right). \tag{1.10}$$

The fields n, $\boldsymbol{u}$, and θ are the macroscopic, or coarse-grained, number density, bulk velocity, and temperature. They correspond to the zeroth, first, and centered second velocity moments; the phrase **first three moments** is sometimes used informally for this collection. The total kinetic-energy density is

$$\omega(\boldsymbol{x}, t) \coloneqq \frac{m}{2}\int_{\mathbb{R}^d} f|\boldsymbol{v}|^2 \,\mathrm{d}\boldsymbol{v}, \tag{1.11}$$

and the internal energy density is defined as

$$e(\boldsymbol{x}, t) \coloneqq \frac{m}{2} \int_{\mathbb{R}^d} f|\boldsymbol{v} - \boldsymbol{u}|^2 \, \mathrm{d}\boldsymbol{v}. \tag{1.12}$$

We observe that the temperature and the internal energy are related via

$$\theta(\boldsymbol{x}, t) = \frac{2}{d} \frac{e(\boldsymbol{x}, t)}{n(\boldsymbol{x}, t)}.$$

By construction, M_f has the same mass, momentum, and total energy as f. Hence the BGK term preserves these collision invariants while driving f toward local Maxwellian equilibrium. If the external acceleration in (1.5) is retained, replacing $Q[f, f]$ by $Q^{BGK}[f]$ gives the Vlasov–BGK equation. The moment-matching and entropy properties of M_f are proved later rather than assumed here.

Proposition 1.1 (Conservation by the BGK Operator)**.** *Whenever the displayed moments are finite,*

$$\int_{\mathbb{R}^d} Q^{BGK}[f] \begin{bmatrix} 1 \\ m\boldsymbol{v} \\ m|\boldsymbol{v}|^2/2 \end{bmatrix} \mathrm{d}\boldsymbol{v} = \boldsymbol{0}. \tag{1.13}$$

Proof. The Maxwellian M_f is chosen to have exactly the same number, momentum, and total energy as f. Each component of the integral is therefore λ times the difference of two equal moments. ■

This proposition illustrates the basic design principle of a reduced collision model: the approximation may simplify the relaxation mechanism, but it must preserve the collision invariants of the original operator.

1.4 Organization

Chapter 2 gives a detailed review of the multispecies Boltzmann equation, including material that is often abbreviated in textbooks. This foundation is important because the BGK-type approximations developed later inherit the collision invariants and entropy structure of the Boltzmann system.

Chapters 3 and 4 connect this kinetic description to continuum models. The first derives the Euler and Navier–Stokes–Fourier equations through the Chapman–Enskog procedure, while the second surveys the general structure and requirements of finite moment closures.

Chapter 5 introduces the single-species BGK equation and derives its conservation laws, information dissipation, and entropy production. The presentation is intended to orient readers who are new to kinetic models without attempting a comprehensive literature review.

Chapter 6 presents finite-volume implicit-explicit (IMEX) Runge–Kutta (RK) methods and representative simulations, including the Sod shock tube and two-stream instability benchmarks. Chapter 7 develops numerical methods for the discrete-velocity models introduced in Chapter 5.

In Chapter 8, we derive the conservative, entropic multispecies BGK model of Haack, Hauck, and Murillo from the Boltzmann collision structure, establish positivity of its pair temperatures, and prove its conservation laws and H-theorem. Chapter 9 describes the asymptotic dynamics of the corresponding space-homogeneous moment equations and summarizes our recent numerical analysis for this model.

Chapter 10 presents a preliminary analysis of unary and multispecies ellipsoidal statistical BGK (ES-BGK) models.

The ES-BGK models can more accurately capture realistic Prandtl numbers of gases than the simpler BGK-type models. As a consequence, they provide more accurate hydrodynamic limits (the Navier–Stokes–Fourier system) than the standard BGK models. At the same time, ES-BGK models retain the computational tractability of their BGK relatives.

Chapter 11 presents dynamical low-rank approximation methods and their application to BGK and high-resolution finite-volume schemes. The report concludes in Chapter 12 with a comparison of the Year-Five advances against the Year Four Report and a statement of near- and long-term directions.

1.5 Terminology and Notation Conventions

Except when reproducing the title of a cited work, this report uses **multispecies** for models or mixtures involving more than one species. Compound modifiers are hyphenated consistently: **single-species**, **space-homogeneous**, **discrete-velocity**, and **finite-volume**. The term **dynamical low-rank approximation** is abbreviated DLRA after its first use. Phase-space configurations are denoted by $1x1v$, $1x2v$, and so on, with the first number giving the spatial dimension and the second the velocity dimension. Microscopic velocity is denoted by $\boldsymbol{v}$ in multiple velocity dimensions and by v in a scalar $1x1v$ setting; the corresponding bulk velocities are $\boldsymbol{u}$ and u. Number density is n, mass density is $\rho = mn$, and temperature in energy units is θ (or $\hat{\theta}_i$ for the species temperatures used in the multispecies moment system). The symbol T is reserved for terminal time and established operator or tensor notation;

dimensional kelvin temperature appears only as T^{dim} in the relation $\theta^{\mathrm{dim}} = k_B T^{\mathrm{dim}}$.

The continuous Maxwellian associated with a distribution f is denoted by $M_f = M[f]$; a pair Maxwellian is $M_{i,j}$; a discrete-velocity Maxwellian is $\mathcal{M}_{\mathcal{J},f}$; and an ES-BGK Gaussian is $\mathcal{G}_\nu$. The report uses the physical entropy sign convention

$$S[f] \coloneqq -H[f], \qquad H[f] \coloneqq \int f \ln f,$$

so entropy is non-decreasing and information is non-increasing. The term **information** is retained for Boltzmann's H-functional and for ordinary descriptive uses, but statements about thermodynamic direction are expressed primarily in terms of entropy production. A symbol defined differently for a specialized model is identified locally. Appendix G provides a consolidated map of densities, velocity frames, energy and temperature variables, stress and flux quantities, and the information/entropy sign convention.

Several appendices are included that contain tedious calculations that we wished to keep out of the main text. A prototype $1x1v$ MATLAB code for the Sod shock tube problem is contained in Appendix F.

1.6 Review Status and Limitations

This cumulative report combines peer-reviewed project results, technical exposition, preliminary research, and prototype computations. The two SIAM articles identified in the Executive Summary have undergone journal peer review; the report as a whole has received internal author review but has not undergone independent external review. Statements that are formal, preliminary, or dependent on additional hypotheses are identified locally.

The numerical results apply to the configurations documented with each test and should not be interpreted as a general validation of the methods outside those settings. The MATLAB listings in Appendix F are research and training prototypes, not production software. Their present relationship to the archived Sod calculation, including the known configuration mismatch, is stated in that appendix and in Section 6.5.3. Users should independently verify the code and its outputs for their intended application.

The authors have checked the report's claims, proofs, equations, citations, and cross-references to the extent practical for this cumulative release and remain responsible for its content. Neither The University of Tennessee nor Oak Ridge National Laboratory can be held responsible for decisions based on the report or on output from the prototype code.

Chapter 2

Multispecies Boltzmann Equations

Our primary objective is to understand the kinetic description of dilute mixtures of particles. The multispecies Boltzmann system is a foundational model for such gas mixtures: instead of tracking individual particles, it evolves one distribution function for each species in position–velocity phase space. Direct deterministic computation is expensive because every binary interaction contributes a high-dimensional collision integral. This motivates reduced collision models, including BGK-type relaxations, but those approximations can be designed and interpreted responsibly only after the conservation, entropy, and equilibrium structure of the Boltzmann operator is understood.

The thesis by Marlies Pirner [62] has been especially useful in preparing this chapter. The books [11, 36] give readable introductions to the multispecies Boltzmann system, while Kremer [48] provides a more detailed modern account. Bressan's notes [9] are a helpful introduction to the unary model and its analysis.

The exposition proceeds from microscopic mechanics to macroscopic closure. After the unary review in the preceding chapter, we construct the elastic binary collision map and use its symmetries to derive weak forms, conservation laws, and the H-theorem for mixtures. After defining velocity moments carefully in both species and barycentric frames, we characterize Maxwellians and their entropy-minimizing property. We conclude this chapter by returning to mixtures in the spatially homogeneous setting and identifying the equilibrium parameters selected by the conserved quantities. The next two chapters then derive the unary Euler and Navier–Stokes–Fourier closures and place that construction within the broader methodology of finite moment closure.

This chapter is written for a reader who has completed a first course in differential equations and is comfortable with multivariable calculus, linear algebra, and elementary partial differential equations. No previous course in kinetic theory or continuum mechanics is as-

sumed. The longer proofs use only changes of variables, symmetry, Taylor's theorem, and Gaussian integration, together with a few explicitly stated formal properties of the collision operator.

There are two useful ways to read this chapter. On a first reading, one may focus on the definitions, the statements of the main propositions, and the explanatory paragraphs following them. The longer weak-form, collision-invariant, and equilibrium proofs can then be studied on a second pass. Throughout, we keep three logical levels separate:

1. **exact identities**, obtained directly from the kinetic equation;
2. **structural conclusions**, obtained from collision symmetries and nondegeneracy assumptions;
3. **closures and asymptotic approximations**, which introduce additional modeling assumptions.

This distinction is especially important in the Chapman–Enskog and moment-closure sections.

Standing Regularity Convention

Unless otherwise stated, the calculations below are understood for kinetic distributions that are sufficiently smooth and rapidly decaying in velocity to justify differentiation under the integral sign, changes in the order of integration, collision-variable transformations, and integrations by parts. We assume that every displayed moment is finite. In discussions involving information or entropy, we additionally assume that the distributions are strictly positive and that $f_i \log f_i$ and the corresponding collision terms are integrable. These conventions justify the pointwise manipulations in a classical setting; alternatively, the identities may be read as formal identities for smooth solutions. Whenever a conclusion depends on a formal asymptotic expansion or on more specific analytic hypotheses, this is stated explicitly.

2.1 The Binary Collision Transformation

To derive the multispecies collision operator, we first isolate the kinematics of a single elastic encounter. The two particles may have different masses, but the collision preserves their total momentum and kinetic energy. The apse direction $\boldsymbol{\sigma}$ specifies the line along which the impulsive exchange occurs; equivalently, only the component of the relative velocity parallel to $\boldsymbol{\sigma}$ is reversed.

Definition 2.1 (Elastic Binary Collision Map)**.** *Let $m_i, m_j > 0$ be the masses of particles of species i and j, respectively, $i, j \in \{1, 2, \ldots, N\}$. Fix a unit vector $\boldsymbol{\sigma} \in \mathbb{S}^2$. The transformation $T^{i,j}_{\boldsymbol{\sigma}}(\boldsymbol{v}, \boldsymbol{v}_\star) = (\boldsymbol{v}', \boldsymbol{v}'_\star)$, which maps $\mathbb{R}^3 \times \mathbb{R}^3$ into $\mathbb{R}^3 \times \mathbb{R}^3$, is defined via*

$$\boldsymbol{v}' = \boldsymbol{v} - [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)] \frac{2m_j}{m_i + m_j} \boldsymbol{\sigma}, \tag{2.1}$$

$$\boldsymbol{v}'_\star = \boldsymbol{v}_\star + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)] \frac{2m_i}{m_i + m_j} \boldsymbol{\sigma}. \tag{2.2}$$

We will write

$$\boldsymbol{v}' = T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star) \quad \textit{and} \quad \boldsymbol{v}'_\star = T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star)$$

for the first and second components of the transformation, respectively. The mapping $T^{i,j}_{\boldsymbol{\sigma}}$ is called a **binary collision transformation** *and $\boldsymbol{\sigma}$ is called the* **apse direction**. *The unprimed coordinates $\boldsymbol{v}$ and $\boldsymbol{v}_\star$ are the* **pre-collisional velocities**; *the primed coordinates $\boldsymbol{v}'$ and $\boldsymbol{v}'_\star$ are the* **post-collisional velocities**. *The labels* ***pre*** *and* ***post*** *are interchangeable because the elastic collision map is an involution, as proved below.*

For equal masses, $m_i = m_j$, the collision map simplifies to

$$\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{\sigma}\big[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)\big], \tag{2.3}$$

$$\boldsymbol{v}'_\star = \boldsymbol{v}_\star + \boldsymbol{\sigma}\big[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)\big]. \tag{2.4}$$

Thus an equal-mass collision exchanges the components of the two velocities parallel to the apse direction, while leaving their components perpendicular to $\boldsymbol{\sigma}$ unchanged. Figure 2.1 illustrates this exchange for the first worked example below. The vectors are drawn in a two-dimensional plane for clarity; they may equivalently be embedded in $\mathbb{R}^3$ by appending a zero third component.

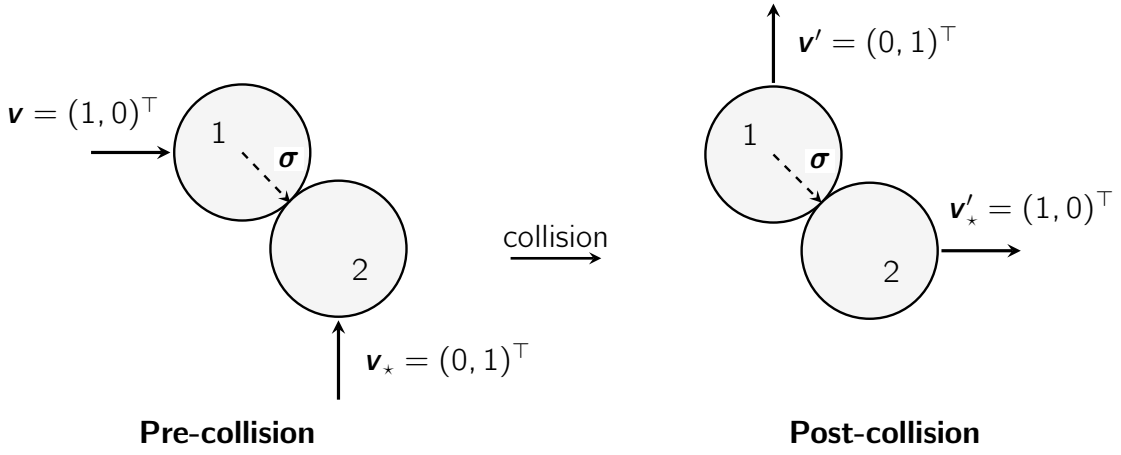


Figure 2.1: An equal-mass binary collision with $\boldsymbol{\sigma} = 2^{-1/2}(1,-1)^\top$. The components parallel to $\boldsymbol{\sigma}$ are exchanged; in this particular configuration, the two velocity vectors are exchanged completely.

Example 2.2 (An oblique apse direction)**.** Suppose that the pre-collisional velocities and apse direction are

$$\boldsymbol{v} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{v}_\star = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\sigma} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Since

$$\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) = \sqrt{2}, \qquad \boldsymbol{\sigma}\big[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)\big] = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

Equations (2.3) to (2.4) give

$$\boldsymbol{v}' = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{v}'_\star = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The particles exchange their complete velocity vectors in this special geometry. Both momentum and kinetic energy are unchanged.

Example 2.3 (A coordinate-aligned apse direction)**.** Keep the same pre-collisional velocities, but choose

$$\boldsymbol{\sigma} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

Now

$$\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) = 1, \qquad \boldsymbol{\sigma}\big[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)\big] = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

and hence

$$\boldsymbol{v}' = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{v}'_\star = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This result seems strange, since one of the post-collision speeds is precisely zero, but is correct. The direction $\boldsymbol{\sigma}$ is vertical, so the collision exchanges the vertical components of the two velocities and leaves their horizontal components unchanged. Particle 1 therefore retains its horizontal velocity and receives the vertical velocity of particle 2, while particle 2 comes to rest. As a direct check,

$$\boldsymbol{v} + \boldsymbol{v}_\star = \boldsymbol{v}' + \boldsymbol{v}'_\star = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad |\boldsymbol{v}|^2 + |\boldsymbol{v}_\star|^2 = |\boldsymbol{v}'|^2 + |\boldsymbol{v}'_\star|^2 = 2.$$

Thus both total momentum and total kinetic energy are conserved.

Proposition 2.4 (Momentum, Energy, and Relative-Speed Conservation)**.** *Let $m_i, m_j > 0$ be the masses of particles of species i and j, respectively, $i, j \in \{1, 2, \ldots, N\}$, and fix an apse direction $\boldsymbol{\sigma} \in \mathbb{S}^2$. Suppose that $(\boldsymbol{v}', \boldsymbol{v}'_\star) = T^{i,j}_{\boldsymbol{\sigma}}(\boldsymbol{v}, \boldsymbol{v}_\star)$. Then,*

$$m_i \boldsymbol{v} + m_j \boldsymbol{v}_\star = m_i \boldsymbol{v}' + m_j \boldsymbol{v}'_\star, \tag{2.5}$$

$$\frac{m_i}{2}|\boldsymbol{v}|^2 + \frac{m_j}{2}|\boldsymbol{v}_\star|^2 = \frac{m_i}{2}|\boldsymbol{v}'|^2 + \frac{m_j}{2}|\boldsymbol{v}'_\star|^2. \tag{2.6}$$

Furthermore,

$$\boldsymbol{\sigma} \cdot (\boldsymbol{v}' - \boldsymbol{v}'_\star) = -\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) \tag{2.7}$$

and

$$|\boldsymbol{v}' - \boldsymbol{v}'_\star| = |\boldsymbol{v} - \boldsymbol{v}_\star|. \tag{2.8}$$

Proof. To get (2.5) add m_i times (2.1) to m_j times (2.2), and the result immediately follows. To get (2.6), square both (2.1) and (2.2) to obtain

$$|\boldsymbol{v}'|^2 = |\boldsymbol{v}|^2 - [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]\frac{4m_j}{m_i + m_j}\boldsymbol{\sigma} \cdot \boldsymbol{v} + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]^2 \frac{4m_j^2}{(m_i + m_j)^2},$$

$$|\boldsymbol{v}'_\star|^2 = |\boldsymbol{v}_\star|^2 + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]\frac{4m_i}{m_i + m_j}\boldsymbol{\sigma} \cdot \boldsymbol{v}_\star + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]^2 \frac{4m_i^2}{(m_i + m_j)^2}.$$

Then,

$$\begin{aligned} m_i|\boldsymbol{v}'|^2 + m_j|\boldsymbol{v}'_\star|^2 = m_i|\boldsymbol{v}|^2 + m_j|\boldsymbol{v}_\star|^2 \\ - [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]\frac{4m_i m_j}{m_i + m_j}\boldsymbol{\sigma} \cdot \boldsymbol{v} + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]\frac{4m_i m_j}{m_i + m_j}\boldsymbol{\sigma} \cdot \boldsymbol{v}_\star \end{aligned}$$

$$
\begin{aligned}
&+ [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]^2 \frac{4m_i m_j^2}{(m_i + m_j)^2} + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]^2 \frac{4m_i^2 m_j}{(m_i + m_j)^2} \\
&= m_i |\boldsymbol{v}|^2 + m_j |\boldsymbol{v}_\star|^2 \\
&\quad - [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]^2 \frac{4m_i m_j}{m_i + m_j} \\
&\quad + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]^2 \frac{4m_i m_j}{(m_i + m_j)^2} (m_i + m_j) \\
&= m_i |\boldsymbol{v}|^2 + m_j |\boldsymbol{v}_\star|^2.
\end{aligned}
$$

Subtracting the two velocity formulas gives

$$\boldsymbol{v}' - \boldsymbol{v}'_\star = (\boldsymbol{v} - \boldsymbol{v}_\star) - 2[\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)]\boldsymbol{\sigma}.$$

Thus the relative velocity is reflected across the plane perpendicular to $\boldsymbol{\sigma}$. Taking its scalar product with $\boldsymbol{\sigma}$ proves (2.7); taking its Euclidean norm proves (2.8), because an orthogonal reflection preserves length. ■

Proposition 2.5 (Linearity and Microscopic Reversibility)**.** *Let $m_i, m_j > 0$ be the masses of particles of species i and j, respectively, $i, j \in \{1, 2, \ldots, N\}$, and fix an apse direction $\boldsymbol{\sigma} \in \mathbb{S}^2$. Then, $T_{\boldsymbol{\sigma}}^{i,j}$ is a linear transformation, that is,*

$$T_{\boldsymbol{\sigma}}^{i,j}\left(\alpha(\boldsymbol{v}, \boldsymbol{v}_\star) + \beta(\boldsymbol{w}, \boldsymbol{w}_\star)\right) = \alpha T_{\boldsymbol{\sigma}}^{i,j}\left(\boldsymbol{v}, \boldsymbol{v}_\star\right) + \beta T_{\boldsymbol{\sigma}}^{i,j}\left(\boldsymbol{w}, \boldsymbol{w}_\star\right), \tag{2.9}$$

for all $\alpha, \beta \in \mathbb{R}$ and all $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{v}_\star, \boldsymbol{w}_\star \in \mathbb{R}^3$. Furthermore,

$$(\boldsymbol{v}, \boldsymbol{v}_\star) = T_{\boldsymbol{\sigma}}^{i,j}(T_{\boldsymbol{\sigma}}^{i,j}(\boldsymbol{v}, \boldsymbol{v}_\star)), \tag{2.10}$$

for all $\boldsymbol{v}, \boldsymbol{v}_\star \in \mathbb{R}^3$. In other words,

$$I = T_{\boldsymbol{\sigma}}^{i,j} \circ T_{\boldsymbol{\sigma}}^{i,j}, \tag{2.11}$$

or, equivalently,

$$\left(T_{\boldsymbol{\sigma}}^{i,j}\right)^{-1} = T_{\boldsymbol{\sigma}}^{i,j}. \tag{2.12}$$

Proof. Each component of $T_{\boldsymbol{\sigma}}^{i,j}$ is a linear combination of $\boldsymbol{v}$ and $\boldsymbol{v}_\star$, so linearity follows directly from the definition. Proposition 2.4 shows that the collision replaces the relative velocity $\boldsymbol{v} - \boldsymbol{v}_\star$ by its reflection across the plane perpendicular to $\boldsymbol{\sigma}$. In particular, $\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)$ changes sign. On applying the same formulas a second time, the increments in (2.1)–(2.2) therefore have the same magnitudes and the opposite signs. Both velocities return to their original values, which proves (2.10). ■

Proposition 2.6 (Compatibility with Particle Exchange)**.** *Let $m_i, m_j > 0$ be the masses of particles of species i and j, respectively, $i, j \in \{1, 2, \ldots, N\}$, and fix an apse direction $\boldsymbol{\sigma} \in \mathbb{S}^2$. Then,*

$$T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v}_\star, \boldsymbol{v}) = T^{j,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star), \tag{2.13}$$

$$T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v}_\star, \boldsymbol{v}) = T^{j,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star). \tag{2.14}$$

Proof. Let $\boldsymbol{v}, \boldsymbol{v}_\star \in \mathbb{R}^3$ be arbitrary. Then,

$$\begin{aligned} T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v}_\star, \boldsymbol{v}) &= \boldsymbol{v}_\star - [\boldsymbol{\sigma} \cdot (\boldsymbol{v}_\star - \boldsymbol{v})] \frac{2m_j}{m_i + m_j} \boldsymbol{\sigma} \\ &= \boldsymbol{v}_\star + [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)] \frac{2m_j}{m_i + m_j} \boldsymbol{\sigma} \\ &= T^{j,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star). \end{aligned}$$

Similarly,

$$\begin{aligned} T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v}_\star, \boldsymbol{v}) &= \boldsymbol{v} + [\boldsymbol{\sigma} \cdot (\boldsymbol{v}_\star - \boldsymbol{v})] \frac{2m_i}{m_i + m_j} \boldsymbol{\sigma} \\ &= \boldsymbol{v} - [\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star)] \frac{2m_i}{m_i + m_j} \boldsymbol{\sigma} \\ &= T^{j,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star). \end{aligned}$$

■

Proposition 2.7 (Unit-Jacobian Change of Variables)**.** *Let $m_i, m_j > 0$ be the masses of particles of species i and j, respectively, $i, j \in \{1, 2, \ldots, N\}$, and fix an apse direction $\boldsymbol{\sigma} \in \mathbb{S}^2$. Suppose that $f : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ is integrable, that is,*

$$\int_{\mathbb{R}^3} \int_{\mathbb{R}^3} |f(\boldsymbol{v}, \boldsymbol{v}_\star)| \, \mathrm{d}\boldsymbol{v} \mathrm{d}\boldsymbol{v}_\star < \infty.$$

Then, $f \circ T^{i,j}_{\boldsymbol{\sigma}} : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ is integrable, and

$$\int_{\mathbb{R}^3} \int_{\mathbb{R}^3} f(\boldsymbol{v}, \boldsymbol{v}_\star) \, \mathrm{d}\boldsymbol{v} \mathrm{d}\boldsymbol{v}_\star = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} f(T^{i,j}_{\boldsymbol{\sigma}}(\boldsymbol{v}, \boldsymbol{v}_\star)) \, \mathrm{d}\boldsymbol{v} \mathrm{d}\boldsymbol{v}_\star. \tag{2.15}$$

Proof. Because $T^{i,j}_{\boldsymbol{\sigma}}$ is an invertible linear map and $(T^{i,j}_{\boldsymbol{\sigma}})^2 = I$, its determinant satisfies $(\det T^{i,j}_{\boldsymbol{\sigma}})^2 = 1$. Hence

$$\left|\det \left(T^{i,j}_{\boldsymbol{\sigma}}\right)\right| = 1.$$

The result now follows from the standard change-of-variables formula. This unit Jacobian is the measure-theoretic expression of microscopic reversibility. ■

2.2 The Boltzmann Collision Operator

Definition 2.8 (Binary Collision Operator and Kernel Symmetries)**.** *Let $m_i, m_j > 0$ be the particle masses of species i and j, respectively. At a fixed position $\boldsymbol{x}$ and time t, suppress that dependence and regard $f_i, f_j : \mathbb{R}^3 \to [0,\infty)$ as functions of velocity. The* ***binary i,j collision operator*** *is*

$$Q_{i,j}[f_i, f_j](\boldsymbol{v}) := \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f_i' f_{j,\star}' - f_i f_{j,\star}\right] B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star, \boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star, \tag{2.16}$$

where

$$f_i' := f_i(\boldsymbol{v}'), \quad f_{j,\star}' := f_j(\boldsymbol{v}_\star'),$$
$$f_i := f_i(\boldsymbol{v}), \quad f_{j,\star} := f_j(\boldsymbol{v}_\star);$$

and the primed coordinates are given precisely by the transformation

$$(\boldsymbol{v}', \boldsymbol{v}_\star') = T_{\boldsymbol{\sigma}}^{i,j}(\boldsymbol{v}, \boldsymbol{v}_\star).$$

$B_{i,j} : \mathbb{R}^3 \times \mathbb{S}^2 \to [0,\infty)$ is the ***collision kernel****. We say that the family of kernels has an* ***admissible symmetry structure*** *if*

1. *for every pair i,j, there is a measurable function $b_{i,j}$ such that, for all $\boldsymbol{\sigma} \in \mathbb{S}^2$ and $\boldsymbol{w} \in \mathbb{R}^3$,*
$$B_{i,j}(\boldsymbol{w}, \boldsymbol{\sigma}) = b_{i,j}(|\boldsymbol{w}\cdot\boldsymbol{\sigma}|, |\boldsymbol{w}|), \tag{2.17}$$
2. *for $i \neq j$, the reciprocity relation*
$$B_{i,j}(\boldsymbol{w}, \boldsymbol{\sigma}) = B_{j,i}(\boldsymbol{w}, \boldsymbol{\sigma}), \tag{2.18}$$
for all $\boldsymbol{w} \in \mathbb{R}^3$ and all $\boldsymbol{\sigma} \in \mathbb{S}^2$.

The first condition makes the kernel invariant under reversal of the relative velocity, reversal of the apse direction, and the pre/post-collisional change of variables. The second says that the physical rate assigned to an i,j encounter is independent of which particle is listed first. These symmetry assumptions are distinct from whatever cutoff, growth, or integrability hypotheses

are needed to make a particular collision integral finite. With the number-density convention $n_i = \int f_i \, \mathrm{d}\boldsymbol{v}$*, no additional factor of* m_i^{-1} *appears in* $Q_{i,j}$*.*

The integrand has the standard **gain-minus-loss** interpretation. The product $f_i' f_{j,\star}'$ counts collisions whose post-collisional state arrives at velocity $\boldsymbol{v}$, whereas $f_i f_{j,\star}$ counts collisions that remove particles from that velocity. The collision kernel weights those events by their relative speed and scattering law. Although the notation is compact, $Q_{i,j}$ remains nonlinear and nonlocal in velocity.

Definition 2.9 (Nondegenerate Kernels and the Interaction Graph)**.** *An* i, j *collision kernel is called* ***nondegenerate****, in the functional sense used below, if it is positive on a sufficiently rich set of collision configurations that equality in the entropy-production integral forces the corresponding collision-invariant relation. This terminology packages a hypothesis rather than prescribing one universal pointwise lower bound; precise coercivity assumptions depend on the kernel class. The* ***species-interaction graph*** *has vertices* $1, \dots, N$ *and an edge between* i *and* j *whenever* $B_{i,j}$ *is nondegenerate. If this graph is connected, collisions can communicate momentum and energy between every pair of species, possibly through intermediate species.*

Proposition 2.10 (Pre/Post Invariance of the Collision Kernel)**.** *Suppose that the collision kernel* $B_{i,j} : \mathbb{R}^3 \times \mathbb{S}^2 \to [0, \infty)$ *has admissible symmetry structure. Let* $\boldsymbol{v}, \boldsymbol{v}_\star \in \mathbb{R}^3$ *be arbitrary and set*

$$(\boldsymbol{v}', \boldsymbol{v}_\star') = T_{\boldsymbol{\sigma}}^{i,j}(\boldsymbol{v}, \boldsymbol{v}_\star),$$

as usual. Then,

$$B_{i,j}(\boldsymbol{v}' - \boldsymbol{v}_\star', \boldsymbol{\sigma}) = B_{i,j}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}). \tag{2.19}$$

Proof. Recall (2.7) and (2.8), respectively,

$$\boldsymbol{\sigma} \cdot (\boldsymbol{v}' - \boldsymbol{v}_\star') = -\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) \quad \text{and} \quad |\boldsymbol{v}' - \boldsymbol{v}_\star'| = |\boldsymbol{v} - \boldsymbol{v}_\star|. \tag{2.20}$$

Then,

$$\begin{aligned} B_{i,j}(\boldsymbol{v}' - \boldsymbol{v}_\star', \boldsymbol{\sigma}) &= b_{i,j}(|(\boldsymbol{v}' - \boldsymbol{v}_\star') \cdot \boldsymbol{\sigma}|, |\boldsymbol{v}' - \boldsymbol{v}_\star'|) \\ &= b_{i,j}(|-(\boldsymbol{v} - \boldsymbol{v}_\star) \cdot \boldsymbol{\sigma}|, |\boldsymbol{v} - \boldsymbol{v}_\star|) \\ &= b_{i,j}(|(\boldsymbol{v} - \boldsymbol{v}_\star) \cdot \boldsymbol{\sigma}|, |\boldsymbol{v} - \boldsymbol{v}_\star|) \\ &= B_{i,j}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}), \end{aligned}$$

which establishes (2.19). ∎

Definition 2.11 (Multispecies Boltzmann System)**.** *Let $\Omega \subset \mathbb{R}^3$ be a bounded open set. We say that the kinetic distributions $f_i : \Omega \times \mathbb{R}^3 \times [0,T] \to [0,\infty)$, $i = 1,\dots,N$, satisfy the* ***multispecies Boltzmann equation*** *if, for all $\boldsymbol{x} \in \Omega$, $\boldsymbol{v} \in \mathbb{R}^3$, and $t \in (0,T]$,*

$$\partial_t f_i(\boldsymbol{x},\boldsymbol{v},t) + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x},\boldsymbol{v},t) = Q_i[f_i,\boldsymbol{f}](\boldsymbol{x},\boldsymbol{v},t), \tag{2.21}$$

where $Q_i[f_i,\boldsymbol{f}]$, the ***cumulative binary collision operator****, is defined as*

$$Q_i[f_i,\boldsymbol{f}](\boldsymbol{x},\boldsymbol{v},t) \coloneqq \sum_{j=1}^{N} Q_{i,j}[f_i,f_j](\boldsymbol{x},\boldsymbol{v},t). \tag{2.22}$$

We often impose a ***specular reflection boundary condition****. Define the incoming phase boundary by*

$$\Gamma_{\text{in}} \coloneqq \left\{(\boldsymbol{x},\boldsymbol{v}) \in \partial\Omega \times \mathbb{R}^3 \;\middle|\; \boldsymbol{v}\cdot\boldsymbol{n}(\boldsymbol{x}) < 0\right\},$$

For every $(\boldsymbol{x},\boldsymbol{v}) \in \Gamma_{\text{in}}$ and $t \in (0,T]$, require

$$f_i(\boldsymbol{x},\boldsymbol{v},t) = f_i(\boldsymbol{x},\boldsymbol{R}(\boldsymbol{x},\boldsymbol{v}),t),$$

where

$$\boldsymbol{R}(\boldsymbol{x},\boldsymbol{v}) \coloneqq \boldsymbol{v} - 2\left[\boldsymbol{v}\cdot\boldsymbol{n}(\boldsymbol{x})\right]\boldsymbol{n}(\boldsymbol{x}),$$

and $\boldsymbol{n}(\boldsymbol{x})$ is the outward-pointing unit normal at $\boldsymbol{x} \in \partial\Omega$. The map $\boldsymbol{R}$ reverses the normal component of velocity and leaves its tangential component unchanged; it is the velocity-space analogue of a perfectly elastic reflection from a smooth wall.

Remark 2.12. *In the following weak-form identities, $\boldsymbol{x}$ and t are fixed and suppressed. This is a notational simplification only: the collision operator is local in physical space and time, so the same identities apply pointwise in $(\boldsymbol{x},t)$ whenever the integrals exist.*

The purpose of a weak form is to transfer the collision operator onto a test function ψ. After averaging over particle exchange and the pre/post-collisional map, the factor multiplying the gain–loss term becomes the change of ψ across one collision. Conservation laws then follow simply by choosing a ψ whose pairwise change vanishes.

Proposition 2.13 (Unary Weak Form)**.** *Suppose that $f_i : \mathbb{R}^3 \to [0,\infty)$ and $\psi : \mathbb{R}^3 \to \mathbb{R}$ are given and integrable. Then, assuming the collision kernel $B_{i,i}$ has admissible symmetry structure,*

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i,f_i](\boldsymbol{v})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}$$

$$= \frac{1}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_i' f_{i,\star}' - f_i f_{i,\star}\right] B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\left[\psi+\psi_\star-\psi'-\psi_\star'\right]\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}, \tag{2.23}$$

where the primed coordinates are determined via

$$(\boldsymbol{v}',\boldsymbol{v}_\star') = T_{\boldsymbol{\sigma}}^{i,i}(\boldsymbol{v},\boldsymbol{v}_\star),$$

and, for shorthand

$$f_i \coloneqq f_i(\boldsymbol{v}),\quad f_{i,\star}\coloneqq f_i(\boldsymbol{v}_\star),\quad f_i'\coloneqq f_i(\boldsymbol{v}'),\quad f_{i,\star}'\coloneqq f_i(\boldsymbol{v}_\star'),$$

and

$$\psi \coloneqq \psi(\boldsymbol{v}),\quad \psi_\star\coloneqq\psi(\boldsymbol{v}_\star),\quad \psi'\coloneqq\psi(\boldsymbol{v}'),\quad \psi_\star'\coloneqq\psi(\boldsymbol{v}_\star'),$$

Proof. The proof uses only two symmetries. First, exchanging the dummy variables $\boldsymbol{v}$ and $\boldsymbol{v}_\star$ replaces ψ by $\psi_\star$ without changing the integral. Second, the unit-Jacobian pre/post change of variables replaces the gain–loss factor by its negative and replaces ψ by ψ'; combining both symmetries gives the corresponding formula with $\psi_\star'$. We record the four versions explicitly and then average them.

Let us set

$$J = \int_{\mathbb{R}^3} Q_{i,i}[f_i,f_i](\boldsymbol{v})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}.$$

Then, by definition,

$$\begin{aligned}
J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_i' f_{i,\star}' - f_i f_{i,\star}\right] B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}\\
&= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T_{\boldsymbol{\sigma},1}^{i,i}(\boldsymbol{v},\boldsymbol{v}_\star))f_i(T_{\boldsymbol{\sigma},2}^{i,i}(\boldsymbol{v},\boldsymbol{v}_\star))B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}\\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v})f_i(\boldsymbol{v}_\star)B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}.
\end{aligned} \tag{2.24}$$

The integration variables are dummy variables and may be interchanged. Doing so, we obtain

$$\begin{aligned}
J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T_{\boldsymbol{\sigma},1}^{i,i}(\boldsymbol{v}_\star,\boldsymbol{v}))f_i(T_{\boldsymbol{\sigma},2}^{i,i}(\boldsymbol{v}_\star,\boldsymbol{v}))B_{i,i}(\boldsymbol{v}_\star-\boldsymbol{v},\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{v}_\star\\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}_\star)f_i(\boldsymbol{v})B_{i,i}(\boldsymbol{v}_\star-\boldsymbol{v},\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{v}_\star\\
&= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T_{\boldsymbol{\sigma},1}^{i,i}(\boldsymbol{v}_\star,\boldsymbol{v}))f_i(T_{\boldsymbol{\sigma},2}^{i,i}(\boldsymbol{v}_\star,\boldsymbol{v}))B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}\\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v})f_i(\boldsymbol{v}_\star)B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v},
\end{aligned}$$

where we have used (2.17). Observe that, using (2.13) and (2.14) (with $i = j$),

$$T^{i,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}_\star, \boldsymbol{v}) = T^{i,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star),$$
$$T^{i,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}_\star, \boldsymbol{v}) = T^{i,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star),$$

so that

$$\begin{aligned}
J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T^{i,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star)) f_i(T^{i,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star)) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}) f_i(\boldsymbol{v}_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}'_\star) f_i(\boldsymbol{v}') B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}) f_i(\boldsymbol{v}_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}') f_i(\boldsymbol{v}'_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}) f_i(\boldsymbol{v}_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f'_i f'_{i,\star} - f_i f_{i,\star}\right] B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}. \qquad (2.25)
\end{aligned}$$

Next, using the change of variables formula in Proposition 2.7 and the fact that $I = T^{i,i}_{\boldsymbol{\sigma}} \circ T^{i,i}_{\boldsymbol{\sigma}}$, where I is the identity transformation, we have

$$\begin{aligned}
J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T^{i,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star)) f_i(T^{i,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star)) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}) f_i(\boldsymbol{v}_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}) f_i(\boldsymbol{v}_\star) B_{i,i}(\boldsymbol{v}' - \boldsymbol{v}'_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}') f_i(\boldsymbol{v}'_\star) B_{i,i}(\boldsymbol{v}' - \boldsymbol{v}'_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}.
\end{aligned}$$

Using (2.19), we find

$$\begin{aligned}
J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}) f_i(\boldsymbol{v}_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}') f_i(\boldsymbol{v}'_\star) B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\
&= -\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f'_i f'_{i,\star} - f_i f_{i,\star}\right] B_{i,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}. \qquad (2.26)
\end{aligned}$$

Finally, with (2.25) as the starting point, using the change of variables formula in Proposition 2.7, using the fact that $I = T^{i,i}_{\boldsymbol{\sigma}} \circ T^{i,i}_{\boldsymbol{\sigma}}$, and using the identity (2.19), we deduce that

$$J = -\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f_i' f_{i,\star}' - f_i f_{i,\star}\right] B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star, \boldsymbol{\sigma})\psi(\boldsymbol{v}_\star')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}. \tag{2.27}$$

Adding identities (2.24)–(2.27) and dividing by four gives the desired factor $\psi + \psi_\star - \psi' - \psi_\star'$. ■

Proposition 2.14 (Cross-Species Weak Form)**.** *For $i \neq j$, if $f_i, f_j : \mathbb{R}^3 \to [0,\infty)$ and $\psi : \mathbb{R}^3 \to \mathbb{R}$ are given and integrable, then, assuming the collision kernel $B_{i,j}$ has admissible symmetry structure,*

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} \\
&= \frac{1}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f_i' f_{j,\star}' - f_i f_{j,\star}\right] B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star, \boldsymbol{\sigma})\left[\psi - \psi'\right]\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T^{i,j}_{\boldsymbol{\sigma}}), \\
&\quad + \frac{1}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f_j' f_{i,\star}' - f_j f_{i,\star}\right] B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star, \boldsymbol{\sigma})\left[\psi_\star - \psi_\star'\right]\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T^{j,i}_{\boldsymbol{\sigma}}),
\end{aligned} \tag{2.28}$$

where, to be precise, for the first integral the primed coordinates are determined via

$$(\boldsymbol{v}', \boldsymbol{v}_\star') = T^{i,j}_{\boldsymbol{\sigma}}(\boldsymbol{v}, \boldsymbol{v}_\star),$$

and, in shorthand,

$$f_i := f_i(\boldsymbol{v}), \quad f_{j,\star} := f_j(\boldsymbol{v}_\star), \quad f_i' := f_i(\boldsymbol{v}'), \quad f_{j,\star}' := f_j(\boldsymbol{v}_\star'), \quad \psi := \psi(\boldsymbol{v}), \quad \psi' := \psi(\boldsymbol{v}'),$$

and, for the second integral the primed coordinates are determined via

$$(\boldsymbol{v}', \boldsymbol{v}_\star') = T^{j,i}_{\boldsymbol{\sigma}}(\boldsymbol{v}, \boldsymbol{v}_\star),$$

and, in shorthand,

$$f_j := f_j(\boldsymbol{v}), \quad f_{i,\star} := f_i(\boldsymbol{v}_\star), \quad f_j' := f_j(\boldsymbol{v}'), \quad f_{i,\star}' := f_i(\boldsymbol{v}_\star') \quad \psi_\star := \psi(\boldsymbol{v}_\star), \quad \psi_\star' := \psi(\boldsymbol{v}_\star').$$

Proof. The same two operations are used as in the unary proof. Particle exchange now also changes the ordered pair (i,j) to (j,i); the reciprocity condition on the kernel makes the two descriptions of the same physical encounter agree. We again display the four equivalent

forms before averaging. Set

$$J = \int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}.$$

Then, by definition,

$$\begin{aligned} J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f_i' f_{j,\star}' - f_i f_{j,\star}\right] B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v},\boldsymbol{v}_\star))f_j(T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v},\boldsymbol{v}_\star))B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v})f_j(\boldsymbol{v}_\star)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T^{i,j}_{\boldsymbol{\sigma}}). \end{aligned} \tag{2.29}$$

Interchanging the integration variables $\boldsymbol{v}$ and $\boldsymbol{v}_\star$, we obtain

$$\begin{aligned} J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v}_\star,\boldsymbol{v}))f_j(T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v}_\star,\boldsymbol{v}))B_{i,j}(\boldsymbol{v}_\star-\boldsymbol{v},\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{v}_\star \\ &\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}_\star)f_j(\boldsymbol{v})B_{i,j}(\boldsymbol{v}_\star-\boldsymbol{v},\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{v}_\star \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_j(T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v}_\star,\boldsymbol{v}))f_i(T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v}_\star,\boldsymbol{v}))B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_j(\boldsymbol{v})f_i(\boldsymbol{v}_\star)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}, \end{aligned}$$

where we have used (2.17). Observe that, using (2.13) and (2.14) (with $i \neq j$) and (2.18), we find that

$$\begin{aligned} J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_j(T^{j,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v},\boldsymbol{v}_\star))f_i(T^{j,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v},\boldsymbol{v}_\star))B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_j(\boldsymbol{v})f_i(\boldsymbol{v}_\star)B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f_j' f_{i,\star}' - f_j f_{i,\star}\right] B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T^{j,i}_{\boldsymbol{\sigma}}). \end{aligned} \tag{2.30}$$

Next, using the change of variables formula in Proposition 2.7 and the fact that $I = T^{i,j}_{\boldsymbol{\sigma}} \circ T^{i,j}_{\boldsymbol{\sigma}}$, where I is the identity transformation, we have

$$\begin{aligned} J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(T^{i,j}_{\boldsymbol{\sigma},1}(\boldsymbol{v},\boldsymbol{v}_\star))f_j(T^{i,j}_{\boldsymbol{\sigma},2}(\boldsymbol{v},\boldsymbol{v}_\star))B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v})f_j(\boldsymbol{v}_\star)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v})f_j(\boldsymbol{v}_\star)B_{i,j}(\boldsymbol{v}'-\boldsymbol{v}_\star',\boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \end{aligned}$$

$$- \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}')f_j(\boldsymbol{v}'_\star)B_{i,j}(\boldsymbol{v}'-\boldsymbol{v}'_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}.$$

Using (2.19), we then find

$$\begin{aligned} J &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v})f_j(\boldsymbol{v}_\star)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &\quad - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_i(\boldsymbol{v}')f_j(\boldsymbol{v}'_\star)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \\ &= -\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f'_i f'_{j,\star} - f_i f_{j,\star}\right] B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T^{i,j}_{\boldsymbol{\sigma}}). \end{aligned} \tag{2.31}$$

Finally, with (2.30) as the starting point, using the change of variables formula in Proposition 2.7, using the fact that $I = T^{j,i}_{\boldsymbol{\sigma}} \circ T^{j,i}_{\boldsymbol{\sigma}}$, and using the identity (2.19), we deduce that

$$J = -\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} \left[f'_j f'_{i,\star} - f_j f_{i,\star}\right] B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\psi(\boldsymbol{v}'_\star)\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T^{j,i}_{\boldsymbol{\sigma}}). \tag{2.32}$$

Adding identities (2.29)–(2.32) and dividing by four gives (2.28). ■

2.3 Collision Invariants

A collision invariant is a velocity function whose total value is unchanged by every binary encounter. When the kinetic equation is integrated against such a function, the collision term disappears. Collision invariants are therefore the bridge from microscopic elastic mechanics to macroscopic conservation laws.

Proposition 2.15 (Unary Collision Moments)**.** *Suppose that $f_i : \mathbb{R}^3 \to [0,\infty)$ is given and integrable. Then, assuming the collision kernel $B_{i,i}$ has admissible symmetry structure, it follows that*

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i,f_i](\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 0, \tag{2.33}$$

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i,f_i](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} = \boldsymbol{0}, \tag{2.34}$$

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i,f_i](\boldsymbol{v})|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v} = 0. \tag{2.35}$$

Proof. These results follow from identity (2.23) in Proposition 2.13 and the transformation invariants (binary collision invariants) in Proposition 2.4 for the case that $i = j$. ■

Definition 2.16 (Unary Collision Invariant)**.** *A function $\psi : \mathbb{R}^3 \to \mathbb{R}$ is called a **species** i*

***collisional invariant**, or a **unary collisional invariant**, if*

$$\psi(\boldsymbol{v}) + \psi(\boldsymbol{v}_\star) = \psi(\boldsymbol{v}') + \psi(\boldsymbol{v}'_\star) \tag{2.36}$$

for every admissible collision $(\boldsymbol{v}, \boldsymbol{v}_\star) \mapsto (\boldsymbol{v}', \boldsymbol{v}'_\star) = T^{i,i}_{\boldsymbol{\sigma}}(\boldsymbol{v}, \boldsymbol{v}_\star)$. Consequently,

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i, f_i](\boldsymbol{v})\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 0 \tag{2.37}$$

for every sufficiently regular distribution f_i for which the integral is defined.

Definition 2.17 (Multispecies Collision Invariant)**.** *A family $(\psi_1, \dots, \psi_N)$ is a **multispecies collisional invariant** if, for every active pair of species i, j,*

$$\psi_i(\boldsymbol{v}) + \psi_j(\boldsymbol{v}_\star) = \psi_i(\boldsymbol{v}') + \psi_j(\boldsymbol{v}'_\star) \tag{2.38}$$

for every admissible i, j collision. For a connected interaction graph, the fundamental invariant families have the form

$$\psi_i(\boldsymbol{v}) = C_i + m_i\boldsymbol{\beta}\cdot\boldsymbol{v} + \frac{Am_i}{2}|\boldsymbol{v}|^2, \tag{2.39}$$

where the constants C_i may depend on the species, while A and $\boldsymbol{\beta}$ are common to all species.

The basic examples are $\psi \equiv 1$, the three components of $\psi(\boldsymbol{v}) = \boldsymbol{v}$, and $\psi(\boldsymbol{v}) = |\boldsymbol{v}|^2$. After multiplication by the appropriate masses, they represent particle number, momentum, and kinetic energy.

Proposition 2.18 (Pairwise Conservation in Cross-Species Collisions)**.** *For $i \neq j$, if $f_i, f_j : \mathbb{R}^3 \to [0, \infty)$ are given and integrable, then, assuming the collision kernel $B_{i,j}$ has admissible symmetry structure, it follows that*

$$\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 0, \tag{2.40}$$

$$m_i\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} + m_j\int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} = \boldsymbol{0}, \tag{2.41}$$

$$m_i\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v} + m_j\int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v})|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v} = 0. \tag{2.42}$$

Proof. The identity (2.40) follows from (2.28) with $\psi \equiv 1$. Using $\psi(\boldsymbol{v}) = \boldsymbol{v}$ in (2.28), we find that

$$m_i\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} + m_j\int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}$$

$$
\begin{aligned}
&= \frac{m_i}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_i' f_{j,\star}' - f_i f_{j,\star}\right] B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\,[\boldsymbol{v}-\boldsymbol{v}']\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j}),\\
&\quad + \frac{m_i}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_j' f_{i,\star}' - f_j f_{i,\star}\right] B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\,[\boldsymbol{v}_\star-\boldsymbol{v}_\star']\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}),\\
&\quad + \frac{m_j}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_j' f_{i,\star}' - f_j f_{i,\star}\right] B_{j,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\,[\boldsymbol{v}-\boldsymbol{v}']\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}),\\
&\quad + \frac{m_j}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_i' f_{j,\star}' - f_i f_{j,\star}\right] B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\,[\boldsymbol{v}_\star-\boldsymbol{v}_\star']\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j})\\
&= \mathbf{0}, \qquad (2.43)
\end{aligned}
$$

where we have used the following binary momentum invariants from Proposition 2.4:

$$m_i\boldsymbol{v} + m_j\boldsymbol{v}_\star = m_i\boldsymbol{v}' + m_j\boldsymbol{v}_\star' \iff m_i(\boldsymbol{v}-\boldsymbol{v}') = m_j(\boldsymbol{v}_\star'-\boldsymbol{v}_\star) \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j}), \qquad (2.44)$$

$$m_j\boldsymbol{v} + m_i\boldsymbol{v}_\star = m_j\boldsymbol{v}' + m_i\boldsymbol{v}_\star' \iff m_j(\boldsymbol{v}-\boldsymbol{v}') = m_i(\boldsymbol{v}_\star'-\boldsymbol{v}_\star) \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}). \qquad (2.45)$$

Using $\psi(\boldsymbol{v}) = |\boldsymbol{v}|^2$ in (2.28) and the appropriate binary kinetic energy invariants from Proposition 2.4, the identity (2.42) immediately follows. ■

Proposition 2.19 (Full-Operator Conservation)**.** *Suppose that $i, j \in \{1, \dots, N\}$ are arbitrary and $f_i, f_j : \mathbb{R}^3 \to [0,\infty)$ are given and integrable. Assume that the collision kernel $B_{i,j}$ has admissible symmetry structure. Recall our definition*

$$Q_i[f_i, \boldsymbol{f}](\boldsymbol{v}) \coloneqq \sum_{j=1}^{N} Q_{i,j}[f_i, f_j](\boldsymbol{v}). \qquad (2.46)$$

Then,

$$\int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 0, \qquad (2.47)$$

$$\sum_{i=1}^{N} m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} = \mathbf{0}, \qquad (2.48)$$

$$\sum_{i=1}^{N} m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{v})|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v} = 0. \qquad (2.49)$$

Proof. We prove only (2.48). The other two identities are proved similarly. Set

$$J = \sum_{i=1}^{N} m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}.$$

Then,

$$\begin{aligned}
J &= \sum_{i=1}^{N} m_i \int_{\mathbb{R}^3} \sum_{j=1}^{N} Q_{i,j}[f_i, f_j](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} m_i \int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} \\
&= \sum_{i=2}^{N}\sum_{j=1}^{i-1} \left(m_i \int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} \right) \\
&\quad + \sum_{i=1}^{N} m_i \int_{\mathbb{R}^3} Q_{i,i}[f_i, f_i](\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} \\
&= \mathbf{0},
\end{aligned}$$

where we have used (2.34) for the diagonal terms and (2.41) for the off-diagonal terms. ■

The preceding examples exhaust the smooth unary collision invariants. The proof has four steps: subtract the quadratic Taylor data at the origin, show that the remainder is radial, use a midpoint identity to show that it is affine in $|\boldsymbol{v}|^2$, and finally use the normalization at the origin to make it vanish.

Proposition 2.20 (Classification of Unary Collision Invariants)**.** *Suppose that $\psi \in C^2(\mathbb{R}^3;\mathbb{R})$ has the property that*

$$\psi(\boldsymbol{v}) + \psi(\boldsymbol{v}_\star) = \psi(T^{i,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v}, \boldsymbol{v}_\star)) + \psi(T^{i,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v}, \boldsymbol{v}_\star)), \tag{2.50}$$

for all $\boldsymbol{v}, \boldsymbol{v}_\star \in \mathbb{R}^3$ and all $\boldsymbol{\sigma} \in \mathbb{S}^2$. Then, there exist constants $A, C \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^3$, such that

$$\psi(\boldsymbol{v}) = \frac{m_i A}{2}|\boldsymbol{v}|^2 + m_i\boldsymbol{\beta}\cdot\boldsymbol{v} + C. \tag{2.51}$$

In other words, ψ is a velocity-space quadratic function.

Proof. This proof is from [9]. See [36] for another. Define

$$\begin{aligned}
\frac{m_i A}{2} &= \frac{1}{6}\Delta\psi(\mathbf{0}), \\
m_i\boldsymbol{\beta} &= \nabla\psi(\mathbf{0}), \\
C &= \psi(\mathbf{0}).
\end{aligned}$$

Set

$$\phi(\boldsymbol{v}) = \psi(\boldsymbol{v}) - \left(\frac{m_i A}{2}|\boldsymbol{v}|^2 + m_i \boldsymbol{\beta}\cdot\boldsymbol{v} + C\right), \quad \forall\, \boldsymbol{v}\in\mathbb{R}^3.$$

Then $\phi \in C^2(\mathbb{R}^3;\mathbb{R})$ is a unary collisional invariant; that is, it satisfies (2.36), with the further property that

$$\phi(\mathbf{0}) = 0, \quad \nabla\phi(\mathbf{0}) = \mathbf{0}, \quad \Delta\phi(\mathbf{0}) = 0.$$

Now, suppose that $\boldsymbol{v}_\star \in \mathbb{R}^3 \setminus \{\mathbf{0}\}$ is fixed but, otherwise, arbitrary. Suppose $\boldsymbol{\sigma} \in \mathbb{S}^2$ is perpendicular to $\boldsymbol{v}_\star$, that is,

$$\boldsymbol{\sigma}\cdot\boldsymbol{v}_\star = 0.$$

Set, for $s \in \mathbb{R}\setminus\{0\}$,

$$\boldsymbol{v} = s\boldsymbol{\sigma}.$$

Then, under the transformation $T^{i,i}_{\boldsymbol{\sigma}}$,

$$\begin{aligned}\boldsymbol{v}' &\coloneqq T^{i,i}_{\boldsymbol{\sigma},1}(\boldsymbol{v},\boldsymbol{v}_\star)\\ &= \boldsymbol{v} - \boldsymbol{\sigma}\cdot(\boldsymbol{v}-\boldsymbol{v}_\star)\frac{2m_i}{m_i+m_i}\boldsymbol{\sigma}\\ &= \boldsymbol{v} - s\boldsymbol{\sigma}\\ &= \mathbf{0}.\end{aligned}$$

Furthermore,

$$\boldsymbol{v}'_\star \coloneqq T^{i,i}_{\boldsymbol{\sigma},2}(\boldsymbol{v},\boldsymbol{v}_\star) = \boldsymbol{v}_\star + s\boldsymbol{\sigma}.$$

Thus,

$$\begin{aligned}\phi(s\boldsymbol{\sigma}) + \phi(\boldsymbol{v}_\star) &= \phi(\boldsymbol{v}) + \phi(\boldsymbol{v}_\star)\\ &= \phi(\boldsymbol{v}') + \phi(\boldsymbol{v}'_\star)\\ &= \phi(\mathbf{0}) + \phi(\boldsymbol{v}_\star + s\boldsymbol{\sigma})\\ &= \phi(\boldsymbol{v}_\star + s\boldsymbol{\sigma}).\end{aligned}$$

Equivalently,

$$\phi(s\boldsymbol{\sigma}) = \phi(\boldsymbol{v}_\star + s\boldsymbol{\sigma}) - \phi(\boldsymbol{v}_\star).$$

Divide this identity by s and let $s\to 0$. Differentiability and the facts $\phi(\mathbf{0}) = 0$ and $\nabla\phi(\mathbf{0}) = \mathbf{0}$ give

$$\boldsymbol{\sigma}\cdot\nabla\phi(\boldsymbol{v}_\star) = 0.$$

Since $\boldsymbol{v}_\star \in \mathbb{R}^3 \setminus \{\mathbf{0}\}$ and $\boldsymbol{\sigma} \in \mathbb{S}^2$ are perpendicular but otherwise arbitrary, it must be that

$$\phi(\boldsymbol{v}) = g(|\boldsymbol{v}|), \quad \forall\, \boldsymbol{v} \in \mathbb{R}^3,$$

that is, ϕ is radially symmetric.

Next, assume that $\boldsymbol{n} \in \mathbb{S}^2$ is arbitrary. Set

$$\boldsymbol{v} = s\boldsymbol{n} \quad \text{and} \quad \boldsymbol{v}_\star = r\boldsymbol{n},$$

where

$$0 \leq s < r.$$

Pick $\boldsymbol{\sigma} \in \mathbb{S}^2$ such that

$$\boldsymbol{\sigma} \cdot \boldsymbol{n} = \pm\frac{1}{\sqrt{2}}.$$

With this particular choice,

$$\boldsymbol{v}' = s\boldsymbol{n} - (s-r)\frac{\pm 1}{\sqrt{2}}\boldsymbol{\sigma},$$
$$\boldsymbol{v}'_\star = r\boldsymbol{n} + (s-r)\frac{\pm 1}{\sqrt{2}}\boldsymbol{\sigma},$$

and

$$|\boldsymbol{v}'|^2 = \frac{s^2+r^2}{2},$$
$$|\boldsymbol{v}'_\star|^2 = \frac{s^2+r^2}{2}.$$

Since

$$\phi(\boldsymbol{v}) + \phi(\boldsymbol{v}_\star) = \phi(\boldsymbol{v}') + \phi(\boldsymbol{v}'_\star),$$

and ϕ is radially symmetric, we have

$$g(s) + g(r) = 2g\left(\sqrt{\frac{s^2+r^2}{2}}\right).$$

Define $h : [0,\infty) \to \mathbb{R}$ by $h(z) = g(\sqrt{z})$. The preceding identity becomes

$$h(a) + h(b) = 2h\left(\frac{a+b}{2}\right), \qquad a, b \geq 0.$$

This is the midpoint-affinity relation. Continuity of h, inherited from ϕ, implies that $h(z) =$

$az + b$ for constants a, b. Therefore

$$\phi(\boldsymbol{v}) = a|\boldsymbol{v}|^2 + b.$$

The normalization $\phi(\mathbf{0}) = 0$ gives $b = 0$, while $\Delta\phi(\mathbf{0}) = 0$ gives $6a = 0$. Hence $a = 0$ and $\phi \equiv 0$, which proves the asserted form of ψ. ■

The classification has a direct physical interpretation: elastic binary collisions have no independent smooth additive invariants beyond particle number, the three components of momentum, and kinetic energy.

2.4 Information, Entropy, and Boltzmann's H-Theorem

Boltzmann's H-functional measures the spread of the velocity distributions. Our sign convention is worth fixing at the outset: information is H, while physical entropy is its negative. The H-theorem will show that collisions decrease H and therefore increase entropy.

Definition 2.21 (Information and Entropy)**.** *Suppose that $f_1, \dots, f_N : \Omega \times \mathbb{R}^3 \times [0, T] \to [0, \infty)$ are kinetic distributions that satisfy the multispecies Boltzmann equation. The quantity $h : \Omega \times [0, T] \to \mathbb{R}$, defined via*

$$h(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) \, \mathrm{d}\boldsymbol{v},$$

*is called the **information density** of the multispecies mixture. The quantity*

$$H(t) = \int_{\Omega} h(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x}$$

*is called the **total information** of the mixture. The quantity $s : \Omega \times [0, T] \to \mathbb{R}$ defined via*

$$s(\boldsymbol{x}, t) \coloneqq -\sum_{i=1}^{N} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) \, \mathrm{d}\boldsymbol{v} = -h(\boldsymbol{x}, t),$$

*is called the **entropy density** of the multispecies mixture. The quantity*

$$S(t) = \int_{\Omega} s(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} = -H(t)$$

*is called the **total entropy** of the mixture.*

Remark 2.22 (Normalization of the logarithm)**.** *Strictly speaking, the logarithm should act on a dimensionless quantity. One may read $f_i \ln f_i$ below as the nondimensionalized expression or replace it by $f_i \ln(f_i/f_{i,\mathrm{ref}})$ for a fixed reference distribution scale. The resulting extra term is linear in f_i and therefore does not alter collision entropy production, because each species mass is conserved. The physical thermodynamic entropy also includes Boltzmann's constant and convention-dependent additive terms; here,* entropy *means the negative of Boltzmann's dimensionless H-functional.*

The sign of entropy production ultimately comes from the following elementary fact.

Lemma 2.23 (Logarithmic Monotonicity)**.** *For any $x, y \in (0, \infty)$,*

$$(\ln(x) - \ln(y))\,(x - y) \geq 0, \tag{2.52}$$

with equality if and only if $x = y$.

Proof. The logarithm is strictly increasing on $(0, \infty)$. Therefore $x - y$ and $\ln x - \ln y$ always have the same sign, proving the inequality. Strict monotonicity shows that equality occurs exactly when $x = y$. ■

Proposition 2.24 (H-Theorem)**.** *Suppose that $f_1, \ldots, f_N : \Omega \times \mathbb{R}^3 \times [0, T] \to [0, \infty)$ satisfy the multispecies Boltzmann equation. Assume that the collision kernels $B_{i,j}$ have admissible symmetry structure. Then, independently of the physical boundary condition,*

$$\partial_t h(\boldsymbol{x}, t) + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{J}_h(\boldsymbol{x}, t) = \sigma_h(\boldsymbol{x}, t) \leq 0, \tag{2.53}$$

*where $\boldsymbol{J}_h$ is the **information flux**, given by*

$$\boldsymbol{J}_h(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \int_{\mathbb{R}^3} \boldsymbol{v} f_i(\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right)\, \mathrm{d}\boldsymbol{v},$$

*and σ_h is the **collisional information rate**, defined by*

$$\sigma_h(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right)\, \mathrm{d}\boldsymbol{v}.$$

If, in addition, the boundary condition produces no net information flux—as happens for specular reflection—then the total information satisfies

$$\mathrm{d}_t H(t) = \int_{\Omega} \sigma_h(\boldsymbol{x}, t)\, \mathrm{d}\boldsymbol{x} \leq 0, \tag{2.54}$$

since, for specular reflection-type boundary conditions, it holds that

$$\boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{J}_h(\boldsymbol{x}, t) = 0, \quad \forall\, \boldsymbol{x} \in \partial\Omega, \tag{2.55}$$

where $\boldsymbol{n}(\boldsymbol{x})$ is the outward-pointing unit normal at $\boldsymbol{x} \in \partial\Omega$. Similarly,

$$\partial_t s(\boldsymbol{x}, t) + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{J}_s(\boldsymbol{x}, t) = \sigma_s(\boldsymbol{x}, t) \geq 0, \tag{2.56}$$

where $\boldsymbol{J}_s$ is the ***entropy flux****, given by*

$$\boldsymbol{J}_s \coloneqq -\sum_{i=1}^{N} \int_{\mathbb{R}^3} \boldsymbol{v} f_i(\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) \, \mathrm{d}\boldsymbol{v} = -\boldsymbol{J}_h,$$

and σ_s is the ***entropy production rate****, defined via*

$$\sigma_s(\boldsymbol{x}, t) \coloneqq -\sum_{i=1}^{N} \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) \, \mathrm{d}\boldsymbol{v} = -\sigma_h(\boldsymbol{x}, t).$$

The total entropy satisfies

$$\mathrm{d}_t S(t) = \int_{\Omega} \sigma_s(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} \geq 0, \tag{2.57}$$

since

$$\boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{J}_s(\boldsymbol{x}, t) = 0, \quad \forall\, \boldsymbol{x} \in \partial\Omega. \tag{2.58}$$

The statement contains a local law and a global consequence. The local inequality is produced entirely by collisions and is independent of the boundary condition. The global inequality additionally requires that information not flow through the boundary; specular reflection has exactly this property.

Proof. There are three steps. We first derive the local balance law by the chain rule. We then use the weak collision forms and logarithmic monotonicity to prove that the collisional rate is nonpositive. Finally, we integrate in space and use the boundary condition.

Multiply the i^{th} equation of the multispecies Boltzmann system (2.21) by $\ln(f_i) + 1$, integrate in velocity, and sum over i. The chain rule gives $\partial_t(f_i \ln f_i) = (\ln f_i + 1)\partial_t f_i$, and the velocity $\boldsymbol{v}$ is independent of $\boldsymbol{x}$. The terms containing 1 vanish by species-mass conservation (2.47), which yields (2.53).

Next, we want to show that

$$\sigma_h(\boldsymbol{x},t)=\sum_{i=1}^{N}\int_{\mathbb{R}^3}Q_i[f_i,\boldsymbol{f}](\boldsymbol{x},\boldsymbol{v},t)\ln\left(f_i(\boldsymbol{x},\boldsymbol{v},t)\right)\,\mathrm{d}\boldsymbol{v}\leq 0,\quad \forall\,\boldsymbol{x}\in\Omega,\quad \forall\,t\geq 0.$$

Let us write this as the sum of sums, a diagonal sum and an off-diagonal sum:

$$\begin{aligned}\sigma_h&=\sum_{i=1}^{N}\int_{\mathbb{R}^3}Q_{i,i}[f_i,f_i]\ln(f_i)\,\mathrm{d}\boldsymbol{v}\\&\quad+\sum_{i=2}^{N}\sum_{j=1}^{i-1}\int_{\mathbb{R}^3}Q_{i,j}[f_i,f_j]\ln(f_i)\,\mathrm{d}\boldsymbol{v}+\sum_{j=2}^{N}\sum_{i=1}^{j-1}\int_{\mathbb{R}^3}Q_{i,j}[f_i,f_j]\ln(f_i)\,\mathrm{d}\boldsymbol{v}\\&=\sum_{i=1}^{N}\int_{\mathbb{R}^3}Q_{i,i}[f_i,f_i]\ln(f_i)\,\mathrm{d}\boldsymbol{v}\\&\quad+\sum_{i=2}^{N}\sum_{j=1}^{i-1}\left[\int_{\mathbb{R}^3}Q_{i,j}[f_i,f_j]\ln(f_i)\,\mathrm{d}\boldsymbol{v}+\int_{\mathbb{R}^3}Q_{j,i}[f_j,f_i]\ln(f_j)\,\mathrm{d}\boldsymbol{v}\right].\end{aligned}\tag{2.59}$$

Let us tackle the diagonal terms first. Observe that, using (2.23),

$$\begin{aligned}&\int_{\mathbb{R}^3}Q_{i,i}[f_i,f_i]\ln(f_i)\,\mathrm{d}\boldsymbol{v}\\&=\frac{1}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\left[f_i'f_{i,\star}'-f_if_{i,\star}\right]B_{i,i}(\boldsymbol{v}-\boldsymbol{v}_\star,\boldsymbol{\sigma})\left[L_i+L_{i,\star}-L_i'-L_{i,\star}'\right]\,\mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}\quad(\prime\leftrightarrow T_{\boldsymbol{\sigma}}^{i,i}),\end{aligned}\tag{2.60}$$

where

$$L_i\coloneqq\ln(f_i),\quad L_{i,\star}\coloneqq\ln(f_{i,\star}),\quad L_i'=\ln(f_i'),\quad L_{i,\star}'=\ln(f_{i,\star}').$$

Now, observe that

$$L_i+L_{i,\star}-L_i'-L_{i,\star}'=\ln(f_if_{i,\star})-\ln(f_i'f_{i,\star}').$$

Using estimate (2.52), it follows that

$$\left[f_i'f_{i,\star}'-f_if_{i,\star}\right]\left[L_i+L_{i,\star}-L_i'-L_{i,\star}'\right]\leq 0,$$

for each $\boldsymbol{x}\in\Omega$, every $\boldsymbol{v}\in\mathbb{R}^3$, and every $t\geq 0$. Since $B_{i,i}(\boldsymbol{w},\boldsymbol{\sigma})\geq 0$, for all $\boldsymbol{w}\in\mathbb{R}^3$ and $\boldsymbol{\sigma}\in\mathbb{S}^2$, it follows that

$$\int_{\mathbb{R}^3}Q_{i,i}[f_i,f_i]\ln(f_i)\,\mathrm{d}\boldsymbol{v}\leq 0$$

for each $\boldsymbol{x} \in \Omega$ and every $t \geq 0$. As a consequence,

$$\sum_{i=1}^{N} \int_{\mathbb{R}^3} Q_{i,i}[f_i, f_i] \ln(f_i)\, \mathrm{d}\boldsymbol{v} \leq 0.$$

Next we address the off-diagonal sum in (2.59). Using (2.28), for $i > j$,

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v}) \ln(f_i(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v} \\
&= \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f_i' f_{j,\star}' - f_i f_{j,\star} \right] B_{i,j}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \left[L_i - L_i' \right] \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j}) \\
&\quad + \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f_j' f_{i,\star}' - f_j f_{i,\star} \right] B_{j,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \left[L_{i,\star} - L_{i,\star}' \right] \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}),
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v}) \ln(f_j(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v} \\
&= \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f_j' f_{i,\star}' - f_j f_{i,\star} \right] B_{j,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \left[L_j - L_j' \right] \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}) \\
&\quad + \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f_i' f_{j,\star}' - f_i f_{j,\star} \right] B_{i,j}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \left[L_{j,\star} - L_{j,\star}' \right] \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j}).
\end{aligned}$$

Thus, for $i > j$,

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v}) \ln(f_i(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v}) \ln(f_j(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v} \\
&= \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f_i' f_{j,\star}' - f_i f_{j,\star} \right] B_{i,j}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \left[L_i - L_i' + L_{j,\star} - L_{j,\star}' \right] \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j}) \\
&\quad + \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \left[f_j' f_{i,\star}' - f_j f_{i,\star} \right] B_{j,i}(\boldsymbol{v} - \boldsymbol{v}_\star, \boldsymbol{\sigma}) \left[L_{i,\star} - L_{i,\star}' + L_j - L_j' \right] \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v} \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}).
\end{aligned} \tag{2.61}$$

Using the logarithmic estimate (2.52) and the fact that $B_{i,j}, B_{j,i} \geq 0$, we conclude that

$$\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j](\boldsymbol{v}) \ln(f_i(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i](\boldsymbol{v}) \ln(f_j(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v} \leq 0.$$

Both terms in (2.59) are nonpositive, and we conclude that $\sigma_h \leq 0$. This proves (2.53).

To prove (2.54), integrate (2.53) over Ω, use the divergence theorem, the no-flux identity (2.55), and the pointwise inequality $\sigma_h \leq 0$. For specular reflection, the no-flux identity follows by pairing each boundary velocity $\boldsymbol{v}$ with its reflected velocity $\boldsymbol{R}(\boldsymbol{x}, \boldsymbol{v})$; the distribution

and velocity measure are unchanged, while the normal velocity changes sign.

The results for the entropy density and the total entropy are simple consequences of the corresponding results for the information density and total information, noting only that $s = -h$ and $S = -H$. ■

Proposition 2.25 (Equilibrium Maxwellians)**.** *Suppose that $f_1, \dots, f_N : \Omega \times \mathbb{R}^3 \times [0, T] \to (0, \infty)$ are sufficiently regular kinetic distributions with finite information. Assume that the collision kernels $B_{i,j}$ have admissible symmetry structure, every unary kernel $B_{i,i}$ is nondegenerate, and the graph formed by the nondegenerate interspecies kernels is connected. Then,*

$$\sigma_h(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{x}, \boldsymbol{v}, t) \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) \, \mathrm{d}\boldsymbol{v} = 0 \tag{2.62}$$

if and only if, for $i = 1, \dots, N$,

$$\ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) = \frac{m_i A(\boldsymbol{x}, t)}{2} |\boldsymbol{v}|^2 + m_i \boldsymbol{\beta}(\boldsymbol{x}, t) \cdot \boldsymbol{v} + C_i(\boldsymbol{x}, t), \tag{2.63}$$

where $A < 0$ and $\boldsymbol{\beta}$ are independent of i and $\boldsymbol{v}$. Equivalently, there exist a common velocity $\boldsymbol{u}$ and common temperature $\theta > 0$ such that

$$f_i(\boldsymbol{x}, \boldsymbol{v}, t) = n_i(\boldsymbol{x}, t) \left(\frac{m_i}{2\pi\theta(\boldsymbol{x}, t)}\right)^{3/2} \exp\left(-\frac{m_i |\boldsymbol{v} - \boldsymbol{u}(\boldsymbol{x}, t)|^2}{2\theta(\boldsymbol{x}, t)}\right). \tag{2.64}$$

Proof. The proof separates the roles of unary and cross-species collisions. Zero production in the unary terms makes each f_i a Maxwellian with its own velocity and temperature parameters. Zero production in every active cross-species term then forces those velocity and temperature parameters to agree along each edge of the interaction graph. Connectivity propagates the agreement to all species.

We have seen from the proof of the last proposition that

$$\begin{aligned} \sigma_h &= \sum_{i=1}^{N} \int_{\mathbb{R}^3} Q_{i,i}[f_i, f_i] \ln(f_i) \, \mathrm{d}\boldsymbol{v} \\ &\quad + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \left[\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j] \ln(f_i) \, \mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i] \ln(f_j) \, \mathrm{d}\boldsymbol{v} \right] \\ &\leq 0. \end{aligned} \tag{2.65}$$

More specifically,

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i, f_i] \ln(f_i)\, \mathrm{d}\boldsymbol{v} \leq 0, \quad i \in \{1, \ldots, N\},$$

and

$$\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j] \ln(f_i)\, \mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i] \ln(f_j)\, \mathrm{d}\boldsymbol{v} \leq 0, \quad i, j \in \{1, \ldots, N\}, \quad i > j.$$

If (2.63) holds, then $\sigma_h = 0$, using (2.33) – (2.35) ($i = j$) and (2.40) – (2.42) ($i > j$). Specifically, for the latter case ($i > j$), observe that

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j] \ln(f_i)\, \mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i] \ln(f_j)\, \mathrm{d}\boldsymbol{v} \\
&\qquad = \frac{A(\boldsymbol{x}, t)}{2} \left(m_i \int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j] |\boldsymbol{v}|^2\, \mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i] |\boldsymbol{v}|^2\, \mathrm{d}\boldsymbol{v} \right) \\
&\qquad\quad + \boldsymbol{\beta}(\boldsymbol{x}, t) \cdot \left(m_i \int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j] \boldsymbol{v}\, \mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i] \boldsymbol{v}\, \mathrm{d}\boldsymbol{v} \right) \\
&\qquad\quad + \left(C_i(\boldsymbol{x}, t) \int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j]\, \mathrm{d}\boldsymbol{v} + C_j(\boldsymbol{x}, t) \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i]\, \mathrm{d}\boldsymbol{v} \right) \\
&\qquad = 0,
\end{aligned}$$

thanks to (2.40) – (2.42).

Conversely, if $\sigma_h(\boldsymbol{x}, t) = 0$ at a particular point $(\boldsymbol{x}, t) \in \Omega \times [0, \infty)$, then it must be that

$$\int_{\mathbb{R}^3} Q_{i,i}[f_i, f_i] \ln(f_i)\, \mathrm{d}\boldsymbol{v} = 0, \quad \forall\, i \in \{1, \ldots, N\}, \tag{2.66}$$

and

$$\int_{\mathbb{R}^3} Q_{i,j}[f_i, f_j] \ln(f_i)\, \mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^3} Q_{j,i}[f_j, f_i] \ln(f_j)\, \mathrm{d}\boldsymbol{v} = 0, \quad \forall\, i, j \in \{1, \ldots, N\}, \quad i > j. \tag{2.67}$$

Using Equation (2.66), Equation (2.60), Lemma 2.23, and nondegeneracy of $B_{i,i}$, it must be that

$$L_i + L_{i,\star} - L_i' - L_{i,\star}' = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,i}),$$

for almost every admissible collision configuration. Under the standing regularity convention, this identity extends to all collision configurations. Invoking Proposition 2.20, it follows that, for $i = 1, \ldots, N$ and all $\boldsymbol{v} \in \mathbb{R}^3$,

$$L_i = \ln\left(f_i(\boldsymbol{x}, \boldsymbol{v}, t)\right) = \frac{m_i A_i(\boldsymbol{x}, t)}{2} |\boldsymbol{v}|^2 + m_i \boldsymbol{\beta}_i(\boldsymbol{x}, t) \cdot \boldsymbol{v} + C_i(\boldsymbol{x}, t). \tag{2.68}$$

Here, A_i and $\boldsymbol{\beta}_i$ are independent of $\boldsymbol{v}$, but we leave open the possibility that A_i and $\boldsymbol{\beta}_i$ depend upon i.

Next, using Equation (2.67), Equation (2.61), Lemma 2.23, and nondegeneracy of the active interspecies kernels, it must be that

$$L_i - L_i' + L_{j,\star} - L_{j,\star}' = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j})$$

and

$$L_{i,\star} - L_{i,\star}' + L_j - L_j' = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}),$$

for all $\boldsymbol{v}$, $\boldsymbol{v}_\star$, and $\boldsymbol{\sigma}$ associated with an active i,j edge. We now insert (2.68)—dropping the $\boldsymbol{x}$ and t dependencies, since these points are fixed—into the last two equations, to obtain

$$\begin{aligned}&\frac{m_i A_i}{2}|\boldsymbol{v}|^2 - \frac{m_i A_i}{2}|\boldsymbol{v}'|^2 + \frac{m_j A_j}{2}|\boldsymbol{v}_\star|^2 - \frac{m_j A_j}{2}|\boldsymbol{v}_\star'|^2 \\ &\quad + m_i\boldsymbol{\beta}_i \cdot \boldsymbol{v} - m_i\boldsymbol{\beta}_i \cdot \boldsymbol{v}' + m_j\boldsymbol{\beta}_j \cdot \boldsymbol{v}_\star - m_j\boldsymbol{\beta}_j \cdot \boldsymbol{v}_\star' = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j})\end{aligned}$$

and

$$\begin{aligned}&\frac{m_i A_i}{2}|\boldsymbol{v}_\star|^2 - \frac{m_i A_i}{2}|\boldsymbol{v}_\star'|^2 + \frac{m_j A_j}{2}|\boldsymbol{v}|^2 - \frac{m_j A_j}{2}|\boldsymbol{v}'|^2 \\ &\quad + m_i\boldsymbol{\beta}_i \cdot \boldsymbol{v}_\star - m_i\boldsymbol{\beta}_i \cdot \boldsymbol{v}_\star' + m_j\boldsymbol{\beta}_j \cdot \boldsymbol{v} - m_j\boldsymbol{\beta}_j \cdot \boldsymbol{v}' = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}),\end{aligned}$$

for all $\boldsymbol{v}$, $\boldsymbol{v}_\star$, and $\boldsymbol{\sigma}$. Now, we invoke the particle transformation invariances, (2.5) and (2.6), to obtain

$$(A_j - A_i)\left(\frac{m_j}{2}|\boldsymbol{v}_\star|^2 - \frac{m_j}{2}|\boldsymbol{v}_\star'|^2\right) + (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\cdot(m_j\boldsymbol{v}_\star - m_j\boldsymbol{v}_\star') = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{i,j}), \tag{2.69}$$

for all $\boldsymbol{v}$, $\boldsymbol{v}_\star$, and $\boldsymbol{\sigma}$, and

$$(A_j - A_i)\left(\frac{m_j}{2}|\boldsymbol{v}|^2 - \frac{m_j}{2}|\boldsymbol{v}'|^2\right) + (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\cdot(m_j\boldsymbol{v} - m_j\boldsymbol{v}') = 0 \quad (' \leftrightarrow T_{\boldsymbol{\sigma}}^{j,i}), \tag{2.70}$$

for all $\boldsymbol{v}$, $\boldsymbol{v}_\star$, and $\boldsymbol{\sigma}$.

Let us examine Equation (2.69). Recall, from the transformation $T_{\boldsymbol{\sigma}}^{i,j}$,

$$\boldsymbol{v}_\star' = \boldsymbol{v}_\star + [\boldsymbol{\sigma}\cdot(\boldsymbol{v} - \boldsymbol{v}_\star)]\mu_{i,j}\boldsymbol{\sigma}, \quad \mu_{i,j} \coloneqq \frac{2m_i}{m_i + m_j}.$$

Let $\boldsymbol{v}_\star$ be an arbitrary, nonzero vector. Choose

$$\boldsymbol{v} = \frac{\mu_{i,j} - 2}{\mu_{i,j}} \boldsymbol{v}_\star, \quad \boldsymbol{\sigma} = \frac{\boldsymbol{v}_\star}{|\boldsymbol{v}_\star|}.$$

Then, it follows that $\boldsymbol{v}'_\star = -\boldsymbol{v}_\star$. Thus, (2.69) indicates that

$$(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) \cdot \boldsymbol{v}_\star = 0.$$

Since $\boldsymbol{v}_\star$ was arbitrary, $\boldsymbol{\beta}_j = \boldsymbol{\beta}_i$ along every active edge. Connectivity of the species-interaction graph therefore implies that $\boldsymbol{\beta}_i = \boldsymbol{\beta}$ for all i, for some species-independent vector $\boldsymbol{\beta}$. Thus, (2.69) is equivalent to

$$(A_j - A_i) \left(\frac{m_j}{2} |\boldsymbol{v}_\star|^2 - \frac{m_j}{2} |\boldsymbol{v}'_\star|^2 \right) = 0 \quad (' \leftrightarrow T^{i,j}_{\boldsymbol{\sigma}}),$$

for all $\boldsymbol{v}$, $\boldsymbol{v}_\star$, and $\boldsymbol{\sigma}$. Again, let $\boldsymbol{v}_\star$ be an arbitrary, nonzero vector. This time choose

$$\boldsymbol{v} = \frac{\mu_{i,j} - 1}{\mu_{i,j}} \boldsymbol{v}_\star, \quad \boldsymbol{\sigma} = \frac{\boldsymbol{v}_\star}{|\boldsymbol{v}_\star|}.$$

Then $\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{v}_\star) = -|\boldsymbol{v}_\star|/\mu_{i,j}$, so the collision formula gives $\boldsymbol{v}'_\star = \boldsymbol{0}$. Thus,

$$(A_j - A_i) |\boldsymbol{v}_\star|^2 = 0.$$

As above, $A_i = A_j$ along every active edge, and connectivity implies that $A_i = A$ for all i. Integrability of each f_i forces $A < 0$. Completing the square in (2.63) gives (2.64) with common velocity $\boldsymbol{u} = -\boldsymbol{\beta}/A$ and common temperature $\theta = -1/A > 0$. ■

2.5 Moments and Moment Equations

Up to this point, we have worked directly with the kinetic distributions. We now compress that information into macroscopic velocity moments: densities, velocities, energies, stresses, and heat fluxes. Taking moments produces exact balance laws, but not a closed fluid system, because the fluxes involve higher moments of the unknown distribution. This is the **closure problem**; the later Chapman–Enskog expansion supplies one systematic approximation.

Mixtures require two velocity frames. Quantities with a hat are measured relative to the velocity $\boldsymbol{u}_i$ of species i; quantities without a hat are measured relative to the barycentric velocity $\boldsymbol{u}$ of the whole mixture. The difference $\boldsymbol{u}_i - \boldsymbol{u}$ is the diffusion velocity. Keeping this

convention in view prevents the most common source of confusion in the moment formulas.

Appendix G collects the complete notation map for these bulk quantities, their frame conversions, and their unary specializations.

Definition 2.26 (Densities and Velocities)**.** *Suppose that* $f_i : \Omega \times \mathbb{R}^3 \times [0,\infty) \to [0,\infty)$, $i = 1, \dots, N$ *are kinetic distributions with the property that*

$$\int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t)|\boldsymbol{v}|^k \,\mathrm{d}\boldsymbol{v} < \infty, \quad k = 0, 1, \tag{2.71}$$

for each $\boldsymbol{x} \in \Omega$ *and every* $t \geq 0$*. Assume that* m_i*, the* ***mass of the*** i^{th} ***particle****, is a positive real number. The quantity*

$$n_i(\boldsymbol{x}, t) \coloneqq \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) \,\mathrm{d}\boldsymbol{v} \tag{2.72}$$

is called the ***number density of the*** i^{th} ***species****, while*

$$\rho_i(\boldsymbol{x}, t) \coloneqq m_i n_i(\boldsymbol{x}, t) \tag{2.73}$$

is called the ***mass density of the*** i^{th} ***species****. The* ***number density of the mixture*** *is denoted*

$$n(\boldsymbol{x}, t) = \sum_{i=1}^{N} n_i(\boldsymbol{x}, t). \tag{2.74}$$

At points where $n_i > 0$*, the quantity*

$$\boldsymbol{u}_i(\boldsymbol{x}, t) \coloneqq \frac{1}{n_i(\boldsymbol{x}, t)} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t)\boldsymbol{v} \,\mathrm{d}\boldsymbol{v} \tag{2.75}$$

is called the ***bulk velocity of the*** i^{th} ***species****. The* ***total mass density of the mixture*** *is defined as*

$$\rho(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \rho_i(\boldsymbol{x}, t). \tag{2.76}$$

At points where $\rho > 0$*, the* ***barycentric velocity*** *(or* ***mixture velocity****) is defined as*

$$\boldsymbol{u}(\boldsymbol{x}, t) \coloneqq \frac{\sum_{i=1}^{N} \rho_i(\boldsymbol{x}, t)\boldsymbol{u}_i(\boldsymbol{x}, t)}{\rho(\boldsymbol{x}, t)}. \tag{2.77}$$

The ***diffusion velocity*** *of species* i *is*

$$\boldsymbol{V}_i(\boldsymbol{x}, t) \coloneqq \boldsymbol{u}_i(\boldsymbol{x}, t) - \boldsymbol{u}(\boldsymbol{x}, t). \tag{2.78}$$

Its mass-weighted sum vanishes, $\sum_i \rho_i \boldsymbol{V}_i = \boldsymbol{0}$*, which is precisely the defining property of the barycentric frame. The* ***peculiar velocity of the*** i^{th} ***species*** *is defined as*

$$\boldsymbol{c}_i(\boldsymbol{x}, \boldsymbol{v}, t) \coloneqq \boldsymbol{v} - \boldsymbol{u}_i(\boldsymbol{x}, t), \tag{2.79}$$

and the ***barycentric peculiar velocity*** *is defined as*

$$\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{v}, t) \coloneqq \boldsymbol{v} - \boldsymbol{u}(\boldsymbol{x}, t). \tag{2.80}$$

Definition 2.27 (Energies: Kinetic and Internal)**.** *Suppose that* $f_i : \Omega \times \mathbb{R}^3 \times [0, \infty) \to [0, \infty)$*,* $i = 1, \ldots, N$ *are kinetic distributions with the property that*

$$\int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) |\boldsymbol{v}|^k \, \mathrm{d}\boldsymbol{v} < \infty, \quad k = 0, 1, 2, \tag{2.81}$$

for each $\boldsymbol{x} \in \Omega$ *and every* $t \geq 0$*. The* ***total kinetic-energy density of the*** i^{th} ***species*** *is defined as*

$$\omega_i(\boldsymbol{x}, t) \coloneqq \frac{m_i}{2} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) |\boldsymbol{v}|^2 \, \mathrm{d}\boldsymbol{v}, \tag{2.82}$$

and the ***mixture total kinetic-energy density*** *is defined as*

$$\omega(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \omega_i(\boldsymbol{x}, t). \tag{2.83}$$

The ***species-frame internal-energy density of the*** i^{th} ***species*** *is defined by*

$$\hat{e}_i(\boldsymbol{x}, t) \coloneqq \frac{m_i}{2} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) |\boldsymbol{c}_i(\boldsymbol{x}, \boldsymbol{v}, t)|^2 \, \mathrm{d}\boldsymbol{v}, \tag{2.84}$$

and the ***total species-frame internal-energy density*** *is*

$$\hat{e}(\boldsymbol{x}, t) \coloneqq \sum_{i=1}^{N} \hat{e}_i(\boldsymbol{x}, t). \tag{2.85}$$

The i^{th} ***barycentric internal-energy contribution*** *is*

$$e_i(\boldsymbol{x}, t) \coloneqq \frac{m_i}{2} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t) |\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{v}, t)|^2 \, \mathrm{d}\boldsymbol{v}, \tag{2.86}$$

and the ***total barycentric internal energy density*** *is*

$$e(\boldsymbol{x}, t) = \sum_{i=1}^{N} e_i(\boldsymbol{x}, t). \tag{2.87}$$

The two internal-energy descriptions are related by

$$e_i = \hat{e}_i + \frac{\rho_i}{2}|\boldsymbol{u}_i - \boldsymbol{u}|^2, \tag{2.88}$$

and hence

$$e = \hat{e} + \frac{1}{2}\sum_{i=1}^{N} \rho_i|\boldsymbol{u}_i - \boldsymbol{u}|^2. \tag{2.89}$$

Thus, barycentric internal energy contains both the thermal energy measured in each species frame and the kinetic energy associated with diffusion of the species relative to the mixture velocity.

There are two natural ways to define temperatures in a gas mixture. Species-frame temperature measures random motion about $\boldsymbol{u}_i$ and excludes diffusion energy. Barycentric temperature measures motion about $\boldsymbol{u}$ and therefore includes the energy of species diffusion.

Definition 2.28 (Temperatures)**.** *Suppose that $f_i : \Omega \times \mathbb{R}^3 \times [0, \infty) \to [0, \infty)$, $i = 1, \ldots, N$ are kinetic distributions with the property that*

$$\int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t)|\boldsymbol{v}|^k \, \mathrm{d}\boldsymbol{v} < \infty, \quad k = 0, 1, 2, \tag{2.90}$$

for each $\boldsymbol{x} \in \Omega$ and every $t \geq 0$. Wherever $n_i > 0$, the ***barycentric temperature of the*** i^{th} ***species*** *and the* ***species-frame temperature of the*** i^{th} ***species*** *are defined, respectively, by*

$$\theta_i(\boldsymbol{x}, t) \coloneqq \frac{2}{3}\frac{e_i(\boldsymbol{x}, t)}{n_i(\boldsymbol{x}, t)} \quad \textit{and} \quad \hat{\theta}_i(\boldsymbol{x}, t) \coloneqq \frac{2}{3}\frac{\hat{e}_i(\boldsymbol{x}, t)}{n_i(\boldsymbol{x}, t)}. \tag{2.91}$$

Wherever $n > 0$, we define the ***species-frame mixture temperature****, $\hat{\theta}$, and the* ***barycentric mixture temperature****, θ, respectively, by*

$$\hat{\theta}(\boldsymbol{x}, t) \coloneqq \frac{2}{3n(\boldsymbol{x}, t)}\hat{e}(\boldsymbol{x}, t) \quad \textit{and} \quad \theta(\boldsymbol{x}, t) \coloneqq \frac{2}{3\,n(\boldsymbol{x}, t)}\, e(\boldsymbol{x}, t). \tag{2.92}$$

In expanded forms, we have

$$\hat{\theta}(\boldsymbol{x}, t) = \frac{1}{n(\boldsymbol{x}, t)}\sum_{i=1}^{N} n_i(\boldsymbol{x}, t)\hat{\theta}_i(\boldsymbol{x}, t) \quad \textit{and} \quad \theta(\boldsymbol{x}, t) = \frac{1}{n(\boldsymbol{x}, t)}\sum_{i=1}^{N} n_i(\boldsymbol{x}, t)\theta_i(\boldsymbol{x}, t). \tag{2.93}$$

Using (2.88)*, we also have*

$$\theta_i = \hat{\theta}_i + \frac{m_i}{3}|\boldsymbol{u}_i - \boldsymbol{u}|^2, \tag{2.94}$$

and

$$\theta = \hat{\theta} + \frac{1}{3n}\sum_{i=1}^{N} \rho_i |\boldsymbol{u}_i - \boldsymbol{u}|^2. \tag{2.95}$$

Remark 2.29. *Directly from the definition,*

$$\theta_i(\boldsymbol{x}, t) = \frac{m_i}{3n_i(\boldsymbol{x}, t)} \int_{\mathbb{R}^3} f_i(\boldsymbol{x}, \boldsymbol{v}, t)|\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{v}, t)|^2 \, \mathrm{d}\boldsymbol{v}. \tag{2.96}$$

Thus θ_i is nonnegative whenever f_i is nonnegative. The same argument applies to each temperature defined above.

Remark 2.30. *The phrase **first three moments** is common but slightly imprecise here. More explicitly, the hydrodynamic data are obtained from the zeroth moment, the vector-valued first moment, and the scalar trace of the second moment. Authors variously parameterize the same data by $(\rho_i, \boldsymbol{u}_i, \hat{\theta}_i)$ or by the conserved densities $(\rho_i, \rho_i \boldsymbol{u}_i, \omega_i)$. These parameterizations are equivalent wherever $\rho_i > 0$.*

Let us define the remaining macroscopic fluxes. We use a hat for quantities measured in the frame of a particular species and no hat for quantities measured in the barycentric frame. This convention will be especially useful when we later specialize to the unary equation, for then the two frames coincide.

Remark 2.31 (Tensor Notation)**.** *For vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the dyadic product $\boldsymbol{a} \otimes \boldsymbol{b}$ is the matrix with entries $(\boldsymbol{a} \otimes \boldsymbol{b})_{k\ell} = a_k b_\ell$. For a matrix field T, the divergence is taken row by row, $(\nabla \cdot \mathsf{T})_k = \sum_\ell \partial_{x_\ell} T_{k\ell}$. The Frobenius contraction used later is $\mathsf{A} : \mathsf{B} = \sum_{k,\ell} A_{k\ell} B_{k\ell}$. These conventions make momentum fluxes matrices and their divergences vectors.*

Definition 2.32 (Momentum and Pressure Tensors)**.** *Suppose that the second velocity moments of f_i are finite. The **momentum-flow tensor of the i^{th} species** and the **total momentum-flow tensor** are*

$$\mathsf{M}_i \coloneqq m_i \int_{\mathbb{R}^3} f_i \, \boldsymbol{v} \otimes \boldsymbol{v} \, \mathrm{d}\boldsymbol{v}, \qquad \mathsf{M} \coloneqq \sum_{i=1}^{N} \mathsf{M}_i. \tag{2.97}$$

*The **species-frame pressure tensor** and the **barycentric pressure contribution** of species i are, respectively,*

$$\widehat{\mathsf{P}}_i \coloneqq m_i \int_{\mathbb{R}^3} f_i \, \boldsymbol{c}_i \otimes \boldsymbol{c}_i \, \mathrm{d}\boldsymbol{v}, \qquad \mathsf{P}_i \coloneqq m_i \int_{\mathbb{R}^3} f_i \, \boldsymbol{c} \otimes \boldsymbol{c} \, \mathrm{d}\boldsymbol{v}. \tag{2.98}$$

The corresponding totals are

$$\widehat{\mathsf{P}} \coloneqq \sum_{i=1}^{N} \widehat{\mathsf{P}}_i, \qquad \mathsf{P} \coloneqq \sum_{i=1}^{N} \mathsf{P}_i. \tag{2.99}$$

Definition 2.33 (Energy and Heat Fluxes)**.** *Suppose that the third velocity moments of f_i are finite. The **energy flux of the** i^{th} **species** and the **total energy flux** are*

$$\boldsymbol{r}_i \coloneqq \frac{m_i}{2} \int_{\mathbb{R}^3} f_i \, \boldsymbol{v} |\boldsymbol{v}|^2 \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{r} \coloneqq \sum_{i=1}^{N} \boldsymbol{r}_i. \tag{2.100}$$

*The **species-frame heat flux** and the **barycentric heat-flux contribution** of species i are, respectively,*

$$\boldsymbol{\chi}_i \coloneqq \frac{m_i}{2} \int_{\mathbb{R}^3} f_i \, \boldsymbol{c}_i |\boldsymbol{c}_i|^2 \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{q}_i \coloneqq \frac{m_i}{2} \int_{\mathbb{R}^3} f_i \, \boldsymbol{c} |\boldsymbol{c}|^2 \, \mathrm{d}\boldsymbol{v}. \tag{2.101}$$

We write

$$\boldsymbol{\chi} \coloneqq \sum_{i=1}^{N} \boldsymbol{\chi}_i, \qquad \boldsymbol{q} \coloneqq \sum_{i=1}^{N} \boldsymbol{q}_i. \tag{2.102}$$

The pressure tensor is the second central moment, while the heat flux is one half of the mass-weighted third central moment. The following identities record precisely which velocity frame is used in each decomposition.

Proposition 2.34 (Macroscopic Flux Decompositions)**.** *Suppose that the velocity moments of every f_i through order three are finite. Then*

$$\omega_i = \frac{\rho_i}{2} |\boldsymbol{u}_i|^2 + \hat{e}_i, \tag{2.103}$$

$$\mathsf{M}_i = \rho_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i + \widehat{\mathsf{P}}_i, \tag{2.104}$$

$$\boldsymbol{r}_i = \omega_i \boldsymbol{u}_i + \widehat{\mathsf{P}}_i \boldsymbol{u}_i + \boldsymbol{\chi}_i. \tag{2.105}$$

Moreover,

$$\mathsf{P}_i = \widehat{\mathsf{P}}_i + \rho_i (\boldsymbol{u}_i - \boldsymbol{u}) \otimes (\boldsymbol{u}_i - \boldsymbol{u}), \tag{2.106}$$

$$\mathsf{M} = \rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}, \tag{2.107}$$

$$\omega = \frac{\rho}{2} |\boldsymbol{u}|^2 + e = \frac{\rho}{2} |\boldsymbol{u}|^2 + \frac{1}{2} \mathrm{tr}(\mathsf{P}), \tag{2.108}$$

$$\boldsymbol{r} = \omega \boldsymbol{u} + \mathsf{P} \boldsymbol{u} + \boldsymbol{q}. \tag{2.109}$$

In particular,

$$\frac{1}{2}\mathrm{tr}(\widehat{\mathsf{P}}_i) = \hat{e}_i, \qquad \frac{1}{2}\mathrm{tr}(\mathsf{P}_i) = e_i. \tag{2.110}$$

Proof. Write $\boldsymbol{v} = \boldsymbol{u}_i + \boldsymbol{c}_i$ in the definitions of ω_i, M_i, and $\boldsymbol{r}_i$. Every term that is odd and linear in $\boldsymbol{c}_i$ vanishes because

$$\int_{\mathbb{R}^3} f_i \boldsymbol{c}_i \,\mathrm{d}\boldsymbol{v} = \boldsymbol{0}. \tag{2.111}$$

This gives (2.103)–(2.105). Next observe that

$$\boldsymbol{c} = \boldsymbol{c}_i + (\boldsymbol{u}_i - \boldsymbol{u}). \tag{2.112}$$

Expanding $\boldsymbol{c} \otimes \boldsymbol{c}$ and again using the vanishing first central moment gives (2.106). Summing over i and using

$$\sum_{i=1}^{N} \rho_i (\boldsymbol{u}_i - \boldsymbol{u}) = \boldsymbol{0} \tag{2.113}$$

gives (2.107). The energy identity follows in the same fashion from $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{c}$. For the final flux identity, expand

$$\boldsymbol{v}|\boldsymbol{v}|^2 = (\boldsymbol{u} + \boldsymbol{c})\big(|\boldsymbol{u}|^2 + 2\boldsymbol{u}\cdot\boldsymbol{c} + |\boldsymbol{c}|^2\big).$$

The terms proportional to $\int f_i \boldsymbol{c}\,\mathrm{d}\boldsymbol{v}$ cancel after summation over the species by (2.113). The remaining terms are exactly $\omega\boldsymbol{u}$, $\mathsf{P}\boldsymbol{u}$, and $\boldsymbol{q}$, proving (2.109). Taking traces in (2.98) gives (2.110). ∎

We can now state the moment equations without ambiguity.

Proposition 2.35 (Exact Multispecies Moment Equations)**.** *Suppose that the kinetic distributions satisfy*

$$\partial_t f_i + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} f_i = Q_i[f_i, \boldsymbol{f}], \qquad Q_i[f_i, \boldsymbol{f}] = \sum_{j=1}^{N} Q_{i,j}[f_i, f_j], \tag{2.114}$$

and suppose that the hypotheses needed for the collisional conservation laws and the preceding integrations are satisfied. Then, for each species,

$$\partial_t \rho_i + \nabla_{\boldsymbol{x}}\cdot(\rho_i \boldsymbol{u}_i) = 0, \tag{2.115}$$

$$\partial_t(\rho_i \boldsymbol{u}_i) + \nabla_{\boldsymbol{x}}\cdot\Big(\rho_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i + \widehat{\mathsf{P}}_i\Big) = m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}]\,\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}, \tag{2.116}$$

$$\partial_t \omega_i + \nabla_{\boldsymbol{x}}\cdot\Big(\omega_i \boldsymbol{u}_i + \widehat{\mathsf{P}}_i \boldsymbol{u}_i + \boldsymbol{\chi}_i\Big) = \frac{m_i}{2}\int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}]|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v}. \tag{2.117}$$

After summing over the species and using conservation of total momentum and total energy in collisions, one obtains the barycentric total equations

$$\partial_t \rho + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u}) = 0, \tag{2.118}$$

$$\partial_t (\rho \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}) = \boldsymbol{0}, \tag{2.119}$$

$$\partial_t \omega + \nabla_{\boldsymbol{x}} \cdot (\omega \boldsymbol{u} + \mathsf{P}\boldsymbol{u} + \boldsymbol{q}) = 0. \tag{2.120}$$

Proof. Multiply the i^{th} equation in (2.114) by m_i, $m_i \boldsymbol{v}$, and $m_i |\boldsymbol{v}|^2/2$, respectively, and integrate over velocity. Moving the spatial derivative outside the velocity integral gives

$$\partial_t \rho_i + \nabla_{\boldsymbol{x}} \cdot (\rho_i \boldsymbol{u}_i) = m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}] \, \mathrm{d}\boldsymbol{v} = 0, \tag{2.121}$$

as well as

$$\partial_t (\rho_i \boldsymbol{u}_i) + \nabla_{\boldsymbol{x}} \cdot \mathsf{M}_i = m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}] \boldsymbol{v} \, \mathrm{d}\boldsymbol{v} \tag{2.122}$$

and

$$\partial_t \omega_i + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{r}_i = \frac{m_i}{2} \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}] |\boldsymbol{v}|^2 \, \mathrm{d}\boldsymbol{v}. \tag{2.123}$$

Equations (2.116) and (2.117) now follow from Equations (2.104) and (2.105). Sum on i, invoke the collisional invariants Equations (2.47) to (2.49), and use Equations (2.107) and (2.109) to obtain Equations (2.118) to (2.120). ■

The species momentum and energy source terms do not generally vanish: collisions exchange momentum and energy between species. Their sums vanish because every exchange appears once as a gain and once as a loss. Thus species mass, total mixture momentum, and total mixture energy are conserved.

Remark 2.36 (Unary Specialization)**.** *When $N = 1$, we have $\boldsymbol{u}_1 = \boldsymbol{u}$ and $\boldsymbol{c}_1 = \boldsymbol{c}$. Consequently,*

$$\hat{e}_1 = e_1 = e, \qquad \hat{\theta}_1 = \theta_1 = \theta, \qquad \widehat{\mathsf{P}}_1 = \mathsf{P}_1 = \mathsf{P}, \qquad \boldsymbol{\chi}_1 = \boldsymbol{q}_1 = \boldsymbol{q}. \tag{2.124}$$

The unary moment equations therefore take the particularly simple form

$$\partial_t \rho + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u}) = 0, \tag{2.125}$$

$$\partial_t (\rho \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}) = \boldsymbol{0}, \tag{2.126}$$

$$\partial_t \omega + \nabla_{\boldsymbol{x}} \cdot (\omega \boldsymbol{u} + \mathsf{P}\boldsymbol{u} + \boldsymbol{q}) = 0. \tag{2.127}$$

These equations are exact, but they are not closed: P *and* $\boldsymbol{q}$ *depend upon the full kinetic distribution. The Chapman–Enskog procedure developed below supplies their first-order hydrodynamic approximations.*

2.6 The Maxwellian and Maximal Entropy States

This section temporarily works in a generic velocity dimension $d \in \mathbb{N}$. A Maxwellian is the Gaussian selected by prescribed number density, mean velocity, and thermal energy. It plays three equivalent roles in the present discussion: collision equilibrium, the target state of BGK relaxation, and the unique entropy maximizer subject to the conserved moments.

Definition 2.37 (Maxwellian Associated with a Distribution)**.** *Suppose that* $\Omega \subset \mathbb{R}^d$ *is an open, bounded set and* $f : \Omega \times \mathbb{R}^d \times [0,\infty) \to [0,\infty)$ *is a kinetic distribution of particles of mass* $m > 0$*, with the property*

$$\int_{\mathbb{R}^d} f(\boldsymbol{x},\boldsymbol{v},t)|\boldsymbol{v}|^k \,\mathrm{d}\boldsymbol{v} < \infty, \quad k = 0,1,2, \tag{2.128}$$

for each $\boldsymbol{x} \in \Omega$ *and every* $t \geq 0$*. At points where* $n > 0$ *and* $\theta > 0$*, the* ***Maxwellian distribution associated with*** f *is*

$$M_f(\boldsymbol{x},\boldsymbol{v},t) \coloneqq M[f](\boldsymbol{x},\boldsymbol{v},t) \coloneqq n\left(\frac{m}{2\pi\theta}\right)^{d/2} \exp\left(-\frac{m|\boldsymbol{v}-\boldsymbol{u}|^2}{2\theta}\right), \tag{2.129}$$

where the hydrodynamic fields are obtained from the zeroth, first, and centered second moments of f*:*

$$n(\boldsymbol{x},t) \coloneqq \int_{\mathbb{R}^d} f \,\mathrm{d}\boldsymbol{v}, \quad \boldsymbol{u}(\boldsymbol{x},t) \coloneqq \frac{1}{n}\int_{\mathbb{R}^d} f\boldsymbol{v} \,\mathrm{d}\boldsymbol{v}, \quad \theta(\boldsymbol{x},t) \coloneqq \frac{m}{d\,n}\int_{\mathbb{R}^d} f|\boldsymbol{v}-\boldsymbol{u}|^2 \,\mathrm{d}\boldsymbol{v}. \tag{2.130}$$

The variable θ is temperature measured in energy units. In dimensional notation, $\theta = k_B T$, so the Gaussian width is determined by the ratio θ/m.

To shorten the formulas in this section only, write

$$\langle \boldsymbol{g} \rangle = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{v}) \,\mathrm{d}\boldsymbol{v}.$$

The brackets therefore mean integration over the entire velocity space and nothing more. In

this notation,

$$\begin{bmatrix} \rho(\boldsymbol{x},t) \\ \rho(\boldsymbol{x},t)\boldsymbol{u}(\boldsymbol{x},t) \\ \omega(\boldsymbol{x},t) \end{bmatrix} = \left\langle f(\boldsymbol{x},\boldsymbol{v},t) \begin{bmatrix} m \\ m\boldsymbol{v} \\ \frac{m}{2}|\boldsymbol{v}|^2 \end{bmatrix} \right\rangle,$$

where $m > 0$ is the mass of the particles with kinetic distribution f.

It is also convenient to collect the conserved moments and their velocity functions into vectors:

$$\boldsymbol{\rho}(\boldsymbol{x},t) \coloneqq \left[\rho, \rho\boldsymbol{u}^\top, \omega\right]^\top : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{d+2}; \tag{2.131}$$

$$\boldsymbol{m}(\boldsymbol{v}) \coloneqq \left[m, m\boldsymbol{v}^\top, \frac{m}{2}|\boldsymbol{v}|^2\right]^\top : \mathbb{R}^d \to \mathbb{R}^{d+2}; \tag{2.132}$$

Then, succinctly,

$$\boldsymbol{\rho} = \langle f\,\boldsymbol{m}\rangle. \tag{2.133}$$

The next result verifies the essential moment-matching property: replacing f by M_f changes the detailed velocity profile but not its mass, momentum, or total energy.

Lemma 2.38 (Moment Matching by the Maxwellian)**.** *Suppose that $f : \mathbb{R}^d \to [0,\infty)$ is a kinetic distribution of particles of mass $m > 0$, with the property*

$$\int_{\mathbb{R}^d} f(\boldsymbol{v})|\boldsymbol{v}|^k \,\mathrm{d}\boldsymbol{v} < \infty, \quad k = 0,1,2. \tag{2.134}$$

The following equality holds:

$$\langle (M_f - f)\,\boldsymbol{m}\rangle = \left\langle \begin{bmatrix} m \\ m\boldsymbol{v} \\ \frac{m}{2}|\boldsymbol{v}|^2 \end{bmatrix} (M_f - f) \right\rangle = \mathbf{0}. \tag{2.135}$$

Proof. Set

$$\boldsymbol{z} = \sqrt{\frac{m}{\theta}}(\boldsymbol{v} - \boldsymbol{u}), \qquad \gamma(\boldsymbol{z}) = (2\pi)^{-d/2} \exp\left(-\frac{|\boldsymbol{z}|^2}{2}\right).$$

Then $M_f(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = m\gamma(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$. The normalized Gaussian is centered and has identity covariance:

$$\int_{\mathbb{R}^d} \gamma \,\mathrm{d}\boldsymbol{z} = 1, \qquad \int_{\mathbb{R}^d} \boldsymbol{z}\gamma \,\mathrm{d}\boldsymbol{z} = \mathbf{0}, \qquad \int_{\mathbb{R}^d} \boldsymbol{z}\otimes\boldsymbol{z}\,\gamma \,\mathrm{d}\boldsymbol{z} = \mathsf{I}. \tag{2.136}$$

The first identity is the Gaussian normalization. The second follows by oddness, and the last follows from independence and the one-dimensional variance identity; Appendix A gives the

corresponding trace calculation. Since

$$\boldsymbol{v} = \boldsymbol{u} + \sqrt{\frac{\theta}{m}}\boldsymbol{z},$$

we obtain

$$\int_{\mathbb{R}^d} M_f \,\mathrm{d}\boldsymbol{v} = n, \tag{2.137}$$

$$\int_{\mathbb{R}^d} \boldsymbol{v} M_f \,\mathrm{d}\boldsymbol{v} = n\boldsymbol{u}, \tag{2.138}$$

$$\int_{\mathbb{R}^d} |\boldsymbol{v} - \boldsymbol{u}|^2 M_f \,\mathrm{d}\boldsymbol{v} = \frac{dn\theta}{m}. \tag{2.139}$$

The cross term in $|\boldsymbol{v}|^2 = |\boldsymbol{u}|^2 + 2\boldsymbol{u}\cdot(\boldsymbol{v}-\boldsymbol{u}) + |\boldsymbol{v}-\boldsymbol{u}|^2$ integrates to zero. Therefore

$$\frac{m}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}|^2 M_f \,\mathrm{d}\boldsymbol{v} = \frac{mn}{2}|\boldsymbol{u}|^2 + \frac{dn\theta}{2}. \tag{2.140}$$

By the definitions of n, $\boldsymbol{u}$, and θ, the same three expressions are obtained from f. Multiplication by the moment vector $\boldsymbol{m}$ proves $\langle (M_f - f)\boldsymbol{m}\rangle = \boldsymbol{0}$. ■

The next orthogonality identity is the key algebraic step in the entropy principle.

Lemma 2.39 (Orthogonality of the Log-Maxwellian)**.** *Suppose that $f : \mathbb{R}^d \to [0,\infty)$ is a kinetic distribution of particles of mass $m > 0$, with the property*

$$\int_{\mathbb{R}^d} f(\boldsymbol{v})|\boldsymbol{v}|^k \,\mathrm{d}\boldsymbol{v} < \infty, \quad k = 0, 1, 2. \tag{2.141}$$

The function $\ln(M_f)$ is orthogonal to the moment-preserving difference $M_f - f$, namely,

$$\int_{\mathbb{R}^d} \ln(M_f(\boldsymbol{v}))\,(M_f(\boldsymbol{v}) - f(\boldsymbol{v}))\,\mathrm{d}\boldsymbol{v} = \langle \ln(M_f)(M_f - f)\rangle = 0.$$

Proof. From the last result, we already know that, for any $C_0, C_2 \in \mathbb{R}$, and any $\boldsymbol{C}_1 \in \mathbb{R}^d$,

$$\langle (C_0 + \boldsymbol{C}_1\cdot\boldsymbol{v} + C_2|\boldsymbol{v}|^2)(M_f - f)\rangle = 0.$$

Expanding the square gives

$$\ln(M_f) = \ln\left[n\left(\frac{m}{2\pi\theta}\right)^{d/2}\right] - \frac{m|\boldsymbol{u}|^2}{2\theta} + \frac{m\boldsymbol{u}\cdot\boldsymbol{v}}{\theta} - \frac{m|\boldsymbol{v}|^2}{2\theta}.$$

This is a linear combination of 1, $\boldsymbol{v}$, and $|\boldsymbol{v}|^2$. The moment-matching lemma therefore gives $\langle \ln(M_f)(M_f - f)\rangle = 0$. ■

The Maxwellian is the least information, or information minimizing (entropy maximizing), state. Consider the following result.

Theorem 2.40 (Maxwellian Entropy Principle)**.** *Suppose that*

$$f_\star : \mathbb{R}^d \to [0, \infty)$$

is a given kinetic distribution with finite information and physically admissible moments satisfying $\rho_\star > 0$ and $\theta_\star > 0$. Let

$$\boldsymbol{\rho}_\star = \left[\rho_\star, \rho_\star \boldsymbol{u}_\star^\top, \omega_\star\right]^\top,$$

be the vector of conserved hydrodynamic moments of $f_\star$, namely,

$$\boldsymbol{\rho}_\star = \begin{bmatrix} \rho_\star \\ \rho_\star \boldsymbol{u}_\star \\ \omega_\star \end{bmatrix} = \left\langle f_\star(\boldsymbol{v}) \begin{bmatrix} m \\ m\boldsymbol{v} \\ \frac{m}{2}|\boldsymbol{v}|^2 \end{bmatrix} \right\rangle.$$

The temperature, $\theta_\star$, is defined via

$$\omega_\star = \frac{\rho_\star}{2}|\boldsymbol{u}_\star|^2 + e_\star = \frac{\rho_\star}{2}|\boldsymbol{u}_\star|^2 + \frac{d}{2} n_\star \theta_\star.$$

The Maxwellian, as usual, is defined using the macroscopic densities as

$$M_\star(\boldsymbol{v}) \coloneqq n_\star \left(\frac{m}{2\pi\theta_\star}\right)^{d/2} \exp\left(-\frac{m|\boldsymbol{v} - \boldsymbol{u}_\star|^2}{2\theta_\star}\right).$$

Consider the admissible class of distribution functions, defined as

$$\mathcal{A}_\star \coloneqq \left\{ g : \mathbb{R}^d \to [0, \infty) \;\middle|\; \langle \boldsymbol{m} g\rangle = \boldsymbol{\rho}_\star \quad \text{and} \quad g \log g \in L^1(\mathbb{R}^d) \right\},$$

where, recall,

$$\boldsymbol{m}(\boldsymbol{v}) \coloneqq \left[m, m\boldsymbol{v}^\top, \frac{m}{2}|\boldsymbol{v}|^2\right]^\top : \mathbb{R}^d \to \mathbb{R}^{d+2}.$$

Then, $M_\star, f_\star \in \mathcal{A}_\star$. In other words, $\mathcal{A}_\star$ is non-empty. For every $g \in \mathcal{A}_\star$,

$$h[M_\star] \leq h[g],$$

where $h[g]$ is the information content of g, defined as

$$h[g] := \int_{\mathbb{R}^d} g(\boldsymbol{v}) \ln(g(\boldsymbol{v}))\, \mathrm{d}\boldsymbol{v}.$$

Moreover,

$$h[g] = h[M_\star] \quad \Longleftrightarrow \quad g = M_\star \quad \textit{almost everywhere.} \tag{2.142}$$

In other words, the Maxwellian is the unique minimum-information distribution among the admissible distributions that give rise to the macroscopic density $\boldsymbol{\rho}_\star$.

Proof. First, $M_\star$ is nonnegative, and Lemma 2.38 gives $\langle \boldsymbol{m} M_\star \rangle = \boldsymbol{\rho}_\star$. Its Gaussian decay also gives $M_\star \log M_\star \in L^1(\mathbb{R}^d)$. Thus $M_\star \in \mathcal{A}_\star$.

Let $g \in \mathcal{A}_\star$ be arbitrary. Since g and $M_\star$ have the same mass, momentum, and energy and since $\log M_\star$ is a linear combination of 1, $\boldsymbol{v}$, and $|\boldsymbol{v}|^2$, Lemma 2.39 gives

$$\int_{\mathbb{R}^d} (g - M_\star) \log M_\star \, \mathrm{d}\boldsymbol{v} = 0. \tag{2.143}$$

Consequently,

$$h[g] - h[M_\star] = \int_{\mathbb{R}^d} g \log\left(\frac{g}{M_\star}\right) \mathrm{d}\boldsymbol{v}. \tag{2.144}$$

The scalar inequality $z \log z - z + 1 \geq 0$ for $z \geq 0$, with the convention $0 \log 0 = 0$, applied with $z = g/M_\star$, yields

$$\int_{\mathbb{R}^d} g \log\left(\frac{g}{M_\star}\right) \mathrm{d}\boldsymbol{v} \geq \int_{\mathbb{R}^d} (g - M_\star)\, \mathrm{d}\boldsymbol{v} = 0. \tag{2.145}$$

Equality holds if and only if $g/M_\star = 1$ almost everywhere. This proves both minimality and uniqueness. ∎

2.7 Space-Homogeneous Model

The space-homogeneous reduction suppresses spatial transport and retains only velocity-space relaxation by collisions. Although it cannot represent gradients, boundaries, or fluid flow through space, it is the cleanest setting in which to see redistribution among species: the mass of each species remains fixed, whereas collisions exchange momentum and energy until a common velocity and temperature are reached.

Definition 2.41 (Space-Homogeneous Multispecies Boltzmann Equation)**.** *We say that the kinetic distributions $f_i : \mathbb{R}^3 \times [0,T] \to [0,\infty)$, $i = 1,\dots,N$, satisfy the* ***space-homogeneous multispecies Boltzmann equation*** *if, for all $\boldsymbol{v} \in \mathbb{R}^3$ and $t \in (0,T]$,*

$$\partial_t f_i(\boldsymbol{v},t) = Q_i[f_i,\boldsymbol{f}](\boldsymbol{v},t), \tag{2.146}$$

where $Q_i[f_i,\boldsymbol{f}]$, the ***cumulative binary collision operator****, is defined as*

$$Q_i[f_i,\boldsymbol{f}](\boldsymbol{v},t) \coloneqq \sum_{j=1}^{N} Q_{i,j}[f_i,f_j](\boldsymbol{v},t). \tag{2.147}$$

For the space-homogeneous model, the H-theorem identifies the equilibrium candidates. It does not, by itself, prove that a solution converges to one of those candidates as $t \to \infty$.

Proposition 2.42 (Homogeneous H-Theorem)**.** *Suppose that $f_1,\dots,f_N : \mathbb{R}^3 \times [0,T] \to [0,\infty)$ are kinetic distributions that satisfy the space-homogeneous multispecies Boltzmann equation. Assume that the collision kernels $B_{i,j}$ have admissible symmetry structure. Then,*

$$\partial_t h(t) = \sigma_h(t) \le 0, \tag{2.148}$$

where h is the ***information content****, defined as*

$$h(t) \coloneqq \sum_{i=1}^{N} \int_{\mathbb{R}^3} f_i(\boldsymbol{v},t) \ln\left(f_i(\boldsymbol{v},t)\right)\, \mathrm{d}\boldsymbol{v},$$

and σ_h is the ***collisional information rate****, defined by*

$$\sigma_h(t) \coloneqq \sum_{i=1}^{N} \int_{\mathbb{R}^3} Q_i[f_i,\boldsymbol{f}](\boldsymbol{v},t) \ln\left(f_i(\boldsymbol{v},t)\right)\, \mathrm{d}\boldsymbol{v}.$$

Proof. The result follows from the proof of Proposition 2.24. ■

The moment equations simplify as well: every spatial-divergence term disappears. Individual species may still exchange momentum and energy, but the mixture totals remain fixed.

Proposition 2.43 (Homogeneous Moment Balances)**.** *Suppose that $f_i : \mathbb{R}^3 \times [0,\infty) \to [0,\infty)$, $i = 1,\dots,N$, are kinetic distributions with the properties*

$$\int_{\mathbb{R}^3} f_i(\boldsymbol{v},t)|\boldsymbol{v}|^k\, \mathrm{d}\boldsymbol{v} < \infty, \quad k = 0,1,2,3, \tag{2.149}$$

for every $t \geq 0$. Assume that the kinetic distributions satisfy the space-homogeneous multispecies Boltzmann system. Then, for all $t \geq 0$, and for each $i = 1, \ldots, N$,

$$\partial_t \rho_i(t) = 0, \tag{2.150}$$

$$\rho_i \partial_t \boldsymbol{u}_i(t) = m_i \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{v}, t)\, \boldsymbol{v}\, \mathrm{d}\boldsymbol{v}, \tag{2.151}$$

and

$$\partial_t \left(\frac{\rho_i}{2} |\boldsymbol{u}_i(t)|^2 + \hat{e}_i(t) \right) = \frac{m_i}{2} \int_{\mathbb{R}^3} Q_i[f_i, \boldsymbol{f}](\boldsymbol{v}, t)\, |\boldsymbol{v}|^2\, \mathrm{d}\boldsymbol{v}. \tag{2.152}$$

Consequently, using the collisional invariants,

$$\partial_t \rho(t) = 0, \tag{2.153}$$

$$\rho\, \partial_t \boldsymbol{u}(t) = 0, \tag{2.154}$$

and

$$\partial_t \omega(t) = \partial_t \sum_{i=1}^{N} \left(\frac{\rho_i}{2} |\boldsymbol{u}_i(t)|^2 + \hat{e}_i(t) \right) = 0. \tag{2.155}$$

Proof. Multiply (2.146) by m_i, $m_i \boldsymbol{v}$, and $m_i |\boldsymbol{v}|^2/2$ and integrate over velocity. This is the space-homogeneous specialization of the exact multispecies moment calculation: there are no spatial flux terms. Species-mass conservation gives $\partial_t \rho_i = 0$, so $\partial_t(\rho_i \boldsymbol{u}_i) = \rho_i \partial_t \boldsymbol{u}_i$. The decomposition (2.103) gives the species energy identity. Finally, sum over i and use (2.47)–(2.49) to obtain conservation of total mass, momentum, and energy. ■

Proposition 2.44 (Equilibrium Parameters Selected by Conservation)**.** *Assume that every unary kernel is nondegenerate and that the graph formed by the nondegenerate interspecies kernels is connected. The unique common-velocity, common-temperature Maxwellian compatible with the conserved species masses, total momentum, and total energy is*

$$f_{i,\infty}(\boldsymbol{v}) = n_i \left(\frac{m_i}{2\pi\theta_\infty} \right)^{3/2} \exp\left(-\frac{m_i |\boldsymbol{v} - \boldsymbol{u}_\infty|^2}{2\theta_\infty} \right), \tag{2.156}$$

where

$$\boldsymbol{u}_\infty = \frac{1}{\rho} \sum_{i=1}^{N} \rho_i \boldsymbol{u}_i(0) \tag{2.157}$$

and

$$\theta_\infty = \frac{2}{3 \sum_{i=1}^{N} n_i} \left[\omega(0) - \frac{\rho}{2} |\boldsymbol{u}_\infty|^2 \right]. \tag{2.158}$$

Here n_i, ρ_i, and ρ are time independent. The formula assumes positive total number density and positive thermal energy, so that $\theta_\infty > 0$. Equivalently,

$$\theta_\infty = \frac{\displaystyle\sum_{i=1}^{N} n_i \hat{\theta}_i(0) + \frac{1}{3}\sum_{i=1}^{N} \rho_i |\boldsymbol{u}_i(0) - \boldsymbol{u}_\infty|^2}{\displaystyle\sum_{i=1}^{N} n_i}. \tag{2.159}$$

Proof. Equation (2.157) follows from conservation of total momentum. At equilibrium,

$$\omega_\infty = \frac{\rho}{2}|\boldsymbol{u}_\infty|^2 + \frac{3}{2}\left(\sum_{i=1}^{N} n_i\right)\theta_\infty. \tag{2.160}$$

Conservation of total energy, $\omega_\infty = \omega(0)$, gives (2.158). Expanding the initial energy as

$$\omega(0) = \sum_{i=1}^{N}\left[\frac{\rho_i}{2}|\boldsymbol{u}_i(0)|^2 + \frac{3}{2}n_i\hat{\theta}_i(0)\right] \tag{2.161}$$

and using the definition of $\boldsymbol{u}_\infty$ gives (2.159). ■

Remark 2.45. *Proposition 2.25 characterizes states with zero entropy-production rate (equivalently, zero dissipation of the information functional), and the preceding proposition identifies the only such state compatible with the conserved quantities. These facts do not alone establish convergence of $f_i(t)$ to $f_{i,\infty}$. Such a conclusion requires additional global-existence, compactness, and coercivity hypotheses. Under suitable assumptions, established convergence results imply*

$$f_i(t) \longrightarrow f_{i,\infty}, \qquad \boldsymbol{u}_i(t) \longrightarrow \boldsymbol{u}_\infty, \qquad \hat{\theta}_i(t) \longrightarrow \theta_\infty. \tag{2.162}$$

Chapter 3

Introduction to the Chapman–Enskog Analysis

The Boltzmann equation describes the evolution of a kinetic distribution in the six-dimensional position–velocity phase space. The equations of fluid mechanics, by contrast, evolve only a small collection of macroscopic space–time fields. The purpose of the Chapman–Enskog procedure is to connect these descriptions when the mean free path is small compared with the length scale over which the macroscopic fields vary. At leading order, the procedure gives the compressible Euler equations. At first order, it gives Newton's law of viscosity and Fourier's law of heat conduction and thereby closes the compressible Navier–Stokes–Fourier equations. Classical accounts may be found in Chapman and Cowling [15], Grad [30], and Struchtrup [74]; the hydrodynamic-limit viewpoint is discussed by Golse [28] and in [12, 5]. For multicomponent systems, Arnault and Guisset [3] provide a modern Chapman–Enskog derivation and discussion of diffusion and heat transport.

We develop the unary, monatomic, force-free case in three spatial dimensions. This is the cleanest setting in which all of the essential ideas appear. The derivation is intentionally detailed because the solvability conditions, rather than the formal power series alone, explain why the fluid equations have their familiar form.

The calculation has four stages. First, nondimensionalization identifies the small Knudsen number ε. Second, rapid collisions select a local Maxwellian and the Euler equations. Third, linearization determines the first non-equilibrium correction. Finally, the stress and heat-flux moments of that correction give Newton's and Fourier's laws. Only the exact moment identities survive without the formal small-ε assumptions.

3.1 Formal Status and Hydrodynamic Scaling

The Chapman–Enskog construction is an asymptotic calculation. We will distinguish identities that follow exactly from the Boltzmann equation from conclusions that depend upon a formal expansion.

Definition 3.1 (Knudsen Number)**.** *Let λ_{mfp} be a characteristic molecular mean free path and let L be a characteristic macroscopic length. The dimensionless ratio*

$$\varepsilon := \frac{\lambda_{\mathrm{mfp}}}{L} \tag{3.1}$$

*is called the **Knudsen number**. The continuum regime corresponds to $0 < \varepsilon \ll 1$.*

We now derive the nondimensional equation rather than merely state it. To distinguish physical quantities from their dimensionless counterparts during this derivation, attach the superscript dim to dimensional variables and use tildes for dimensionless variables. Begin with

$$\partial_{t^{\mathrm{dim}}} f^{\mathrm{dim}} + \boldsymbol{v}^{\mathrm{dim}} \cdot \nabla_{\boldsymbol{x}^{\mathrm{dim}}} f^{\mathrm{dim}} = Q_{B^{\mathrm{dim}}}[f^{\mathrm{dim}}, f^{\mathrm{dim}}].$$

Choose a macroscopic length L, a reference speed V_0, a reference number density n_0, and a characteristic collision-kernel magnitude B_0. Introduce

$$\boldsymbol{x}^{\mathrm{dim}} = L\widetilde{\boldsymbol{x}}, \qquad \boldsymbol{v}^{\mathrm{dim}} = V_0\widetilde{\boldsymbol{v}}, \qquad t^{\mathrm{dim}} = \frac{L}{V_0}\widetilde{t},$$

and scale the distribution and collision kernel by

$$f^{\mathrm{dim}} = \frac{n_0}{V_0^3}\widetilde{f}, \qquad B^{\mathrm{dim}} = B_0\widetilde{B}.$$

The factor n_0/V_0^3 is selected so that $\int f^{\mathrm{dim}}\,\mathrm{d}\boldsymbol{v}^{\mathrm{dim}} = n_0\int \widetilde{f}\,\mathrm{d}\widetilde{\boldsymbol{v}}$. In the present convention, B_0 has the dimensions of volume divided by time; for hard spheres, for example, it is proportional to a reference cross-sectional area times V_0. Since $\mathrm{d}\boldsymbol{v}_\star^{\mathrm{dim}} = V_0^3\mathrm{d}\widetilde{\boldsymbol{v}}_\star$, the collision operator scales as

$$Q_{B^{\mathrm{dim}}}[f^{\mathrm{dim}}, f^{\mathrm{dim}}] = \frac{n_0^2 B_0}{V_0^3}\widetilde{Q}_{\widetilde{B}}[\widetilde{f}, \widetilde{f}] = \frac{n_0}{V_0^3}\frac{1}{\tau_c}\widetilde{Q}_{\widetilde{B}}[\widetilde{f}, \widetilde{f}], \qquad \tau_c := \frac{1}{n_0 B_0}.$$

Here τ_c is the reference collision time; numerical angular factors may be absorbed into the

choice of B_0. The transport operator scales according to

$$\partial_{t^{\dim}} f^{\dim} + \boldsymbol{v}^{\dim} \cdot \nabla_{x^{\dim}} f^{\dim} = \frac{n_0}{V_0^3}\frac{V_0}{L}\left(\partial_{\widetilde{t}}\widetilde{f} + \widetilde{\boldsymbol{v}} \cdot \nabla_{\widetilde{x}}\widetilde{f}\right).$$

Cancellation of the common distribution scale gives

$$\partial_{\widetilde{t}}\widetilde{f} + \widetilde{\boldsymbol{v}} \cdot \nabla_{\widetilde{x}}\widetilde{f} = \frac{L}{V_0\tau_c}\widetilde{Q}_{\widetilde{B}}[\widetilde{f}, \widetilde{f}].$$

If $\lambda_{\text{mfp}} \coloneqq V_0\tau_c$, then

$$\frac{L}{V_0\tau_c} = \frac{L}{\lambda_{\text{mfp}}} = \frac{1}{\varepsilon}.$$

Thus the nondimensional equation is

$$\varepsilon\left(\partial_{\widetilde{t}}\widetilde{f} + \widetilde{\boldsymbol{v}} \cdot \nabla_{\widetilde{x}}\widetilde{f}\right) = \widetilde{Q}_{\widetilde{B}}[\widetilde{f}, \widetilde{f}].$$

The macroscopic moments must be scaled consistently with this kinetic normalization. Choose a reference particle mass m_0 and set

$$m^{\dim} = m_0\widetilde{m}, \qquad n^{\dim} = n_0\widetilde{n}, \qquad \boldsymbol{u}^{\dim} = V_0\widetilde{\boldsymbol{u}}, \qquad \theta^{\dim} = m_0V_0^2\widetilde{\theta}.$$

The last relation uses the energy-temperature convention $\theta^{\dim} = k_B T^{\dim}$. It follows directly from the moment definitions that

$$\rho^{\dim} = m_0n_0\widetilde{\rho}, \qquad e^{\dim} = m_0n_0V_0^2\widetilde{e}, \qquad \omega^{\dim} = m_0n_0V_0^2\widetilde{\omega},$$

and

$$p^{\dim} = m_0n_0V_0^2\widetilde{p}, \qquad \mathsf{P}^{\dim} = m_0n_0V_0^2\widetilde{\mathsf{P}}, \qquad \boldsymbol{q}^{\dim} = m_0n_0V_0^3\widetilde{\boldsymbol{q}}.$$

Thus every quantity appearing below has a specified dimensionless normalization. In particular, the Maxwellian and all moment identities retain exactly the same algebraic form in the tilded variables.

From this point onward, every variable and operator in this Chapman–Enskog chapter is dimensionless unless a superscript dim is displayed. We now drop the tildes and the collision kernel subscript to avoid burdening the notation. The force-free unary Boltzmann equation in hydrodynamic scaling is therefore

$$\varepsilon\left(\partial_t f + \boldsymbol{v} \cdot \nabla_x f\right) = Q[f, f]. \tag{3.2}$$

Equivalently, the collision term in the unmultiplied equation is $Q[f, f]/\varepsilon$. Thus, on a dimensionless macroscopic time interval of order one, the distribution experiences order ε^{-1} collisions. The rapid relaxation in the continuum regime, $0 < \varepsilon \ll 1$, is the mechanism that confines f close to the manifold of local Maxwellians.

Remark 3.2 (Formal Standing Hypotheses)**.** *Throughout this chapter, unless explicitly stated otherwise, we make the following* formal *assumptions.*

1. *The macroscopic fields and all kinetic corrections are smooth, and they decay sufficiently rapidly in velocity to justify every differentiation, integration, and interchange of limits used below.*

2. *The number density and temperature remain strictly positive.*

3. *The collision kernel is elastic, nondegenerate, and invariant under orthogonal changes of velocity coordinates and common translations of both collision velocities. It has sufficient integrability that the linearized collision operator is well defined in the weighted spaces used below. Here nondegeneracy includes the following rigidity property: if a smooth function satisfies the collision-invariant identity for almost every collision on which $B > 0$, then it satisfies that identity for every elastic collision.*

4. *After its null space has been identified, the restriction of the linearized collision operator to the orthogonal complement of the unary collision invariants is formally invertible, with suitable coercivity.*

5. *The Chapman–Enskog series may be manipulated as a formal series: after substitution, coefficients of equal powers of ε may be matched, and the coefficients may be differentiated term by term.*

In particular, an $\mathcal{O}(\varepsilon^2)$ term below initially means that terms through order ε have been matched; it is not, by itself, an error bound for an exact solution. Coercivity depends upon the collision kernel; representative results for broad classes of kernels are given in [57]. Rigorous justifications of hydrodynamic limits or compressible Navier–Stokes approximations require substantially more analysis; see, for example, [5, 52].

Remark 3.3 (Compressible versus Incompressible Limits)**.** *Equation* (3.2) *is the scaling used here to obtain the compressible Euler equations and their first-order Navier–Stokes–Fourier correction. It does not produce an incompressible Navier–Stokes limit. That limit requires an additional amplitude, low-Mach-number, or diffusive scaling [5].*

3.2 Unary Macroscopic Variables and the Closure Problem

Let $m > 0$ denote the dimensionless particle-mass parameter m^{dim}/m_0. In the unary case one may choose $m_0 = m^{\mathrm{dim}}$, which makes $m = 1$; we retain m so that the mass dependence remains visible in the formulas. The species velocity and the barycentric velocity coincide, so we write

$$\boldsymbol{c} \coloneqq \boldsymbol{v} - \boldsymbol{u} \tag{3.3}$$

for the peculiar velocity. Recall the macroscopic fields

$$n \coloneqq \int_{\mathbb{R}^3} f \,\mathrm{d}\boldsymbol{v}, \qquad \rho \coloneqq mn, \tag{3.4}$$

$$\rho\boldsymbol{u} \coloneqq m\int_{\mathbb{R}^3} f\boldsymbol{v} \,\mathrm{d}\boldsymbol{v}, \qquad e \coloneqq \frac{m}{2}\int_{\mathbb{R}^3} f|\boldsymbol{c}|^2 \,\mathrm{d}\boldsymbol{v}, \tag{3.5}$$

$$\theta \coloneqq \frac{2e}{3n}, \qquad \omega \coloneqq \frac{m}{2}\int_{\mathbb{R}^3} f|\boldsymbol{v}|^2 \,\mathrm{d}\boldsymbol{v}. \tag{3.6}$$

The dimensional temperature, expressed in energy units, satisfies $\theta^{\mathrm{dim}} = k_B T^{\mathrm{dim}}$; after the scaling and notational convention above, the symbol θ denotes the dimensionless ratio $\theta^{\mathrm{dim}}/(m_0 V_0^2)$. Consequently,

$$e = \frac{3}{2}n\theta, \qquad \omega = \frac{\rho}{2}|\boldsymbol{u}|^2 + \frac{3}{2}n\theta. \tag{3.7}$$

The pressure tensor and heat flux are

$$\mathsf{P} \coloneqq m\int_{\mathbb{R}^3} f\,\boldsymbol{c}\otimes\boldsymbol{c} \,\mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{q} \coloneqq \frac{m}{2}\int_{\mathbb{R}^3} f\,\boldsymbol{c}|\boldsymbol{c}|^2 \,\mathrm{d}\boldsymbol{v}. \tag{3.8}$$

Definition 3.4 (Scalar and Deviatoric Pressure)**.** *The* ***scalar pressure****, or mean normal pressure, and the* ***deviatoric pressure tensor*** *are defined by*

$$p \coloneqq \frac{1}{3}\mathrm{tr}(\mathsf{P}), \tag{3.9}$$

and

$$\Pi \coloneqq \mathsf{P} - p\mathsf{I}. \tag{3.10}$$

Thus $\mathsf{P} = p\mathsf{I} + \Pi$ *and* $\mathrm{tr}(\Pi) = 0$.

This definition does not require f to be a Maxwellian. Indeed, since

$$\mathrm{tr}(\boldsymbol{c}\otimes\boldsymbol{c}) = |\boldsymbol{c}|^2, \tag{3.11}$$

the definitions of P, e, and θ give

$$p = \frac{m}{3}\int_{\mathbb{R}^3} f|\boldsymbol{c}|^2\,\mathrm{d}\boldsymbol{v} = \frac{2}{3}e = n\theta = \frac{\rho\theta}{m}. \tag{3.12}$$

Thus p is determined by the density and temperature; it is not an additional hydrodynamic unknown. The relation $p = n\theta$ is the ideal monatomic-gas pressure law when temperature is measured in energy units. In dimensional variables, where $\theta^{\mathrm{dim}} = k_B T^{\mathrm{dim}}$, it becomes $p^{\mathrm{dim}} = n^{\mathrm{dim}} k_B T^{\mathrm{dim}}$.

It is important, however, that knowing p determines only the trace of P, not the entire tensor. The five hydrodynamic fields determine p, but for a general nonequilibrium distribution they do not determine Π or $\boldsymbol{q}$. These are the genuinely unclosed fluxes.

Proposition 3.5 (Exact Unary Balance Laws)**.** *Suppose that f is a sufficiently regular solution of* (3.2)*. Then*

$$\partial_t \rho + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u}) = 0, \tag{3.13}$$

$$\partial_t (\rho \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}) = \boldsymbol{0}, \tag{3.14}$$

$$\partial_t \omega + \nabla_{\boldsymbol{x}} \cdot (\omega \boldsymbol{u} + \mathsf{P}\boldsymbol{u} + \boldsymbol{q}) = 0. \tag{3.15}$$

Proof. Multiply (3.2) successively by m, $m\boldsymbol{v}$, and $m|\boldsymbol{v}|^2/2$, and integrate over velocity. The factor ε cancels from each identity, since the corresponding collision moments vanish. The flux decompositions proved in the preceding section give the displayed forms. ■

Equations (3.13) to (3.15) are exact but not closed. The fields ρ, $\boldsymbol{u}$, and θ comprise five scalar fields: one density, three velocity components, and one temperature. They determine the scalar pressure through Equation (3.12), but they do not determine the trace-free pressure Π or the heat flux $\boldsymbol{q}$. The closure problem is to express these remaining fluxes, exactly or approximately, in terms of the five hydrodynamic fields and their derivatives.

3.3 Local Maxwellians and the Euler Closure

For given $n > 0$, $\boldsymbol{u} \in \mathbb{R}^3$, and $\theta > 0$, define the local Maxwellian

$$M[n, \boldsymbol{u}, \theta](\boldsymbol{v}) \coloneqq n\left(\frac{m}{2\pi\theta}\right)^{3/2} \exp\left(-\frac{m|\boldsymbol{v}-\boldsymbol{u}|^2}{2\theta}\right). \tag{3.16}$$

Lemma 3.6 (Moments of a Local Maxwellian)**.** *Let $M = M[n, \boldsymbol{u}, \theta]$, and let $p = n\theta$ be the scalar pressure defined by* (3.9)–(3.12)*. Then*

$$\int_{\mathbb{R}^3} M \,\mathrm{d}\boldsymbol{v} = n, \qquad m\int_{\mathbb{R}^3} M\boldsymbol{v} \,\mathrm{d}\boldsymbol{v} = \rho\boldsymbol{u}, \tag{3.17}$$

$$\frac{m}{2}\int_{\mathbb{R}^3} M|\boldsymbol{v}|^2 \,\mathrm{d}\boldsymbol{v} = \frac{\rho}{2}|\boldsymbol{u}|^2 + \frac{3}{2}n\theta, \qquad m\int_{\mathbb{R}^3} M\boldsymbol{c}\otimes\boldsymbol{c} \,\mathrm{d}\boldsymbol{v} = p\mathsf{I}, \tag{3.18}$$

$$\frac{m}{2}\int_{\mathbb{R}^3} M\boldsymbol{c}|\boldsymbol{c}|^2 \,\mathrm{d}\boldsymbol{v} = \boldsymbol{0}. \tag{3.19}$$

Proof. The first three identities were established in Section 2.6. For the fourth, the off-diagonal components vanish by oddness, and rotational symmetry makes the three diagonal components equal. Taking the trace gives

$$\operatorname{tr}\left(m\int_{\mathbb{R}^3} M\boldsymbol{c}\otimes\boldsymbol{c} \,\mathrm{d}\boldsymbol{v}\right) = m\int_{\mathbb{R}^3} M|\boldsymbol{c}|^2 \,\mathrm{d}\boldsymbol{v} = 3n\theta, \tag{3.20}$$

so each diagonal component equals $n\theta = p$. The heat-flux integrand is odd in $\boldsymbol{c}$, which proves the last identity. ∎

If $f = M$, the exact fluxes close as

$$\mathsf{P} = p\mathsf{I}, \qquad \boldsymbol{q} = \boldsymbol{0}. \tag{3.21}$$

In other words, Maxwellian isotropy sets the undetermined trace-free part Π equal to zero, while oddness sets the undetermined heat flux equal to zero. No sixth unknown is introduced: $p = n\theta = \rho\theta/m$ is already determined by the five hydrodynamic fields. Substitution into the exact balance laws therefore gives a closed system of five scalar equations for those five fields, namely, the compressible Euler equations. This is a statement about the moments of a Maxwellian. A spatially varying local Maxwellian need not itself solve the Boltzmann equation exactly: free transport generally moves it away from local equilibrium. In the Chapman–Enskog calculation, the Euler system instead governs the leading-order Maxwellian while a small kinetic correction compensates for that transport.

Proposition 3.7 (Compressible Euler Equations)**.** *Suppose that the distribution is a local Maxwellian. Then its moments satisfy*

$$\partial_t\rho + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}) = 0, \tag{3.22}$$

$$\partial_t(\rho\boldsymbol{u}) + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}\otimes\boldsymbol{u}) + \nabla_{\boldsymbol{x}}p = \boldsymbol{0}, \tag{3.23}$$

$$\partial_t\omega + \nabla_{\boldsymbol{x}}\cdot\big((\omega+p)\boldsymbol{u}\big) = 0, \tag{3.24}$$

where

$$p = n\theta = \frac{\rho\theta}{m} = \frac{2}{3}\left(\omega - \frac{\rho}{2}|\boldsymbol{u}|^2\right), \qquad \omega = \frac{\rho}{2}|\boldsymbol{u}|^2 + \frac{3}{2}n\theta. \tag{3.25}$$

Consequently, when the system is written in the conservative variables ρ, $\rho\boldsymbol{u}$, and ω, one first recovers $\boldsymbol{u} = (\rho\boldsymbol{u})/\rho$ and then computes p from the displayed formula. Thus every flux in the conservative system is determined by the five conservative unknowns wherever $\rho > 0$. Equivalently, with the material derivative

$$D_t \coloneqq \partial_t + \boldsymbol{u}\cdot\nabla_{\boldsymbol{x}}, \tag{3.26}$$

the primitive-variable form is

$$D_t n + n\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u} = 0, \tag{3.27}$$

$$\rho D_t\boldsymbol{u} + \nabla_{\boldsymbol{x}} p = \mathbf{0}, \tag{3.28}$$

$$D_t\theta + \frac{2}{3}\theta\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u} = 0. \tag{3.29}$$

Proof. The conservative equations follow immediately from (3.13)–(3.15) and (3.21). Expanding the mass equation gives (3.27). Subtracting $\boldsymbol{u}$ times the mass equation from the momentum equation gives (3.28). Finally, subtract the kinetic-energy equation, obtained by taking the scalar product of (3.28) with $\boldsymbol{u}$, from the total-energy equation. Using $e = 3n\theta/2$ gives

$$D_t e + (e+p)\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u} = 0, \tag{3.30}$$

which is equivalent to (3.29). ■

3.4 The Linearized Collision Operator

The first non-equilibrium correction is determined by linearizing the collision operator about a local Maxwellian. This is the infinite-dimensional analogue of replacing a nonlinear function by its derivative near a known state. We use a sign convention for which the linearized operator is nonnegative.

Definition 3.8 (Linearized Collision Operator)**.** *Fix a local Maxwellian $M = M[n, \boldsymbol{u}, \theta]$. For a relative perturbation $\phi = \phi(\boldsymbol{v})$, define the* ***linearized collision operator*** *via*

$$\mathcal{L}_M\phi \coloneqq -\frac{1}{M}\left(Q[M, M\phi] + Q[M\phi, M]\right). \tag{3.31}$$

We equip velocity functions with the Maxwellian-weighted inner product

$$(\phi, \psi)_M := \int_{\mathbb{R}^3} M\phi\psi \, \mathrm{d}\boldsymbol{v}, \qquad \|\phi\|_M^2 := (\phi, \phi)_M. \tag{3.32}$$

Definition 3.9 (Maxwellian-Weighted L^2 Space)**.** *The notation $L^2(M\mathrm{d}\boldsymbol{v})$ denotes the corresponding* ***Maxwellian-weighted*** *L^2* ***space****:*

$$L^2(M\mathrm{d}\boldsymbol{v}) := \left\{ \phi : \mathbb{R}^3 \to \mathbb{R} \;\middle|\; \phi \text{ is measurable and } \int_{\mathbb{R}^3} M(\boldsymbol{v})|\phi(\boldsymbol{v})|^2 \, \mathrm{d}\boldsymbol{v} < \infty \right\}. \tag{3.33}$$

As usual in an L^2 space, functions that agree almost everywhere are identified. Because $M(\boldsymbol{v}) > 0$ for every $\boldsymbol{v}$, "almost everywhere" with respect to the weighted measure $M\mathrm{d}\boldsymbol{v}$ is equivalent to "almost everywhere" with respect to Lebesgue measure $\mathrm{d}\boldsymbol{v}$. The inner product and norm on this space are precisely those in (3.32). Thus, for example,

$$\phi \perp \mathcal{H}_M \tag{3.34}$$

means that $(\phi, \psi)_M = 0$ for every $\psi \in \mathcal{H}_M$.

This weighting is natural for the relative perturbation used above. If the actual perturbation of the Maxwellian is written as $g = M\phi$, then

$$\|\phi\|_M^2 = \int_{\mathbb{R}^3} M|\phi|^2 \, \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^3} \frac{|g|^2}{M} \, \mathrm{d}\boldsymbol{v}. \tag{3.35}$$

Hence $L^2(M\mathrm{d}\boldsymbol{v})$ measures the size of the relative perturbation $\phi = g/M$, or equivalently measures the additive perturbation g with the reciprocal weight M^{-1}. It is also the space in which the sign convention in (3.31) makes $\mathcal{L}_M$ formally symmetric and nonnegative.

The solvability condition below is familiar from finite-dimensional linear algebra. A symmetric matrix equation $A\boldsymbol{x} = \boldsymbol{b}$ can be solved only when $\boldsymbol{b}$ is orthogonal to every vector in $\ker A$. Here $\mathcal{L}_M$ plays the role of A, and the five collision invariants span its null space.

Proposition 3.10 (Formal Structure of $\mathcal{L}_M$)**.** *Under the collision-kernel hypotheses stated above, $\mathcal{L}_M$ is symmetric and nonnegative in the inner product* (3.32)*. On the formal smooth domain specified in the standing hypotheses, its null space is*

$$\mathcal{H}_M := \ker \mathcal{L}_M = \operatorname{span}\left\{1, c_1, c_2, c_3, |\boldsymbol{c}|^2\right\}. \tag{3.36}$$

Consequently, the equation

$$\mathcal{L}_M\phi = S \tag{3.37}$$

can be solved only if

$$(S, \psi)_M = 0 \qquad \text{for every } \psi \in \mathcal{H}_M. \tag{3.38}$$

Subject to the normalization $\phi \perp \mathcal{H}_M$, the solution is formally unique whenever $\mathcal{L}_M$ is invertible on $\mathcal{H}_M^{\perp}$.

Proof. Insert the weak form of the collision operator into

$$(\phi, \mathcal{L}_M \psi)_M = -\int_{\mathbb{R}^3} \phi \left(Q[M, M\psi] + Q[M\psi, M] \right) \mathrm{d}\boldsymbol{v}. \tag{3.39}$$

We first explain the identity $MM_\star = M'M'_\star$. A collision is local in space and time, so all four Maxwellian factors have the same parameters n, $\boldsymbol{u}$, and θ; only their velocity arguments differ. Directly from (3.16),

$$MM_\star = n^2 \left(\frac{m}{2\pi\theta} \right)^3 \exp\left(-\frac{m}{2\theta} \left(|\boldsymbol{v} - \boldsymbol{u}|^2 + |\boldsymbol{v}_\star - \boldsymbol{u}|^2 \right) \right). \tag{3.40}$$

The quantity in the exponential can be expanded as

$$|\boldsymbol{v} - \boldsymbol{u}|^2 + |\boldsymbol{v}_\star - \boldsymbol{u}|^2 = |\boldsymbol{v}|^2 + |\boldsymbol{v}_\star|^2 - 2\boldsymbol{u} \cdot (\boldsymbol{v} + \boldsymbol{v}_\star) + 2|\boldsymbol{u}|^2. \tag{3.41}$$

For unary elastic collisions, conservation of momentum and kinetic energy gives

$$\boldsymbol{v} + \boldsymbol{v}_\star = \boldsymbol{v}' + \boldsymbol{v}'_\star, \qquad |\boldsymbol{v}|^2 + |\boldsymbol{v}_\star|^2 = |\boldsymbol{v}'|^2 + |\boldsymbol{v}'_\star|^2. \tag{3.42}$$

Consequently, the right-hand side of (3.41) is unchanged when the unprimed velocities are replaced by the primed velocities. The prefactor in (3.40) is also unchanged, and therefore

$$M(\boldsymbol{v})M(\boldsymbol{v}_\star) = M(\boldsymbol{v}')M(\boldsymbol{v}'_\star), \qquad \text{or briefly} \qquad MM_\star = M'M'_\star. \tag{3.43}$$

This is the detailed-balance identity for a Maxwellian: the product assigned to a colliding pair is the same before and after the collision. Using this identity and the pre- and post-collisional changes of variables gives the symmetric Dirichlet form

$$(\phi, \mathcal{L}_M \psi)_M = \frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} MM_\star (\phi' + \phi'_\star - \phi - \phi_\star)(\psi' + \psi'_\star - \psi - \psi_\star) B \, \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v}_\star \mathrm{d}\boldsymbol{v}. \tag{3.44}$$

This expression is symmetric in ϕ and ψ and is nonnegative when $\phi = \psi$. It remains to prove the null-space statement; we do this in both directions.

First suppose

$$\phi \in \operatorname{span}\{1, c_1, c_2, c_3, |\boldsymbol{c}|^2\}.$$

Because $\boldsymbol{c} = \boldsymbol{v} - \boldsymbol{u}$ and $\boldsymbol{u}$ is fixed during the velocity-space calculation, the same space can be written as

$$\operatorname{span}\{1, v_1, v_2, v_3, |\boldsymbol{v}|^2\}.$$

Mass, momentum, and kinetic energy are conserved in every unary elastic collision. Therefore

$$\phi' + \phi'_\star - \phi - \phi_\star = 0$$

for every collision. Taking this ϕ in (3.44) shows that $(\psi, \mathcal{L}_M\phi)_M = 0$ for every smooth test function ψ in the operator domain. Hence $\mathcal{L}_M\phi = 0$, and so

$$\operatorname{span}\{1, c_1, c_2, c_3, |\boldsymbol{c}|^2\} \subseteq \ker \mathcal{L}_M.$$

Conversely, let $\phi \in \ker \mathcal{L}_M$. Substituting $\psi = \phi$ in (3.44) gives

$$0 = (\phi, \mathcal{L}_M\phi)_M = \frac{1}{4} \iiint MM_\star B\left(\phi' + \phi'_\star - \phi - \phi_\star\right)^2 \mathrm{d}\boldsymbol{\sigma}\mathrm{d}\boldsymbol{v}_\star\mathrm{d}\boldsymbol{v}. \tag{3.45}$$

The weight $MM_\star$ is strictly positive and $B \geq 0$. The integrand in (3.45) is therefore nonnegative, so a zero integral implies

$$\phi' + \phi'_\star = \phi + \phi_\star \tag{3.46}$$

for almost every collision on which $B > 0$. The nondegeneracy assumption on the collision kernel and the smoothness of ϕ extend (3.46) to every unary elastic collision. Thus ϕ is a unary collision invariant. Proposition 2.20 then gives

$$\phi(\boldsymbol{v}) = A_0 + \boldsymbol{\beta} \cdot \boldsymbol{v} + A_2|\boldsymbol{v}|^2$$

for constants $A_0, A_2 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^3$. Replacing $\boldsymbol{v}$ by $\boldsymbol{c} + \boldsymbol{u}$ shows that this is exactly an element of $\operatorname{span}\{1, c_1, c_2, c_3, |\boldsymbol{c}|^2\}$. This proves the reverse inclusion and hence (3.36).

Finally, if $\psi \in \ker \mathcal{L}_M$, symmetry gives

$$(\mathcal{L}_M\phi, \psi)_M = (\phi, \mathcal{L}_M\psi)_M = 0,$$

which proves the stated orthogonality condition for the range. These arguments establish the formal smooth-domain result. Identifying the kernel of a particular closed realization

of $\mathcal{L}_M$ on all of $L^2(M\mathrm{d}\boldsymbol{v})$, and proving coercivity there, require kernel-dependent functional analysis [12, 57]. ■

Definition 3.11 (Hydrodynamic and Kinetic Projections)**.** *Let $\mathcal{P}_M$ denote the orthogonal projection in $L^2(M\mathrm{d}\boldsymbol{v})$ onto $\mathcal{H}_M$, and let*

$$\mathcal{P}_M^{\perp} := I - \mathcal{P}_M. \tag{3.47}$$

*The range of $\mathcal{P}_M$ is called the **hydrodynamic subspace**; its orthogonal complement is called the **kinetic subspace**.*

3.5 The Chapman–Enskog Ansatz

Let

$$U := (n, \boldsymbol{u}, \theta) \tag{3.48}$$

denote the five hydrodynamic fields. A general distribution $f(\boldsymbol{x}, \boldsymbol{v}, t)$ contains considerably more information than these five moments. For example, many different velocity profiles have the same density, mean velocity, and temperature. In a general kinetic solution, this additional information may evolve independently of the five moments.

Definition 3.12 (Normal Solution)**.** *A **normal solution** is a kinetic distribution whose dependence on space and time occurs only through the hydrodynamic fields and their spatial derivatives. In schematic form,*

$$f(\boldsymbol{x}, \boldsymbol{v}, t) = \mathcal{F}\left(\boldsymbol{v}; U(\boldsymbol{x}, t), \nabla_x U(\boldsymbol{x}, t), \nabla_x^2 U(\boldsymbol{x}, t), \ldots\right). \tag{3.49}$$

Thus $\mathcal{F}$ has no independent dependence upon $\boldsymbol{x}$ or t. All space and time dependence of f is inherited from the hydrodynamic fields and their spatial derivatives. In this sense, the non-hydrodynamic part of the distribution is *slaved* to the hydrodynamic fields rather than evolving as independent kinetic data.

Why should the first term be a Maxwellian? To answer this without assuming the conclusion, temporarily write an unspecified expansion $f = f^{(0)} + \varepsilon f^{(1)} + \cdots$. The order-zero part of (3.2) is $Q[f^{(0)}, f^{(0)}] = 0$. For a positive leading term and a nondegenerate elastic kernel, the equilibrium characterization proved earlier identifies $f^{(0)}$ as a local Maxwellian. This argument is pointwise in $(\boldsymbol{x}, t)$: collisions select the velocity profile but do not yet determine how its parameters evolve. The next order supplies that evolution through a solvability condition.

The Chapman–Enskog ansatz seeks this functional $\mathcal{F}$ as an asymptotic expansion. One writes

$$f = M\left(1 + \varepsilon\phi^{(1)} + \varepsilon^2\phi^{(2)} + \cdots\right), \tag{3.50}$$

where $M = M[U] = M[n, \boldsymbol{u}, \theta]$. The coefficients are not arbitrary functions of $(\boldsymbol{x}, t)$. Rather, they are differential functionals of U; schematically,

$$\phi^{(1)} = \Phi^{(1)}\left(\boldsymbol{v}; U, \nabla_{\boldsymbol{x}} U\right), \tag{3.51}$$

$$\phi^{(2)} = \Phi^{(2)}\left(\boldsymbol{v}; U, \nabla_{\boldsymbol{x}} U, \nabla_{\boldsymbol{x}}^2 U\right), \tag{3.52}$$

The second expression may contain products of first derivatives. More explicitly, write $U = (U_1, \ldots, U_5) = (n, u_1, u_2, u_3, \theta)$. After the lower-order macroscopic equations have been used to eliminate time derivatives, the second correction may be taken in the form

$$\phi^{(2)} = \sum_{\alpha=1}^{5}\sum_{i,j=1}^{3} a_{\alpha ij}(\boldsymbol{v}; U)\,\partial_{x_i x_j} U_\alpha + \sum_{\alpha,\beta=1}^{5}\sum_{i,j=1}^{3} b_{\alpha\beta ij}(\boldsymbol{v}; U)\,(\partial_{x_i} U_\alpha)(\partial_{x_j} U_\beta). \tag{3.53}$$

Here $a_{\alpha ij}$ and $b_{\alpha\beta ij}$ are velocity-dependent coefficient functions of the local fields. The first sum is linear in second spatial derivatives; the second is quadratic in first spatial derivatives. This formula specifies the derivative structure, not the coefficients, which require a second-order collision problem. Higher orders may involve correspondingly higher spatial derivatives.

The Maxwellian M is required to have the same mass, momentum, and energy as the full distribution f. Thus all corrections describe only the remaining kinetic information; they do not alter the five hydrodynamic fields. This requirement is enforced order by order by the normalization conditions

$$\int_{\mathbb{R}^3} M\phi^{(r)} \begin{bmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2 \end{bmatrix} \mathrm{d}\boldsymbol{v} = \boldsymbol{0}, \qquad r \geq 1. \tag{3.54}$$

Equivalently,

$$\phi^{(r)} \in \mathcal{H}_M^{\perp}, \qquad r \geq 1. \tag{3.55}$$

Without this normalization, a component of $\phi^{(r)}$ in $\mathcal{H}_M$ could be absorbed into a change of n, $\boldsymbol{u}$, or θ; the decomposition into a Maxwellian and a kinetic correction would not be unique.

The normal-solution assumption is not an identity satisfied by every solution of the Boltzmann equation. General initial data may contain non-hydrodynamic modes that are not determined by U. Formally, collisions damp these modes on the fast kinetic time scale, after

which the solution is expected to lie near a slow hydrodynamic manifold described by (3.49). An initial kinetic layer may therefore precede the regime described by the Chapman–Enskog expansion.

In the Chapman–Enskog analysis, the time derivative of U is organized as an asymptotic expansion of operators. We now explain precisely what that means. A simple ordinary differential equation contains the main idea. Suppose that a scalar $y(t)$ obeys

$$\frac{\mathrm{d}y}{\mathrm{d}t} = F_0(y) + \varepsilon F_1(y). \tag{3.56}$$

There is only one time variable t. If H is a differentiable function, the ordinary chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} H(y(t)) = H'(y)F_0(y) + \varepsilon H'(y)F_1(y). \tag{3.57}$$

We could call the first term the order-zero contribution to the time derivative of H and the second term the order-one contribution. This naming does not create two different time variables or change the ordinary derivative $\mathrm{d}/\mathrm{d}t$.

For the Boltzmann equation, the role of y is played by the five-component field $U = (n, \boldsymbol{u}, \theta)$. Their evolution is governed by the exact moment equations, not by a separate assumption about time. We record those equations in primitive variables for use below; this is only a reformulation of Proposition 3.5.

Lemma 3.13 (Exact Primitive-Variable Evolution)**.** *For every sufficiently regular solution of the scaled unary Boltzmann equation with $n, \theta > 0$,*

$$\partial_t n = -\boldsymbol{u} \cdot \nabla_x n - n \nabla_x \cdot \boldsymbol{u}, \tag{3.58}$$

$$\partial_t \boldsymbol{u} = -(\boldsymbol{u} \cdot \nabla_x)\boldsymbol{u} - \frac{1}{\rho}\left(\nabla_x p + \nabla_x \cdot \Pi\right), \tag{3.59}$$

$$\partial_t \theta = -\boldsymbol{u} \cdot \nabla_x \theta - \frac{2}{3}\theta \nabla_x \cdot \boldsymbol{u} - \frac{2}{3n}\left(\Pi : \nabla_x \boldsymbol{u} + \nabla_x \cdot \boldsymbol{q}\right). \tag{3.60}$$

Here $p = n\theta$, $\rho = mn$, and $\Pi = \mathsf{P} - p\mathsf{I}$.

Proof. The first two identities follow from the mass and momentum balances in Proposition 3.5. Subtracting the bulk kinetic-energy balance from its total-energy balance gives the internal-energy equation; using $\mathsf{P} = p\mathsf{I} + \Pi$ and $p = n\theta$ gives the third identity. ∎

These equations are exact but not closed: Π and $\boldsymbol{q}$ still depend on the velocity distribution. The normal-solution ansatz supplies formal expansions for these fluxes. Since the Maxwellian

has $\Pi = \mathbf{0}$ and $\boldsymbol{q} = \mathbf{0}$, write

$$\Pi = \varepsilon \mathrm{P}^{(1)} + \varepsilon^2 \mathrm{P}^{(2)} + \cdots, \qquad \boldsymbol{q} = \varepsilon \boldsymbol{q}^{(1)} + \varepsilon^2 \boldsymbol{q}^{(2)} + \cdots. \tag{3.61}$$

The coefficients are moments of the corrections in (3.50):

$$\mathrm{P}^{(r)} = m \int_{\mathbb{R}^3} M \phi^{(r)} \boldsymbol{c} \otimes \boldsymbol{c} \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{q}^{(r)} = \frac{m}{2} \int_{\mathbb{R}^3} M \phi^{(r)} \boldsymbol{c} |\boldsymbol{c}|^2 \, \mathrm{d}\boldsymbol{v}. \tag{3.62}$$

The normalization (3.54) makes every $\mathrm{P}^{(r)}$ trace free. At this point the coefficients are not yet known; (3.61) records how they will enter the macroscopic equations.

Let us carry out this substitution explicitly. The mass equation contains neither Π nor $\boldsymbol{q}$, so it remains

$$\partial_t n = -\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} n - n \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}. \tag{3.63}$$

In the velocity equation, only Π is expanded. Since $p = n\theta$ and $\rho = mn$ are functions of the unexpanded field U, we obtain

$$\begin{aligned} \partial_t \boldsymbol{u} &= -(\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}})\boldsymbol{u} - \frac{1}{\rho} \nabla_{\boldsymbol{x}} p - \frac{1}{\rho} \nabla_{\boldsymbol{x}} \cdot \left(\varepsilon \mathrm{P}^{(1)} + \varepsilon^2 \mathrm{P}^{(2)} + \cdots \right) \\ &= \left[-(\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}})\boldsymbol{u} - \frac{1}{\rho} \nabla_{\boldsymbol{x}} p \right] + \varepsilon \left[-\frac{1}{\rho} \nabla_{\boldsymbol{x}} \cdot \mathrm{P}^{(1)} \right] + \varepsilon^2 \left[-\frac{1}{\rho} \nabla_{\boldsymbol{x}} \cdot \mathrm{P}^{(2)} \right] + \cdots. \end{aligned} \tag{3.64}$$

The temperature equation contains both unclosed fluxes. Substituting their series and using the linearity of tensor contraction and divergence gives

$$\begin{aligned} \partial_t \theta &= -\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \theta - \frac{2}{3} \theta \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} - \frac{2}{3n} \left(\varepsilon \mathrm{P}^{(1)} + \varepsilon^2 \mathrm{P}^{(2)} + \cdots \right) : \nabla_{\boldsymbol{x}} \boldsymbol{u} \\ &\quad - \frac{2}{3n} \nabla_{\boldsymbol{x}} \cdot \left(\varepsilon \boldsymbol{q}^{(1)} + \varepsilon^2 \boldsymbol{q}^{(2)} + \cdots \right) \\ &= \left[-\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \theta - \frac{2}{3} \theta \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} \right] + \varepsilon \left[-\frac{2}{3n} \left(\mathrm{P}^{(1)} : \nabla_{\boldsymbol{x}} \boldsymbol{u} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{q}^{(1)} \right) \right] \\ &\quad + \varepsilon^2 \left[-\frac{2}{3n} \left(\mathrm{P}^{(2)} : \nabla_{\boldsymbol{x}} \boldsymbol{u} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{q}^{(2)} \right) \right] + \cdots. \end{aligned} \tag{3.65}$$

Thus the terms without an explicit factor of ε give the order-zero evolution of U; the coefficients of ε, ε^2, and so on give its successive corrections. In particular, the mass equation has no term with an explicit positive power of ε. Collecting the three equations gives a formal evolution law of the form

$$\partial_t U \sim \mathcal{G}^{(0)}[U] + \varepsilon \mathcal{G}^{(1)}[U] + \varepsilon^2 \mathcal{G}^{(2)}[U] + \cdots. \tag{3.66}$$

Each $\mathcal{G}^{(r)}[U]$ is a spatial expression involving U and its derivatives. The symbol $\sim$ means that we are matching coefficients formally; it does not assert convergence of the series. We do not expand the field itself as $U^{(0)} + \varepsilon U^{(1)} + \cdots$. All of the expressions $\mathcal{G}^{(r)}$ are evaluated at the same field U. Define

$$\partial_t^{(r)} U := \mathcal{G}^{(r)}[U]. \tag{3.67}$$

Then (3.66) may be written as

$$\partial_t U \sim \sum_{r=0}^{\infty} \varepsilon^r \partial_t^{(r)} U. \tag{3.68}$$

We can now identify the operators in the expansion:

$$\partial_t^{(0)} n = -\boldsymbol{u} \cdot \nabla_x n - n \nabla_x \cdot \boldsymbol{u}, \tag{3.69}$$

$$\partial_t^{(0)} \boldsymbol{u} = -(\boldsymbol{u} \cdot \nabla_x)\boldsymbol{u} - \frac{1}{\rho} \nabla_x p, \tag{3.70}$$

$$\partial_t^{(0)} \theta = -\boldsymbol{u} \cdot \nabla_x \theta - \frac{2}{3} \theta \nabla_x \cdot \boldsymbol{u}. \tag{3.71}$$

These are the Euler evolution equations from Proposition 3.7. At each order $r \geq 1$, the same coefficient comparison gives

$$\partial_t^{(r)} n = 0, \tag{3.72}$$

$$\partial_t^{(r)} \boldsymbol{u} = -\frac{1}{\rho} \nabla_x \cdot \mathsf{P}^{(r)}, \tag{3.73}$$

$$\partial_t^{(r)} \theta = -\frac{2}{3n} \left(\mathsf{P}^{(r)} : \nabla_x \boldsymbol{u} + \nabla_x \cdot \boldsymbol{q}^{(r)} \right). \tag{3.74}$$

In particular, $\partial_t^{(1)} n = 0$ means that there is no separate first-order term in the continuity equation. It does not mean that a solution n is independent of ε: its evolution is coupled to the velocity field, which does receive higher-order corrections.

To apply these rules to a quantity depending on U, we use the ordinary chain rule separately at each order. For a differentiable function $A = A(\boldsymbol{v}; U)$ with no independent time dependence, and with $\boldsymbol{v}$ held fixed, define

$$\partial_t^{(r)} A := \frac{\partial A}{\partial n} \partial_t^{(r)} n + \sum_{j=1}^{3} \frac{\partial A}{\partial u_j} \partial_t^{(r)} u_j + \frac{\partial A}{\partial \theta} \partial_t^{(r)} \theta. \tag{3.75}$$

If A also depends on spatial derivatives of U, apply the chain rule to those arguments as well, using

$$\partial_t^{(r)} \left(\nabla_x^k U \right) := \nabla_x^k \partial_t^{(r)} U. \tag{3.76}$$

Consequently, for a normal-solution functional $A[U]$, the time-derivative expansion means

$$\partial_t A[U] \sim \sum_{r=0}^{\infty} \varepsilon^r \partial_t^{(r)} A[U]. \tag{3.77}$$

There is still only one physical time variable. The symbols $\partial_t^{(r)}$ are coefficient-wise chain-rule operations on normal-solution functionals, not ordinary time derivatives with respect to separate time variables.

For example, to find $\partial_t^{(0)} M[U]$, differentiate the Maxwellian with respect to n, the three components of $\boldsymbol{u}$, and θ at fixed $\boldsymbol{v}$, and substitute (3.69)–(3.71) into (3.75). No formula for $\mathsf{P}^{(1)}$ or $\boldsymbol{q}^{(1)}$ is needed for this calculation.

The factor ε in front of transport in (3.2) makes the order counting particularly important. Using (3.50) and (3.77) gives, formally,

$$\varepsilon \partial_t f = \varepsilon \partial_t^{(0)} M + \varepsilon^2 \left(\partial_t^{(1)} M + \partial_t^{(0)} (M\phi^{(1)}) \right) + \mathcal{O}(\varepsilon^3), \tag{3.78}$$

$$\varepsilon \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f = \varepsilon \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M + \varepsilon^2 \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} (M\phi^{(1)}) + \mathcal{O}(\varepsilon^3). \tag{3.79}$$

The order symbols here mean formal coefficient matching, not norm estimates. The key observation is that $\partial_t^{(1)} M$ first contributes at order ε^2. Therefore the first kinetic correction requires only the already determined Euler evolution $\partial_t^{(0)} U$.

Proposition 3.14 (Formal Equations at the First Two Orders)**.** *Under the formal standing hypotheses, substitution of* (3.50) *and the time-derivative rule* (3.77) *into* (3.2) *gives, at orders ε^0 and ε^1, respectively,*

$$Q[M, M] = 0, \tag{3.80}$$

$$\mathcal{L}_M \phi^{(1)} = -\frac{1}{M} \left(\partial_t^{(0)} M + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M \right). \tag{3.81}$$

The five solvability conditions for (3.81) *are precisely the compressible Euler equations. Under the assumed invertibility of $\mathcal{L}_M$ on $\mathcal{H}_M^{\perp}$, these conditions determine a unique normalized $\phi^{(1)}$.*

Proof. We substitute the distribution expansion into each side of (3.2) and retain the terms needed through order ε. On the transport side, the time-derivative rule gives

$$\begin{aligned} \varepsilon \partial_t f &= \varepsilon \partial_t M + \varepsilon^2 \partial_t (M\phi^{(1)}) + \cdots \\ &= \varepsilon \partial_t^{(0)} M + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{3.82}$$

because $\partial_t M = \partial_t^{(0)} M + \varepsilon \partial_t^{(1)} M + \cdots$. Here and below, $\mathcal{O}(\varepsilon^2)$ refers only to formal order counting. Spatial differentiation similarly gives

$$\varepsilon \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f = \varepsilon \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M + \mathcal{O}(\varepsilon^2). \tag{3.83}$$

Thus the complete left-hand side is

$$\varepsilon(\partial_t f + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f) = \varepsilon \left(\partial_t^{(0)} M + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M \right) + \mathcal{O}(\varepsilon^2). \tag{3.84}$$

On the collision side, bilinearity gives

$$\begin{aligned} Q[f, f] &= Q[M + \varepsilon M \phi^{(1)} + \cdots, M + \varepsilon M \phi^{(1)} + \cdots] \\ &= Q[M, M] + \varepsilon \left(Q[M, M\phi^{(1)}] + Q[M\phi^{(1)}, M] \right) + \mathcal{O}(\varepsilon^2) \\ &= Q[M, M] - \varepsilon M \mathcal{L}_M \phi^{(1)} + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{3.85}$$

where the last equality uses (3.31). At order ε^0, $Q[M, M] = 0$ because M is a collision equilibrium. At order ε^1, we obtain

$$\partial_t^{(0)} M + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M = -M \mathcal{L}_M \phi^{(1)}. \tag{3.86}$$

Division by the positive Maxwellian M yields (3.81). Notice that neither $\partial_t^{(1)} M$ nor a derivative of $M\phi^{(1)}$ enters this order: the prefactor ε on transport moves both to order ε^2.

We next check solvability. Since $\mathcal{L}_M$ has range in $\mathcal{H}_M^{\perp}$, the right-hand side of (3.81) must be orthogonal to each of the five collision invariants $1, v_1, v_2, v_3, |\boldsymbol{v}|^2$. Multiplying that right-hand side by M shows that the corresponding conditions are the five moments of $\partial_t^{(0)} M + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M$. For mass, use $\int M \, \mathrm{d}\boldsymbol{v} = n$ and $\int \boldsymbol{v} M \, \mathrm{d}\boldsymbol{v} = n\boldsymbol{u}$ to obtain

$$0 = \int_{\mathbb{R}^3} \left(\partial_t^{(0)} M + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M \right) \mathrm{d}\boldsymbol{v} = \partial_t^{(0)} n + \nabla_{\boldsymbol{x}} \cdot (n\boldsymbol{u}). \tag{3.87}$$

For momentum, the Maxwellian second moment is

$$m \int_{\mathbb{R}^3} \boldsymbol{v} \otimes \boldsymbol{v} \, M \, \mathrm{d}\boldsymbol{v} = \rho \boldsymbol{u} \otimes \boldsymbol{u} + p \mathsf{I}. \tag{3.88}$$

Therefore the three component conditions combine into

$$\boldsymbol{0} = m \int_{\mathbb{R}^3} \boldsymbol{v} \left(\partial_t^{(0)} M + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} M \right) \mathrm{d}\boldsymbol{v} = \partial_t^{(0)} (\rho \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + p \mathsf{I}), \tag{3.89}$$

where $\partial_t^{(0)}(\rho \boldsymbol{u})$ means the chain-rule derivative induced by $\partial_t^{(0)} U$. Finally, the Maxwellian energy and energy flux satisfy

$$\frac{m}{2}\int_{\mathbb{R}^3} |\boldsymbol{v}|^2 M \,\mathrm{d}\boldsymbol{v} = \omega, \qquad \frac{m}{2}\int_{\mathbb{R}^3} |\boldsymbol{v}|^2 \boldsymbol{v} M \,\mathrm{d}\boldsymbol{v} = (\omega + p)\boldsymbol{u}. \tag{3.90}$$

The fifth condition is consequently

$$0 = \frac{m}{2}\int_{\mathbb{R}^3} |\boldsymbol{v}|^2 \left(\partial_t^{(0)} M + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} M\right) \mathrm{d}\boldsymbol{v} = \partial_t^{(0)}\omega + \nabla_{\boldsymbol{x}}\cdot((\omega + p)\boldsymbol{u}). \tag{3.91}$$

Multiplying the mass condition by m identifies it with (3.22); the momentum and energy conditions are exactly (3.23) and (3.24). Conversely, these Euler equations make all five moments vanish, so the right-hand side of (3.81) lies in $\mathcal{H}_M^{\perp}$. The normalization (3.54) requires $\phi^{(1)} \in \mathcal{H}_M^{\perp}$, removing the null-space ambiguity. The standing invertibility assumption then gives a unique solution in that subspace. ■

The Euler equations are therefore not inserted merely to simplify a calculation. They are the compatibility conditions that make the first kinetic correction solvable. The projections defined above separate those conditions from the kinetic equation:

$$\mathcal{P}_M\left(\frac{\partial_t^{(0)} M + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} M}{M}\right) = 0, \tag{3.92}$$

$$\mathcal{L}_M \phi^{(1)} = -\mathcal{P}_M^{\perp}\left(\frac{\partial_t^{(0)} M + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} M}{M}\right). \tag{3.93}$$

The first equation is the five-component Euler system; the second determines the normalized kinetic correction. This is the familiar solvability idea from finite-dimensional linear algebra applied, at each point $(\boldsymbol{x}, t)$, to an operator on velocity functions.

The order-by-order construction is now explicit:

1. The order-zero collision equation selects a local Maxwellian $M[U]$.
2. Its macroscopic moments give the Euler evolution $\partial_t^{(0)} U$, which makes the order-one kinetic equation solvable.
3. Solving that equation gives $\phi^{(1)}$. Its moments give $\mathrm{P}^{(1)}$ and $\boldsymbol{q}^{(1)}$, and hence the first macroscopic correction $\partial_t^{(1)} U$.
4. Only at the next kinetic order do we use $\partial_t^{(1)} M$ to determine $\phi^{(2)}$.

This organization is recursive rather than circular. In particular, it does not require a second physical time variable. It differs from a direct Hilbert expansion by keeping the macroscopic field U unexpanded and eliminating its time derivatives order by order [15, 74].

3.6 The First-Order Source

We now compute the right-hand side of (3.81). Define the symmetric trace-free velocity-gradient tensor

$$\mathsf{D}^{\circ}(\boldsymbol{u}) \coloneqq \frac{1}{2}\left(\nabla_{\boldsymbol{x}}\boldsymbol{u} + (\nabla_{\boldsymbol{x}}\boldsymbol{u})^{\top}\right) - \frac{1}{3}(\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u})\mathsf{I}. \tag{3.94}$$

Also define the velocity tensor and velocity vector

$$\mathsf{A}(\boldsymbol{c}) \coloneqq \boldsymbol{c}\otimes\boldsymbol{c} - \frac{1}{3}|\boldsymbol{c}|^2\mathsf{I}, \tag{3.95}$$

$$\boldsymbol{B}(\boldsymbol{c}) \coloneqq \left(\frac{m|\boldsymbol{c}|^2}{2\theta} - \frac{5}{2}\right)\boldsymbol{c}. \tag{3.96}$$

Lemma 3.15 (Euler-Reduced Maxwellian Derivative)**.** *Let*

$$\mathcal{D}^{(0)} \coloneqq \partial_t^{(0)} + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}}. \tag{3.97}$$

If n, $\boldsymbol{u}$, and θ satisfy the Euler equations, then

$$\frac{\mathcal{D}^{(0)}M}{M} = \frac{m}{\theta}\mathsf{A}(\boldsymbol{c}) : \mathsf{D}^{\circ}(\boldsymbol{u}) + \boldsymbol{B}(\boldsymbol{c})\cdot\nabla_{\boldsymbol{x}}\log\theta. \tag{3.98}$$

Proof. Since

$$\log M = \log n + \frac{3}{2}\log\left(\frac{m}{2\pi\theta}\right) - \frac{m|\boldsymbol{c}|^2}{2\theta}, \tag{3.99}$$

differentiation gives

$$\frac{\mathcal{D}^{(0)}M}{M} = \mathcal{D}^{(0)}\log n - \frac{3}{2}\mathcal{D}^{(0)}\log\theta + \frac{m}{\theta}\boldsymbol{c}\cdot\mathcal{D}^{(0)}\boldsymbol{u} + \frac{m|\boldsymbol{c}|^2}{2\theta}\mathcal{D}^{(0)}\log\theta. \tag{3.100}$$

Write

$$\mathcal{D}^{(0)} = D_t^{(0)} + \boldsymbol{c}\cdot\nabla_{\boldsymbol{x}}, \qquad D_t^{(0)} \coloneqq \partial_t^{(0)} + \boldsymbol{u}\cdot\nabla_{\boldsymbol{x}}. \tag{3.101}$$

The primitive Euler equations give

$$D_t^{(0)}\log n = -\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u}, \tag{3.102}$$

$$D_t^{(0)}\boldsymbol{u} = -\frac{1}{\rho}\nabla_{\boldsymbol{x}}(n\theta) = -\frac{\theta}{m}\nabla_{\boldsymbol{x}}\log n - \frac{\theta}{m}\nabla_{\boldsymbol{x}}\log\theta, \tag{3.103}$$

$$D_t^{(0)}\log\theta = -\frac{2}{3}\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u}. \tag{3.104}$$

Substitute these identities into (3.100). The two density-gradient terms cancel. The scalar terms $D_t^{(0)}\log n$ and $-(3/2)D_t^{(0)}\log\theta$ also cancel. The trace part of

$$\frac{m}{\theta}\boldsymbol{c}\otimes\boldsymbol{c} : \nabla_{\boldsymbol{x}}\boldsymbol{u} \tag{3.105}$$

is canceled by the contribution of (3.104) to the last term in (3.100). The antisymmetric part of $\nabla_{\boldsymbol{x}}\boldsymbol{u}$ has zero contraction with $\boldsymbol{c}\otimes\boldsymbol{c}$. What remains is

$$\frac{m}{\theta}\left(\boldsymbol{c}\otimes\boldsymbol{c} - \frac{1}{3}|\boldsymbol{c}|^2\mathsf{I}\right) : \mathrm{D}^{\circ}(\boldsymbol{u}) + \left(\frac{m|\boldsymbol{c}|^2}{2\theta} - \frac{5}{2}\right)\boldsymbol{c}\cdot\nabla_{\boldsymbol{x}}\log\theta, \tag{3.106}$$

which is (3.98). ■

The cancellation in this lemma is fundamental. For a unary monatomic gas, the first-order source contains neither an independent density-gradient term nor a scalar multiple of $\nabla\cdot\boldsymbol{u}$. It contains only shear deformation and temperature variation.

Lemma 3.16 (Solvability of the Tensor and Vector Problems)**.** *Every component of* A *and* $\boldsymbol{B}$ *lies in* $\mathcal{H}_M^{\perp}$.

Proof. The tensor A is even in $\boldsymbol{c}$, symmetric, and trace free. Its inner products with 1 and $|\boldsymbol{c}|^2$ vanish by isotropy and tracelessness, and its inner products with the components of $\boldsymbol{c}$ vanish by oddness. The vector $\boldsymbol{B}$ is odd, so it is orthogonal to 1 and $|\boldsymbol{c}|^2$. Its coefficient $5/2$ is selected so that it is also orthogonal to each component of $\boldsymbol{c}$. Indeed, the Gaussian identities

$$\int_{\mathbb{R}^3} M c_i c_j \,\mathrm{d}\boldsymbol{v} = n\frac{\theta}{m}\delta_{ij}, \tag{3.107}$$

and

$$\int_{\mathbb{R}^3} M|\boldsymbol{c}|^2 c_i c_j \,\mathrm{d}\boldsymbol{v} = 5n\left(\frac{\theta}{m}\right)^2\delta_{ij} \tag{3.108}$$

give

$$\int_{\mathbb{R}^3} M B_i c_j \,\mathrm{d}\boldsymbol{v} = 0. \tag{3.109}$$

■

Definition 3.17 (Auxiliary Collision Problems)**.** *The* ***auxiliary collision solutions*** *are the tensor*

$\mathsf{X} = \mathsf{X}(\boldsymbol{c})$ *and the vector* $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{c})$ *determined by*

$$\mathcal{L}_M \mathsf{X} = \mathsf{A}, \qquad \mathsf{X} \perp \mathcal{H}_M, \tag{3.110}$$

$$\mathcal{L}_M \boldsymbol{Y} = \boldsymbol{B}, \qquad \boldsymbol{Y} \perp \mathcal{H}_M, \tag{3.111}$$

where the operator acts componentwise. These equations are interpreted on the kinetic subspace $\mathcal{H}_M^{\perp}$.

Under our formal invertibility assumption,

$$\mathsf{X} = \mathcal{L}_M^{-1}\mathsf{A}, \qquad \boldsymbol{Y} = \mathcal{L}_M^{-1}\boldsymbol{B}. \tag{3.112}$$

No explicit formulas are available for a general collision kernel. These two velocity-space problems contain all of the collision-model dependence of the first-order transport coefficients.

Proposition 3.18 (Formal First Chapman–Enskog Correction)**.** *The normalized solution of the first-order equation is*

$$\phi^{(1)} = -\frac{m}{\theta}\mathsf{X} : \mathsf{D}^{\circ}(\boldsymbol{u}) - \boldsymbol{Y} \cdot \nabla_{\boldsymbol{x}} \log \theta. \tag{3.113}$$

Proof. Combine (3.81) with Lemma 3.15. More precisely, using (3.98), the first-order equation becomes

$$\mathcal{L}_M \phi^{(1)} = -\frac{m}{\theta}\mathsf{A} : \mathsf{D}^{\circ}(\boldsymbol{u}) - \boldsymbol{B} \cdot \nabla_{\boldsymbol{x}} \log \theta. \tag{3.114}$$

Equations (3.110) and (3.111) give Equation (3.113). The normalization follows from $\mathsf{X}, \boldsymbol{Y} \perp \mathcal{H}_M$. ∎

3.7 Viscosity and Heat Conduction

We now insert the approximation

$$f = M\left(1 + \varepsilon\phi^{(1)}\right) + \mathcal{O}(\varepsilon^2) \tag{3.115}$$

into the definitions of the pressure tensor and heat flux. Rotational invariance implies that a symmetric trace-free tensor can produce only a scalar multiple of the same tensor, and that a vector can produce only a scalar multiple of the same vector. Parity eliminates the cross terms: the tensor problem is even in $\boldsymbol{c}$, while the vector problem is odd.

Definition 3.19 (Unary Transport Coefficients)**.** *Define the **shear viscosity** μ and **thermal conductivity** κ, relative to the energy temperature θ, by*

$$\mu := \frac{m^2}{10\theta}\int_{\mathbb{R}^3} M\,\mathsf{A} : \mathsf{X}\,\mathrm{d}\boldsymbol{v} = \frac{m^2}{10\theta}(\mathsf{A}, \mathcal{L}_M^{-1}\mathsf{A})_M, \tag{3.116}$$

$$\kappa := \frac{1}{3}\int_{\mathbb{R}^3} M\,\boldsymbol{B}\cdot\boldsymbol{Y}\,\mathrm{d}\boldsymbol{v} = \frac{1}{3}(\boldsymbol{B}, \mathcal{L}_M^{-1}\boldsymbol{B})_M. \tag{3.117}$$

For tensor- and vector-valued functions, the weighted inner product denotes the sum of the componentwise inner products.

Remark 3.20 (Restoring Physical Transport Coefficients)**.** *The coefficients μ and κ in this section are dimensionless. If μ^{dim} is the physical dynamic viscosity and $\kappa_\theta^{\mathrm{dim}}$ is the physical conductivity relative to the energy temperature, so that $\boldsymbol{q}^{\mathrm{dim}} = -\kappa_\theta^{\mathrm{dim}}\nabla_{\boldsymbol{x}^{\mathrm{dim}}}\theta^{\mathrm{dim}}$, then the scalings above give*

$$\mu^{\mathrm{dim}} = m_0 n_0 V_0^2 \tau_c\,\mu, \qquad \kappa_\theta^{\mathrm{dim}} = n_0 V_0^2 \tau_c\,\kappa.$$

If Fourier's law is instead written in terms of temperature in kelvin, $\boldsymbol{q}^{\mathrm{dim}} = -\kappa_T^{\mathrm{dim}}\nabla_{\boldsymbol{x}^{\mathrm{dim}}}T^{\mathrm{dim}}$, then $\kappa_T^{\mathrm{dim}} = k_B\kappa_\theta^{\mathrm{dim}}$.

Proposition 3.21 (First-Order Constitutive Laws)**.** *Under the formal Chapman–Enskog assumptions,*

$$\mathsf{P} = p\mathsf{I} - 2\varepsilon\mu\mathsf{D}^\circ(\boldsymbol{u}) + \mathcal{O}(\varepsilon^2), \tag{3.118}$$

$$\boldsymbol{q} = -\varepsilon\kappa\nabla_{\boldsymbol{x}}\theta + \mathcal{O}(\varepsilon^2). \tag{3.119}$$

Moreover, $\mu \geq 0$ and $\kappa \geq 0$. Under strict coercivity on $\mathcal{H}_M^{\perp}$, both coefficients are strictly positive.

Proof. First consider the pressure tensor. Substitution of (3.113) into (3.8) gives

$$\mathsf{P} = p\mathsf{I} - \varepsilon\frac{m^2}{\theta}\int_{\mathbb{R}^3} M\,\boldsymbol{c}\otimes\boldsymbol{c}(\mathsf{X} : \mathsf{D}^\circ(\boldsymbol{u}))\,\mathrm{d}\boldsymbol{v} + \mathcal{O}(\varepsilon^2). \tag{3.120}$$

The $\boldsymbol{Y}$ term is odd and therefore integrates to zero. Since D° is trace free, $\boldsymbol{c}\otimes\boldsymbol{c}$ may be replaced by A. Isotropy on the five-dimensional space of symmetric trace-free tensors gives

$$\int_{\mathbb{R}^3} M\,\mathsf{A}(\mathsf{X} : \mathsf{D}^\circ)\,\mathrm{d}\boldsymbol{v} = \frac{1}{5}\left(\int_{\mathbb{R}^3} M\,\mathsf{A} : \mathsf{X}\,\mathrm{d}\boldsymbol{v}\right)\mathsf{D}^\circ. \tag{3.121}$$

Equation (3.116) now gives Equation (3.118).

For the heat flux, the X term is odd and integrates to zero. Hence

$$\boldsymbol{q} = -\varepsilon \frac{m}{2} \int_{\mathbb{R}^3} M\, \boldsymbol{c}|\boldsymbol{c}|^2 (\boldsymbol{Y} \cdot \nabla_{\boldsymbol{x}} \log\theta)\, \mathrm{d}\boldsymbol{v} + \mathcal{O}(\varepsilon^2). \tag{3.122}$$

Isotropy gives

$$\int_{\mathbb{R}^3} M\, \boldsymbol{c}|\boldsymbol{c}|^2 \otimes \boldsymbol{Y}\, \mathrm{d}\boldsymbol{v} = \frac{1}{3} \left(\int_{\mathbb{R}^3} M|\boldsymbol{c}|^2 \boldsymbol{c} \cdot \boldsymbol{Y}\, \mathrm{d}\boldsymbol{v} \right) \mathsf{I}. \tag{3.123}$$

The normalization $\boldsymbol{Y} \perp \mathcal{H}_M$ implies

$$\int_{\mathbb{R}^3} M\boldsymbol{c} \cdot \boldsymbol{Y}\, \mathrm{d}\boldsymbol{v} = 0. \tag{3.124}$$

Using Equation (3.96), we therefore have

$$\int_{\mathbb{R}^3} M|\boldsymbol{c}|^2 \boldsymbol{c} \cdot \boldsymbol{Y}\, \mathrm{d}\boldsymbol{v} = \frac{2\theta}{m} \int_{\mathbb{R}^3} M\boldsymbol{B} \cdot \boldsymbol{Y}\, \mathrm{d}\boldsymbol{v}. \tag{3.125}$$

Since $\nabla \log\theta = \nabla\theta/\theta$, Equation (3.117) gives Equation (3.119).

Finally, nonnegativity of $\mathcal{L}_M$ gives

$$(\mathsf{A}, \mathcal{L}_M^{-1}\mathsf{A})_M = (\mathsf{X}, \mathcal{L}_M\mathsf{X})_M \geq 0, \tag{3.126}$$

and

$$(\boldsymbol{B}, \mathcal{L}_M^{-1}\boldsymbol{B})_M = (\boldsymbol{Y}, \mathcal{L}_M\boldsymbol{Y})_M \geq 0. \tag{3.127}$$

Strict coercivity and the fact that A and $\boldsymbol{B}$ are nonzero give strict positivity. ■

Remark 3.22 (Why the Unary Bulk Viscosity Vanishes)**.** *The Newtonian viscous stress for a general compressible fluid may contain both a trace-free shear contribution and a scalar bulk-viscosity contribution. In the present monatomic elastic model, the scalar $\nabla \cdot \boldsymbol{u}$ term cancels from the first-order source in Lemma 3.15. In addition, the moment normalization fixes the trace of* P *through the internal energy. Thus, no independent scalar dissipative mode remains, and the Chapman–Enskog bulk viscosity is zero. This conclusion changes for gases with internal molecular degrees of freedom or additional relaxation mechanisms [15, 74].*

Remark 3.23 (Kernel Dependence)**.** *Equations* (3.116) *and* (3.117) *are exact formal definitions of the first Chapman–Enskog transport coefficients for the chosen collision kernel and normalization. Closed numerical constants require a specified interaction law and, in most cases, an additional approximation such as a Sonine-polynomial expansion. Quoting hard-sphere or Maxwell-molecule constants without first fixing the normalization of B would be ambiguous.*

Remark 3.24 (A Solvable BGK Check)**.** *The BGK model introduced earlier makes the two auxiliary problems concrete. For this illustration, choose its dimensionless relaxation rate to be one, so that* $Q^{BGK}[f] = M_f - f$*. If* $\phi \perp \mathcal{H}_M$*, then* $M\phi$ *has zero mass, momentum, and energy moments. The Maxwellian associated with* $M + \delta M\phi$ *is therefore* M*, to first order in* δ*; in fact it is* M *whenever that perturbed distribution is admissible. Hence the linearized BGK operator acts as the identity on* $\mathcal{H}_M^{\perp}$*, and the auxiliary solutions are simply* $\mathsf{X} = \mathsf{A}$ *and* $\boldsymbol{Y} = \boldsymbol{B}$*. The Gaussian moments*

$$\int_{\mathbb{R}^3} M|\boldsymbol{c}|^2 \,\mathrm{d}\boldsymbol{v} = 3n\frac{\theta}{m}, \qquad \int_{\mathbb{R}^3} M|\boldsymbol{c}|^4 \,\mathrm{d}\boldsymbol{v} = 15n\left(\frac{\theta}{m}\right)^2, \qquad \int_{\mathbb{R}^3} M|\boldsymbol{c}|^6 \,\mathrm{d}\boldsymbol{v} = 105n\left(\frac{\theta}{m}\right)^3 \tag{3.128}$$

give $\int M\,\mathsf{A} : \mathsf{A}\,\mathrm{d}\boldsymbol{v} = 10n(\theta/m)^2$ *and* $\int M\,\boldsymbol{B}\cdot\boldsymbol{B}\,\mathrm{d}\boldsymbol{v} = (15/2)n\theta/m$*. Thus* (3.116)–(3.117) *reduce to* $\mu = n\theta = p$ *and* $\kappa = 5n\theta/(2m)$*. These values check the algebra of the method for a simple relaxation model; they are not the transport coefficients of the full Boltzmann operator.*

3.8 The Compressible Navier–Stokes–Fourier System

Define the viscous stress tensor

$$\boldsymbol{\tau} \coloneqq 2\mu\mathsf{D}^{\circ}(\boldsymbol{u}) = \mu\left(\nabla_{\boldsymbol{x}}\boldsymbol{u} + (\nabla_{\boldsymbol{x}}\boldsymbol{u})^{\top} - \frac{2}{3}(\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u})\mathsf{I}\right). \tag{3.129}$$

Then the first-order constitutive relations are

$$\mathsf{P} = p\mathsf{I} - \varepsilon\boldsymbol{\tau}, \qquad \boldsymbol{q} = -\varepsilon\kappa\nabla_{\boldsymbol{x}}\theta, \tag{3.130}$$

up to errors formally of order ε^2.

Proposition 3.25 (Formal Compressible Navier–Stokes–Fourier Approximation)**.** *Truncating the Chapman–Enskog expansion after the first correction closes the exact balance laws as*

$$\partial_t\rho + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}) = 0, \tag{3.131}$$

$$\partial_t(\rho\boldsymbol{u}) + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}\otimes\boldsymbol{u}) + \nabla_{\boldsymbol{x}}p = \varepsilon\nabla_{\boldsymbol{x}}\cdot\boldsymbol{\tau}, \tag{3.132}$$

$$\partial_t\omega + \nabla_{\boldsymbol{x}}\cdot\big((\omega+p)\boldsymbol{u}\big) = \varepsilon\nabla_{\boldsymbol{x}}\cdot(\boldsymbol{\tau}\boldsymbol{u}) + \varepsilon\nabla_{\boldsymbol{x}}\cdot(\kappa\nabla_{\boldsymbol{x}}\theta), \tag{3.133}$$

where

$$p = n\theta = \frac{\rho\theta}{m}, \qquad \omega = \frac{\rho}{2}|\boldsymbol{u}|^2 + \frac{3}{2}n\theta, \tag{3.134}$$

and $\boldsymbol{\tau}$ *is given by Equation* (3.129)*. Equations* (3.131) *to* (3.133) *are the compressible Navier–Stokes–Fourier equations for a monatomic ideal gas, with transport coefficients determined by the unary Boltzmann collision operator.*

Proof. Insert Equation (3.130) into the exact momentum balance in Equation (3.14):

$$\partial_t(\rho \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + p\mathsf{I} - \varepsilon \boldsymbol{\tau}) = \mathbf{0}. \tag{3.135}$$

This is (3.132). Similarly, the exact total-energy flux becomes

$$\omega \boldsymbol{u} + \mathsf{P}\boldsymbol{u} + \boldsymbol{q} = \omega \boldsymbol{u} + (p\mathsf{I} - \varepsilon \boldsymbol{\tau})\,\boldsymbol{u} - \varepsilon\kappa \nabla_{\boldsymbol{x}}\theta \tag{3.136}$$

$$= (\omega + p)\boldsymbol{u} - \varepsilon \boldsymbol{\tau}\boldsymbol{u} - \varepsilon\kappa \nabla_{\boldsymbol{x}}\theta. \tag{3.137}$$

Substitution into (3.15) gives (3.133). ■

Proposition 3.26 (Temperature Equation)**.** *The internal-energy form of the first-order system is*

$$\frac{3}{2} n D_t\theta + p\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} = \varepsilon \boldsymbol{\tau} : \nabla_{\boldsymbol{x}}\boldsymbol{u} + \varepsilon \nabla_{\boldsymbol{x}} \cdot (\kappa \nabla_{\boldsymbol{x}}\theta). \tag{3.138}$$

Since $\boldsymbol{\tau} = 2\mu \mathsf{D}^{\circ}(\boldsymbol{u})$,

$$\boldsymbol{\tau} : \nabla_{\boldsymbol{x}}\boldsymbol{u} = 2\mu |\mathsf{D}^{\circ}(\boldsymbol{u})|^2 \geq 0. \tag{3.139}$$

Proof. Take the scalar product of the momentum equation with $\boldsymbol{u}$ to obtain the bulk kinetic-energy equation. Subtract it from the total-energy equation, and use the mass equation together with $e = 3n\theta/2$. The resulting pressure-work, viscous-work, and heat-conduction terms give (3.138). ■

Proposition 3.27 (Formal Residual of the First-Order Approximation)**.** *Let* $U = (n, \boldsymbol{u}, \theta)$ *be a smooth solution of* (3.131)–(3.133) *on a fixed time interval, and let* $\phi^{(1)}$ *be given by* (3.113)*. Define*

$$f_{\mathrm{app}} := M[U]\big(1 + \varepsilon \phi^{(1)}[U]\big). \tag{3.140}$$

Assume that n *and* θ *stay bounded away from zero and that the derivatives and collision expressions below remain bounded independently of* ε *in a suitable velocity norm. Then the residual of the* scaled *Boltzmann equation satisfies*

$$\varepsilon(\partial_t + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}}) f_{\mathrm{app}} - Q[f_{\mathrm{app}}, f_{\mathrm{app}}] = \mathcal{O}(\varepsilon^2). \tag{3.141}$$

Furthermore, the mass, momentum, and energy moments of this residual vanish exactly.

Proof. Write $M = M[U]$ and $\phi = \phi^{(1)}[U]$. Bilinearity, $Q[M, M] = 0$, and (3.31) give the exact algebraic identity

$$\begin{aligned}&\varepsilon(\partial_t + \boldsymbol{v}\cdot\nabla_x)f_{\text{app}} - Q[f_{\text{app}}, f_{\text{app}}]\\ &\qquad = \varepsilon\left((\partial_t + \boldsymbol{v}\cdot\nabla_x)M + M\mathcal{L}_M\phi\right) + \varepsilon^2\left((\partial_t + \boldsymbol{v}\cdot\nabla_x)(M\phi) - Q[M\phi, M\phi]\right). \qquad (3.142)\end{aligned}$$

Because U solves the truncated first-order fluid equations, its evolution has the exact form $\partial_t U = \mathcal{G}^{(0)}[U] + \varepsilon\mathcal{G}^{(1)}[U]$, with the coefficients defined by (3.66). Consequently, the chain rule gives $\partial_t M = \partial_t^{(0)} M + \varepsilon\partial_t^{(1)} M$. Equation (3.81) cancels the order-ε terms in (3.142); all remaining terms have the factor ε^2 and are bounded by hypothesis.

For the moment statement, the normalization (3.54) makes the five moments of f_{app} exactly those of $M[U]$. Since pressure and heat flux are linear moments of f_{app}, they satisfy the truncated constitutive laws (3.130) exactly. Taking collision-invariant moments of the residual therefore gives ε times the left-hand sides of the first-order fluid balance laws, which vanish because U solves those laws. ■

The proposition establishes formal *consistency*, not convergence. In particular, a small residual does not control the difference between an exact kinetic solution and f_{app} without stability estimates and compatible initial and boundary data. Nor does the polynomial correction guarantee $f_{\text{app}} \geq 0$ for every velocity. These are separate questions from the coefficient-matching calculation.

3.9 What Has and Has Not Been Proved

It is useful to summarize the logical status of the preceding calculation.

1. The moment identities and balance laws are exact for sufficiently regular solutions of the Boltzmann equation.

2. The characterization of the leading local equilibrium as a Maxwellian follows from the collision invariants and the H-theorem, under the nondegeneracy hypotheses stated earlier.

3. The symmetry, null space, and nonnegativity of the linearized collision operator are structural facts. Precise domains, spectral gaps, and coercivity estimates depend upon the collision kernel [12, 57].

4. The expansion (3.50), the term-by-term differentiation, and the neglect of the $\mathcal{O}(\varepsilon^2)$ remainder are formal steps.

5. The compressible Navier–Stokes–Fourier equations are therefore a first-order approximation to the Boltzmann dynamics, not the direct $\varepsilon \to 0$ limiting system in this scaling. The direct leading limit is Euler. Rigorous approximation theorems require well-prepared data, regularity, stability, and quantitative remainder estimates; one such line of analysis is represented by [52].

6. Boundary and initial layers are absent from the interior calculation above. Kinetic boundary conditions generally produce velocity-slip and temperature-jump corrections that require a separate half-space analysis.

7. Continuing the expansion produces Burnett and super-Burnett corrections. Those higher-order truncations introduce additional analytic and stability difficulties and should not be assumed to improve the model automatically [30, 74].

The essential chain of reasoning can now be read from the equations:

$$\begin{aligned} &\text{rapid collisions} \Longrightarrow \text{local Maxwellian} \Longrightarrow \text{Euler solvability conditions,} \\ &\text{Euler reduction} \Longrightarrow \text{first kinetic correction} \Longrightarrow \text{Newton and Fourier laws,} \\ &\text{Newton and Fourier laws} \Longrightarrow \text{Navier–Stokes–Fourier closure.} \end{aligned} \tag{3.143}$$

This unary derivation supplies the template for the multispecies calculation. In a mixture, the scalar operator $\mathcal{L}_M$ is replaced by a coupled operator matrix, the leading equilibrium retains separate species densities but has a common velocity and temperature, and the first-order kinetic correction additionally generates diffusion and thermal-diffusion fluxes.

Chapter 4

Moment Closure Methods

The preceding chapter exhibited two closures of the exact unary balance laws. The local-Maxwellian approximation produced the compressible Euler equations. The first Chapman–Enskog correction produced the compressible Navier–Stokes–Fourier equations. We now place those constructions within the broader methodology of moment closure.

The central difficulty is simple to state. The distribution $f(\boldsymbol{x}, \boldsymbol{v}, t)$ depends upon three velocity variables, whereas a fluid model retains only finitely many functions of $(\boldsymbol{x}, t)$. Velocity integration can compress the kinetic equation into finitely many balance laws, but it cannot, by itself, recover the velocity dependence that was discarded. A closure supplies this missing information. The choice of closure is therefore a mathematical modeling decision, not an additional identity hidden inside the Boltzmann equation.

For clarity, most of this chapter concerns the scaled unary problem in Equation (3.2). The same construction applies componentwise to a mixture; the additional conservation requirements are discussed at the end.

Every moment method answers the same three questions: Which velocity averages will be retained? How will a distribution be reconstructed from those averages? Which structural properties of the kinetic equation survive that reconstruction? The first question determines the unknowns, the second closes their equations, and the third determines whether the resulting model is mathematically and physically credible.

4.1 Finite Collections of Moments

Definition 4.1 (Velocity Moment)**.** *Let $\psi : \mathbb{R}^3 \to \mathbb{R}$ be a prescribed function of velocity. Whenever the integral is finite, the* ***moment of*** *f* ***associated with*** *ψ is*

$$u_\psi(\boldsymbol{x}, t) \coloneqq \int_{\mathbb{R}^3} f(\boldsymbol{x}, \boldsymbol{v}, t)\psi(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}. \tag{4.1}$$

The function ψ is called the ***moment function*** *or* ***velocity test function****.*

For example, the choice 1 gives number density, the three scalar entries of $m\boldsymbol{v}$ give momentum density, and $m|\boldsymbol{v}|^2/2$ gives total-energy density. The entries of $m\boldsymbol{v} \otimes \boldsymbol{v}$ give the momentum-flow tensor, while those of $m\boldsymbol{v}|\boldsymbol{v}|^2/2$ give the energy flux. Thus the macroscopic quantities introduced earlier are all velocity moments in the sense of (4.1).

Proposition 4.2 (Exact Equation for a General Moment)**.** *Suppose that f satisfies the scaled unary Boltzmann equation in Equation* (3.2)*, and let ψ be independent of $\boldsymbol{x}$ and t. Then*

$$\varepsilon\partial_t u_\psi + \varepsilon\nabla_{\boldsymbol{x}} \cdot \left(\int_{\mathbb{R}^3} f\, \boldsymbol{v}\psi\,\mathrm{d}\boldsymbol{v} \right) = \int_{\mathbb{R}^3} Q[f, f]\psi\,\mathrm{d}\boldsymbol{v}. \tag{4.2}$$

If ψ is a unary collision invariant, the right-hand side vanishes.

Proof. Multiply (3.2) by ψ and integrate over velocity. Since ψ is independent of $\boldsymbol{x}$ and t, differentiation may be moved outside the velocity integral. The transport term becomes

$$\int_{\mathbb{R}^3} (\boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f)\psi\,\mathrm{d}\boldsymbol{v} = \nabla_{\boldsymbol{x}} \cdot \left(\int_{\mathbb{R}^3} f\, \boldsymbol{v}\psi\,\mathrm{d}\boldsymbol{v} \right). \tag{4.3}$$

The final statement is the defining property of a collision invariant. ∎

Now choose finitely many linearly independent moment functions

$$\psi_1, \ldots, \psi_K \tag{4.4}$$

and retain the corresponding fields $u_{\psi_1}, \ldots, u_{\psi_K}$. The preceding proposition is written componentwise, so it immediately gives K exact balance laws. These equations are generally not closed: their fluxes and collision terms still depend upon the entire distribution f.

Lemma 4.3 (The Polynomial Moment Hierarchy)**.** *Suppose that ψ is a polynomial in $\boldsymbol{v}$ of*

total degree r. The transport flux in its exact moment equation is

$$\boldsymbol{F}_\psi \coloneqq \int_{\mathbb{R}^3} f(\boldsymbol{x}, \boldsymbol{v}, t)\, \boldsymbol{v}\psi(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} = \begin{pmatrix} \displaystyle\int_{\mathbb{R}^3} f\, v_1 \psi\, \mathrm{d}\boldsymbol{v} \\ \displaystyle\int_{\mathbb{R}^3} f\, v_2 \psi\, \mathrm{d}\boldsymbol{v} \\ \displaystyle\int_{\mathbb{R}^3} f\, v_3 \psi\, \mathrm{d}\boldsymbol{v} \end{pmatrix}. \tag{4.5}$$

Each component of $\boldsymbol{F}_\psi$ is a linear combination of moments of degree at most $r+1$, and the degree-r part of ψ produces moments of degree $r+1$.

Proof. Write the polynomial in multi-index notation as

$$\psi(\boldsymbol{v}) = \sum_{|\alpha| \le r} a_\alpha \boldsymbol{v}^\alpha, \qquad \boldsymbol{v}^\alpha = v_1^{\alpha_1} v_2^{\alpha_2} v_3^{\alpha_3}.$$

For $j = 1, 2, 3$, the jth component of (4.5) is

$$(\boldsymbol{F}_\psi)_j = \sum_{|\alpha| \le r} a_\alpha \int_{\mathbb{R}^3} f(\boldsymbol{x}, \boldsymbol{v}, t)\, \boldsymbol{v}^{\alpha + \boldsymbol{e}_j}\, \mathrm{d}\boldsymbol{v},$$

where $\boldsymbol{e}_j$ is the jth coordinate multi-index. Since $|\alpha + \boldsymbol{e}_j| = |\alpha| + 1$, every term with $|\alpha| = r$ is a moment of degree $r+1$. Substitution of $\boldsymbol{F}_\psi$ into (4.2) makes this higher-moment dependence explicit. ■

This elementary observation explains the infinite hierarchy. The mass equation involves momentum. The momentum equation involves the second moment M. The energy equation involves the third moment $\boldsymbol{r}$. Equations for those moments involve still higher moments. Moreover, except for collision invariants and certain special collision models, the collision production term cannot be evaluated from a prescribed finite list of moments alone [29, 51].

4.2 Reconstruction and Closure

Definition 4.4 (Moment Reconstruction and Moment Closure)**.** *Fix the moment functions in* (4.4)*. A* ***moment reconstruction*** *is a rule that assigns a velocity function*

$$\widehat{f} = \widehat{f}[u_{\psi_1}, \ldots, u_{\psi_K}] \tag{4.6}$$

such that

$$\int_{\mathbb{R}^3} \widehat{f}\,\psi_k \,\mathrm{d}\boldsymbol{v} = u_{\psi_k}, \qquad k = 1, \dots, K. \tag{4.7}$$

The associated ***moment closure*** *replaces the unknown distribution in every unresolved flux and collision term by* $\widehat{f}$*. Thus*

$$\int_{\mathbb{R}^3} f\,\boldsymbol{v}\psi_k \,\mathrm{d}\boldsymbol{v} \approx \int_{\mathbb{R}^3} \widehat{f}\,\boldsymbol{v}\psi_k \,\mathrm{d}\boldsymbol{v}, \tag{4.8}$$

$$\int_{\mathbb{R}^3} Q[f,f]\psi_k \,\mathrm{d}\boldsymbol{v} \approx \int_{\mathbb{R}^3} Q[\widehat{f},\widehat{f}]\psi_k \,\mathrm{d}\boldsymbol{v}. \tag{4.9}$$

The reconstruction is called ***positive*** *if* $\widehat{f} \geq 0$*.*

The word "reconstruction" should not suggest that $\widehat{f}$ is the unique distribution having the prescribed moments. In general, no such uniqueness is possible.

Lemma 4.5 (Finite Moment Data Do Not Determine the Distribution)**.** *Let* $f \geq 0$*, and suppose that a bounded, nonzero, compactly supported function* h *satisfies*

$$\int_{\mathbb{R}^3} h\psi_k \,\mathrm{d}\boldsymbol{v} = 0, \qquad k = 1, \dots, K. \tag{4.10}$$

Assume that $f \geq \eta > 0$ *on the support of* h*. Then, for every sufficiently small* $\delta > 0$*, the two distinct nonnegative distributions*

$$f_+ := f + \delta h, \qquad f_- := f - \delta h \tag{4.11}$$

have exactly the same retained moments as f*.*

Proof. Choose δ so that $\delta\|h\|_{L^\infty} \leq \eta$. Then $f_\pm \geq 0$. Equation (4.10) gives

$$\int_{\mathbb{R}^3} f_\pm \psi_k \,\mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^3} f\psi_k \,\mathrm{d}\boldsymbol{v}, \qquad k = 1, \dots, K. \tag{4.12}$$

Since $h \neq 0$, the two distributions are distinct. ■

There are infinitely many velocity directions orthogonal to a fixed finite collection of moment functions, so perturbations of the form in this lemma are abundant. A closure selects one representative from a large family of distributions that share the same retained moments [51, 74].

Definition 4.6 (Realizable Moment Data)**.** *A collection of values* $u_{\psi_1}, \dots, u_{\psi_K}$ *is called* ***real-***

izable *if there exists a nonnegative distribution g such that*

$$\int_{\mathbb{R}^3} g\psi_k \,\mathrm{d}\boldsymbol{v} = u_{\psi_k}, \qquad k = 1, \dots, K. \tag{4.13}$$

The set of all such collections is called the ***realizable set*** *for the chosen moment functions.*

Realizability is the moment-level remnant of the kinetic condition $f \geq 0$. It is not automatic. For example, a proposed density must be nonnegative, a temperature must be nonnegative, and a second-moment matrix must satisfy corresponding positive-semidefiniteness conditions. Higher-order moment sets obey additional inequalities that may be less obvious.

Proposition 4.7 (Inheritance of Collision Conservation)**.** *Suppose that ψ_k is a collision invariant and that the closed collision source is evaluated by* (4.9) *using the original Boltzmann collision operator. Then*

$$\int_{\mathbb{R}^3} Q[\widehat{f}, \widehat{f}]\psi_k \,\mathrm{d}\boldsymbol{v} = 0 \tag{4.14}$$

whenever the integral is defined.

Proof. The weak collision identity for a collision invariant holds for every admissible distribution. Apply it with $f = \widehat{f}$. ■

This proposition is useful but limited. A practical moment model may approximate the collision source further, in which case conservation must be verified again. Likewise, conservation alone does not guarantee positivity, nonnegative entropy production, hyperbolicity, or accuracy.

4.3 The Five-Moment Maxwellian Closure

The simplest important closure retains only mass, momentum, and total energy. In three dimensions these comprise five scalar fields. The local Maxwellian is the natural reconstruction because it has precisely these moments and minimizes information subject to them [10, 51].

Proposition 4.8 (Euler as a Five-Moment Closure)**.** *Retain the five hydrodynamic fields $(n, \boldsymbol{u}, \theta)$ and choose*

$$\widehat{f} = M[n, \boldsymbol{u}, \theta]. \tag{4.15}$$

Then the reconstruction is positive and moment consistent. Moreover,

$$\mathsf{P} = p\mathsf{I}, \qquad \boldsymbol{q} = \boldsymbol{0}, \qquad p = n\theta = \frac{2}{3}e. \tag{4.16}$$

The resulting five-moment system is the compressible Euler system (3.22)–(3.24).

Proof. Moment consistency and the identities in (4.16) follow from Lemma 3.6, while positivity follows directly from the definition of the local Maxwellian in (3.16). Substitution into the exact unary balance laws gives the Euler equations, as proved in Proposition 3.7. ■

The scalar pressure p is not an additional unknown. It is fixed by the retained energy through $p = 2e/3$, or equivalently by n and θ through $p = n\theta$. What the Maxwellian assumption removes is not the scalar part of the pressure tensor, whose trace is already fixed by energy, but its five-component trace-free part together with the heat flux and all higher kinetic moments.

4.4 Grad's Polynomial Moment Method

The Euler closure is accurate near local equilibrium when non-equilibrium moments relax rapidly. In more rarefied regimes, the stress and heat flux may possess dynamics that are not well represented by instantaneous Newton and Fourier laws. Grad's idea is to retain additional moments as independent fields and reconstruct the distribution by a finite polynomial correction to a Maxwellian [29, 30, 74].

For a monatomic gas in three dimensions, the classical thirteen fields are

$$n, \qquad \boldsymbol{u}, \qquad \theta, \qquad \Pi, \qquad \boldsymbol{q}. \tag{4.17}$$

There are $1 + 3 + 1 + 5 + 3 = 13$ scalar components because Π is symmetric and trace free.

Definition 4.9 (Grad Thirteen-Moment Reconstruction)**.** *Recall from* (3.96) *that, with* $\boldsymbol{c} = \boldsymbol{v} - \boldsymbol{u}$*, the vector-valued velocity polynomial is*

$$\boldsymbol{B}(\boldsymbol{c}) = \left(\frac{m|\boldsymbol{c}|^2}{2\theta} - \frac{5}{2} \right) \boldsymbol{c}.$$

Here $M = M[n, \boldsymbol{u}, \theta]$ *is the local Maxwellian defined in* (3.16)*. Using this polynomial, define the* ***Grad thirteen-moment reconstruction***

$$f_{13} \coloneqq M \left[1 + \frac{m}{2n\theta^2} \Pi : (\boldsymbol{c} \otimes \boldsymbol{c}) + \frac{2m}{5n\theta^2} \boldsymbol{q} \cdot \boldsymbol{B}(\boldsymbol{c}) \right]. \tag{4.18}$$

The numerical coefficients in (4.18) are not arbitrary. They enforce exact agreement with all thirteen retained fields.

Proposition 4.10 (Moment Matching for the Grad Reconstruction)**.** *Suppose that $n > 0$ and $\theta > 0$, that P is symmetric, and that $\operatorname{tr}(\mathsf{P}) = 3n\theta$. Then f_{13} has number density n, velocity $\boldsymbol{u}$, temperature θ, pressure tensor P, and heat flux $\boldsymbol{q}$.*

Proof. The deviatoric pressure tensor Π is trace free. Consequently, its contribution to the mass and energy moments vanishes by isotropy, and its contribution to the momentum moment vanishes by oddness. The $\boldsymbol{q}$ correction is odd, so it does not change the mass, energy, or pressure moments. Its momentum contribution vanishes because $\boldsymbol{B}$ is orthogonal to the components of $\boldsymbol{c}$, as proved in the solvability lemma preceding (3.110).

It remains to verify the two non-equilibrium moments. The Gaussian fourth-moment identity gives

$$m \int_{\mathbb{R}^3} M\, \boldsymbol{c} \otimes \boldsymbol{c} \frac{m}{2n\theta^2} \Pi : (\boldsymbol{c} \otimes \boldsymbol{c})\, \mathrm{d}\boldsymbol{v} = \Pi. \tag{4.19}$$

For the heat flux, isotropy and the Gaussian identities

$$\int_{\mathbb{R}^3} M c_i c_j |\boldsymbol{c}|^2\, \mathrm{d}\boldsymbol{v} = 5n \left(\frac{\theta}{m}\right)^2 \delta_{ij}, \tag{4.20}$$

$$\int_{\mathbb{R}^3} M c_i c_j |\boldsymbol{c}|^4\, \mathrm{d}\boldsymbol{v} = 35n \left(\frac{\theta}{m}\right)^3 \delta_{ij} \tag{4.21}$$

give

$$\frac{m}{2} \int_{\mathbb{R}^3} M\, \boldsymbol{c} |\boldsymbol{c}|^2 \frac{2m}{5n\theta^2} \boldsymbol{q} \cdot \boldsymbol{B}(\boldsymbol{c})\, \mathrm{d}\boldsymbol{v} = \boldsymbol{q}. \tag{4.22}$$

Adding the Maxwellian contributions completes the proof. ■

Substitution of f_{13} into the moment equations for the thirteen fields expresses the unresolved fourth-order moments and collision productions in terms of those fields. The deviatoric pressure and heat flux now have their own evolution equations; they are not prescribed instantaneously by gradients of $\boldsymbol{u}$ and θ. This is the principal conceptual difference from the five-field Chapman–Enskog closure.

The elementary polynomial form also exposes an important defect.

Proposition 4.11 (Loss of Positivity in the Grad Reconstruction)**.** *If either $\Pi \neq 0$ or $\boldsymbol{q} \neq \boldsymbol{0}$, then f_{13} takes negative values for some velocities. Hence the classical Grad reconstruction is not a positive reconstruction away from local equilibrium.*

Proof. If $\boldsymbol{q} \neq \boldsymbol{0}$, the highest-degree part of the bracket in (4.18) is an odd cubic polynomial in $\boldsymbol{c}$. It tends to $-\infty$ along one of the two directions parallel to $\boldsymbol{q}$. If $\boldsymbol{q} = \boldsymbol{0}$ but $\Pi \neq 0$, the symmetric trace-free matrix Π has a negative eigenvalue. Along a corresponding eigenvector,

the quadratic correction tends to $-\infty$. Since $M > 0$, the reconstructed distribution is negative wherever the bracket is negative. ■

This failure does not make Grad's method useless: near equilibrium, it can give a systematic and computationally tractable description of non-equilibrium moments. It does mean that positivity and realizability cannot be inferred from the polynomial ansatz. Classical Grad systems may also lose hyperbolicity away from equilibrium. Regularized moment systems modify the closure of the highest retained equations, often with Chapman–Enskog or order-of-magnitude arguments, to improve stability and asymptotic behavior [74, 75].

4.5 Maximum-Entropy Closure

An entropy-based closure selects a distribution by a variational principle rather than by truncating a polynomial expansion. It therefore makes explicit which representative of the moment class has been chosen.

Definition 4.12 (Entropy Reconstruction)**.** *Let $u_{\psi_1}, \dots, u_{\psi_K}$ be realizable moment data. A **maximum-entropy reconstruction** is a nonnegative minimizer of the information functional*

$$H[g] = \int_{\mathbb{R}^3} g \log g \, \mathrm{d}\boldsymbol{v} \tag{4.23}$$

among all nonnegative g satisfying

$$\int_{\mathbb{R}^3} g\psi_k \, \mathrm{d}\boldsymbol{v} = u_{\psi_k}, \qquad k = 1, \dots, K. \tag{4.24}$$

As usual, $0 \log 0$ is interpreted as zero.

The terminology comes from physical entropy, which is the negative of the information functional up to dimensional constants. Strict convexity implies that there is at most one minimizer. Existence is a separate question and depends upon the moment functions and upon whether the prescribed data lie in the appropriate part of the realizable set.

Proposition 4.13 (Exponential Form of an Entropy Reconstruction)**.** *Suppose that a strictly positive minimizer $\widehat{f}$ of (4.23)–(4.24) exists and that the moment constraints are regular enough for the Lagrange multiplier rule to apply. Then there are real numbers $\lambda_1, \dots, \lambda_K$ such that, for almost every $\boldsymbol{v} \in \mathbb{R}^3$,*

$$\widehat{f} = \exp\left(\sum_{k=1}^{K} \lambda_k \psi_k - 1 \right). \tag{4.25}$$

The multipliers are selected so that the moment-matching conditions in Equation (4.24) *hold.*

Proof. Introduce the Lagrangian

$$\mathcal{L}[g,\lambda] = H[g] - \sum_{k=1}^{K} \lambda_k \left(\int_{\mathbb{R}^3} g\psi_k \,\mathrm{d}\boldsymbol{v} - u_{\psi_k} \right).$$

At a constrained minimizer, formal stationarity with respect to an arbitrary variation h gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{L}[\widehat{f} + \varepsilon h, \lambda]\Big|_{\varepsilon=0} = \int_{\mathbb{R}^3} \left(\log \widehat{f} + 1 - \sum_{k=1}^{K} \lambda_k \psi_k \right) h \,\mathrm{d}\boldsymbol{v}.$$

Because h is arbitrary, the factor multiplying it must vanish:

$$\log \widehat{f} + 1 - \sum_{k=1}^{K} \lambda_k \psi_k = 0. \tag{4.26}$$

Thus exponentiation yields (4.25). ■

When the retained functions are the five collision invariants, the exponential in (4.25) is a quadratic function of velocity with a negative definite quadratic part. Completing the square gives exactly the local Maxwellian. Thus the Euler closure is the first member of the entropy-closure hierarchy [51].

Entropy closures also possess a useful differential structure.

Theorem 4.14 (Formal Symmetric Hyperbolicity of Entropy Closure)**.** *Assume that the moment functions are linearly independent, that the exponential reconstruction* (4.25) *and all required moments are finite, and that the multipliers depend smoothly upon the retained moments. In the variables* $\lambda_1, \ldots, \lambda_K$*, the transport part of the closed moment system is symmetric hyperbolic.*

Proof. Write $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ and regard the reconstructed distribution as a function of these variables:

$$\widehat{f}(\boldsymbol{\lambda}) = \exp\left(\sum_{k=1}^{K} \lambda_k \psi_k - 1 \right). \tag{4.27}$$

Define the retained-moment vector $\boldsymbol{u}(\boldsymbol{\lambda})$ and its flux in the spatial direction x_ℓ componentwise by

$$u_i(\boldsymbol{\lambda}) = \int_{\mathbb{R}^3} \psi_i \widehat{f} \,\mathrm{d}\boldsymbol{v}, \qquad (F_\ell)_i(\boldsymbol{\lambda}) = \int_{\mathbb{R}^3} v_\ell \psi_i \widehat{f} \,\mathrm{d}\boldsymbol{v}. \tag{4.28}$$

The key observation is

$$\frac{\partial \widehat{f}}{\partial \lambda_j} = \psi_j \widehat{f}.$$

It follows that the Jacobian of the moments and the Jacobian of the x_ℓ-flux are

$$G_{ij} \coloneqq \frac{\partial u_i}{\partial \lambda_j} = \int_{\mathbb{R}^3} \psi_i \psi_j \widehat{f}\, \mathrm{d}\boldsymbol{v}, \qquad (A_\ell)_{ij} \coloneqq \frac{\partial (F_\ell)_i}{\partial \lambda_j} = \int_{\mathbb{R}^3} v_\ell \psi_i \psi_j \widehat{f}\, \mathrm{d}\boldsymbol{v}. \tag{4.29}$$

Both matrices are symmetric because their integrands are unchanged when i and j are interchanged. Moreover, for every nonzero vector $\boldsymbol{a} = (a_1, \ldots, a_K)$,

$$\boldsymbol{a}^\top G \boldsymbol{a} = \int_{\mathbb{R}^3} \left(\sum_{k=1}^{K} a_k \psi_k \right)^2 \widehat{f}\, \mathrm{d}\boldsymbol{v} > 0, \tag{4.30}$$

because $\widehat{f} > 0$ and the functions $\psi_1, \ldots, \psi_K$ are linearly independent. Thus G is positive definite and, in particular, invertible. Locally, $\boldsymbol{\lambda}$ may therefore be used in place of the retained moments $\boldsymbol{u}$.

The closed moment equations, before this change of variables, have the vector form

$$\varepsilon \partial_t \boldsymbol{u} + \varepsilon \sum_{\ell=1}^{3} \partial_{x_\ell} \boldsymbol{F}_\ell = \boldsymbol{S},$$

where $\boldsymbol{S}$ is the closed collision source. Substituting $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{\lambda})$ and $\boldsymbol{F}_\ell = \boldsymbol{F}_\ell(\boldsymbol{\lambda})$ and applying the chain rule gives

$$\varepsilon G \partial_t \boldsymbol{\lambda} + \varepsilon \sum_{\ell=1}^{3} A_\ell \partial_{x_\ell} \boldsymbol{\lambda} = \boldsymbol{S}(\boldsymbol{\lambda}). \tag{4.31}$$

By definition, a first-order system is symmetric hyperbolic when the matrix multiplying the time derivative is symmetric positive definite and the matrices multiplying the spatial derivatives are symmetric. Here these matrices are εG and εA_ℓ, respectively, so the system has precisely this form. The source contains no derivatives and therefore does not affect the conclusion. ■

If high-degree polynomial moments are constrained, the exponential in (4.25) need not be integrable for arbitrary multipliers. Near the boundary of the realizable set, the multiplier problem may become singular or poorly conditioned. Maximum-entropy closure exchanges the explicit simplicity of Grad's polynomial formula for positivity, convexity, and a stronger structural connection to entropy.

4.6 Quadrature and Discrete-Velocity Closures

A different strategy represents the velocity distribution by finitely many nonnegative weights concentrated at velocity nodes. Formally, one writes

$$\widehat{f}(\boldsymbol{v}) = \sum_{\ell=1}^{L} w_\ell \delta(\boldsymbol{v} - \boldsymbol{v}_\ell), \qquad w_\ell \geq 0, \tag{4.32}$$

so that

$$u_{\psi_k} = \sum_{\ell=1}^{L} w_\ell \psi_k(\boldsymbol{v}_\ell). \tag{4.33}$$

The weights, the nodes, or both are chosen from the retained moments. Equation (4.33) converts velocity integrals into finite sums and makes nonnegativity transparent when all weights are nonnegative [26, 76].

There is an important distinction between a quadrature closure for transport moments and a complete discretization of the Boltzmann collision operator. Collisions typically send a pair of velocity nodes to post-collisional velocities that are not nodes of the chosen quadrature. A numerical method must therefore add a conservative projection, interpolation, or separate collision quadrature. Mass, momentum, energy, and entropy properties must then be checked for that discrete collision procedure; they do not follow from (4.32) alone [20, 54].

4.7 Local, Differential, and Extended Closures

The preceding examples can be organized by what information is retained.

Definition 4.15 (Three Forms of Closure)**.** *A closure is called*

(i) ***local algebraic*** *if the unresolved moments at* $(\boldsymbol{x}, t)$ *depend only upon the retained moments at the same point;*

(ii) ***differential*** *if they also depend upon spatial derivatives of the retained moments;*

(iii) ***extended*** *if selected non-equilibrium moments are retained as independent unknowns with their own evolution equations.*

The Maxwellian Euler closure is local algebraic. Entropy and quadrature closures are also local algebraic once their full set of retained moments has been selected. Grad's thirteen-moment system is extended relative to the five hydrodynamic variables because the deviatoric pressure and heat flux are independent fields.

Chapman–Enskog closure is differential. Under the normal-solution assumption, kinetic information is expressed as a functional of the hydrodynamic fields and their spatial derivatives; see Equation (3.49). At first order, the Newton and Fourier laws derived above give

$$\Pi = -2\varepsilon\mu \mathrm{D}^{\circ}(\boldsymbol{u}) + \mathcal{O}(\varepsilon^2), \qquad \boldsymbol{q} = -\varepsilon\kappa\nabla_{\boldsymbol{x}}\theta + \mathcal{O}(\varepsilon^2). \tag{4.34}$$

Thus the stress and heat flux are not evolved independently; they are slaved, to first order, to gradients of the five hydrodynamic fields.

Proposition 4.16 (Relation Between Extended Moments and Chapman–Enskog Closure)**.** *Consider a consistent extended moment model that retains n, $\boldsymbol{u}$, θ, Π, and $\boldsymbol{q}$. Suppose that its non-equilibrium moments relax to zero on the kinetic time scale and that a formal expansion in ε is valid. Then its leading equilibrium reduction is the Euler closure. If its first-order asymptotics agree with the Boltzmann equation, its first non-equilibrium terms must reproduce* (4.34) *with the Boltzmann viscosity and thermal conductivity.*

Proof. At leading order, rapid collisional relaxation sets the non-equilibrium stress and heat flux equal to zero. The exact trace relation leaves $\mathsf{P} = p\mathsf{I}$, and the five balance laws reduce to Euler. At first order, agreement with the Boltzmann Chapman–Enskog expansion requires equality of the first-order constitutive moments. Those moments were computed uniquely from the normalized solution of the linearized collision problems (3.110)–(3.111), and they are precisely the two relations in (4.34). ■

This proposition explains both the relation and the distinction. Chapman–Enskog analysis eliminates the rapidly relaxing moments order by order. An extended moment method keeps some of them as dynamical variables. One may nevertheless apply a Chapman–Enskog expansion to the extended system itself. Recovery of the correct Euler and Navier–Stokes–Fourier limits is then an essential consistency test for the moment model [30, 74, 75].

4.8 What Should a Closure Preserve?

Moment matching is necessary, but it is not enough. A reliable closure should be assessed against several separate requirements. The criteria below collect standard structural concerns in moment-closure theory [51, 74].

1. **Consistency.** The reconstruction must reproduce every retained moment exactly. Otherwise the variables appearing in the closed equations do not have their stated kinetic meanings.

2. **Collision invariants.** The closed collision sources must preserve mass, total momentum, and total energy. For a mixture, each species mass is separately conserved, whereas momentum and energy are generally exchanged between species and conserved only after summation.

3. **Equilibrium.** The closure should contain the appropriate local Maxwellians and should not introduce spurious equilibria near the physical regime of interest.

4. **Realizability and positivity.** Ideally, admissible moment data should correspond to a nonnegative distribution, and the closed evolution should remain inside the realizable set. Grad's polynomial closure shows that moment matching alone does not ensure this property.

5. **Entropy structure.** The nonnegative kinetic entropy-production law should descend, exactly or approximately, to the closed system. Equivalently, the associated information functional should be non-increasing. This structure often provides both physical guidance and mathematical stability.

6. **Well-posedness.** The closed partial differential equations should have an appropriate hyperbolic, parabolic, or hyperbolic–relaxation structure. A formally accurate closure is of limited use if its initial-value problem is unstable.

7. **Asymptotic consistency.** In the small-Knudsen-number regime, the model should recover Euler at leading order and the correct Navier–Stokes–Fourier transport coefficients at first order.

8. **Symmetry.** Rotational and Galilean invariance should be inherited from the kinetic equation. Tensorial moments and their closures must transform accordingly.

9. **Boundary compatibility.** A finite moment system requires boundary conditions derived from, or at least consistent with, the kinetic boundary condition. Interior closure accuracy does not by itself resolve Knudsen layers, velocity slip, or temperature jump.

These requirements can conflict. A low-order polynomial closure may be explicit and inexpensive but fail positivity. An entropy closure may preserve positivity and hyperbolicity but require a nonlinear optimization at every point. A large moment system may describe stronger non-equilibrium effects but introduce many variables, difficult collision integrals, and nontrivial boundary conditions. There is no universally best closure independent of the physical regime and computational purpose.

4.9 Extension to Multispecies Systems

For a mixture, one chooses moment functions for each f_i and reconstructs the family of distributions simultaneously. Equations (2.115) to (2.117) show what must be closed: species pressure tensors and heat fluxes, as well as interspecies momentum- and energy-exchange terms. The conservation structure is more delicate than in the unary problem.

If the reconstructed distributions are inserted into the original binary operators $Q_{i,j}$ without further approximation, the weak collision identities automatically preserve each species mass and preserve total momentum and energy after summation over species. If the collision productions are replaced by algebraic relaxation terms, these cancellations must be imposed and proved for the approximate sources.

The equilibrium manifold also changes in a precise way. Interspecies collisions permit distinct species densities, but they drive the mixture toward a common bulk velocity and a common temperature. A multispecies closure should therefore distinguish between fields that remain independent at equilibrium and fields, such as diffusion velocities and temperature differences, that relax through interspecies exchange. At first Chapman–Enskog order, the closure additionally contains diffusion and thermal-diffusion laws. These are the mixture analogues of the unary viscous-stress and heat-flux laws [62, 3].

4.10 Summary of the Methodology

The general moment-closure procedure can now be stated without ambiguity. The following workflow synthesizes the standard moment-method framework developed in [29, 51, 74]:

1. choose the velocity moments to be retained;
2. derive their exact balance laws from the kinetic equation;
3. identify every unresolved flux and collision production;
4. prescribe a reconstruction or constitutive rule for those quantities;
5. prove moment consistency and conservation;
6. determine the realizable set and examine positivity, entropy, and well-posedness;
7. compare the small-Knudsen-number expansion with the Chapman–Enskog closure;
8. derive boundary conditions consistent with the underlying kinetic model.

The first three steps are exact consequences of the Boltzmann equation. The fourth is the closure assumption. Every property claimed after that point must be proved for the chosen closure rather than attributed automatically to the kinetic equation.

Chapter 5

Single-Species BGK Kinetic Models

In this chapter, we describe some basic theory for the single-species BGK equation [6, 74], namely,

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f = \lambda \left(M_f - f \right). \tag{5.1}$$

The appeal of the BGK model is computational as well as structural. Its collision term is much less expensive to evaluate than the Boltzmann collision integral, while still driving solutions toward Maxwellian states.

The existence of non-negative solutions to the single-species BGK equation on $\boldsymbol{x} \in \mathbb{R}^d$ was proved by Perthame [60]. Uniqueness of mild solutions in the spatially periodic case was shown by Perthame and Pulvirenti [59], and extended to the full space by Mischler [56].

Beyond questions about existence and uniqueness of solutions, it is important to have a firm grasp of the properties of solutions to BGK-type equations, since it is vital to build numerical approximation schemes that respect analogous features at the fully discrete level. Herein, we review conservation, information dissipation, and entropy production, and, later, we discuss how these are used in the design of numerical schemes.

5.1 Basic Properties of Solutions

Since we are examining the single-species case first, that is, the unary case, we can simplify our discussion slightly by setting the mass to be one, $m = 1$. We recall that the Maxwellian is constructed using the first three integral moments of the kinetic distribution, f, that is,

$$n(\boldsymbol{x}, t) \coloneqq \int_{\mathbb{R}^d} f \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u}(\boldsymbol{x}, t) \coloneqq \frac{1}{n} \int_{\mathbb{R}^d} f \boldsymbol{v} \, \mathrm{d}\boldsymbol{v}, \qquad \theta(\boldsymbol{x}, t) \coloneqq \frac{1}{nd} \int_{\mathbb{R}^d} f |\boldsymbol{v} - \boldsymbol{u}|^2 \, \mathrm{d}\boldsymbol{v}, \tag{5.2}$$

and is defined as

$$M_f(\boldsymbol{x}, \boldsymbol{v}, t) \coloneqq M[f](\boldsymbol{x}, \boldsymbol{v}, t) \coloneqq n \left(\frac{1}{2\pi\theta}\right)^{d/2} \exp\left(-\frac{|\boldsymbol{v}-\boldsymbol{u}|^2}{2\theta}\right). \tag{5.3}$$

The space-time fields n, $\boldsymbol{u}$, and θ, are the macroscopic (coarse grained) number density, bulk velocity, and temperature, respectively. The macroscopic kinetic energy density is given by

$$\omega(\boldsymbol{x}, t) \coloneqq \frac{1}{2}\int_{\mathbb{R}^d} f|\boldsymbol{v}|^2 \,\mathrm{d}\boldsymbol{v}, \tag{5.4}$$

and the internal energy density is defined as

$$e(\boldsymbol{x}, t) \coloneqq \frac{1}{2}\int_{\mathbb{R}^d} f|\boldsymbol{v}-\boldsymbol{u}|^2 \,\mathrm{d}\boldsymbol{v}. \tag{5.5}$$

We observe that the temperature and the internal energy are related via

$$\theta(\boldsymbol{x}, t) = \frac{2}{d}\frac{e(\boldsymbol{x}, t)}{n(\boldsymbol{x}, t)}.$$

5.1.1 Collision Invariants

Again, we define $\langle\cdot\rangle\colon L^1(\mathbb{R}^d;\mathbb{R}^p) \to \mathbb{R}^p$ by

$$\langle \boldsymbol{g} \rangle = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{v}) \,\mathrm{d}\boldsymbol{v}.$$

Therefore,

$$\begin{bmatrix} n(\boldsymbol{x}, t) \\ n(\boldsymbol{x}, t)\boldsymbol{u}(\boldsymbol{x}, t) \\ \omega(\boldsymbol{x}, t) \end{bmatrix} = \left\langle f(\boldsymbol{x}, \boldsymbol{v}, t) \begin{bmatrix} 1 \\ \boldsymbol{v} \\ \frac{1}{2}|\boldsymbol{v}|^2 \end{bmatrix} \right\rangle.$$

Thus, the macroscopic number density n, momentum density $n\boldsymbol{u}$, and total energy density ω may be viewed as the first three moments of the particle distribution function f. The first three moments of the collision operator are equal to zero. This gives us some conservation properties, as we show below.

For convenience, let us define

$$\boldsymbol{\rho}(\boldsymbol{x}, t) \coloneqq \left[n, n\boldsymbol{u}^\top, \omega\right]^\top : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{d+2}; \tag{5.6}$$

$$\boldsymbol{m}(\boldsymbol{v}) \coloneqq \left[1, \boldsymbol{v}^\top, \frac{1}{2}|\boldsymbol{v}|^2\right]^\top : \mathbb{R}^d \to \mathbb{R}^{d+2}; \tag{5.7}$$

and

$$\boldsymbol{\alpha}(\boldsymbol{x},t) \coloneqq \left[\ln\left(\frac{n}{(2\pi\theta)^{d/2}}\right) - \frac{|\boldsymbol{u}|^2}{2\theta}, \frac{\boldsymbol{u}^\top}{\theta}, -\frac{1}{\theta}\right]^\top : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{d+2}. \tag{5.8}$$

Then, succinctly,

$$\boldsymbol{\rho} = \langle f\boldsymbol{m}\rangle, \qquad \text{and} \qquad M_f = \exp\left(\boldsymbol{\alpha}\cdot\boldsymbol{m}(\boldsymbol{v})\right). \tag{5.9}$$

We summarize some facts that we have already established in the previous chapter.

Lemma 5.1. *The following equality holds:*

$$\langle (M_f - f)\,\boldsymbol{m}\rangle = \left\langle \begin{bmatrix} 1 \\ \boldsymbol{v} \\ \frac{|\boldsymbol{v}|^2}{2} \end{bmatrix} (M_f - f) \right\rangle = \boldsymbol{0}. \tag{5.10}$$

Definition 5.2 (BGK Collision Invariant)**.** *We say that a function $g(\boldsymbol{v})$ is a **BGK collision invariant** iff*

$$\langle g(\boldsymbol{v})\,(M_f - f)\rangle = \int_{\mathbb{R}^d} g(\boldsymbol{v})\,(M_f - f)\,\mathrm{d}\boldsymbol{v} = 0.$$

From the last lemma, we observe that 1, $\boldsymbol{v}$, and $|\boldsymbol{v}|^2$ are BGK collision invariants. Of course, any linear combination of these functions will also be collision invariant. We will use this fact later.

Multiplying the BGK Equation by the vector $\boldsymbol{m} = [1, \boldsymbol{v}^\top, \frac{1}{2}|\boldsymbol{v}|^2]^\top$ and integrating, we get expressions for the conservation of mass, momentum, and energy, respectively.

Lemma 5.3. *Suppose that f is a non-negative solution to the BGK model in Equation* (5.1)*. The following conservation equations hold:*

$$\frac{\partial}{\partial t} n + \nabla_{\boldsymbol{x}}\cdot(n\boldsymbol{u}) = 0, \tag{5.11}$$

$$\frac{\partial}{\partial t}(n\boldsymbol{u}) + \nabla_{\boldsymbol{x}}\cdot(n\boldsymbol{u}\otimes\boldsymbol{u} + \mathrm{P}) = \boldsymbol{0}, \tag{5.12}$$

$$\frac{\partial}{\partial t}\omega + \nabla_{\boldsymbol{x}}\cdot(\omega\boldsymbol{u} + \mathrm{P}\boldsymbol{u} + \boldsymbol{q}) = 0, \tag{5.13}$$

where

$$\mathrm{P} \coloneqq \int_{\mathbb{R}^d} (\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u}) f\,\mathrm{d}\boldsymbol{v}$$

is the pressure tensor and

$$\boldsymbol{q} \coloneqq \frac{1}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}-\boldsymbol{u}|^2(\boldsymbol{v}-\boldsymbol{u}) f\,\mathrm{d}\boldsymbol{v}$$

is the heat flux.

Proof. We follow the moment calculation used for the multispecies Boltzmann equation in the previous chapter. Multiplying the BGK equation by 1 and integrating gives

$$\int_{\mathbb{R}^d} \left[\frac{\partial f}{\partial t} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{v} f)\right] \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} \lambda(M_f - f)\,\mathrm{d}\boldsymbol{v} \tag{5.14}$$

$$\implies \frac{\partial}{\partial t}\left(\int_{\mathbb{R}^d} f\,\mathrm{d}\boldsymbol{v}\right) + \nabla_{\boldsymbol{x}} \cdot \left(\int_{\mathbb{R}^d} \boldsymbol{v} f\,\mathrm{d}\boldsymbol{v}\right) = 0 \tag{5.15}$$

$$\implies \frac{\partial}{\partial t} n + \nabla_{\boldsymbol{x}} \cdot (n\boldsymbol{u}) = 0. \tag{5.16}$$

This verifies the first equation, which is an expression for the conservation of mass.

Next, multiplying by $\boldsymbol{v}$ and integrating, we have

$$\int_{\mathbb{R}^d} \left[\boldsymbol{v}\frac{\partial f}{\partial t} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{v} \otimes \boldsymbol{v} f)\right] \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} \boldsymbol{v}\lambda(M_f - f)\,\mathrm{d}\boldsymbol{v} \tag{5.17}$$

$$\implies \frac{\partial}{\partial t}\left(\int_{\mathbb{R}^d} \boldsymbol{v} f\,\mathrm{d}\boldsymbol{v}\right) + \nabla_{\boldsymbol{x}} \cdot \left(\int_{\mathbb{R}^d} \boldsymbol{v} \otimes \boldsymbol{v} f\,\mathrm{d}\boldsymbol{v}\right) = \boldsymbol{0}. \tag{5.18}$$

The first term is equal to $\partial_t(n\boldsymbol{u})$, as desired. It remains to show that

$$\int_{\mathbb{R}^d} \boldsymbol{v} \otimes \boldsymbol{v} f\,\mathrm{d}\boldsymbol{v} = n\boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}.$$

Writing $\boldsymbol{v} = \boldsymbol{u} + (\boldsymbol{v} - \boldsymbol{u})$, gives

$$\begin{aligned}
\int_{\mathbb{R}^d} \boldsymbol{v} \otimes \boldsymbol{v} f\,\mathrm{d}\boldsymbol{v} &= \int_{\mathbb{R}^d} [\boldsymbol{u} + (\boldsymbol{v} - \boldsymbol{u})] \otimes [\boldsymbol{u} + (\boldsymbol{v} - \boldsymbol{u})]\, f\,\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}^d} [\boldsymbol{u} \otimes \boldsymbol{u} + \boldsymbol{u} \otimes (\boldsymbol{v} - \boldsymbol{u}) + (\boldsymbol{v} - \boldsymbol{u}) \otimes \boldsymbol{u} + (\boldsymbol{v} - \boldsymbol{u}) \otimes (\boldsymbol{v} - \boldsymbol{u})]\, f\,\mathrm{d}\boldsymbol{v} \\
&= n\boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P} + \int_{\mathbb{R}^d} \boldsymbol{u} \otimes (\boldsymbol{v} - \boldsymbol{u}) f\,\mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^d} (\boldsymbol{v} - \boldsymbol{u}) \otimes \boldsymbol{u} f\,\mathrm{d}\boldsymbol{v}.
\end{aligned} \tag{5.19}$$

The last two terms in the above expression are equal to zero:

$$\boldsymbol{u} \otimes \int_{\mathbb{R}^d} (\boldsymbol{v} - \boldsymbol{u}) f\,\mathrm{d}\boldsymbol{v} = \boldsymbol{u} \otimes \left[\int_{\mathbb{R}^d} \boldsymbol{v} f\,\mathrm{d}\boldsymbol{v} - \boldsymbol{u}\int_{\mathbb{R}^d} f\,\mathrm{d}\boldsymbol{v}\right] = \boldsymbol{u} \otimes [n\boldsymbol{u} - \boldsymbol{u}n] = \boldsymbol{0}.$$

Putting it all together, we obtain the second equation, an expression for the conservation of momentum.

Finally, multiplying by $\frac{1}{2}|\boldsymbol{v}|^2$ and integrating, gives

$$\int_{\mathbb{R}^d} \frac{1}{2}|\boldsymbol{v}|^2 \left(\frac{\partial f}{\partial t} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{v} f)\right) \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} \frac{1}{2}|\boldsymbol{v}|^2 \lambda(M_f - f)\,\mathrm{d}\boldsymbol{v} \tag{5.20}$$

$$\implies \frac{\partial}{\partial t}\left(\int_{\mathbb{R}^d}\frac{1}{2}|\boldsymbol{v}|^2 f\,\mathrm{d}\boldsymbol{v}\right)+\nabla_{\boldsymbol{x}}\cdot\left(\int_{\mathbb{R}^d}\frac{1}{2}|\boldsymbol{v}|^2\boldsymbol{v}f\,\mathrm{d}\boldsymbol{v}\right)=0. \tag{5.21}$$

The first term is equal to $\partial_t\omega$, as desired. It remains to show that

$$\int_{\mathbb{R}^d}\frac{1}{2}|\boldsymbol{v}|^2\boldsymbol{v}f\,\mathrm{d}\boldsymbol{v}=\omega\boldsymbol{u}+\mathrm{P}\boldsymbol{u}+\boldsymbol{q}.$$

Writing $\boldsymbol{v}=\boldsymbol{u}+(\boldsymbol{v}-\boldsymbol{u})$, gives

$$\begin{aligned}
&\int_{\mathbb{R}^d}\frac{1}{2}|\boldsymbol{v}|^2\boldsymbol{v}f\,\mathrm{d}\boldsymbol{v}=\frac{1}{2}\int_{\mathbb{R}^d}|\boldsymbol{v}|^2\left[\boldsymbol{u}+(\boldsymbol{v}-\boldsymbol{u})\right]f\,\mathrm{d}\boldsymbol{v}\\
&=\boldsymbol{u}\int_{\mathbb{R}^d}\frac{1}{2}|\boldsymbol{v}|^2 f\,\mathrm{d}\boldsymbol{v}+\frac{1}{2}\int_{\mathbb{R}^d}|\boldsymbol{v}|^2(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\\
&=\omega\boldsymbol{u}+\frac{1}{2}\int_{\mathbb{R}^d}\left[\boldsymbol{u}+(\boldsymbol{v}-\boldsymbol{u})\right]^{\top}\left[\boldsymbol{u}+(\boldsymbol{v}-\boldsymbol{u})\right](\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\\
&=\omega\boldsymbol{u}+\frac{1}{2}\int_{\mathbb{R}^d}\left[|\boldsymbol{u}|^2+2\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})+|\boldsymbol{v}-\boldsymbol{u}|^2\right](\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\\
&=\omega\boldsymbol{u}+\frac{|\boldsymbol{u}|^2}{2}\int_{\mathbb{R}^d}(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}+\int_{\mathbb{R}^d}\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}+\frac{1}{2}\int_{\mathbb{R}^d}|\boldsymbol{v}-\boldsymbol{u}|^2(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\\
&=\omega\boldsymbol{u}+\boldsymbol{q}+\frac{1}{2}|\boldsymbol{u}|^2\left(\int_{\mathbb{R}^d}\boldsymbol{v}f\,\mathrm{d}\boldsymbol{v}-\boldsymbol{u}\int_{\mathbb{R}^d}f\,\mathrm{d}\boldsymbol{v}\right)+\int_{\mathbb{R}^d}\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\\
&=\omega\boldsymbol{u}+\boldsymbol{q}+\frac{1}{2}|\boldsymbol{u}|^2\left(n\boldsymbol{u}-\boldsymbol{u}n\right)+\int_{\mathbb{R}^d}\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\\
&=\omega\boldsymbol{u}+\boldsymbol{q}+\int_{\mathbb{R}^d}\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}.
\end{aligned}\tag{5.22}$$

The remaining contraction satisfies

$$\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})=\left[(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})\right]\boldsymbol{u}.$$

Thus, the final term in the above Equation is

$$\begin{aligned}
\int_{\mathbb{R}^d}\boldsymbol{u}^{\top}(\boldsymbol{v}-\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}&=\int_{\mathbb{R}^d}\left[(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})\right]\boldsymbol{u}f\,\mathrm{d}\boldsymbol{v}\\
&=\left(\int_{\mathbb{R}^d}(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})f\,\mathrm{d}\boldsymbol{v}\right)\boldsymbol{u}\\
&=\mathrm{P}\boldsymbol{u}.
\end{aligned}\tag{5.23}$$

Putting it all together gives the third equation, which is an expression for the conservation of energy. The proof is complete. ■

The conservation laws/properties above are reminiscent of those involved with the derivations of the Navier–Stokes and Euler equations. In fact, using the Chapman–Enskog expansion method one can show that in the limit as $\lambda \to \infty$, or $\tau \to 0$, one recovers macroscopic Navier–Stokes and/or Euler equations as formal limits, under certain assumptions [64, 81, 79, 74]. These limits can even guide in the design of stable numerical methods for the macroscopic models [81].

5.1.2 Space-Homogeneous Problem

Suppose that the density function is spatially homogeneous, so that $f = f(\boldsymbol{v}, t)$. In this case, the distribution function satisfies the space-homogeneous problem

$$\frac{\partial f}{\partial t} = \lambda(M_f - f). \tag{5.24}$$

This is a first order integro-differential equation (IDE), which can be solved using the integrating factor method.

Lemma 5.4. *Suppose that f is a solution to the space-homogeneous problem* (5.24)*. Then*

$$f(\boldsymbol{v}, t) = e^{-\lambda t} f(\boldsymbol{v}, 0) + \left(1 - e^{-\lambda t}\right) M_f(\boldsymbol{v}, 0). \tag{5.25}$$

Proof. We begin by proving that $\frac{\partial n}{\partial t} = 0 = \frac{\partial \theta}{\partial t}$, $\frac{\partial \boldsymbol{u}}{\partial t} = \boldsymbol{0}$, and $\frac{\partial M_f}{\partial t} = 0$. We will make frequent use of Lemma 5.1.

First, we show that $\frac{\partial n}{\partial t} = 0$. Utilizing the IDE,

$$\frac{\partial}{\partial t} n = \frac{\partial}{\partial t} \int_{\mathbb{R}^d} f \, \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} \frac{\partial f}{\partial t} \, \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} \lambda(M_f - f) \, \mathrm{d}\boldsymbol{v} \overset{\text{Lemma 5.1}}{=} 0. \tag{5.26}$$

Next, we show that $\frac{\partial \boldsymbol{u}}{\partial t} = \boldsymbol{0}$. Using the quotient rule on the definition of $\boldsymbol{u} = \frac{1}{n} \int_{\mathbb{R}^d} \boldsymbol{v} f \, \mathrm{d}\boldsymbol{v}$, and the fact that $\frac{\partial n}{\partial t} = 0$ gives

$$\frac{\partial \boldsymbol{u}}{\partial t} = \frac{\partial}{\partial t} \frac{1}{n} \int_{\mathbb{R}^d} f \boldsymbol{v} \, \mathrm{d}\boldsymbol{v} = \frac{(n) \int_{\mathbb{R}^d} \frac{\partial f}{\partial t} \boldsymbol{v} \, \mathrm{d}\boldsymbol{v} - 0}{n^2} = \frac{1}{n} \int_{\mathbb{R}^d} \lambda(M_f - f) \boldsymbol{v} \, \mathrm{d}\boldsymbol{v} \overset{\text{Lemma 5.1}}{=} \boldsymbol{0}. \tag{5.27}$$

Next, we show that $\frac{\partial \theta}{\partial t} = 0$. Recall that

$$\frac{d}{2} n\theta = \omega - \frac{1}{2} n |\boldsymbol{u}|^2 \quad \iff \quad \theta = \frac{2\omega}{nd} - \frac{1}{d} |\boldsymbol{u}|^2.$$

Using the quotient rule, and the fact that $\frac{\partial n}{\partial t} = 0$ and $\frac{\partial \boldsymbol{u}}{\partial t} = \mathbf{0}$, gives

$$\frac{\partial \theta}{\partial t} = \frac{2}{d}\frac{\left(\frac{\partial \omega}{\partial t}\right)(n) - 0}{n^2} - \frac{2}{d}\boldsymbol{u}^\top \frac{\partial \boldsymbol{u}}{\partial t} = \frac{2}{nd}\int_{\mathbb{R}^d} \frac{|\boldsymbol{v}|^2}{2}\lambda(M_f - f)\,\mathrm{d}\boldsymbol{v} \overset{\text{Lemma 5.1}}{=} 0. \tag{5.28}$$

Next, to show $\frac{\partial M_f}{\partial t} = 0$, we use the temporal invariance of n, $\boldsymbol{u}$, and θ, as shown above,

$$\begin{aligned}
\frac{\partial}{\partial t}M_f(\boldsymbol{v}, t) &= (2\pi)^{-\frac{d}{2}}\frac{\partial}{\partial t}\left[n(t)\theta(t)^{-\frac{d}{2}}\exp\left(-\frac{1}{2}|\boldsymbol{v} - \boldsymbol{u}(t)|^2\theta(t)^{-1}\right)\right] \\
&= (2\pi)^{-\frac{d}{2}}\left[\frac{\partial n}{\partial t}\theta^{-\frac{d}{2}}\exp\left(-\frac{1}{2}|\boldsymbol{v} - \boldsymbol{u}|^2\theta^{-1}\right)\right. \\
&\quad + n(t)\left(-\frac{d}{2}\right)\theta^{-\frac{d+2}{2}}\frac{\partial \theta}{\partial t}\exp\left(-\frac{1}{2}|\boldsymbol{v} - \boldsymbol{u}|^2\theta^{-1}\right) \\
&\quad \left. + n\theta^{-\frac{d}{2}}\exp\left(-\frac{1}{2}|\boldsymbol{v} - \boldsymbol{u}|^2\theta^{-1}\right)\frac{\partial}{\partial t}\left[-\frac{1}{2}|\boldsymbol{v} - \boldsymbol{u}|^2\theta^{-1}\right]\right] \\
&= 0.
\end{aligned} \tag{5.29}$$

So far, we have $\frac{\partial n}{\partial t} = 0 = \frac{\partial \theta}{\partial t}$, $\frac{\partial \boldsymbol{u}}{\partial t} = \mathbf{0}$, and $\frac{\partial M_f}{\partial t} = 0$. To finish, we use these properties to solve the IDE. Rearranging (5.24), gives

$$\frac{\mathrm{d}f}{\mathrm{d}t} + \lambda f = \lambda M_f.$$

Multiplying by the integrating factor, $e^{\lambda t}$, and applying the product rule in reverse, gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\lambda t}f\right) = \lambda M_f e^{\lambda t}.$$

Integrating with respect to t, we have

$$e^{\lambda t}f = C + \lambda\int e^{\lambda t}M_f(\boldsymbol{v}, t)\,\mathrm{d}t,$$

Since $M_f = M_f(\boldsymbol{v}, t)$ is constant with respect to t, $M_f(\boldsymbol{v}, t) = M_f(\boldsymbol{v}, 0)$. Hence, it can be pulled through the integral sign:

$$\begin{aligned}
e^{\lambda t}f &= C + \lambda M_f(\boldsymbol{v}, 0)\int e^{\lambda t}\,\mathrm{d}t \\
&= C + M_f(\boldsymbol{v}, 0)e^{\lambda t}.
\end{aligned} \tag{5.30}$$

Plugging in $t = 0$, and rearranging gives an expression for the constant term:

$$e^0 f(\boldsymbol{v}, 0) = C + e^0 M_f(\boldsymbol{v}, 0) \tag{5.31}$$

$$\implies C = f(\boldsymbol{v}, 0) - M_f(\boldsymbol{v}, 0). \tag{5.32}$$

Putting everything together, and multiplying by $e^{-\lambda t}$, we have the following solution to the space-homogeneous problem:

$$\begin{aligned} f(\boldsymbol{v}, t) &= e^{-\lambda t} \left(f(\boldsymbol{v}, 0) - M_f(\boldsymbol{v}, 0) \right) + M_f(\boldsymbol{v}, 0) \\ &= e^{-\lambda t} f(\boldsymbol{v}, 0) + \left(1 - e^{-\lambda t}\right) M_f(\boldsymbol{v}, 0). \end{aligned} \tag{5.33}$$

■

Since we have the true solution to this space-homogeneous problem, this allows us to test the accuracy of the code on the right hand side source term. Initial tests are performed in Section 6.5.1.

5.1.3 The H-Theorem

In this section, we discuss a central solution property for the BGK equation: physical entropy is produced, while Boltzmann's H-functional is dissipated. This stability structure should, in some form, be preserved in numerical approximations. We start with a definition.

Definition 5.5 (Information Functional and Entropy Production)**.** *Suppose that* $f : \Omega \times \mathbb{R}^d \times [0, \infty) \to [0, \infty)$ *is a particle distribution function. The object*

$$H[f](\boldsymbol{x}, t) \coloneqq \int_{\mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{v}, t) \ln(f(\boldsymbol{x}, \boldsymbol{v}, t)) \, \mathrm{d}\boldsymbol{v}$$

is called the H***-functional*** *or* ***information functional****. The corresponding total physical entropy is*

$$S[f](t) \coloneqq -\int_{\Omega} H[f](\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x}.$$

We say that a flow is ***entropy-producing*** *when, for every* $0 \leq t_1 \leq t_2 < \infty$*,*

$$S[f](t_2) \geq S[f](t_1),$$

or, equivalently, when the total information $\int_\Omega H[f] \, \mathrm{d}\boldsymbol{x}$ *is non-increasing.*

Remark 5.6 (Entropy and information)**.** *The physical entropy convention is used throughout this report: entropy $S = -H$ is non-decreasing. The historical quantity $H = \int f \ln f$ is retained because it is standard in kinetic theory, but it is called information and is non-increasing. Thus the H-theorem and the entropy-production law are the same statement with opposite signs.*

Remark 5.7. *The entropy-production property is an important marker for the design of numerical methods. In particular, a numerical approximation should, if possible, produce entropy, or equivalently dissipate information, at the discrete level. Proving this property for a fully discrete method can be challenging and may impose significant computational cost [54, 55]; nevertheless, checking it numerically in benchmark simulations remains worthwhile.*

Theorem 5.8. *Suppose that $f : \Omega \times \mathbb{R}^d \times [0,\infty) \to [0,\infty)$ is a solution to the spatially homogeneous BGK problem, that is,*

$$\frac{\partial f}{\partial t} = \frac{1}{\tau}\left(M_f - f\right),$$

where $\tau > 0$ is a constant. Then the information is non-increasing and the physical entropy is non-decreasing:

$$\frac{\mathrm{d}}{\mathrm{d}t}H[f] \le 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}S[f] \ge 0, \qquad S[f] \coloneqq -H[f].$$

Proof. Observe that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H[f](t) &= \int_{\mathbb{R}^d} \frac{\partial f}{\partial t}\left(\ln(f)+1\right)\,\mathrm{d}\boldsymbol{v} \\ &= \frac{1}{\tau}\int_{\mathbb{R}^d}\left(M_f - f\right)\left(\ln(f)+1\right)\,\mathrm{d}\boldsymbol{v} \\ &= \frac{1}{\tau}\int_{\mathbb{R}^d}\ln(f)\left(M_f - f\right)\,\mathrm{d}\boldsymbol{v}\end{aligned}$$

To finish the proof, we use that fact that $\ln(M_f)$ is a BGK collision invariant, that is

$$\int_{\mathbb{R}^d}\ln(M_f)\left(M_f - f\right)\mathrm{d}\boldsymbol{v} = 0.$$

Thus

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H[f](t) &= \frac{1}{\tau}\int_{\mathbb{R}^d}\left(\ln(f) - \ln(M_f)\right)\left(M_f - f\right)\,\mathrm{d}\boldsymbol{v} \\ &= -\frac{1}{\tau}\int_{\mathbb{R}^d}\left(\ln(M_f) - \ln(f)\right)\left(M_f - f\right)\,\mathrm{d}\boldsymbol{v} \le 0,\end{aligned}$$

where in the last step we used Lemma 2.23. ■

More generally, we have

Theorem 5.9. *Suppose that $f : \Omega \times \mathbb{R}^d \times [0,\infty) \to [0,\infty)$ is an Ω-periodic (spatially periodic) solution to the BGK equation, that is,*

$$\partial_t f + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f = \frac{1}{\tau}\left(M_f - f\right),$$

where $\tau > 0$ is constant. Then

$$\partial_t \langle f \ln(f)\rangle + \nabla_{\boldsymbol{x}} \cdot \langle \boldsymbol{v} f \ln(f)\rangle \leq 0,$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega H[f]\,\mathrm{d}\boldsymbol{x} \leq 0.$$

Equivalently, the total physical entropy $S[f](t) = -\int_\Omega H[f](\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}$ satisfies $\mathrm{d}S[f]/\mathrm{d}t \geq 0$. In other words, solutions of the BGK equation are entropy-producing.

Proof. Similar to the last proof, but using Lemma 5.3, we have

$$\begin{aligned}
\frac{\partial}{\partial t}H[f](t) &= \int_{\mathbb{R}^d} \frac{\partial f}{\partial t}\left(\ln(f)+1\right)\,\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}^d} \frac{\partial f}{\partial t}\ln(f)\,\mathrm{d}\boldsymbol{v} + \frac{\partial n}{\partial t} \\
&= \frac{1}{\tau}\int_{\mathbb{R}^d}(M_f - f)\ln(f)\,\mathrm{d}\boldsymbol{v} - \int_{\mathbb{R}^d}\boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} f\ln(f)\,\mathrm{d}\boldsymbol{v} + \frac{\partial n}{\partial t} \\
&= \frac{1}{\tau}\int_{\mathbb{R}^d}(M_f - f)\ln(f)\mathrm{d}\boldsymbol{v} - \nabla_{\boldsymbol{x}}\cdot\int_{\mathbb{R}^d}\boldsymbol{v} f\ln(f)\,\mathrm{d}\boldsymbol{v} + \int_{\mathbb{R}^d}\boldsymbol{v} f\cdot\nabla_{\boldsymbol{x}}(\ln(f))\,\mathrm{d}\boldsymbol{v} + \frac{\partial n}{\partial t} \\
&= \frac{1}{\tau}\int_{\mathbb{R}^d}\ln(f)\left(M_f - f\right)\,\mathrm{d}\boldsymbol{v} - \nabla_{\boldsymbol{x}}\cdot\int_{\mathbb{R}^d}\boldsymbol{v} f\ln(f)\,\mathrm{d}\boldsymbol{v} + \frac{\partial n}{\partial t} + \nabla_{\boldsymbol{x}}\cdot(n\boldsymbol{u}) \\
&= \frac{1}{\tau}\int_{\mathbb{R}^d}\ln(f)\left(M_f - f\right)\,\mathrm{d}\boldsymbol{v} - \nabla_{\boldsymbol{x}}\cdot\int_{\mathbb{R}^d}\boldsymbol{v} f\ln(f)\,\mathrm{d}\boldsymbol{v}.
\end{aligned}$$

Therefore,

$$\partial_t\int_{\mathbb{R}^d} f\ln(f)\,\mathrm{d}\boldsymbol{v} + \nabla_{\boldsymbol{x}}\cdot\int_{\mathbb{R}^d}\boldsymbol{v} f\ln(f)\,\mathrm{d}\boldsymbol{v} = -\frac{1}{\tau}\int_{\mathbb{R}^d}(f - M_f)(\ln(f) - \ln(M_f))\,\mathrm{d}\boldsymbol{v} \leq 0.$$

In other words,

$$\partial_t\langle f\ln(f)\rangle + \nabla_{\boldsymbol{x}}\cdot\langle\boldsymbol{v} f\ln(f)\rangle \leq 0.$$

Using the Ω-periodicity of f and integrating over Ω, we have

$$\mathrm{d}_t \int_\Omega \int_{\mathbb{R}^d} f \ln(f)\, \mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{x} = -\frac{1}{\tau}\int_\Omega \int_{\mathbb{R}^d} (f - M_f)(\ln(f) - \ln(M_f))\, \mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{x} \leq 0,$$

and the proof is complete. ■

Finally, we have

Theorem 5.10. *Suppose that*

$$f_\star : \Omega \times \mathbb{R}^d \to [0, \infty)$$

is a given Ω-periodic distribution. Let

$$\boldsymbol{\rho}_\star = \left[n_\star, n_\star \boldsymbol{u}_\star^\top, \omega_\star\right]^\top,$$

be the vector comprised of the first three moments of $f_\star$, that is,

$$\boldsymbol{\rho}_\star(\boldsymbol{x}) = \begin{bmatrix} n_\star(\boldsymbol{x}) \\ n_\star(\boldsymbol{x})\boldsymbol{u}_\star(\boldsymbol{x}) \\ \omega_\star(\boldsymbol{x}) \end{bmatrix} = \left\langle f_\star(\boldsymbol{x}, \boldsymbol{v}) \begin{bmatrix} 1 \\ \boldsymbol{v} \\ \frac{1}{2}|\boldsymbol{v}|^2 \end{bmatrix} \right\rangle, \quad \forall\, \boldsymbol{x} \in \Omega.$$

The Ω-periodic temperature, $\theta_\star$, is defined via

$$\omega_\star(\boldsymbol{x}) = \frac{1}{2} n_\star(\boldsymbol{x})|\boldsymbol{u}_\star(\boldsymbol{x})|^2 + \frac{d}{2} n_\star(\boldsymbol{x})\theta_\star(\boldsymbol{x}).$$

The Maxwellian, as usual, is defined using the macroscopic densities as

$$M_\star(\boldsymbol{x}, \boldsymbol{v}) \coloneqq n_\star \left(\frac{1}{2\pi\theta_\star}\right)^{d/2} \exp\left(-\frac{|\boldsymbol{v} - \boldsymbol{u}_\star|^2}{2\theta_\star}\right).$$

Consider the admissible class of distribution functions defined as

$$\mathcal{A}_\star := \left\{ g : \Omega \times \mathbb{R}^d \to \mathbb{R} \;\middle|\; g \text{ is } \Omega\text{-periodic}, \quad g \geq 0, \quad \text{and} \quad \langle \boldsymbol{m} g \rangle = \boldsymbol{\rho}_\star \right\},$$

where, recall,

$$\boldsymbol{m}(\boldsymbol{v}) \coloneqq \left[1, \boldsymbol{v}^\top, \frac{1}{2}|\boldsymbol{v}|^2\right]^\top : \mathbb{R}^d \to \mathbb{R}^{d+2}.$$

Then, $M_\star, f_\star \in \mathcal{A}_\star$. In other words, $\mathcal{A}_\star$ is non-empty. For every $g \in \mathcal{A}_\star$,

$$H[M_\star](\boldsymbol{x}) \leq H[g](\boldsymbol{x}), \quad \forall\, \boldsymbol{x} \in \Omega.$$

In other words, the Maxwellian is the minimum information distribution out of the set of admissible distributions that gives rise to the macroscopic density $\boldsymbol{\rho}_\star$.

Proof. This is the unary specialization of the Maxwellian minimum-information principle proved in Theorem 2.40. The admissible class here fixes the same mass, momentum, and energy moments, so the hypotheses and conclusion of that theorem apply pointwise in $\boldsymbol{x}$. ■

5.2 Discrete-Velocity BGK Models

Discrete velocity BGK models are constructed under the assumption that the velocity space is partitioned into discrete, allowable states. This discretized velocity space could be finite or countably infinite, in principle. But, typically the velocity space is constructed so that it is finite and approximates well the density of particles, f, for a given physical case, meaning that the velocity points might clustered around particular velocity values. To make the discussion simple, we will restrict our attention to the $d = 1$ case. The general case is constructed similarly.

5.2.1 Discrete-Velocity BGK Model in One Space Dimension

Let us make an analogous model in one space dimension ($d = 1$) with discrete, rather than continuous, velocities, along the lines of [54, 55]. Suppose that $h_v \in (0, 1)$ and $J \in \mathbb{N}$ with $J \geq 2$. Set

$$\mathcal{V}_{\mathcal{J}} \coloneqq \left\{ v_{\mathrm{j}} = h_v \cdot \left(\mathrm{j} - \frac{1}{2} \right) \;\middle|\; \mathrm{j} \in \mathcal{J} \right\},$$

where

$$\mathcal{J} \coloneqq \{-J + 1, \ldots, J\}.$$

Note that the simple structure of $\mathcal{V}_{\mathcal{J}}$, with uniformly spaced points arranged symmetrically about 0, is used only for simplicity, and we may choose it in a more general fashion. In any case, set

$$\boldsymbol{m}_{\mathcal{J}}(\mathrm{j}) \coloneqq \left[1, v_{\mathrm{j}}, \frac{1}{2} |v_{\mathrm{j}}|^2 \right]^\top, \quad \forall \mathrm{j} \in \mathcal{J}.$$

Define, for $\boldsymbol{g} : \mathcal{J} \to \mathbb{R}^d$

$$\langle \boldsymbol{g} \rangle_{\mathcal{J}} \coloneqq \sum_{\mathrm{j} \in \mathcal{J}} \boldsymbol{g}_{\mathrm{j}} h_v.$$

Let us define the admissible class of solutions.

Definition 5.11 (Admissible Discrete-Velocity Distributions)**.** *Suppose that* $\Omega = [a, b] \subset \mathbb{R}$ *and*

$$f_{\mathcal{J},\star} : \Omega \times \mathcal{J} \to (0, \infty)$$

is a given Ω*-periodic discrete-velocity distribution. Let*

$$\boldsymbol{\rho}_{\mathcal{J},\star} = \left[n_{\mathcal{J},\star}, n_{\mathcal{J},\star} u_{\mathcal{J},\star}, \omega_{\mathcal{J},\star}\right]^\top,$$

be the vector comprised of the first three discrete-velocity moments of $f_{\mathcal{J},\star}$*, that is,*

$$\boldsymbol{\rho}_{\mathcal{J},\star}(x) = \begin{bmatrix} n_{\mathcal{J},\star}(x) \\ n_{\mathcal{J},\star}(x) u_{\mathcal{J},\star}(x) \\ \omega_{\mathcal{J},\star}(x) \end{bmatrix} = \langle f_{\mathcal{J},\star}(x, \cdot) \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}}, \quad \forall\, x \in \Omega. \tag{5.34}$$

The ***admissible class of discrete-velocity distribution functions****, subordinate to* $\boldsymbol{\rho}_{\mathcal{J},\star}$ *or* $f_{\mathcal{J},\star}$ *is defined as the set*

$$\mathcal{A}_{\mathcal{J},\star} := \left\{ g_{\mathcal{J}} : \Omega \times \mathcal{J} \to \mathbb{R} \;\middle|\; g_{\mathcal{J}} \text{ is } \Omega\text{-periodic}, \quad g_{\mathcal{J}} \geq 0, \quad \text{and} \quad \langle \boldsymbol{m}_{\mathcal{J}} g_{\mathcal{J}} \rangle_{\mathcal{J}} = \boldsymbol{\rho}_{\mathcal{J},\star} \right\}.$$

We need the following technical lemma.

Lemma 5.12. *Suppose that* $J \geq 2$ *and that the moment vector* $\boldsymbol{\rho}_{\mathcal{J},\star}$ *is realized by the strictly positive* Ω*-periodic discrete-velocity density in the preceding definition. Fix* $x \in \Omega = [a, b]$*, and define the function* $K : \mathbb{R}^3 \to \mathbb{R}$ *via*

$$K(\boldsymbol{\beta}) = K_x(\boldsymbol{\beta}) := \langle \exp(\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}) \rangle_{\mathcal{J}} - \boldsymbol{\beta} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x), \quad \forall \boldsymbol{\beta} \in \mathbb{R}^3.$$

$K(\,\cdot\,)$ *is smooth, strictly convex, and coercive and, therefore, has a unique minimizer,* $\boldsymbol{\beta}_\star \in \mathbb{R}^3$*, which satisfies the nonlinear equation*

$$\langle \exp(\boldsymbol{\beta}_\star \cdot \boldsymbol{m}_{\mathcal{J}}) \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} = \boldsymbol{\rho}_{\mathcal{J},\star}(x).$$

Proof. Because K is a finite sum of exponentials of affine functions, it is smooth. Its convexity follows from the first two derivatives. The first derivative is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s} K(\boldsymbol{\beta} + s\boldsymbol{\eta}) &= \frac{\mathrm{d}}{\mathrm{d}s} \langle \exp(\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}) \exp(s\boldsymbol{\eta} \cdot \boldsymbol{m}_{\mathcal{J}}) \rangle_{\mathcal{J}} - \frac{\mathrm{d}}{\mathrm{d}s} (\boldsymbol{\beta} + s\boldsymbol{\eta}) \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x) \\ &= \langle \exp(\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}) \boldsymbol{\eta} \cdot \boldsymbol{m}_{\mathcal{J}} \exp(s\boldsymbol{\eta} \cdot \boldsymbol{m}_{\mathcal{J}}) \rangle_{\mathcal{J}} - \boldsymbol{\eta} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x). \end{aligned}$$

Thus,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}K(\boldsymbol{\beta}+s\boldsymbol{\eta})\bigg|_{s=0} &= \langle\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}} - \boldsymbol{\eta}\cdot\boldsymbol{\rho}_{\mathcal{J},\star}(x)\\ &= \boldsymbol{\eta}\cdot\left\{\langle\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}} - \boldsymbol{\rho}_{\mathcal{J},\star}(x)\right\},\end{aligned}$$

and we can write

$$\nabla K(\boldsymbol{\beta}) = \langle\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}} - \boldsymbol{\rho}_{\mathcal{J},\star}(x).$$

Further,

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}s^2}K(\boldsymbol{\beta}+s\boldsymbol{\eta}) &= \frac{\mathrm{d}}{\mathrm{d}s}\langle\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}}\exp(s\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}}\\ &= \langle\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\left(\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}}\right)^2\exp(s\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}}\\ &= \boldsymbol{\eta}^{\top}\left[\langle\boldsymbol{m}_{\mathcal{J}}\boldsymbol{m}_{\mathcal{J}}^{\top}\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\exp(s\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}}\right]\boldsymbol{\eta}.\end{aligned}$$

Thus,

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}K(\boldsymbol{\beta}+s\boldsymbol{\eta})\bigg|_{s=0} = \boldsymbol{\eta}^{\top}\left[\langle\boldsymbol{m}_{\mathcal{J}}\boldsymbol{m}_{\mathcal{J}}^{\top}\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}}\right]\boldsymbol{\eta},$$

and we write

$$H_K(\boldsymbol{\beta}) = \langle\boldsymbol{m}_{\mathcal{J}}\boldsymbol{m}_{\mathcal{J}}^{\top}\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}},$$

which is the 3×3 Hessian matrix. This matrix is positive definite because

$$\boldsymbol{\eta}^{\top}H_K(\boldsymbol{\beta})\boldsymbol{\eta} = \langle\left(\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}}\right)^2\exp(\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}} > 0,$$

for every $\boldsymbol{\beta}\in\mathbb{R}^3$ and every $\boldsymbol{\eta}\in\mathbb{R}^3_\star := \mathbb{R}^3\backslash\{\mathbf{0}\}$. Indeed, when $J\geq 2$, the quadratic polynomial $\boldsymbol{\eta}\cdot\boldsymbol{m}_{\mathcal{J}}(\mathrm{j})$ cannot vanish at every one of the distinct velocity nodes unless $\boldsymbol{\eta}=\mathbf{0}$. It follows that K is strictly convex.

Now, let us prove that K is coercive. Assume that $\boldsymbol{\beta}\in S^2 := \{\boldsymbol{x}\in\mathbb{R}^3 \mid \|\boldsymbol{x}\|_2 = 1\}$. To be precise, we will show the following: for every $M>0$, there exists a number $R=R(M)>0$, independent of $\boldsymbol{\beta}$, such that if s is any number satisfying $s>R$, then it follows that

$$K(s\boldsymbol{\beta}) > M,$$

where

$$K(s\boldsymbol{\beta}) := \langle\exp(s\boldsymbol{\beta}\cdot\boldsymbol{m}_{\mathcal{J}})\rangle_{\mathcal{J}} - s\boldsymbol{\beta}\cdot\boldsymbol{\rho}_{\mathcal{J},\star}(x).$$

Fix a vector $\boldsymbol{\beta} \in S^2$, and suppose that for this vector, for some $\mathrm{j}_0 \in \mathcal{J}$,

$$\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}_0) > 0.$$

Then,

$$K(s\boldsymbol{\beta}) \geq h_v \exp(s\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}_0)) - s\boldsymbol{\beta} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x) \overset{s\to+\infty}{\longrightarrow} +\infty.$$

In fact, by continuity, there is an $\varepsilon_{\boldsymbol{\beta}} > 0$ and a constant $C_1 > 0$, such that, if $\boldsymbol{\alpha} \in B(\boldsymbol{\beta}, \varepsilon_{\boldsymbol{\beta}})$ then,

$$\boldsymbol{\alpha} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}_0) \geq C_1 > 0,$$

so that

$$K(s\,\boldsymbol{\alpha}) \geq h_v \exp(sC_1) - s\,\boldsymbol{\alpha} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x) \overset{s\to+\infty}{\longrightarrow} +\infty.$$

To sum up, given any $M > 0$, there is an $R_{\boldsymbol{\beta}} = R_{\boldsymbol{\beta}}(M)$, such that, if $s > R_{\boldsymbol{\beta}}$, then

$$K(s\,\boldsymbol{\alpha}) > M,$$

for all $\boldsymbol{\alpha} \in B(\boldsymbol{\beta}, \varepsilon_{\boldsymbol{\beta}})$.

Again, fix a vector $\boldsymbol{\beta} \in S^2$, and suppose that for this vector, for all $\mathrm{j} \in \mathcal{J}$,

$$\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}) \leq 0.$$

Now, we know that

$$\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}) = 0, \quad \forall\, \mathrm{j} \in \mathcal{J} \quad \iff \quad \boldsymbol{\beta} = \mathbf{0},$$

but $\boldsymbol{\beta} \neq \mathbf{0}$, since $\|\boldsymbol{\beta}\|_2 = 1$. Therefore, for some $\mathrm{j}_1 \in \mathcal{J}$,

$$\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}_1) < 0.$$

It follows from (5.34) that

$$\boldsymbol{\beta} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x) < 0.$$

Consequently,

$$K(s\boldsymbol{\beta}) \geq h_v \exp(s\boldsymbol{\beta} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}_1)) - s\boldsymbol{\beta} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x) \overset{s\to+\infty}{\longrightarrow} +\infty.$$

In fact, there is an $\varepsilon_{\boldsymbol{\beta}} > 0$, such that, if $\boldsymbol{\alpha} \in B(\boldsymbol{\beta}, \varepsilon_{\boldsymbol{\beta}})$, then

$$\boldsymbol{\alpha} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j}_1) \leq C_2 < 0,$$

and

$$\boldsymbol{\alpha} \cdot \boldsymbol{\rho}_{\mathcal{J},\star}(x) \leq C_3 < 0.$$

Hence,

$$K(s\,\boldsymbol{\alpha}) \geq -sC_3 \overset{s\to+\infty}{\longrightarrow} +\infty.$$

To sum up this second case, given any $M > 0$, there is an $R_{\boldsymbol{\beta}} = R_{\boldsymbol{\beta}}(M)$, such that, if $s > R_{\boldsymbol{\beta}}$, then

$$K(s\,\boldsymbol{\alpha}) > M,$$

for all $\boldsymbol{\alpha} \in B(\boldsymbol{\beta}, \varepsilon_{\boldsymbol{\beta}})$.

Observe that

$$S^2 \subset \bigcup_{\boldsymbol{\beta}\in S^2} B(\boldsymbol{\beta}, \varepsilon_{\boldsymbol{\beta}}),$$

the object on the right being an open cover. Since S^2 is compact, there is a finite sub-cover. In other words, there are finite points $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N \in S^2$ such that

$$S^2 \subset \bigcup_{i=1}^{N} B(\boldsymbol{\beta}_i, \varepsilon_{\boldsymbol{\beta}_i}).$$

Let $M > 0$ be given. Set

$$R(M) = \max_{i=1}^{N} R_{\boldsymbol{\beta}_i}(M).$$

If $s > R(M)$, it follows that

$$K(s\boldsymbol{\beta}) > M,$$

for all $\boldsymbol{\beta} \in S^2$. Since $R(M)$ is independent of $\boldsymbol{\beta}$, K is coercive.

Since K is smooth, strictly convex, and coercive, it has a unique minimizer, $\boldsymbol{\beta}_\star \in \mathbb{R}^3$ that satisfies

$$\nabla K(\boldsymbol{\beta}_\star) = \langle \exp(\boldsymbol{\beta}_\star \cdot \boldsymbol{m}_{\mathcal{J}})\boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} - \boldsymbol{\rho}_{\mathcal{J},\star}(x) = \mathbf{0}.$$

The proof is complete. ■

Remark 5.13. *The compactness argument is required above because, without it, it is not clear that*

$$R(M) = \sup_{\boldsymbol{\beta}\in S^2} R_{\boldsymbol{\beta}}$$

is finite, which is needed for our definition of coercivity.

Lemma 5.14. *Suppose that* $\Omega = [a, b] \subset \mathbb{R}$ *and*

$$f_{\mathcal{J}} : \Omega \times \mathcal{J} \to [0, \infty)$$

is a given Ω*-periodic discrete-velocity distribution. Let*

$$\boldsymbol{\rho}_{\mathcal{J}} = \left[n_{\mathcal{J}}, n_{\mathcal{J}} u_{\mathcal{J}}, \omega_{\mathcal{J}}\right]^{\top},$$

be the vector comprised of the first three discrete-velocity moments of $f_{\mathcal{J}}$*, that is,*

$$\boldsymbol{\rho}_{\mathcal{J}}(x) = \begin{bmatrix} n_{\mathcal{J}}(x) \\ n_{\mathcal{J}}(x) u_{\mathcal{J}}(x) \\ \omega_{\mathcal{J}}(x) \end{bmatrix} = \langle f_{\mathcal{J}}(x, \cdot) \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}}, \quad \forall\, x \in \Omega. \tag{5.35}$$

Fix $x \in [a, b]$*, and define*

$$\mathcal{M}_{\mathcal{J},f}(x, \mathrm{j}) = \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x) \cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j})), \quad \forall\, \mathrm{j} \in \mathcal{J},$$

where $\boldsymbol{\alpha}_{\mathcal{J},f}(x) \in \mathbb{R}^3$ *is the solution to*

$$\langle \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x) \cdot \boldsymbol{m}_{\mathcal{J}}) \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} = \boldsymbol{\rho}_{\mathcal{J}}(x),$$

which is guaranteed to exist, uniquely, by Lemma 5.12. Then

$$0 = \sum_{\mathrm{j} \in \mathcal{J}} h_v \big(\mathcal{M}_{\mathcal{J},f}(x, \mathrm{j}) - f_{\mathcal{J}}(x, \mathrm{j})\big), \tag{5.36}$$

$$0 = \sum_{\mathrm{j} \in \mathcal{J}} h_v v_{\mathrm{j}} \big(\mathcal{M}_{\mathcal{J},f}(x, \mathrm{j}) - f_{\mathcal{J}}(x, \mathrm{j})\big), \tag{5.37}$$

$$0 = \sum_{\mathrm{j} \in \mathcal{J}} h_v \frac{1}{2} |v_{\mathrm{j}}|^2 \big(\mathcal{M}_{\mathcal{J},f}(x, \mathrm{j}) - f_{\mathcal{J}}(x, \mathrm{j})\big). \tag{5.38}$$

This may be expressed succinctly as

$$\big\langle \boldsymbol{m}_{\mathcal{J}} \big(\mathcal{M}_{\mathcal{J},f}(x, \cdot) - f_{\mathcal{J}}(x, \cdot)\big) \big\rangle_{\mathcal{J}} = \mathbf{0}. \tag{5.39}$$

Furthermore, the function $\ln(\mathcal{M}_{\mathcal{J},f})$ *is a discrete-velocity collision invariant, that is*

$$\langle \ln\left(\mathcal{M}_{\mathcal{J},f}\right) \left(\mathcal{M}_{\mathcal{J},f} - f_{\mathcal{J}}\right) \rangle_{\mathcal{J}} = 0. \tag{5.40}$$

Proof. Identities (5.36) – (5.38) follow immediately from the definition. Since

$$\ln\left(\mathcal{M}_{\mathcal{J},f}(x,\mathrm{j})\right) = \ln\left(\exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x)\cdot\boldsymbol{m}_{\mathcal{J}}(\mathrm{j}))\right) = \boldsymbol{\alpha}_{\mathcal{J},f}(x)\cdot\boldsymbol{m}_{\mathcal{J}}(\mathrm{j}),$$

identity (5.40) also follows. ■

Theorem 5.15. *Define, for all $f : \Omega\times\mathcal{J}\to[0,\infty)$, the discrete-velocity information functional*

$$H_{\mathcal{J}}[f](\,\cdot\,) \coloneqq \langle f\ln f\rangle_{\mathcal{J}} : \Omega\to\mathbb{R}.$$

Let $\mathcal{A}_{\mathcal{J},\star}$ be the admissible class of discrete-velocity distribution functions subordinate to a strictly positive distribution $f_{\mathcal{J},\star}:\Omega\times\mathcal{J}\to(0,\infty)$, so that $\mathcal{A}_{\mathcal{J},\star}$ is non-empty. Then there is a unique function $\mathcal{M}_{\mathcal{J},\star}\in\mathcal{A}_{\mathcal{J},\star}$ such that

$$H_{\mathcal{J}}\left[\mathcal{M}_{\mathcal{J},\star}\right](x) \le H_{\mathcal{J}}[g](x), \tag{5.41}$$

for all $g\in\mathcal{A}_{\mathcal{J},\star}$, for all $x\in\Omega$. Furthermore, at every point $x\in\Omega$, there is a unique vector $\boldsymbol{\alpha}_{\mathcal{J},\star}(x)\in\mathbb{R}^3$ such that

$$\mathcal{M}_{\mathcal{J},\star}(x,\mathrm{j}) = \exp(\boldsymbol{\alpha}_{\mathcal{J},\star}(x)\cdot\boldsymbol{m}_{\mathcal{J}}(\mathrm{j})),\quad \forall\,\mathrm{j}\in\mathcal{J}.$$

Proof. Fix $x\in[a,b]$. Define

$$\mathcal{M}_{\mathcal{J},\star}(x,\mathrm{j}) = \exp(\boldsymbol{\alpha}_{\mathcal{J},\star}(x)\cdot\boldsymbol{m}_{\mathcal{J}}(\mathrm{j})),\quad \forall\,\mathrm{j}\in\mathcal{J},$$

where $\boldsymbol{\alpha}_{\mathcal{J},\star}(x)\in\mathbb{R}^3$ is the solution to

$$\langle\exp(\boldsymbol{\alpha}_{\mathcal{J},\star}(x)\cdot\boldsymbol{m}_{\mathcal{J}})\boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}} = \boldsymbol{\rho}_{\mathcal{J},\star}(x),$$

which is guaranteed to exist, uniquely, by Lemma 5.12. This definition shows that the function $\mathcal{M}_{\mathcal{J},\star}\in\mathcal{A}_{\mathcal{J},\star}$. Thus, to conclude, we need only show that

$$H_{\mathcal{J}}\left[\mathcal{M}_{\mathcal{J},\star}\right](x) \le H_{\mathcal{J}}[g](x),$$

for all $g\in\mathcal{A}_{\mathcal{J},\star}$. The proof follows that of the continuous-velocity case.

Let $g\in\mathcal{A}_{\mathcal{J},\star}$ be arbitrary. We may write $\mathcal{M}_{\mathcal{J},\star}=\mathcal{M}_{\mathcal{J},g}$, since g gives rise to the same macroscopic, discrete-velocity fields as $f_{\mathcal{J},\star}$. Define $h(s)\coloneqq s\ln(s)$. This function h is convex

on $(0,\infty)$, and it follows that

$$\begin{aligned} h(g) &\geq h(\mathcal{M}_{\mathcal{J},g}) + h'(\mathcal{M}_{\mathcal{J},g})(g - \mathcal{M}_{\mathcal{J},g}) \\ &= h(\mathcal{M}_{\mathcal{J},g}) + \ln(\mathcal{M}_{\mathcal{J},g})(g - \mathcal{M}_{\mathcal{J},g}) + g - \mathcal{M}_{\mathcal{J},g}. \end{aligned}$$

By Lemma 5.14,

$$\langle h(g)\rangle_{\mathcal{J}} \geq \langle h(\mathcal{M}_{\mathcal{J},g})\rangle_{\mathcal{J}} + \langle \ln(\mathcal{M}_{\mathcal{J},g})(g - \mathcal{M}_{\mathcal{J},g})\rangle_{\mathcal{J}} + \langle g - \mathcal{M}_{\mathcal{J},g}\rangle_{\mathcal{J}} = \langle h(\mathcal{M}_{\mathcal{J},g})\rangle_{\mathcal{J}}\,.$$

This shows that, point-wise in Ω,

$$H_{\mathcal{J}}[g] \geq H_{\mathcal{J}}[\mathcal{M}_{\mathcal{J},g}] = H_{\mathcal{J}}[\mathcal{M}_{\mathcal{J},\star}],$$

and the result is proven. ■

Definition 5.16 (1x1v Discrete-Velocity BGK Model)**.** *The* 1x1v ***discrete-velocity BGK model*** *is defined as follows: given an* $[a,b]$*-periodic function* $f_0 : [a,b]\times\mathcal{J}\to(0,\infty)$*, find* $f : [a,b]\times\mathcal{J}\times[0,T]\to(0,\infty)$ *such that* $f(\,\cdot\,,\,\cdot\,,0) = f_0$ *and*

$$\partial_t f(x,\mathrm{j},t) + v_{\mathrm{j}}\partial_x f(x,\mathrm{j},t) = \lambda\left(\mathcal{M}_{\mathcal{J},f}(x,\mathrm{j},t) - f(x,\mathrm{j},t)\right), \quad \forall \mathrm{j}\in\mathcal{J}, \quad \forall x\in[a,b], \quad (5.42)$$

where

$$\mathcal{M}_{\mathcal{J},f}(x,\mathrm{j},t) = \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x,t)\cdot\boldsymbol{m}_{\mathcal{J}}(\mathrm{j})), \quad \forall \mathrm{j}\in\mathcal{J},$$

and $\boldsymbol{\alpha}_{\mathcal{J},f}(x,t)\in\mathbb{R}^3$ *is the solution to*

$$\langle \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x,t)\cdot\boldsymbol{m}_{\mathcal{J}})\,\boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}} = \langle f(x,\,\cdot\,,t)\,\boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}}.$$

We have the following result concerning the solutions of the discrete-velocity model.

Theorem 5.17. *Suppose that* $\Omega = [a,b]\subset\mathbb{R}$ *and*

$$f : \Omega\times\mathcal{J}\times[0,\infty)\to[0,\infty)$$

is a given Ω*-periodic solution to the discrete-velocity BGK model. Define the discrete-velocity information and physical entropy densities by*

$$H_{\mathcal{J}}[f](x,t) \coloneqq \langle f(x,\,\cdot\,,t)\ln f(x,\,\cdot\,,t)\rangle_{\mathcal{J}}, \qquad S_{\mathcal{J}}[f](x,t) \coloneqq -H_{\mathcal{J}}[f](x,t).$$

Then, for all $x \in [a, b]$ and $t \geq 0$,

$$\frac{\partial}{\partial t} \langle \boldsymbol{m}_{\mathcal{J}} f(x, \cdot, t) \rangle_{\mathcal{J}} + \frac{\partial}{\partial x} \langle v(\cdot) \boldsymbol{m}_{\mathcal{J}} f(x, \cdot, t) \rangle_{\mathcal{J}} = \mathbf{0}; \tag{5.43}$$

and

$$\frac{\partial}{\partial t} \langle f(x, \cdot, t) \ln(f(x, \cdot, t)) \rangle_{\mathcal{J}} + \frac{\partial}{\partial x} \langle v(\cdot) f(x, \cdot, t) \ln(f(x, \cdot, t)) \rangle_{\mathcal{J}} \leq 0. \tag{5.44}$$

Consequently,

$$\frac{d}{dt} \int_{\Omega} H_{\mathcal{J}}[f](x, t)\, dx \leq 0. \tag{5.45}$$

Equivalently, $\frac{d}{dt} \int_{\Omega} S_{\mathcal{J}}[f](x, t)\, \mathrm{d}x \geq 0$.

Proof. The proofs are analogous to the continuous-velocity cases. Equation (5.43) follows directly from (5.39). To prove inequality (5.44), observe that

$$\begin{aligned}
\frac{\partial}{\partial t} H_{\mathcal{J}}[f] &= \left\langle \frac{\partial f}{\partial t} (\ln(f) + 1) \right\rangle_{\mathcal{J}} \\
&= \left\langle \frac{\partial f}{\partial t} \ln(f) \right\rangle_{\mathcal{J}} + \frac{\partial}{\partial t} \langle f \rangle_{\mathcal{J}} \\
&= \lambda \langle (\mathcal{M}_{\mathcal{J},f} - f) \ln(f) \rangle_{\mathcal{J}} - \langle v \partial_x f \ln(f) \rangle_{\mathcal{J}} - \partial_x \langle v f \rangle_{\mathcal{J}} \\
&= \lambda \langle \ln(f) (\mathcal{M}_{\mathcal{J},f} - f) \rangle_{\mathcal{J}} - \partial_x \langle v f \ln(f) \rangle_{\mathcal{J}} \\
&= \lambda \langle (\ln(f) - \ln(\mathcal{M}_{\mathcal{J},f})) (\mathcal{M}_{\mathcal{J},f} - f) \rangle_{\mathcal{J}} - \partial_x \langle v f \ln(f) \rangle_{\mathcal{J}} \\
&= -\lambda \langle (\ln(\mathcal{M}_{\mathcal{J},f}) - \ln(f)) (\mathcal{M}_{\mathcal{J},f} - f) \rangle_{\mathcal{J}} - \partial_x \langle v f \ln(f) \rangle_{\mathcal{J}} \\
&\leq -\partial_x \langle v f \ln(f) \rangle_{\mathcal{J}},
\end{aligned}$$

where we used Lemma 2.23. Finally, using the Ω-periodicity of f and integrating over Ω, gives (5.45). ■

5.2.2 Space-Homogeneous Discrete-Velocity Model

Similar to the continuous velocity case, we can make the following definition:

Definition 5.18 (Homogeneous Discrete-Velocity BGK Model)**.** *The **space-homogeneous, 1-dimensional discrete-velocity BGK model** is defined as follows: given the function $f_0 : \mathcal{J} \to (0, \infty)$, find $f : \mathcal{J} \times [0, T] \to (0, \infty)$ such that $f(\cdot, 0) = f_0$ and*

$$\partial_t f(\mathrm{j}, t) = \lambda (\mathcal{M}_{\mathcal{J},f}(\mathrm{j}, t) - f(\mathrm{j}, t)), \quad \forall \mathrm{j} \in \mathcal{J}, \tag{5.46}$$

where

$$\mathcal{M}_{\mathcal{J},f}(\mathrm{j},t) = \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(t)\cdot \boldsymbol{m}_{\mathcal{J}}(\mathrm{j})), \quad \forall\, \mathrm{j}\in\mathcal{J}, \tag{5.47}$$

and $\boldsymbol{\alpha}_{\mathcal{J},f}(t)\in\mathbb{R}^3$ *is the solution to*

$$\langle \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(t)\cdot \boldsymbol{m}_{\mathcal{J}})\, \boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}} = \langle f(\,\cdot\,,t)\, \boldsymbol{m}_{\mathcal{J}}\rangle_{\mathcal{J}}. \tag{5.48}$$

Theorem 5.19. *Suppose that*

$$f : \mathcal{J}\times[0,\infty)\to(0,\infty)$$

is a given solution to the space-homogeneous discrete-velocity BGK model. Define the discrete-velocity information and physical entropy by

$$H_{\mathcal{J}}[f](t) \coloneqq \langle f(\,\cdot\,,t)\ln f(\,\cdot\,,t)\rangle_{\mathcal{J}}, \qquad S_{\mathcal{J}}[f](t) \coloneqq -H_{\mathcal{J}}[f](t).$$

Then, for all $t\geq 0$*,*

$$\frac{\partial}{\partial t}\langle \boldsymbol{m}_{\mathcal{J}} f(\,\cdot\,,t)\rangle_{\mathcal{J}} = \mathbf{0}; \tag{5.49}$$

and

$$\frac{\partial}{\partial t}\langle f(\,\cdot\,,t)\ln\left(f(\,\cdot\,,t)\right)\rangle_{\mathcal{J}} \leq 0. \tag{5.50}$$

Equivalently, $\partial_t S_{\mathcal{J}}[f](t)\geq 0$*. Consequently,*

$$\frac{\partial}{\partial t}\mathcal{M}_{\mathcal{J},f}(\mathrm{j},t) = 0, \quad \forall\, \mathrm{j}\in\mathcal{J}. \tag{5.51}$$

Proof. The proofs are again analogous to the continuous-velocity cases. Equation (5.49) follows directly from (5.39). To prove inequality (5.50), observe that

$$\begin{aligned}
\frac{\partial}{\partial t}H_{\mathcal{J}}[f] &= \left\langle \frac{\partial f}{\partial t}\left(\ln(f)+1\right)\right\rangle_{\mathcal{J}} \\
&= \left\langle \frac{\partial f}{\partial t}\ln(f)\right\rangle_{\mathcal{J}} + \frac{\partial}{\partial t}\langle f\rangle_{\mathcal{J}} \\
&= \lambda\,\langle(\mathcal{M}_{\mathcal{J},f}-f)\ln(f)\rangle_{\mathcal{J}} \\
&= \lambda\,\langle \ln(f)\,(\mathcal{M}_{\mathcal{J},f}-f)\rangle_{\mathcal{J}} \\
&= \lambda\,\langle(\ln(f)-\ln(\mathcal{M}_{\mathcal{J},f}))\,(\mathcal{M}_{\mathcal{J},f}-f)\rangle_{\mathcal{J}} \\
&= -\lambda\,\langle(\ln(\mathcal{M}_{\mathcal{J},f})-\ln(f))\,(\mathcal{M}_{\mathcal{J},f}-f)\rangle_{\mathcal{J}} \\
&\leq 0,
\end{aligned}$$

where we used Lemma 2.23. Finally, (5.51) follows from (5.47) and (5.48), using (5.49). ■

Chapter 6

Numerics for the Single-Species BGK Model

In this chapter, we consider the numerical approximation of the BGK model for the $d = 1$ case. Therefore, phase space is two-dimensional, with one dimension for physical space and one for velocity space. It is not necessary to equate the dimensions of velocity and physical space, but this is a common practice, and, for the purpose of describing the numerical methods, one does not lose much generality using such simplifying assumptions. We call the present case the $1x1v$ case. The methods that we describe in this section are scalable, meaning that, as more phase-space dimensions are added, the methods themselves change only slightly. The biggest obstacle for high-dimensional numerical simulation is the added number of degrees of freedom that accompany an increase in phase-space dimensions. Of course, this increase can be significant, since real-world phase space has 6 dimensions in the model.

6.1 Finite Volume Space and Velocity Discretization

Since the Vlasov–BGK equation is a nonlinear conservation-like law, shocks (discontinuities) and rarefaction waves can form and propagate in the solution. Thus, the integro-differential equation does not necessarily hold in the classical (strong) sense. In this case, finite volume methods, which are based on the integral form of the differential equation, are typically more appropriate and simpler to use.

First, we truncate the velocity space so that it is finite in size: $V = [-v_{\max}, v_{\max}]$, where $v_{\max} > 0$. This limits the range of velocities that may be approximated, but, as we will see, does not generally affect the accuracy of approximation as long as $v_{\max}$ is chosen sufficiently large. We note that it is not necessary to make a symmetric truncation about zero velocity,

and sometimes it is not advantageous to do so. This is only done for simplicity of presentation. In similar fashion, let us assume that $\Omega = [-L, L]$, with $L > 0$.

Let us define

$$h_x \coloneqq \frac{2L}{N_x} \quad \text{and} \quad h_v \coloneqq \frac{2v_{\max}}{N_v},$$

and then set

$$x_{\mathrm{i}} \coloneqq -L + (\mathrm{i} - 1/2)h_x \quad \text{and} \quad v_{\mathrm{j}} \coloneqq -v_{\max} + (\mathrm{j} - 1/2)h_v,$$

where i and j take the appropriate integer or half-integer values for cell centers and interfaces. To discretize phase space, we break the rectangular $1x1v$ domain, $\Omega \times V$, into a two-dimensional grid of cells with finite volume, and approximate the cell average of the function. $\Omega \times V \subset \mathbb{R}^2$ can be written as the union of the cells:

$$\Omega \times V = \bigcup_{\mathrm{i,j}=1}^{N_x,N_v} C_{\mathrm{i,j}} = \bigcup_{\mathrm{i,j}=1}^{N_x,N_v} C_{\mathrm{i}}^x \times C_{\mathrm{j}}^v = \bigcup_{\mathrm{i,j}=1}^{N_x,N_v} \left[x_{\mathrm{i}-\frac{1}{2}}, x_{\mathrm{i}+\frac{1}{2}}\right] \times \left[v_{\mathrm{j}-\frac{1}{2}}, v_{\mathrm{j}+\frac{1}{2}}\right], \tag{6.1}$$

with the point $(x_{\mathrm{i}}, v_{\mathrm{j}})$ at the center of cell $C_{\mathrm{i,j}} = C_{\mathrm{i}}^x \times C_{\mathrm{j}}^v$. Next, we define

$$\overline{f}_{\mathrm{i,j}}(t) \coloneqq \overline{f}(x_{\mathrm{i}}, v_{\mathrm{j}}, t) \coloneqq \frac{1}{|C_{\mathrm{i}}^x \times C_{\mathrm{j}}^v|} \int_{C_{\mathrm{i}}^x \times C_{\mathrm{j}}^v} f(x, v, t)\,\mathrm{d}x\mathrm{d}v. \tag{6.2}$$

The integral form of the Vlasov–BGK equation is

$$\partial_t \overline{f}_{\mathrm{i,j}}(t) + \frac{1}{|C_{\mathrm{i,j}}|} \int_{C_{\mathrm{i,j}}} \left(\partial_x(vf) + \partial_v(a(x,t)f)\right)\,\mathrm{d}x\mathrm{d}v = \frac{\lambda}{|C_{\mathrm{i,j}}|} \int_{C_{\mathrm{i,j}}} M_f(x, v, t)\,\mathrm{d}x\mathrm{d}v - \lambda \overline{f}_{\mathrm{i,j}}(t), \tag{6.3}$$

where we have assumed that the acceleration of particles due to the external field, a, is independent of velocity. Let us define the flux function, $\boldsymbol{F}$, via

$$\boldsymbol{F}(x, v, t) \coloneqq [vf(x, v, t), a(x, t)f(x, v, t)]^\top \coloneqq [F(x, v, t), G(x, v, t)]^\top,$$

where

$$F(x, v, t) \coloneqq vf(x, v, t), \quad G(x, v, t) \coloneqq a(x, t)f(x, v, t).$$

Define

$$\overline{M}_{\mathrm{i,j}}(t) \coloneqq \frac{1}{|C_{\mathrm{i,j}}|} \int_{C_{\mathrm{i,j}}} M_f(x, v, t)\,\mathrm{d}x\mathrm{d}v.$$

Applying the Divergence Theorem,

$$\partial_t \overline{f}_{i,j}(t) + \lambda \overline{f}_{i,j}(t) - \lambda \overline{M}_{i,j}(t) = -\frac{1}{h_x h_v} \int_{C_j^v} \left(F(x_{i+1/2}, v, t) - F(x_{i-1/2}, v, t) \right) \, dv - \frac{1}{h_x h_v} \int_{C_i^x} \left(G(x, v_{j+1/2}, t) - G(x, v_{j-1/2}, t) \right) \, dx. \quad (6.4)$$

Now, (6.4) is exact. To gain a practical numerical method, we must make some approximations. Let us first approximate the flux integrals:

$$\frac{1}{h_x h_v} \int_{C_j^v} \left(F(x_{i+1/2}, v, t) - F(x_{i-1/2}, v, t) \right) \, dv \approx \frac{F(x_{i+1/2}, v_j, t) - F(x_{i-1/2}, v_j, t)}{h_x}, \quad (6.5)$$

$$\frac{1}{h_x h_v} \int_{C_i^x} \left(G(x, v_{j+1/2}, t) - G(x, v_{j-1/2}, t) \right) \, dx \approx \frac{G(x_i, v_{j+1/2}, t) - G(x_i, v_{j-1/2}, t)}{h_v}. \quad (6.6)$$

To approximate the density function on the cell $C_{i,j}$, we use piecewise linear reconstructions:

$$p_{i,j}(x, v, t) = \overline{f}_{i,j}(t) + \sigma_{i,j}^x (x - x_i) + \sigma_{i,j}^v (v - v_j),$$

with the θ-minmod slope limiter,

$$\sigma_{i,j}^x = \text{minmod}\left(\left(\frac{\overline{f}_{i+1,j} - \overline{f}_{i-1,j}}{2h_x} \right), \theta^x \left(\frac{\overline{f}_{i,j} - \overline{f}_{i-1,j}}{h_x} \right), \theta^x \left(\frac{\overline{f}_{i+1,j} - \overline{f}_{i,j}}{h_x} \right) \right), \quad (6.7)$$

$$\sigma_{i,j}^v = \text{minmod}\left(\left(\frac{\overline{f}_{i,j+1} - \overline{f}_{i,j-1}}{2h_v} \right), \theta^v \left(\frac{\overline{f}_{i,j} - \overline{f}_{i,j-1}}{h_v} \right), \theta^v \left(\frac{\overline{f}_{i,j+1} - \overline{f}_{i,j}}{h_v} \right) \right), \quad (6.8)$$

where $\theta^x, \theta^v \in [1, 2]$. Note that for three *real* arguments, the minmod function is defined by

$$\text{minmod}(x, y, z) := \begin{cases} \min\{x, y, z\} & \text{if} \quad x, y, z \geq 0 \\ \max\{x, y, z\} & \text{if} \quad x, y, z \leq 0 \\ 0 & \text{otherwise} \end{cases}.$$

The reconstructions make the density approximation multi-valued at the cell edges, a fact that we use to our advantage in the numerical approximation. Using the reconstructions, we define the density at the midpoints of each cell edge as follows:

$$f_{i+1/2,j}^-(t) := p_{i,j}(x_{i+1/2}, v_j, t), \quad f_{i+1/2,j}^+(t) := p_{i+1,j}(x_{i+1/2}, v_j, t),$$

and, likewise,

$$f^-_{i,j+1/2}(t) \coloneqq p_{i,j}(x_i, v_{j+1/2}, t), \quad f^+_{i,j+1/2}(t) \coloneqq p_{i,j+1}(x_i, v_{j+1/2}, t).$$

Next, the exact fluxes are replaced by numerical fluxes of the form

$$F(x_{i+1/2}, v_j, t) \approx \widetilde{F}(f^+_{i+1/2,j}(t), f^-_{i+1/2,j}(t)) =: \widehat{F}_{i+1/2,j}(t)$$

and

$$G(x_i, v_{j+1/2}, t) \approx \widetilde{G}(f^+_{i,j+1/2}(t), f^-_{i,j+1/2}(t)) =: \widehat{G}_{i,j+1/2}(t).$$

We use a simple upwind strategy to construct the numerical fluxes:

$$\widehat{F}_{i+1/2,j}(t) = \widetilde{F}(f^+_{i+1/2,j}(t), f^-_{i+1/2,j}(t)) = \begin{cases} v_j f^-_{i+1/2,j}(t) & \text{if} \quad v_j \geq 0 \\ v_j f^+_{i+1/2,j}(t) & \text{if} \quad v_j < 0 \end{cases}, \tag{6.9}$$

$$\widehat{G}_{i,j+1/2}(t) = \widetilde{G}(f^+_{i,j+1/2}(t), f^-_{i,j+1/2}(t)) = \begin{cases} a(x_i, t) f^-_{i,j+1/2}(t) & \text{if} \quad a(x_i, t) \geq 0 \\ a(x_i, t) f^+_{i,j+1/2}(t) & \text{if} \quad a(x_i, t) < 0 \end{cases}. \tag{6.10}$$

We can write the approximation scheme to this point as follows:

$$\begin{aligned} \partial_t \overline{f}_{i,j}(t) + \lambda \overline{f}_{i,j}(t) - \lambda \overline{M}_{i,j}(t) = &- \frac{\widehat{F}_{i+1/2,j}(t) - \widehat{F}_{i-1/2,j}(t)}{h_x} \\ &- \frac{\widehat{G}_{i,j+1/2}(t) - \widehat{G}_{i,j-1/2}(t)}{h_v} + \widetilde{E}_{i,j}(t), \end{aligned} \tag{6.11}$$

where $\widetilde{E}_{i,j}(t)$ is a local truncation (approximation) error.

Finally, to complete the spatial discretization of the Vlasov–BGK equation, we need to approximate the Maxwellian. For this we use the following:

$$n_i(t) = h_v \sum_{j=1}^{N_v} \overline{f}_{i,j}(t), \tag{6.12}$$

$$u_i(t) = \frac{h_v}{n_i(t)} \sum_{j=1}^{N_v} v_j \overline{f}_{i,j}(t), \tag{6.13}$$

$$\theta_i(t) = \frac{h_v}{n_i(t)} \sum_{j=1}^{N_v} |v_j|^2 \overline{f}_{i,j}(t) - |u_i(t)|^2, \tag{6.14}$$

$$\overline{M}_{i,j}(t) \approx \widehat{M}_{i,j}(t) = \frac{n_i(t)}{(2\pi\theta_i(t))^{1/2}} \exp\left(-\frac{(v_j - u_i(t))^2}{2\theta_i(t)} \right). \tag{6.15}$$

Since we are replacing integrations by midpoint quadratures, we lose the collision invariances that we enjoyed at the continuum level. But, for the purpose of approximation, we will continue to assume that these invariances still exist at the discrete level.

We conclude this section by stating the spatially discrete approximation to the Vlasov–BGK equation:

$$\mathrm{d}_t \overline{f}_{\mathrm{i,j}}(t) + \lambda \overline{f}_{\mathrm{i,j}}(t) - \lambda \widehat{M}_{\mathrm{i,j}}(t) = -\frac{\widehat{F}_{\mathrm{i+1/2,j}}(t) - \widehat{F}_{\mathrm{i-1/2,j}}(t)}{h_x} - \frac{\widehat{G}_{\mathrm{i,j+1/2}}(t) - \widehat{G}_{\mathrm{i,j-1/2}}(t)}{h_v} + \widehat{E}_{\mathrm{i,j}}(t), \tag{6.16}$$

where $\widehat{E}_{\mathrm{i,j}}(t)$ is a local truncation (approximation) error.

6.2 Implicit-Explicit Runge–Kutta Time Stepping

In this section we introduce an implicit-explicit (IMEX) Runge–Kutta (RK) method for integration in time that was proposed in [61]. The convection part is treated using an explicit method, while the collision part is solved using a diagonally implicit method. We will work with the phase-space continuous problem first, in order to take advantage of the collision invariants, before moving to the phase-space discrete problem. Let us rewrite the original Vlasov–BGK equation as

$$\frac{\partial}{\partial t} f(x, v, t) = \mathcal{T}[f](x, v, t) + Q[f](x, v, t)$$

where $\mathcal{T}$ represents the transport term, and Q represents the collision term:

$$\mathcal{T}[f](x, v, t) \coloneqq -v\partial_x f(x, v, t) - a(x, t)\partial_v f(x, v, t), \tag{6.17}$$

$$Q[f](x, v, t) \coloneqq \lambda(M_f(x, v, t) - f(x, v, t)). \tag{6.18}$$

Remark 6.1 (BGK Collision Notation)**.** *When no confusion with the bilinear Boltzmann collision operator can arise, we use $Q[f]$ as shorthand for the BGK collision approximation $Q^{\mathrm{BGK}}[f] = \lambda(M_f - f)$. Thus, $Q[f]$ denotes a generally nonlinear unary operator, whereas $Q[f, f]$ denotes the bilinear Boltzmann collision operator.*

The general ν-stage IMEX Runge–Kutta scheme (diagonally implicit) is one of the form

$$y^{(s)} = y^{\mathfrak{n}} + \Delta t \sum_{r=1}^{s-1} \widetilde{a}_{s,r} \mathcal{T}[y^{(r)}] + \Delta t \sum_{r=1}^{s} a_{s,r} Q[y^{(r)}], \quad s \in \{1, \dots, \nu\}, \tag{6.19}$$

$$y^{\mathfrak{n}+1} = y^{\mathfrak{n}} + \Delta t \sum_{s=1}^{\nu} \widetilde{b}_s \mathcal{T}[y^{(s)}] + \Delta t \sum_{s=1}^{\nu} b_s Q[y^{(s)}], \tag{6.20}$$

where $\widetilde{a}_{s,r}$, $\widetilde{b}_s$, $a_{s,r}$, b_s are taken from the following Butcher tables, respectively:

$$\begin{array}{c|c} \widetilde{\boldsymbol{c}} & \widetilde{\mathsf{A}} \\ \hline & \widetilde{\boldsymbol{b}}^{\top} \end{array} \qquad\qquad \begin{array}{c|c} \boldsymbol{c} & \mathsf{A} \\ \hline & \boldsymbol{b}^{\top} \end{array}. \tag{6.21}$$

The matrix $\widetilde{\mathsf{A}}$, for the explicit part, is strictly lower triangular, and A is lower triangular.

To run the algorithm, for each stage, we first calculate

$$\widehat{y}^{(s-1)} := y^{\mathfrak{n}} + \Delta t \sum_{r=1}^{s-1} \widetilde{a}_{s,r} \mathcal{T}[y^{(r)}] + \Delta t \sum_{r=1}^{s-1} a_{s,r} Q[y^{(r)}]. \tag{6.22}$$

We then rearrange the terms in the final sum to find an expression for $y^{(s)}$:

$$\begin{aligned} y^{(s)} &= \widehat{y}^{(s-1)} + \Delta t a_{s,s} Q[y^{(s)}] \\ &= \widehat{y}^{(s-1)} + \Delta t a_{s,s} \lambda (M_{y^{(s)}} - y^{(s)}) && (6.23) \\ \iff (1 + \Delta t a_{s,s} \lambda) y^{(s)} &= \widehat{y}^{(s-1)} + \Delta t a_{s,s} \lambda M_{y^{(s)}} && (6.24) \\ \iff y^{(s)} &= \frac{1}{1 + \Delta t a_{s,s} \lambda} \widehat{y}^{(s-1)} + \frac{\Delta t a_{s,s} \lambda}{1 + \Delta t a_{s,s} \lambda} M_{y^{(s)}}. && (6.25) \end{aligned}$$

The Maxwellian $M_{y^{(s)}}$ appears to depend implicitly on the current stage. The collision invariance in Lemma 5.1 removes that dependence. In particular,

$$\begin{aligned} \int_{\mathbb{R}^d} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2 \end{pmatrix} y^{(s)} \, \mathrm{d}\boldsymbol{v} &= \int_{\mathbb{R}^d} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2 \end{pmatrix} \widehat{y}^{(s-1)} \, \mathrm{d}\boldsymbol{v} + \Delta t a_{s,s} \lambda \int_{\mathbb{R}^d} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2 \end{pmatrix} (M_{y^{(s)}} - y^{(s)}) \, \mathrm{d}\boldsymbol{v} \\ &= \int_{\mathbb{R}^d} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2 \end{pmatrix} \widehat{y}^{(s-1)} \, \mathrm{d}\boldsymbol{v}. \end{aligned} \tag{6.26}$$

Since the first, second, and third moments of $\widehat{y}^{(s-1)}$ are equal to those of $y^{(s)}$, the two

Maxwellians at these stages are equal:

$$M_{y^{(s)}} = M_{\widehat{y}^{(s-1)}}.$$

Therefore, we have the explicit update formula

$$y^{(s)} = \frac{1}{1+\Delta t a_{s,s}\lambda}\widehat{y}^{(s-1)} + \frac{\Delta t a_{s,s}\lambda}{1+\Delta t a_{s,s}\lambda} M_{\widehat{y}^{(s-1)}}. \tag{6.27}$$

With the above expression for $y^{(s)}$, a convex combination of $\widehat{y}^{(s-1)}$ and $M_{\widehat{y}^{(s-1)}}$, the IMEX-RK scheme can be completed without any complicated inversions.

6.3 Fully Discrete Scheme

When we discretize velocity space, we lose the collision invariances at the discrete level; this is because velocity integrals are replaced with midpoint rule quadrature, in addition to the fact that the velocity space is truncated. It follows that (6.27) breaks down. Nevertheless, we will assume that the phase-space discrete analog to (6.27) holds so that stages may be updated in an explicit fashion.

Suppose that $f_{\mathrm{i,j}}^{\mathfrak{n}}$ is an approximation of the cell average of the density field over cell $C_{\mathrm{i,j}}$ at time $t^{\mathfrak{n}} := t^{\mathfrak{n}-1} + \Delta t$. Similarly, we denote by $f_{\mathrm{i,j}}^{(\ell)}$ the ℓ^{th} stage of the IMEX-RK scheme with respect to $f_{\mathrm{i,j}}^{\mathfrak{n}}$. Define

$$\mathcal{T}_{\mathrm{i,j}}^{\square} := -\frac{\widehat{F}_{\mathrm{i+1/2,j}}^{\square} - \widehat{F}_{\mathrm{i-1/2,j}}^{\square}}{h_x} - \frac{\widehat{G}_{\mathrm{i,j+1/2}}^{\square} - \widehat{G}_{\mathrm{i,j-1/2}}^{\square}}{h_v}, \tag{6.28}$$

$$Q_{\mathrm{i,j}}^{\square} := \lambda(M_{\mathrm{i,j}}^{\square} - f_{\mathrm{i,j}}^{\square}), \tag{6.29}$$

where the fluxes $\widehat{F}_{\mathrm{i+1/2,j}}^{\square}$ and $\widehat{G}_{\mathrm{i,j+1/2}}^{\square}$ are computed with respect to the cell-centered approximation $f_{\mathrm{i,j}}^{\square}$ and $\square$ is a stage iteration or a time step index. The fully discrete Maxwellian is computed via

$$n_{\mathrm{i}}^{\square} = h_v \sum_{\mathrm{j}=1}^{N_v} f_{\mathrm{i,j}}^{\square}, \tag{6.30}$$

$$u_{\mathrm{i}}^{\square} = \frac{h_v}{n_{\mathrm{i}}^{\square}} \sum_{\mathrm{j}=1}^{N_v} v_{\mathrm{j}} f_{\mathrm{i,j}}^{\square}, \tag{6.31}$$

$$\theta_{\mathrm{i}}^{\square} = \frac{h_v}{n_{\mathrm{i}}^{\square}} \sum_{\mathrm{j}=1}^{N_v} |v_{\mathrm{j}}|^2 f_{\mathrm{i,j}}^{\square} - |u_{\mathrm{i}}^{\square}|^2, \tag{6.32}$$

$$M_{i,j}^{\square} = \frac{n_i^{\square}}{(2\pi\theta_i^{\square})^{1/2}} \exp\left(-\frac{(v_j - u_i^{\square})^2}{2\theta_i^{\square}}\right). \tag{6.33}$$

Then we compute

$$\widehat{f}_{i,j}^{(s-1)} := f_{i,j}^{n} + \Delta t \sum_{r=1}^{s-1} \widetilde{a}_{s,r} \mathcal{T}_{i,j}^{(r)} + \Delta t \sum_{r=1}^{s-1} a_{s,r} Q_{i,j}^{(r)}, \tag{6.34}$$

$$f_{i,j}^{(s)} = \frac{1}{1 + \Delta t a_{s,s} \lambda} \widehat{f}_{i,j}^{(s-1)} + \frac{\Delta t a_{s,s} \lambda}{1 + \Delta t a_{s,s} \lambda} \widehat{M}_{i,j}^{(s-1)}, \tag{6.35}$$

$$f_{i,j}^{n+1} = f_{i,j}^{n} + \Delta t \sum_{s=1}^{\nu} \widetilde{b}_s \mathcal{T}_{i,j}^{(s)} + \Delta t \sum_{s=1}^{\nu} b_s Q_{i,j}^{(s)}, \tag{6.36}$$

where $\widehat{M}_{i,j}^{(s-1)}$ is the discrete Maxwellian computed with respect to the fully discrete approximation $\widehat{f}_{i,j}^{(s-1)}$. This completes the description of the fully discrete scheme.

6.4 Poisson Solver for Vlasov–Poisson–BGK Equation

For the Vlasov–Poisson–BGK equation the acceleration a is determined by an electric field as follows:

$$a(x, t) = -\chi \partial_x \Phi,$$

where $\chi > 0$ is a physical constant and Φ is the electric potential determined via

$$-\partial_{xx} \Phi(x, t) = n(x, t),$$

subject to appropriate boundary conditions.

First let us consider *Dirichlet boundary conditions*:

$$\Phi(x_{1/2} = -L) = \alpha, \qquad \Phi(x_{N_x+1/2} = L) = \beta. \tag{6.37}$$

Using the standard three-point stencil approximation, we have, for $i = 2, \ldots, N_x - 1$,

$$-\frac{1}{h_x^2} \left[\Phi_{i+1} - 2\Phi_i + \Phi_{i-1}\right] = n_i. \tag{6.38}$$

The boundary conditions are specified at the peripheral edges of the domain, $-L = x_{1/2}$ and $L = x_{N_x+1/2}$. As is standard, we use the average of the ghost cell and neighboring interior

cell. That is, we use the approximations

$$\frac{\Phi_0 + \Phi_1}{2} = \alpha, \quad \implies \Phi_0 = 2\alpha - \Phi_1, \tag{6.39}$$

and we cancel the term Φ_0 in the stencil approximation. The first equation in the approximation thus becomes

$$-\Phi_2 + 3\Phi_1 = h_x^2 n_1 + 2\alpha. \tag{6.40}$$

Using a similar procedure for right boundary condition, we have

$$\Phi_{N_x+1} = 2\beta - \Phi_{N_x}, \implies -\Phi_{N_x-1} + 3\Phi_{N_x} = h_x^2 n_{N_x} + 2\beta. \tag{6.41}$$

Putting this all into a matrix system, we must solve

$$\begin{pmatrix} 3 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 3 \end{pmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_{N_x-1} \\ \Phi_{N_x} \end{pmatrix} = \begin{pmatrix} h_x^2 n_1 + 2\alpha \\ h_x^2 n_2 \\ \vdots \\ h_x^2 n_{N_x-1} \\ h_x^2 n_{N_x} + 2\beta \end{pmatrix}. \tag{6.42}$$

This system is symmetric positive definite (SPD) and always has a unique solution.

Next, we turn our attention to the case of *Neumann boundary conditions*:

$$\frac{d\Phi}{dx}(-L) = \alpha, \qquad \frac{d\Phi}{dx}(L) = \beta. \tag{6.43}$$

In this case, using a second-order approximation at the boundary, we must solve matrix system

$$\begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_{N_x-1} \\ \Phi_{N_x} \end{pmatrix} = \begin{pmatrix} h_x^2 n_1 - h_x \alpha \\ h_x^2 n_2 \\ \vdots \\ h_x^2 n_{N_x-1} \\ h_x^2 n_{N_x} + h_x \beta \end{pmatrix}. \tag{6.44}$$

This is a symmetric positive semi-definite (SPSD) system. It has a unique mean-zero solution,

that is, a solution satisfying

$$h_x \sum_{i=1}^{N_x} \Phi_i = 0, \qquad \text{if and only if} \qquad h_x \sum_{i=1}^{N_x} n_i = \alpha - \beta.$$

This is the discrete analog of the standard continuous compatibility conditions for unique solvability.

Given $f_{i,j}^{\square}$, we compute the accompanying macroscopic density,

$$n_i^{\square} = h_v \sum_{j=1}^{N_v} f_{i,j}^{\square},$$

and then the associated electric potential $\Phi_i^{\square}$, as described above. Once the discrete potential is available, the acceleration may be approximated via

$$a_{i+1/2}^{\square} = -\chi \frac{\Phi_{i+1}^{\square} - \Phi_i^{\square}}{h_x}.$$

6.5 Sample Computations and Accuracy Tests

In this section, we report on several numerical tests showing accuracy of the numerical implementation.

6.5.1 Relaxation Test 1

This first test is designed to confirm that the BGK operator is calculated correctly for the single-species case. We consider, in particular, the space-homogeneous case: the IPDE becomes the following IODE:

$$\begin{cases} \frac{df}{dt} = \lambda(M_f - f), & (x, v, t) \in \Omega \times V \times [0, \infty), \\ f(x, v, 0) = f_0(x, v), & (x, v, t) \in \Omega \times V \times \{t = 0\}. \end{cases} \tag{6.45}$$

Recall that as shown in Section 5.1.2, the exact solution is

$$f(x, v, t) = e^{-\lambda t} f(x, v, 0) + \left(1 - e^{-\lambda t}\right) M_f(x, v, 0). \tag{6.46}$$

For the first test, we chose a function of the form $f_0(x, v) = b(v)g(x)$, where $b(v)$ is a compactly supported C^∞ function, with nonzero values on $v \in (-2, 2)$. In particular, we use

the function $g(x) = e^{-|x|}$. Consider the function

$$f(x, v, 0) = \begin{cases} e^{-|x|} \frac{5\left(v^2+4\right) \exp\left(\frac{5v}{v^2-4}\right)}{(v^2-4)^2\left(1+\exp\left(\frac{5v}{v^2-4}\right)\right)^2}, & \text{if } v \in (-2, 2), \\ 0, & \text{otherwise.} \end{cases} \tag{6.47}$$

Proposition 6.2 (Density and Velocity of the Relaxation-Test Data)**.** *For the initial distribution in Equation* (6.47)*, with symmetric velocity domain* $V = [-2, 2]$*, the number density and bulk velocity are*

$$n(x, 0) = e^{-|x|}, \qquad u(x, 0) = 0.$$

Proof. Write $f_0(x, v) = e^{-|x|} b(v)$ and set $z(v) = 5v/(v^2 - 4)$. On $(-2, 2)$,

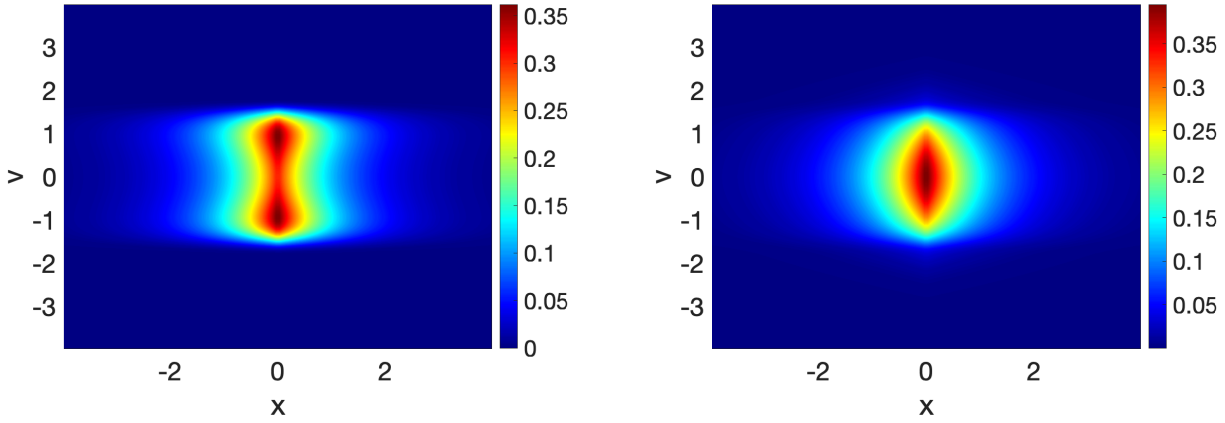
$$b(v) = \frac{\mathrm{d}}{\mathrm{d}v} \left(\frac{1}{1 + e^{z(v)}} \right).$$

The expression in parentheses tends to 0 as $v \to -2^+$ and to 1 as $v \to 2^-$, so $\int_{-2}^{2} b(v)\,\mathrm{d}v = 1$. Hence $n(x, 0) = e^{-|x|}$. Moreover, $z(-v) = -z(v)$ and a direct substitution shows that $b(-v) = b(v)$. Thus $vb(v)$ is odd, and integration over the symmetric interval gives $u(x, 0) = n(x, 0)^{-1} \int_V v f_0(x, v)\,\mathrm{d}v = 0$. ■

The remaining energy integral is evaluated numerically:

$$\omega = \frac{1}{2} \int_V v^2 f\,\mathrm{d}v \approx 0.3713094964845 e^{-|x|}.$$

The code is run with $\lambda = 1$ up to a final time $t_f = 1$. The error is calculated by taking the difference of the true solution minus the computed solution. The initial distribution is given in Figure 6.1a, and the computational solution is given in Figure 6.1b. The error in Figure 6.2 is calculated by taking the difference of the true solution minus the computed solution. The errors are on the order of 10^{-5} when a mesh of size $N_x = N_v = 128$ is used.



(a) Initial conditions. **(b)** Numerical solution at final time $t_f = 1$.

Figure 6.1: Relaxation Test 1 for the space-homogeneous BGK model (5.24): initial data (6.47) (left) and the numerical solution at $t_f = 1$ (right). Here $\lambda = 1$ and $N_x = N_v = 128$.

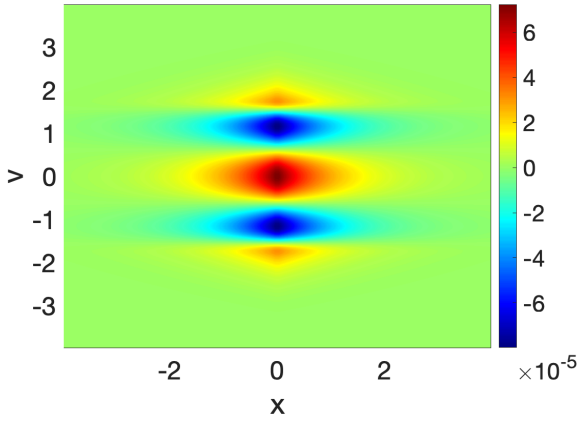


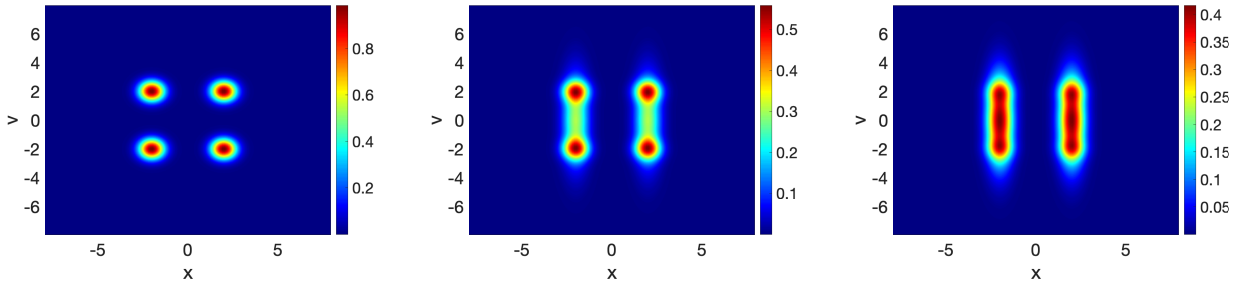
Figure 6.2: Relaxation Test 1: signed pointwise error $f_{\text{exact}} - f_{\text{num}}$ at $t_f = 1$ for $\lambda = 1$ and $N_x = N_v = 128$. The displayed error has magnitude $\mathcal{O}(10^{-5})$.

6.5.2 Relaxation Test 2

For the second test, we set $f_0(x, v)$ to be the sum of four Gaussians, centered at $(x, v) \in \{(\pm 2, \pm 2)\}$:

$$f_0(x, v) = e^{-2(x-2)^2-2(v-2)^2} + e^{-2(x-2)^2-2(v+2)^2} + e^{-2(x+2)^2-2(v-2)^2} + e^{-2(x+2)^2-2(v+2)^2}. \quad (6.48)$$

See Figure 6.3a. The trapezoidal rule is used for the initial moments. The code is run with $\lambda = 1$ up to time $t_f = 2$. The error is calculated by taking the difference of the true solution minus the computed solution. The initial and final computed solutions are given in Figure 6.3, and the errors are shown in Figure 6.4.

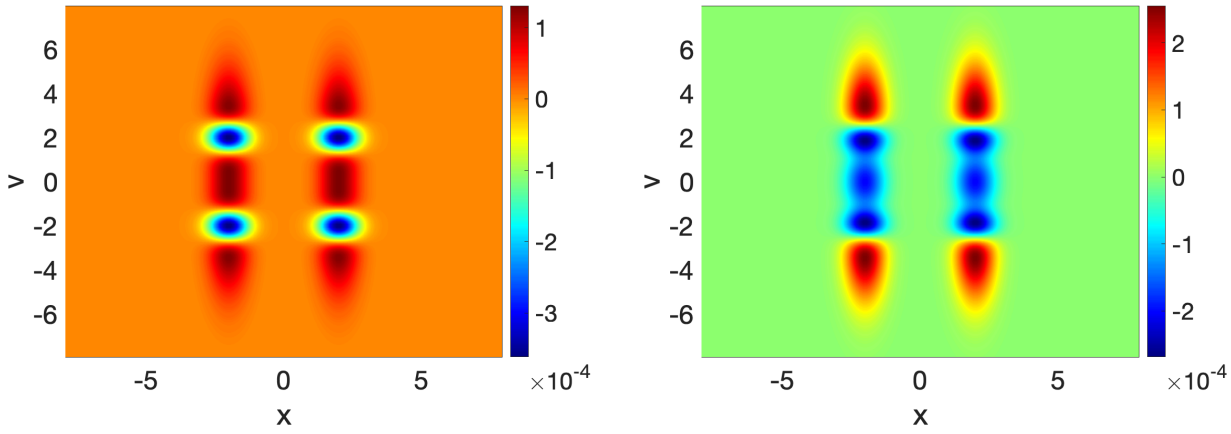


(a) Initial conditions.

(b) Numerical solution at final time $t = 1$.

(c) Numerical solution at final time $t_f = 2$.

Figure 6.3: Relaxation Test 2 for the space-homogeneous BGK model (5.24): four-Gaussian initial data (6.48) and numerical solutions at $t = 1$ and $t = 2$. Here $\lambda = 1$, and the initial moments are evaluated with the trapezoidal rule.



(a) Errors at $t = 1$.

(b) Errors at final time $t_f = 2$.

Figure 6.4: Relaxation Test 2: signed pointwise errors $f_{\text{exact}} - f_{\text{num}}$ at $t = 1$ and $t = 2$ for $\lambda = 1$ and the four-Gaussian initial data (6.48).

6.5.3 Sod Shock Tube (Euler Equation Limit of BGK)

The Sod shock tube test is a standard test. Formally, in the collision limit as $\lambda \to \infty$, the BGK equation is asymptotically equivalent to the Euler Equations. (That is, f converges to the Maxwellian M_f, and the moments $n, \boldsymbol{u}, \theta$ follow the Euler equations.) Thus, we may test the ability of the code to solve the Sod shock tube problem, by letting λ be large (or equivalently, letting $\tau = \lambda^{-1}$ be small), and setting the initial conditions appropriately. For the test, we set $\lambda = 10^4$ ($\tau = 10^{-4}$).

The $1x1v$ BGK equation is

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \lambda(M_f - f). \tag{6.49}$$

To test our IMEX code, we compared the number density ($n = \int f \, \mathrm{d}v$), bulk velocity ($u = n^{-1}\int vf \, \mathrm{d}v$), pressure ($p = n\theta$), and specific internal energy ($e_{\text{sp}} = e/n = \theta/2$) with the theoretical solution derived from Toro [77]. We set the phase-space domain to $\Omega \times V = [-0.75, 0.75] \times [-10, 10]$. Because the exact moments are defined on $\mathbb{R}$, the truncated velocity interval $V = [-10, 10]$ must capture the distribution tails accurately. In particular, we require

$$\int_{\mathbb{R}} f(x, v, t) \, \mathrm{d}v \approx \int_{V} f(x, v, t) \, \mathrm{d}v,$$

where V is the truncated velocity domain. Using $[-10, 10]$ gives a reasonable approximation, as we show.

The setup for the Sod problem is a contact discontinuity separating gases of differing density and temperature, and zero velocity. Thus, for the Sod problem, the initial condition for the particle density is a piecewise Maxwellian with the following values:

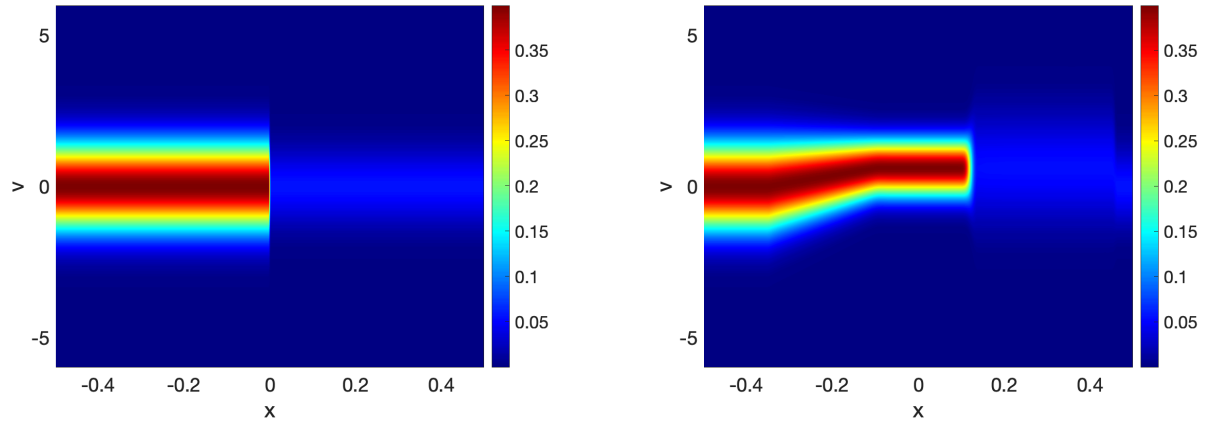
$$\begin{pmatrix} n_L \\ u_L \\ \theta_L \end{pmatrix} = \begin{pmatrix} 1.0 \\ 0.0 \\ 1.0 \end{pmatrix}, \quad x \in (-0.75, 0), \qquad \begin{pmatrix} n_R \\ u_R \\ \theta_R \end{pmatrix} = \begin{pmatrix} 0.125 \\ 0.0 \\ 0.8 \end{pmatrix}, \quad x \in (0, 0.75). \tag{6.50}$$

That is, the initial condition function (contour plot shown in Figure 6.5a) is given by

$$f_0(x, v) = \begin{cases} \frac{1.0}{\sqrt{2\pi(1.0)}} \exp\left(-\frac{|v-0.0|^2}{2(1.0)}\right), & x \in [-0.75, 0], v \in [-10, 10], \\ \frac{0.125}{\sqrt{2\pi(0.8)}} \exp\left(-\frac{|v-0.0|^2}{2(0.8)}\right), & x \in (0, 0.75], v \in [-10, 10]. \end{cases} \tag{6.51}$$

Figure 6.6 shows the moments of the numerical/computed solution at time $t_{\text{f}} = 0.20$. The profiles of the macroscopic quantities follow the correct values (as computed using

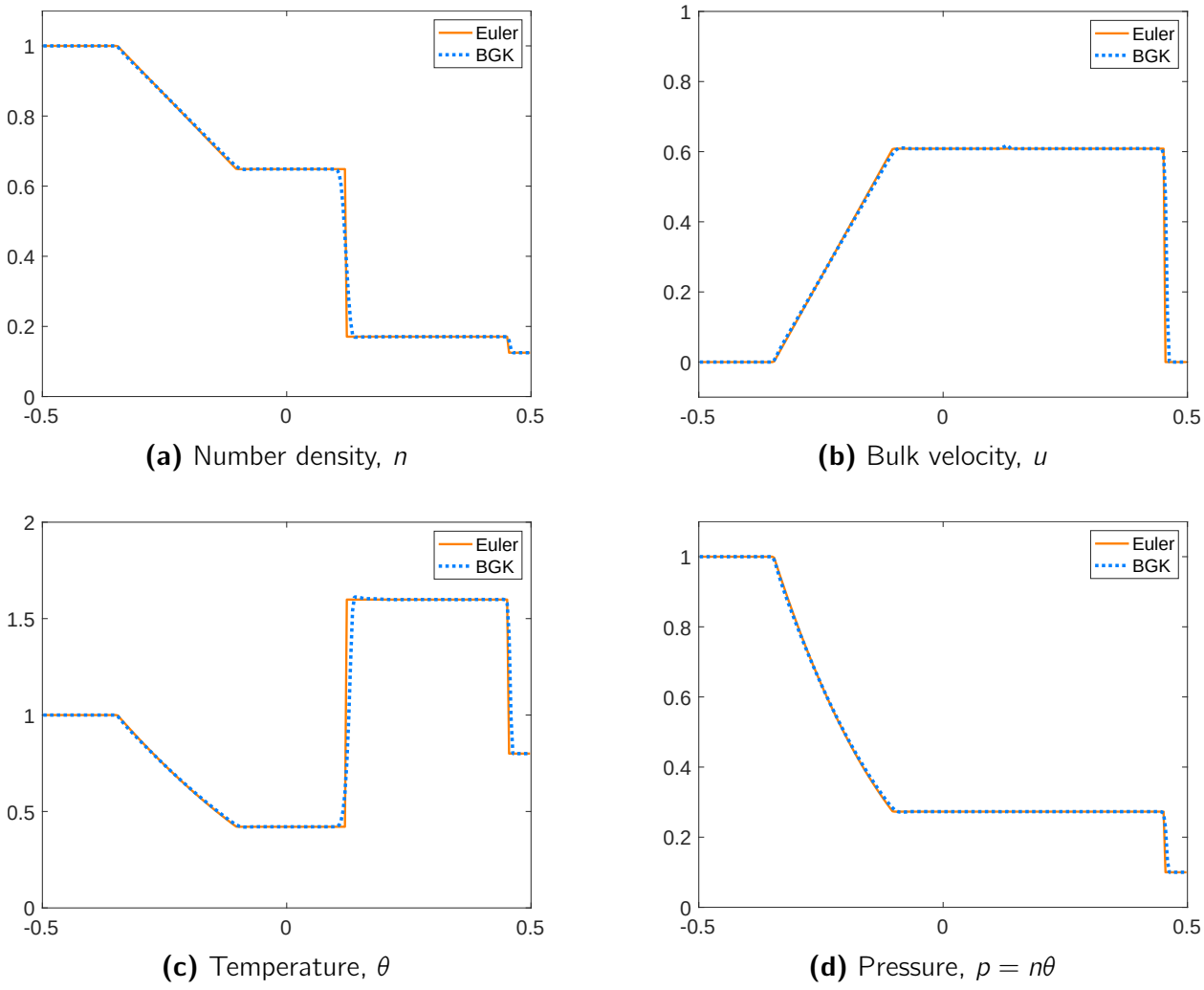
[77] for reference). The final distribution solution is shown in Figure 6.5b. Appendix F contains the prototype associated with this Sod workflow. The listing uses the same $1x1v$ BGK/IMEX method and Euler reference calculation, but its currently archived spatial interval differs from the interval recorded for these figures; the appendix therefore documents an associated implementation rather than claiming exact figure reproduction.



(a) Initial conditions for Sod test.

(b) Contour plot of final numerical solution.

Figure 6.5: Sod shock tube: initial distribution and computed phase-space distribution at $t_f = 0.2$ for the BGK model with $\tau = 10^{-4}$ on $\Omega \times V = [-0.75, 0.75] \times [-10, 10]$.



(a) Number density, n

(b) Bulk velocity, u

(c) Temperature, θ

(d) Pressure, $p = n\theta$

Figure 6.6: Sod shock tube at $t_f = 0.2$ with $\tau = 10^{-4}$: BGK moments (number density, bulk velocity, temperature, and pressure) compared with the exact Euler Riemann solution. The agreement tests the highly collisional Euler limit; a norm-based mesh study is reserved for the code-verification pass described in Table 6.1.

6.5.4 Square Pulse Rotation

Consider the $1x1v$ Vlasov equation:

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + a\frac{\partial f}{\partial v} = 0 \tag{6.52}$$

Setting the acceleration term to $a = -x$ results in a system where an initial distribution is advected counterclockwise, along circular characteristics around the origin. For constructing the characteristics curves, we have the system of ODEs:

$$\frac{dt}{ds} = 1 \tag{6.53}$$
$$\frac{dx}{ds} = v \tag{6.54}$$
$$\frac{dv}{ds} = -x. \tag{6.55}$$

The first equation gives $t = s$, and the remaining two form a system of two coupled ODEs:

$$\dot{x} = v \tag{6.56}$$
$$\dot{v} = -x. \tag{6.57}$$

Taking the derivative of the first equation, and plugging in the second equation, we have

$$\ddot{x} = \dot{v} = -x \iff \ddot{x} + x = 0. \tag{6.58}$$

The solution of this equation is

$$x(t) = c_1\cos(t) + c_2\sin(t). \tag{6.59}$$

Thus,

$$\dot{v} = -x = -c_1\cos(t) - c_2\sin(t). \tag{6.60}$$

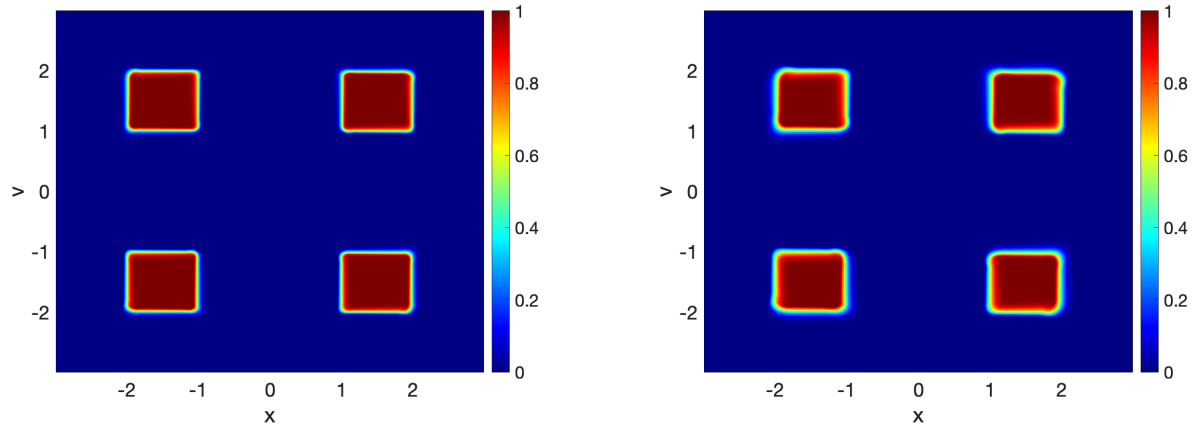
Integrating gives

$$v(t) = -c_1\sin(t) + c_2\cos(t). \tag{6.61}$$

Putting these together, we have

$$x(t) = c_1\cos(t) + c_2\sin(t) \tag{6.62}$$
$$v(t) = -c_1\sin(t) + c_2\cos(t). \tag{6.63}$$

This is a circle of radius $\sqrt{c_1^2 + c_2^2}$, traversed clockwise around the origin in the (x, v)-plane. Figure 6.7 shows an image of four square pulses rotated counterclockwise to time $t_f = \frac{\pi}{2}$ (quarter rotation) and $t_f = 2\pi$ (full rotation), using a $N_x \times N_v = 400 \times 400$ grid on $\Omega \times V = [-3, 3] \times [-3, 3]$.



(a) $t_f = \frac{\pi}{2}$ **(b)** $t_f = 2\pi$

Figure 6.7: Rotation Problem: $\frac{\partial}{\partial t} f + v \frac{\partial}{\partial x} f - x \frac{\partial}{\partial v} f = 0$. Grid: $N_v = Nx = 400$.

6.5.5 Two-Stream Instability (Vlasov–Poisson)

The next test that we performed is known as the *Two-Stream Instability Test*, and the present version is taken from Section 5.1.2 of [27]. In this test, two streams of electrons interact and create a highly filamented vortex. The test is designed to assess the code's ability to capture fine structure and examines only the advection piece of our equation (the Vlasov–Poisson equation):

$$\partial_t f + v\partial_x f + E\partial_v f = 0\ , \qquad (x, v, t) \in [-2\pi, 2\pi] \times [-2\pi, 2\pi] \times [0, \infty) \tag{6.64}$$

where

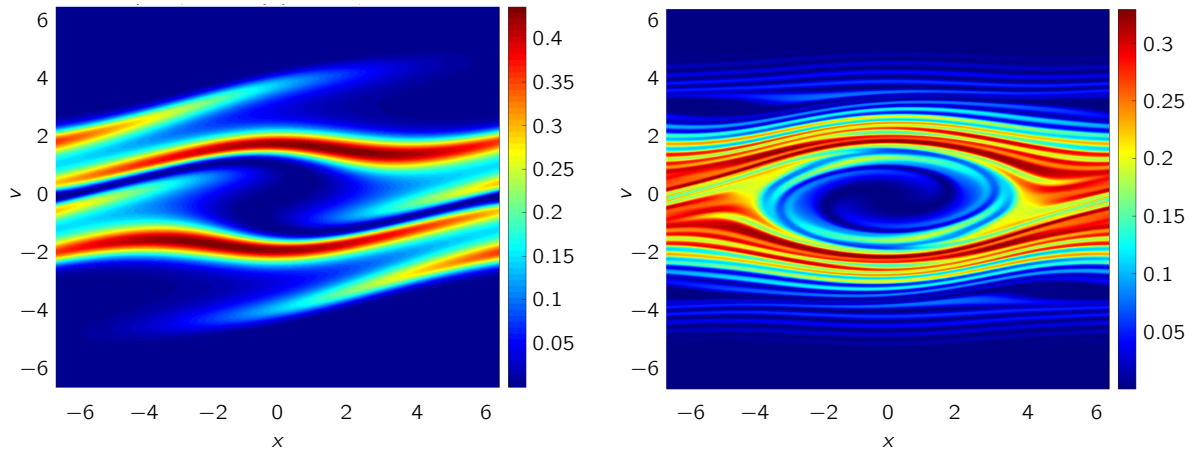
$$-\partial_{xx}\Phi = \frac{e}{\epsilon_0}(n - \overline{n}), \quad E = -\partial_x \Phi, \quad n = \int_V f\, \mathrm{d}v, \quad \overline{n} = \frac{1}{|V|}\int_V f\, \mathrm{d}v. \tag{6.65}$$

For our test runs, we used $\frac{e}{\epsilon_0} = 1$. The initial condition function for the test is

$$f(x, v, 0) = \frac{v^2}{\sqrt{8\pi}}\left(2 - \cos\left(\frac{x}{2}\right)\right) e^{-\frac{v^2}{2}}. \tag{6.66}$$

The test uses periodic boundary conditions in x, and zero flow boundary conditions in v.

Figure 6.8 shows the results of running the code until $t_\mathrm{f} = 5$ and $t_\mathrm{f} = 45$. For these plots, we used 800^2 grid points; this is twice the number of grid points as in the paper (400^2). We find good agreement with the computed solutions.



(a) $t_\mathrm{f} = 5$ **(b)** $t_\mathrm{f} = 45$

Figure 6.8: Two-Stream instability. The numerical solution is shown at $t_\mathrm{f} = 5$ (left) and $t_\mathrm{f} = 45$ (right). The numerical grid has $N_v \times N_x = 800 \times 800$ cells on the grid $\Omega \times V = [-2\pi, 2\pi] \times [-2\pi, 2\pi]$. The fine filamentation of the particle streams is adequately resolved and captured.

6.5.6 Landau Damping (Vlasov–Poisson)

This test is from Section 5.1.3 of [27]. To perform the test, the Vlasov–Poisson equation is solved and the L^2 norm of the electric field is computed at each time step. The initial condition function is

$$f(x, v, 0) = \frac{1}{\sqrt{2\pi}} \left(1 + \alpha \cos(kx)\right) e^{-\frac{v^2}{2}}, \tag{6.67}$$

where $\alpha = 0.01$, $k = 0.5$. According to [27], with these parameters, the L^2 norm of the electric field should decay exponentially at a rate of -0.1533. Figure 6.9 demonstrates that the methods developed in the preceding sections successfully recover the correct decay rate of the L^2 norm of the electric field.

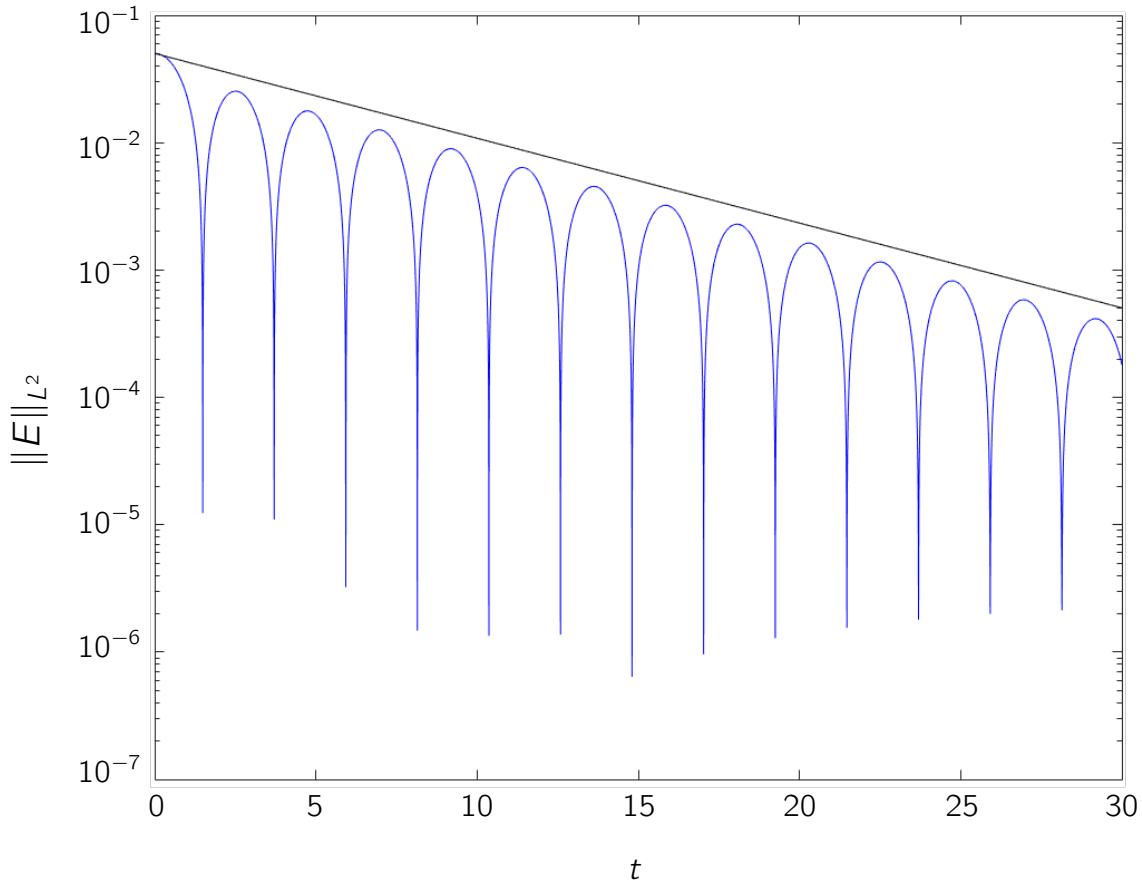


Figure 6.9: Landau damping for $\alpha = 0.01$ and $k = 0.5$: the computed L^2 norm of the electric field follows the reference exponential decay rate -0.1533 over the displayed interval. A regression estimate and fit interval will be reported when the archived output data are regenerated.

6.5.7 Quantitative Evidence and Reproducibility Status

Table 6.1 separates numerical quantities supported by the archived report material from comparisons that are currently graphical. This distinction prevents a visual comparison from being presented as a measured error norm. The next code-verification pass will add L^1, L^2, and L^∞ errors, observed refinement rates, conservation residuals, and positivity diagnostics from regenerated output data. For shock problems, the primary convergence measure will be the L^1 error because pointwise high-order convergence is not expected at discontinuities.

Test	**Reference quantity**	**Archived configuration**	**Evidence presently supported**
Relaxation 1	Exact solution of the homogeneous BGK equation	$\lambda = 1$, $t_f = 1$, $N_x = N_v = 128$	Signed pointwise error plot; magnitude $\mathcal{O}(10^{-5})$.
Relaxation 2	Exact solution of the homogeneous BGK equation	$\lambda = 1$, $t = 1, 2$	Signed pointwise error plots; norm and refinement data not archived.
Sod shock tube	Exact Euler Riemann solution in the collisional limit	$\tau = 10^{-4}$, $t_f = 0.2$, $V = [-10, 10]$	Graphical moment comparison; L^1 error study pending code verification.
Square-pulse rotation	Exact rotation through $\pi/2$ and 2π	400×400 grid	Graphical shape preservation; error and mass drift not archived.
Two-stream instability	Published benchmark morphology	800×800 grid, $t = 5, 45$	Resolved filamentation shown graphically; invariant drift not archived.
Landau damping	Reference electric-field decay rate -0.1533	$\alpha = 0.01$, $k = 0.5$	Semilog decay plot; fitted rate and fit interval not archived.

Table 6.1: Audit of the quantitative evidence retained with the unary BGK and Vlasov–Poisson test cases. Entries identified as pending require regenerated solver output and are not inferred from figures.

Chapter 7

Discrete-Velocity Numerical Methods for the Single-Species BGK Problem

In this chapter, we will describe numerical methods based on the discrete-velocity BGK model that we introduced in a previous chapter. Let us start with the standard discrete-velocity BGK problem first. After that, we will consider the space-homogeneous case.

7.1 A Fully Explicit Scheme

Let us again consider the $d = 1$ case. In other words, for simplicity, we will assume that the dimension of physical space is one and the dimension of velocity space is one. We will work under the assumption that velocity space is discrete and the discrete-velocity Maxwellian is computed via optimization. In other words, our starting model is the following:

$$\partial_t f(x, \mathrm{j}, t) + v_\mathrm{j} \partial_x f(x, \mathrm{j}, t) = \lambda \left(\mathcal{M}_{\mathcal{J},f}(x, \mathrm{j}, t) - f(x, \mathrm{j}, t) \right), \quad \forall \mathrm{j} \in \mathcal{J}, \quad \forall x \in [a, b], \quad (7.1)$$

where

$$\mathcal{M}_{\mathcal{J},f}(x, \mathrm{j}, t) = \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x, t) \cdot \boldsymbol{m}_\mathcal{J}(\mathrm{j})), \quad \forall \mathrm{j} \in \mathcal{J},$$

and $\boldsymbol{\alpha}_{\mathcal{J},f}(x, t) \in \mathbb{R}^3$ is the solution to

$$\langle \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(x, t) \cdot \boldsymbol{m}_\mathcal{J}) \, \boldsymbol{m}_\mathcal{J} \rangle_\mathcal{J} = \langle f(x, \cdot, t) \, \boldsymbol{m}_\mathcal{J} \rangle_\mathcal{J}.$$

Throughout this chapter, $\tau \coloneqq \lambda^{-1}$ is the relaxation time. For simplicity, we assume that the discrete velocities satisfy

$$v_{\mathfrak{j}} = h_v \cdot \left(\mathfrak{j} - \frac{1}{2}\right), \quad \mathfrak{j} \in \mathcal{J}, \quad \mathcal{J} \coloneqq \{-J+1, \ldots, J\}.$$

Definition 7.1 (First-Order Explicit Scheme)**.** *The **first-order explicit discrete-velocity scheme** is defined as follows:*

$$f_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}+1} = f_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}} - \frac{\Delta t}{h_x}\left(\widehat{F}(x_{\mathfrak{i}+1/2}, v_{\mathfrak{j}}, t^{\mathfrak{n}}) - \widehat{F}(x_{\mathfrak{i}-1/2}, v_{\mathfrak{j}}, t^{\mathfrak{n}})\right) + \frac{\Delta t}{\tau}\left(\mathcal{M}_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}} - f_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}}\right)$$

where

$$\mathcal{M}_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}} = \exp(\boldsymbol{\alpha}_{\mathfrak{i}}^{\mathfrak{n}} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathfrak{j})),$$

and $\boldsymbol{\alpha}_{\mathfrak{i}}^{\mathfrak{n}} \in \mathbb{R}^3$ is the solution to

$$\langle \exp(\boldsymbol{\alpha}_{\mathfrak{i}}^{\mathfrak{n}} \cdot \boldsymbol{m}_{\mathcal{J}})\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} = \langle f_{\mathfrak{i},\cdot}^{\mathfrak{n}}\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}}.$$

The first-order numerical flux is obtained via a simple upwind methodology:

$$\widehat{F}(x_{\mathfrak{i}+1/2}, v_{\mathfrak{j}}, t^{\mathfrak{n}}) = \begin{cases} v_{\mathfrak{j}} f_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}} & \text{if } v_{\mathfrak{j}} \geq 0, \\ v_{\mathfrak{j}} f_{\mathfrak{i}+1,\mathfrak{j}}^{\mathfrak{n}} & \text{if } v_{\mathfrak{j}} < 0. \end{cases}$$

Based on our construction, we have the following result:

Theorem 7.2. *Suppose that the spatial grid is periodic and $f_{\mathfrak{i},\mathfrak{j}}^{0} > 0$ for all $\mathfrak{j} \in \mathcal{J}$ and $1 \leq \mathfrak{i} \leq N_x$. Then, for all $\mathfrak{n} \in \mathbb{N}$,*

$$h_x \sum_{\mathfrak{i}=1}^{N_x} \left\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\mathfrak{i},\cdot}^{\mathfrak{n}} \right\rangle_{\mathcal{J}} = h_x \sum_{\mathfrak{i}=1}^{N_x} \left\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\mathfrak{i},\cdot}^{0} \right\rangle_{\mathcal{J}}. \tag{7.2}$$

Additionally, provided that the CFL condition

$$\Delta t \left(\frac{1}{\tau} + \max_{\mathfrak{j}\in\mathcal{J}} \frac{|v_{\mathfrak{j}}|}{h_x}\right) \leq 1$$

holds, it follows that, for all $\mathfrak{n} \in \mathbb{N}$, $\mathfrak{j} \in \mathcal{J}$, and $\mathfrak{i} \in \{1, 2, \ldots, N_x\}$, $f_{\mathfrak{i},\mathfrak{j}}^{\mathfrak{n}} > 0$. Furthermore, for all $\mathfrak{n} \in \mathbb{N}$,

$$h_x \sum_{\mathfrak{i}=1}^{N_x} \left\langle \ln\left(f_{\mathfrak{i},\cdot}^{\mathfrak{n}}\right) f_{\mathfrak{i},\cdot}^{\mathfrak{n}} \right\rangle_{\mathcal{J}} \leq h_x \sum_{\mathfrak{i}=1}^{N_x} \left\langle \ln\left(f_{\mathfrak{i},\cdot}^{\mathfrak{n}-1}\right) f_{\mathfrak{i},\cdot}^{\mathfrak{n}-1} \right\rangle_{\mathcal{J}}. \tag{7.3}$$

In other words, the scheme dissipates information and therefore produces physical entropy.

Proof. For $v_j > 0$, the scheme can be written as

$$\begin{aligned} f_{i,j}^{n+1} &= f_{i,j}^n - \frac{\Delta t v_j}{h_x}\left(f_{i,j}^n - f_{i-1,j}^n\right) + \frac{\Delta t}{\tau}\left(\mathcal{M}_{i,j}^n - f_{i,j}^n\right) \\ &= \left(1 - \frac{\Delta t v_j}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,j}^n + \frac{\Delta t v_j}{h_x} f_{i-1,j}^n + \frac{\Delta t}{\tau}\mathcal{M}_{i,j}^n \\ &= \left(1 - \frac{\Delta t |v_j|}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,j}^n + \frac{\Delta t |v_j|}{h_x} f_{i-1,j}^n + \frac{\Delta t}{\tau}\mathcal{M}_{i,j}^n. \end{aligned}$$

For $v_j < 0$, the scheme is

$$\begin{aligned} f_{i,j}^{n+1} &= f_{i,j}^n - \frac{\Delta t v_j}{h_x}\left(f_{i+1,j}^n - f_{i,j}^n\right) + \frac{\Delta t}{\tau}\left(\mathcal{M}_{i,j}^n - f_{i,j}^n\right) \\ &= \left(1 + \frac{\Delta t v_j}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,j}^n - \frac{\Delta t v_j}{h_x} f_{i+1,j}^n + \frac{\Delta t}{\tau}\mathcal{M}_{i,j}^n \\ &= \left(1 - \frac{\Delta t |v_j|}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,j}^n + \frac{\Delta t |v_j|}{h_x} f_{i+1,j}^n + \frac{\Delta t}{\tau}\mathcal{M}_{i,j}^n. \end{aligned}$$

The composite scheme can be written as

$$f_{i,j}^{n+1} = \left(1 - \frac{\Delta t |v_j|}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,j}^n + \frac{\Delta t(|v_j| - v_j)}{2h_x} f_{i+1,j}^n + \frac{\Delta t(|v_j| + v_j)}{2h_x} f_{i-1,j}^n + \frac{\Delta t}{\tau}\mathcal{M}_{i,j}^n,$$

which is a convex combination of $f_{i,j}^n$, $f_{i-1,j}^n$, $f_{i+1,j}^n$, and $\mathcal{M}_{i,j}^n$. If all of these values are positive, then $f_{i,j}^{n+1}$ is positive. Since the function $x\ln(x)$ is convex, Jensen's inequality applies to this four-term convex combination. Thus,

$$\begin{aligned} f_{i,j}^{n+1} \ln\left(f_{i,j}^{n+1}\right) \leq{}& \left(1 - \frac{\Delta t |v_j|}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,j}^n \ln\left(f_{i,j}^n\right) + \frac{\Delta t(|v_j| - v_j)}{2h_x} f_{i+1,j}^n \ln\left(f_{i+1,j}^n\right) \\ &+ \frac{\Delta t(|v_j| + v_j)}{2h_x} f_{i-1,j}^n \ln\left(f_{i-1,j}^n\right) + \frac{\Delta t}{\tau}\mathcal{M}_{i,j}^n \ln\left(\mathcal{M}_{i,j}^n\right). \end{aligned}$$

Summing over the velocity index j, we have

$$\begin{aligned} \left\langle f_{i,\cdot}^{n+1} \ln\left(f_{i,\cdot}^{n+1}\right)\right\rangle_{\mathcal{J}} \leq{}& \left\langle \left(1 - \frac{\Delta t |v_\cdot|}{h_x} - \frac{\Delta t}{\tau}\right) f_{i,\cdot}^n \ln\left(f_{i,\cdot}^n\right)\right\rangle_{\mathcal{J}} \\ &+ \left\langle \frac{\Delta t(|v_\cdot| - v_\cdot)}{2h_x} f_{i+1,\cdot}^n \ln\left(f_{i+1,\cdot}^n\right)\right\rangle_{\mathcal{J}} \\ &+ \left\langle \frac{\Delta t(|v_\cdot| + v_\cdot)}{2h_x} f_{i-1,\cdot}^n \ln\left(f_{i-1,\cdot}^n\right)\right\rangle_{\mathcal{J}} + \left\langle \frac{\Delta t}{\tau}\mathcal{M}_{i,\cdot}^n \ln\left(\mathcal{M}_{i,\cdot}^n\right)\right\rangle_{\mathcal{J}}, \end{aligned}$$

We next invoke the minimization property in (5.41): for all $\mathfrak{i}$,

$$\begin{aligned}
\left\langle f_{\mathfrak{i},\cdot}^{\mathfrak{n}+1} \ln\left(f_{\mathfrak{i},\cdot}^{\mathfrak{n}+1}\right)\right\rangle_{\mathcal{J}} &\leq \left\langle \left(1 - \frac{\Delta t|v_\cdot|}{h_x} - \frac{\Delta t}{\tau}\right) f_{\mathfrak{i},\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i},\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}} \\
&\quad + \left\langle \frac{\Delta t(|v_\cdot| - v_\cdot)}{2h_x} f_{\mathfrak{i}+1,\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i}+1,\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}} \\
&\quad + \left\langle \frac{\Delta t(|v_\cdot| + v_\cdot)}{2h_x} f_{\mathfrak{i}-1,\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i}-1,\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}} + \left\langle \frac{\Delta t}{\tau} f_{\mathfrak{i},\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i},\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}} \\
&= \left\langle \left(1 - \frac{\Delta t|v_\cdot|}{h_x}\right) f_{\mathfrak{i},\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i},\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}} \\
&\quad + \left\langle \frac{\Delta t(|v_\cdot| - v_\cdot)}{2h_x} f_{\mathfrak{i}+1,\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i}+1,\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}} \\
&\quad + \left\langle \frac{\Delta t(|v_\cdot| + v_\cdot)}{2h_x} f_{\mathfrak{i}-1,\cdot}^{\mathfrak{n}} \ln\left(f_{\mathfrak{i}-1,\cdot}^{\mathfrak{n}}\right)\right\rangle_{\mathcal{J}},
\end{aligned}$$

Summing on $\mathfrak{i}$ and using spatial periodicity completes the proof of (7.3). For moment conservation, multiply the update by $\boldsymbol{m}_{\mathcal{J}}(\mathfrak{j})$ and sum over $\mathfrak{j}$ and $\mathfrak{i}$. The collision contribution vanishes by discrete moment matching, while the two numerical-flux sums telescope on the periodic grid. This proves (7.2). ■

7.2 An Explicit Scheme for the Space-Homogeneous Case

In this section, we will explore an explicit scheme for the space-homogeneous, discrete-velocity BGK problem. Recall that the one-dimensional, space-homogeneous, discrete-velocity model is as follows:

$$\partial_t f(\mathfrak{j}, t) = \lambda\left(\mathcal{M}_{\mathcal{J},f}(\mathfrak{j}, t) - f(\mathfrak{j}, t)\right), \quad \forall \mathfrak{j} \in \mathcal{J}, \tag{7.4}$$

where

$$\mathcal{M}_{\mathcal{J},f}(\mathfrak{j}, t) = \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(t) \cdot \boldsymbol{m}_{\mathcal{J}}(\mathfrak{j})), \quad \forall \mathfrak{j} \in \mathcal{J},$$

and $\boldsymbol{\alpha}_{\mathcal{J},f}(t) \in \mathbb{R}^3$ is the solution to

$$\langle \exp(\boldsymbol{\alpha}_{\mathcal{J},f}(t) \cdot \boldsymbol{m}_{\mathcal{J}})\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} = \langle f(\,\cdot\,, t)\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}}.$$

As before, it may be useful to suppose that the discrete velocities satisfy

$$v_{\mathfrak{j}} = h_v \cdot \left(\mathfrak{j} - \frac{1}{2}\right), \quad \mathfrak{j} \in \mathcal{J}, \quad \mathcal{J} \coloneqq \{-J+1, \ldots, J\},$$

though this can be generalized without any difficulty. The explicit scheme for the space-homogeneous case is given by the following:

Definition 7.3 (Explicit Space-Homogeneous Scheme)**.** *The **first-order explicit discrete-velocity scheme** is defined as follows:*

$$f_{\mathfrak{j}}^{\mathfrak{n}+1} = f_{\mathfrak{j}}^{\mathfrak{n}} + \frac{\Delta t}{\tau}\left(\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}} - f_{\mathfrak{j}}^{\mathfrak{n}}\right),$$

where

$$\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}} = \exp(\boldsymbol{\alpha}^{\mathfrak{n}} \cdot \boldsymbol{m}_{\mathcal{J}}(\mathfrak{j})), \tag{7.5}$$

and $\boldsymbol{\alpha}^{\mathfrak{n}} \in \mathbb{R}^3$ is the solution to

$$\langle \exp(\boldsymbol{\alpha}^{\mathfrak{n}} \cdot \boldsymbol{m}_{\mathcal{J}})\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} = \langle f_{\cdot}^{\mathfrak{n}}\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}}. \tag{7.6}$$

We have the following result, which shows that an explicit scheme will always have a time step restriction owing to the stiffness of the BGK problem:

Theorem 7.4. *Suppose that $f_{\mathfrak{j}}^0 > 0$, for all $\mathfrak{j} \in \mathcal{J}$. Then, for all $\mathfrak{n} \in \mathbb{N}$,*

$$\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{\mathfrak{n}} \rangle_{\mathcal{J}} = \left\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{0} \right\rangle_{\mathcal{J}}. \tag{7.7}$$

Additionally, provided that the condition

$$\frac{\Delta t}{\tau} \leq 1$$

holds, it follows that, for all $\mathfrak{n} \in \mathbb{N}$ and $\mathfrak{j} \in \mathcal{J}$, $f_{\mathfrak{j}}^{\mathfrak{n}} > 0$. Furthermore, for all $\mathfrak{n} \in \mathbb{N}$,

$$\langle \ln\left(f_{\cdot}^{\mathfrak{n}}\right) f_{\cdot}^{\mathfrak{n}} \rangle_{\mathcal{J}} \leq \left\langle \ln\left(f_{\cdot}^{\mathfrak{n}-1}\right) f_{\cdot}^{\mathfrak{n}-1} \right\rangle_{\mathcal{J}}. \tag{7.8}$$

In other words, the scheme dissipates information and therefore produces physical entropy. Finally,

$$\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}} = \mathcal{M}_{\mathfrak{j}}^{0}, \quad \forall\, \mathfrak{j} \in \mathcal{J}, \tag{7.9}$$

for all $\mathfrak{n} \in \mathbb{N}$.

Proof. The scheme can be written as

$$f_{\mathfrak{j}}^{\mathfrak{n}+1} = \left(1 - \frac{\Delta t}{\tau}\right) f_{\mathfrak{j}}^{\mathfrak{n}} + \frac{\Delta t}{\tau} \mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}},$$

which is a convex combination of f_j^n and $\mathcal{M}_j^n$. The CFL condition makes every coefficient in this convex combination nonnegative, so f_j^{n+1} is positive. Since the function $x\ln(x)$ is convex, it follows that

$$f_j^{n+1}\ln\left(f_j^{n+1}\right) \leq \left(1-\frac{\Delta t}{\tau}\right) f_j^n \ln\left(f_j^n\right) + \frac{\Delta t}{\tau}\mathcal{M}_j^n \ln\left(\mathcal{M}_j^n\right).$$

Summing on j, we have

$$\left\langle f_\cdot^{n+1}\ln\left(f_\cdot^{n+1}\right)\right\rangle_{\mathcal{J}} \leq \left\langle \left(1-\frac{\Delta t}{\tau}\right) f_\cdot^n \ln\left(f_\cdot^n\right)\right\rangle_{\mathcal{J}} + \left\langle \frac{\Delta t}{\tau}\mathcal{M}_\cdot^n \ln\left(\mathcal{M}_\cdot^n\right)\right\rangle_{\mathcal{J}},$$

Using the minimization property in (5.41), we find that,

$$\begin{aligned}\left\langle f_\cdot^{n+1}\ln\left(f_\cdot^{n+1}\right)\right\rangle_{\mathcal{J}} &\leq \left\langle \left(1-\frac{\Delta t}{\tau}\right) f_\cdot^n \ln\left(f_\cdot^n\right)\right\rangle_{\mathcal{J}} + \left\langle \frac{\Delta t}{\tau}\mathcal{M}_\cdot^n \ln\left(\mathcal{M}_\cdot^n\right)\right\rangle_{\mathcal{J}} \\ &\leq \left\langle \left(1-\frac{\Delta t}{\tau}\right) f_\cdot^n \ln\left(f_\cdot^n\right)\right\rangle_{\mathcal{J}} + \left\langle \frac{\Delta t}{\tau} f_\cdot^n \ln\left(f_\cdot^n\right)\right\rangle_{\mathcal{J}} \\ &= \left\langle f_\cdot^n \ln\left(f_\cdot^n\right)\right\rangle_{\mathcal{J}}.\end{aligned}$$

Multiplying the update by $\boldsymbol{m}_{\mathcal{J}}(j)$ and summing over j proves (7.7), because the discrete Maxwellian matches the moments of f^n. The identity (7.9) then follows from (7.7), (7.5), and (7.6). ■

7.3 An Implicit Scheme for the Space-Homogeneous Case

In the last section, we learned that an explicit scheme for the space-homogeneous case still requires a potentially restrictive time step constraint for stability. But, this can be easily overcome in the space-homogeneous case.

Definition 7.5 (Implicit Space-Homogeneous Scheme)**.** *The **first-order implicit discrete-velocity scheme** is defined as follows:*

$$f_j^{n+1} = f_j^n + \frac{\Delta t}{\tau}\left(\mathcal{M}_j^n - f_j^{n+1}\right),$$

where

$$\mathcal{M}_j^n = \exp(\boldsymbol{\alpha}^n \cdot \boldsymbol{m}_{\mathcal{J}}(j)), \tag{7.10}$$

and $\boldsymbol{\alpha}^{\mathfrak{n}} \in \mathbb{R}^3$ *is the solution to*

$$\langle \exp(\boldsymbol{\alpha}^{\mathfrak{n}} \cdot \boldsymbol{m}_{\mathcal{J}})\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}} = \langle f_{\cdot}^{\mathfrak{n}}\, \boldsymbol{m}_{\mathcal{J}} \rangle_{\mathcal{J}}. \tag{7.11}$$

Technically, the last scheme is an IMEX scheme, but we will stick to our definition, since we can show that, actually, $\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}} = \mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}+1}$. Specifically, we have the following result:

Theorem 7.6. *Suppose that* $f_{\mathfrak{j}}^0 > 0$*, for all* $\mathfrak{j} \in \mathcal{J}$*. Then, for all* $\mathfrak{n} \in \mathbb{N}$*,*

$$\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{\mathfrak{n}} \rangle_{\mathcal{J}} = \left\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{0} \right\rangle_{\mathcal{J}}. \tag{7.12}$$

Additionally, for all $\mathfrak{n} \in \mathbb{N}$ *and* $\mathfrak{j} \in \mathcal{J}$*,* $f_{\mathfrak{j}}^{\mathfrak{n}} > 0$*, and for all* $\mathfrak{n} \in \mathbb{N}$*,*

$$\langle \ln\left(f_{\cdot}^{\mathfrak{n}}\right) f_{\cdot}^{\mathfrak{n}} \rangle_{\mathcal{J}} \leq \left\langle \ln\left(f_{\cdot}^{\mathfrak{n}-1}\right) f_{\cdot}^{\mathfrak{n}-1} \right\rangle_{\mathcal{J}}. \tag{7.13}$$

In other words, the scheme dissipates information and therefore produces physical entropy. Finally,

$$\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}} = \mathcal{M}_{\mathfrak{j}}^{0}, \quad \forall\, \mathfrak{j} \in \mathcal{J}, \tag{7.14}$$

for all $\mathfrak{n} \in \mathbb{N}$*. All of these properties hold unconditionally.*

Proof. The scheme can be written as

$$\left(1 + \frac{\Delta t}{\tau}\right) f_{\mathfrak{j}}^{\mathfrak{n}+1} = f_{\mathfrak{j}}^{\mathfrak{n}} + \frac{\Delta t}{\tau} \mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}},$$

or, equivalently, as

$$f_{\mathfrak{j}}^{\mathfrak{n}+1} = \frac{1}{\left(1 + \frac{\Delta t}{\tau}\right)} f_{\mathfrak{j}}^{\mathfrak{n}} + \frac{\frac{\Delta t}{\tau}}{\left(1 + \frac{\Delta t}{\tau}\right)} \mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}}, \tag{7.15}$$

which is again a convex combination of $f_{\mathfrak{j}}^{\mathfrak{n}}$ and $\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}}$. Both weights are positive and sum to one. Hence $f_{\mathfrak{j}}^{\mathfrak{n}+1}$ is positive whenever $f_{\mathfrak{j}}^{\mathfrak{n}}$ and $\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}}$ are positive. Let us write, for simplicity,

$$\alpha \coloneqq \frac{1}{1 + \frac{\Delta t}{\tau}} \quad \text{and} \quad \beta \coloneqq \frac{\frac{\Delta t}{\tau}}{1 + \frac{\Delta t}{\tau}}.$$

Since the function $x \ln(x)$ is convex, it follows that

$$f_{\mathfrak{j}}^{\mathfrak{n}+1} \ln\left(f_{\mathfrak{j}}^{\mathfrak{n}+1}\right) \leq \alpha f_{\mathfrak{j}}^{\mathfrak{n}} \ln\left(f_{\mathfrak{j}}^{\mathfrak{n}}\right) + \beta \mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}} \ln\left(\mathcal{M}_{\mathfrak{j}}^{\mathfrak{n}}\right).$$

Summing on $\mathfrak{j}$, we have

$$\left\langle f_{\cdot}^{n+1} \ln\left(f_{\cdot}^{n+1}\right)\right\rangle_{\mathcal{J}} \leq \langle \alpha f_{\cdot}^{n} \ln\left(f_{\cdot}^{n}\right)\rangle_{\mathcal{J}} + \langle \beta \mathcal{M}_{\cdot}^{n} \ln\left(\mathcal{M}_{\cdot}^{n}\right)\rangle_{\mathcal{J}},$$

Using the minimization property in (5.41), we find that,

$$\begin{aligned}\left\langle f_{\cdot}^{n+1} \ln\left(f_{\cdot}^{n+1}\right)\right\rangle_{\mathcal{J}} &\leq \langle \alpha f_{\cdot}^{n} \ln\left(f_{\cdot}^{n}\right)\rangle_{\mathcal{J}} + \langle \beta \mathcal{M}_{\cdot}^{n} \ln\left(\mathcal{M}_{\cdot}^{n}\right)\rangle_{\mathcal{J}} \\ &\leq \langle \alpha f_{\cdot}^{n} \ln\left(f_{\cdot}^{n}\right)\rangle_{\mathcal{J}} + \langle \beta f_{\cdot}^{n} \ln\left(f_{\cdot}^{n}\right)\rangle_{\mathcal{J}} \\ &= \langle f_{\cdot}^{n} \ln\left(f_{\cdot}^{n}\right)\rangle_{\mathcal{J}}.\end{aligned}$$

Next, summing $\boldsymbol{m}_{\mathcal{J}}(\,\cdot\,)$ times (7.15) on $\mathfrak{j}$, we have

$$\begin{aligned}\left\langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{n+1}\right\rangle_{\mathcal{J}} &= \langle \alpha \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{n}\rangle_{\mathcal{J}} + \langle \beta \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) \mathcal{M}_{\cdot}^{n}\rangle_{\mathcal{J}} \\ &= \langle \alpha \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{n}\rangle_{\mathcal{J}} + \langle \beta \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{n}\rangle_{\mathcal{J}} \\ &= \langle \boldsymbol{m}_{\mathcal{J}}(\,\cdot\,) f_{\cdot}^{n}\rangle_{\mathcal{J}},\end{aligned}$$

which proves (7.12), and, finally, (7.14) follows directly from (7.12). ■

Chapter 8

Multispecies BGK Equations

This chapter presents the conservative, entropic multispecies BGK model introduced by Haack, Hauck, and Murillo [32]. The construction and the proofs in Sections 8.1–8.5 are adapted from Sections 2 and 3 of that paper, but have been rewritten in the notation and entropy convention of Chapter 2. In particular, $\boldsymbol{v} \in \mathbb{R}^d$ is microscopic velocity, n_i and $\rho_i = m_i n_i$ are species number and mass densities, $\boldsymbol{u}_i$ is the species velocity, and $\hat{\theta}_i$ is the species temperature in energy units. We use the physical entropy $S = -H$, where $H = \sum_i \int f_i \ln f_i \, \mathrm{d}\boldsymbol{v}$ is information; hence $\mathrm{d}S/\mathrm{d}t \geq 0$ and $\mathrm{d}H/\mathrm{d}t \leq 0$.

The purpose here is to derive the particular model used throughout this report, not to reproduce every topic in [32]. We therefore omit the paper's Section 4 catalog of collision-frequency formulas: the structural results below require only $\lambda_{i,j} > 0$ and do not depend on a particular cross section. We retain the leading fluid-limit consequence of its Section 5 in Section 8.6, and omit its comparison and numerical sections (Sections 6 onward).

8.1 Boltzmann Structure to Be Preserved

For species $i \in \mathcal{S} \coloneqq \{1, \ldots, N\}$, the multispecies Boltzmann equation is

$$\partial_t f_i + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f_i + \boldsymbol{a}_i \cdot \nabla_{\boldsymbol{v}} f_i = \sum_{j \in \mathcal{S}} Q_{i,j}[f_i, f_j]. \tag{8.1}$$

The collision maps and collision operators were developed in Chapter 2. The properties needed here are species-mass conservation and pairwise conservation of momentum and energy:

$$\int_{\mathbb{R}^d} Q_{i,j}[f_i, f_j] \mathrm{d}\boldsymbol{v} = 0, \tag{8.2a}$$

$$m_i \int_{\mathbb{R}^d} \boldsymbol{v} Q_{i,j}[f_i, f_j] \mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^d} \boldsymbol{v} Q_{j,i}[f_j, f_i] \mathrm{d}\boldsymbol{v} = \boldsymbol{0}, \tag{8.2b}$$

$$m_i \int_{\mathbb{R}^d} |\boldsymbol{v}|^2 Q_{i,j}[f_i, f_j] \mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^d} |\boldsymbol{v}|^2 Q_{j,i}[f_j, f_i] \mathrm{d}\boldsymbol{v} = 0. \tag{8.2c}$$

The Boltzmann H-theorem additionally requires

$$\sum_{i,j} \int_{\mathbb{R}^d} Q_{i,j}[f_i, f_j] \ln f_i \mathrm{d}\boldsymbol{v} \leq 0, \tag{8.3}$$

with equality, for a connected interacting mixture, only when every species is Maxwellian with a common velocity and temperature.

We recall the species moments

$$n_i = \int_{\mathbb{R}^d} f_i \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u}_i = \frac{1}{n_i} \int_{\mathbb{R}^d} \boldsymbol{v} f_i \, \mathrm{d}\boldsymbol{v}, \qquad \hat{\theta}_i = \frac{m_i}{d n_i} \int_{\mathbb{R}^d} |\boldsymbol{v} - \boldsymbol{u}_i|^2 f_i \, \mathrm{d}\boldsymbol{v}. \tag{8.4}$$

For later use, define the barycentric velocity and mixture temperature by

$$\rho \boldsymbol{u} = \sum_i \rho_i \boldsymbol{u}_i, \quad \frac{d}{2} n\theta = \sum_i \left(\frac{d}{2} n_i \hat{\theta}_i + \frac{1}{2} \rho_i |\boldsymbol{u}_i - \boldsymbol{u}|^2 \right), \quad n = \sum_i n_i, \quad \rho = \sum_i \rho_i. \tag{8.5}$$

Thus $\hat{\theta}_i$ measures thermal energy in the species frame, whereas θ includes both species thermal energy and diffusion kinetic energy relative to $\boldsymbol{u}$.

8.2 Formal Ad Hoc Construction from the Boltzmann Operator

The following argument is a formal modeling construction, not an asymptotically controlled approximation. To keep its logical status clear, we use equality for exact algebraic identities and $\approx$ only for modeling closures. There are three distinct steps: approximate the gain term by pair Maxwellians, replace two kernel-weighted rates by one common rate, and then remove the remaining velocity dependence from that rate.

Fix an ordered pair (i, j) and suppress the dependence on $\boldsymbol{x}$ and t. The Boltzmann operator is

$$Q_{i,j}[f_i, f_j](\boldsymbol{v}) = \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} \left(f_i' f_{j,*}' - f_i f_{j,*} \right) B_{i,j}(\boldsymbol{v} - \boldsymbol{v}_*, \boldsymbol{\sigma}) \, \mathrm{d}\boldsymbol{\sigma} \, \mathrm{d}\boldsymbol{v}_*. \tag{8.6}$$

Here $f_i = f_i(\boldsymbol{v})$, $f_{j,*} = f_j(\boldsymbol{v}_*)$, and the primes denote the collision-related velocities from Chapter 2. We first define the target distributions used in the construction.

Definition 8.1 (Pair Maxwellians)**.** *Fix $i, j \in \mathcal{S}$. Let $n_i > 0$ and $m_i > 0$, and let $\boldsymbol{u}_{i,j} \in \mathbb{R}^d$ and $\hat{\theta}_{i,j} > 0$. The* ***pair Maxwellian for species*** *i* ***in collisions with species*** *j is*

$$M_{i,j}(\boldsymbol{v}) = n_i \left(\frac{m_i}{2\pi\hat{\theta}_{i,j}}\right)^{d/2} \exp\left(-\frac{m_i|\boldsymbol{v}-\boldsymbol{u}_{i,j}|^2}{2\hat{\theta}_{i,j}}\right). \tag{8.7}$$

For an unordered interacting pair $\{i, j\}$, the two ordered targets $M_{i,j}$ and $M_{j,i}$ are called a ***compatible pair of Maxwellians*** *when they share their mean velocity and temperature:*

$$\boldsymbol{u}_{i,j} = \boldsymbol{u}_{j,i}, \qquad \hat{\theta}_{i,j} = \hat{\theta}_{j,i}. \tag{8.8}$$

Remark 8.2. *Although the pair parameters are shared, the two Maxwellians generally differ because $M_{i,j}$ has species-i density n_i and mass m_i, whereas $M_{j,i}$ has species-j density n_j and mass m_j. In particular,*

$$\int_{\mathbb{R}^d} M_{i,j}\,\mathrm{d}\boldsymbol{v} = n_i, \qquad \int_{\mathbb{R}^d} \boldsymbol{v} M_{i,j}\,\mathrm{d}\boldsymbol{v} = n_i\boldsymbol{u}_{i,j}, \qquad \int_{\mathbb{R}^d} |\boldsymbol{v}|^2 M_{i,j}\,\mathrm{d}\boldsymbol{v} = n_i|\boldsymbol{u}_{i,j}|^2 + \frac{dn_i}{m_i}\hat{\theta}_{i,j}. \tag{8.9}$$

When $i = j$, choosing $\boldsymbol{u}_{i,i} = \boldsymbol{u}_i$ and $\hat{\theta}_{i,i} = \hat{\theta}_i$ recovers the ordinary species Maxwellian.

Step 1: Approximate the gain product. The near-equilibrium ansatz replaces the two distributions in the gain term, but not yet those in the loss term:

$$f_i' f_{j,*}' \approx M_{i,j}(\boldsymbol{v}')M_{j,i}(\boldsymbol{v}_*'). \tag{8.10}$$

The following proposition isolates the exact calculation that permits the primed velocities to be removed. Thus the proposition is not itself a closure assumption; the approximation has already entered through (8.10).

Proposition 8.3 (Collision Invariance of Compatible Pair Maxwellians)**.** *Let $m_i, m_j, n_i, n_j > 0$. Suppose that $M_{i,j}$ and $M_{j,i}$ are the compatible pair Maxwellians of Definition 8.1, with common velocity $\boldsymbol{u}_{ij}^{\mathrm{p}} \in \mathbb{R}^d$ and common temperature $\hat{\theta}_{ij}^{\mathrm{p}} > 0$. If $(\boldsymbol{v}', \boldsymbol{v}_*')$ is obtained from $(\boldsymbol{v}, \boldsymbol{v}_*)$ by an elastic binary collision that conserves the pair's total momentum and kinetic energy, then*

$$M_{i,j}(\boldsymbol{v}')M_{j,i}(\boldsymbol{v}_*') = M_{i,j}(\boldsymbol{v})M_{j,i}(\boldsymbol{v}_*). \tag{8.11}$$

Proof. The Maxwellian normalization factors are independent of the velocities. Consequently,

up to an additive constant, the logarithm of their product is

$$-\frac{1}{2\hat{\theta}_{ij}^{\mathrm{p}}}\left(m_i|\boldsymbol{v}-\boldsymbol{u}_{ij}^{\mathrm{p}}|^2+m_j|\boldsymbol{v}_*-\boldsymbol{u}_{ij}^{\mathrm{p}}|^2\right). \tag{8.12}$$

Expanding the quadratic expression gives

$$\begin{aligned} &m_i|\boldsymbol{v}-\boldsymbol{u}_{ij}^{\mathrm{p}}|^2+m_j|\boldsymbol{v}_*-\boldsymbol{u}_{ij}^{\mathrm{p}}|^2 \\ &\quad = m_i|\boldsymbol{v}|^2+m_j|\boldsymbol{v}_*|^2-2\boldsymbol{u}_{ij}^{\mathrm{p}}\cdot(m_i\boldsymbol{v}+m_j\boldsymbol{v}_*)+(m_i+m_j)|\boldsymbol{u}_{ij}^{\mathrm{p}}|^2. \end{aligned} \tag{8.13}$$

The first two terms on the right are, respectively, the pair's total kinetic energy and the common velocity dotted with its total momentum. Both are unchanged by the elastic collision; the final term is constant. Hence the centered quadratic expression has the same value at $(\boldsymbol{v}',\boldsymbol{v}_*')$ and at $(\boldsymbol{v},\boldsymbol{v}_*)$. Exponentiating proves (8.11). ■

Applying Proposition 8.3 to the right-hand side of (8.10) transforms the exact Boltzmann operator into the two-rate approximation

$$Q_{i,j}[f_i,f_j](\boldsymbol{v})\approx M_{i,j}(\boldsymbol{v})\nu_{i,j}^{M}(\boldsymbol{v})-f_i(\boldsymbol{v})\nu_{i,j}^{f}(\boldsymbol{v}), \tag{8.14}$$

where

$$\nu_{i,j}^{M}(\boldsymbol{v}) \coloneqq \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} M_{j,i}(\boldsymbol{v}_*)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_*,\boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{v}_*, \tag{8.15}$$

$$\nu_{i,j}^{f}(\boldsymbol{v}) \coloneqq \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} f_j(\boldsymbol{v}_*)B_{i,j}(\boldsymbol{v}-\boldsymbol{v}_*,\boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{v}_*. \tag{8.16}$$

At this point the gain and loss coefficients are generally different.

Step 2: Impose a common kernel-weighted rate. The next closure assumes that f_j is sufficiently close to $M_{j,i}$ in the kernel-weighted average that

$$\nu_{i,j}^{f}(\boldsymbol{v})\approx\nu_{i,j}^{M}(\boldsymbol{v})=:\nu_{i,j}(\boldsymbol{v}). \tag{8.17}$$

Only after this assumption does (8.14) acquire the BGK relaxation structure

$$Q_{i,j}[f_i,f_j](\boldsymbol{v})\approx\nu_{i,j}(\boldsymbol{v})\big(M_{i,j}(\boldsymbol{v})-f_i(\boldsymbol{v})\big). \tag{8.18}$$

Step 3: Replace the velocity-dependent rate by a constant. For a general collision kernel, $\nu_{i,j}(\boldsymbol{v})$ is both difficult to compute and incompatible with the simple closed moment

calculations used below. The model therefore makes the additional replacement

$$\nu_{i,j}(\boldsymbol{v}) \;\longrightarrow\; \lambda_{i,j} > 0, \tag{8.19}$$

which yields the pairwise BGK operator

$$Q_{i,j}^{\mathrm{BGK}}[f_i, f_j] = \lambda_{i,j}(M_{i,j} - f_i). \tag{8.20}$$

This last replacement is independent of the common-rate assumption in Step 2. It compresses the velocity dependence and detailed cross-section information into the effective parameter $\lambda_{i,j}$, which must be specified or calibrated separately. In general, the ordered rates $\lambda_{i,j}$ and $\lambda_{j,i}$ need not be equal. These steps motivate the relaxation form only; they do not determine $\boldsymbol{u}_{i,j}$ or $\hat{\theta}_{i,j}$. The next section selects those parameters by enforcing pairwise momentum and energy conservation.

The resulting model is

$$\partial_t f_i + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f_i + \boldsymbol{a}_i \cdot \nabla_{\boldsymbol{v}} f_i = \sum_{j \in \mathcal{S}} \lambda_{i,j}(M_{i,j} - f_i), \qquad i \in \mathcal{S}. \tag{8.21}$$

Self-collisions use the ordinary species Maxwellian:

$$\boldsymbol{u}_{i,i} = \boldsymbol{u}_i, \qquad \hat{\theta}_{i,i} = \hat{\theta}_i. \tag{8.22}$$

8.3 Pair Parameters and Their Positivity

The common pair velocity and temperature are determined by enforcing the pairwise momentum and energy balances in (8.2b)–(8.2c). The next result makes precise the hypotheses under which these constraints determine the parameters uniquely.

Proposition 8.4 (Unique Pair Parameters Imposed by Conservation)**.** *Fix distinct species $i, j \in \mathcal{S}$. Assume that $m_i, m_j > 0$, that f_i and f_j have positive densities n_i, n_j, finite second moments, and nonnegative temperatures $\hat{\theta}_i, \hat{\theta}_j$, and that $\lambda_{i,j}, \lambda_{j,i} > 0$. Among all pair candidates that share a common velocity $\boldsymbol{u}_{i,j} = \boldsymbol{u}_{j,i}$ and a common temperature $\hat{\theta}_{i,j} = \hat{\theta}_{j,i}$, the pairwise momentum and energy constraints determine a unique algebraic pair of parameters, namely,*

$$\boldsymbol{u}_{i,j} = \frac{\lambda_{i,j}\rho_i \boldsymbol{u}_i + \lambda_{j,i}\rho_j \boldsymbol{u}_j}{\lambda_{i,j}\rho_i + \lambda_{j,i}\rho_j} \tag{8.23}$$

and

$$\begin{aligned}\hat{\theta}_{i,j} = &\frac{\lambda_{i,j} n_i \hat{\theta}_i + \lambda_{j,i} n_j \hat{\theta}_j}{\lambda_{i,j} n_i + \lambda_{j,i} n_j} \\ &+ \frac{\lambda_{i,j}\rho_i(|\boldsymbol{u}_i|^2 - |\boldsymbol{u}_{i,j}|^2) + \lambda_{j,i}\rho_j(|\boldsymbol{u}_j|^2 - |\boldsymbol{u}_{i,j}|^2)}{d(\lambda_{i,j} n_i + \lambda_{j,i} n_j)}.\end{aligned} \tag{8.24}$$

Species mass is conserved automatically. Proposition 8.6 below gives the precise admissibility condition for the resulting pair temperature.

Proof. The normalization in (8.9) first gives

$$\int_{\mathbb{R}^d} Q_{i,j}^{\mathrm{BGK}} \,\mathrm{d}\boldsymbol{v} = \lambda_{i,j}\left(\int_{\mathbb{R}^d} M_{i,j}\,\mathrm{d}\boldsymbol{v} - \int_{\mathbb{R}^d} f_i\,\mathrm{d}\boldsymbol{v}\right) = 0. \tag{8.25}$$

For momentum, compatibility gives $\boldsymbol{u}_{j,i} = \boldsymbol{u}_{i,j}$. Therefore the pairwise momentum production is

$$m_i \int_{\mathbb{R}^d} \boldsymbol{v} Q_{i,j}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^d} \boldsymbol{v} Q_{j,i}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} = \lambda_{i,j}\rho_i(\boldsymbol{u}_{i,j} - \boldsymbol{u}_i) + \lambda_{j,i}\rho_j(\boldsymbol{u}_{i,j} - \boldsymbol{u}_j). \tag{8.26}$$

Setting this expression equal to zero and collecting the terms containing $\boldsymbol{u}_{i,j}$ gives

$$(\lambda_{i,j}\rho_i + \lambda_{j,i}\rho_j)\boldsymbol{u}_{i,j} = \lambda_{i,j}\rho_i \boldsymbol{u}_i + \lambda_{j,i}\rho_j \boldsymbol{u}_j. \tag{8.27}$$

The coefficient on the left is positive, so this vector equation has exactly one solution, namely (8.23).

For the energy balance, the definition of $\hat{\theta}_i$ implies

$$\int_{\mathbb{R}^d} |\boldsymbol{v}|^2 f_i\,\mathrm{d}\boldsymbol{v} = n_i|\boldsymbol{u}_i|^2 + \frac{d n_i}{m_i}\hat{\theta}_i. \tag{8.28}$$

Combining this identity with the corresponding pair-Maxwellian moment in (8.9), and using $\hat{\theta}_{j,i} = \hat{\theta}_{i,j}$, yields

$$\begin{aligned}0 &= m_i \int_{\mathbb{R}^d} |\boldsymbol{v}|^2 Q_{i,j}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} + m_j \int_{\mathbb{R}^d} |\boldsymbol{v}|^2 Q_{j,i}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} \\ &= \lambda_{i,j}\rho_i\big(|\boldsymbol{u}_{i,j}|^2 - |\boldsymbol{u}_i|^2\big) + \lambda_{j,i}\rho_j\big(|\boldsymbol{u}_{i,j}|^2 - |\boldsymbol{u}_j|^2\big) \\ &\quad + d\lambda_{i,j} n_i(\hat{\theta}_{i,j} - \hat{\theta}_i) + d\lambda_{j,i} n_j(\hat{\theta}_{i,j} - \hat{\theta}_j).\end{aligned} \tag{8.29}$$

Collecting the terms containing $\hat{\theta}_{i,j}$ gives

$$d(\lambda_{i,j} n_i + \lambda_{j,i} n_j)\hat{\theta}_{i,j} = d\lambda_{i,j} n_i \hat{\theta}_i + d\lambda_{j,i} n_j \hat{\theta}_j$$

$$+ \lambda_{i,j}\rho_i\big(|\boldsymbol{u}_i|^2 - |\boldsymbol{u}_{i,j}|^2\big)$$
$$+ \lambda_{j,i}\rho_j\big(|\boldsymbol{u}_j|^2 - |\boldsymbol{u}_{i,j}|^2\big). \tag{8.30}$$

The coefficient of the term on the left, $d(\lambda_{i,j}n_i + \lambda_{j,i}n_j)$, is also positive. Thus the scalar energy balance has exactly one solution, which is (8.24). This proves existence and uniqueness of the algebraic parameters, while (8.25) proves the last assertion. ■

Remark 8.5. *Equation* (8.24) *contains* $|\boldsymbol{u}_j|^2$ *in its second kinetic-energy term. This corrects a typographical error in Equation (36) of the published paper, where that occurrence is printed as* $|\boldsymbol{u}_i|^2$.

Proposition 8.6 (Positivity of the Pair Temperature)**.** *If* $n_i, n_j, \lambda_{i,j}, \lambda_{j,i} > 0$ *and* $\hat{\theta}_i, \hat{\theta}_j \geq 0$, *then* $\hat{\theta}_{i,j} \geq 0$. *It is strictly positive if either species temperature is positive or* $\boldsymbol{u}_i \neq \boldsymbol{u}_j$.

Proof. Set $a = \lambda_{i,j}\rho_i$ and $b = \lambda_{j,i}\rho_j$. By (8.23), $\boldsymbol{u}_{i,j} = (a\boldsymbol{u}_i + b\boldsymbol{u}_j)/(a+b)$. Hence the weighted variance identity gives

$$a\big(|\boldsymbol{u}_i|^2 - |\boldsymbol{u}_{i,j}|^2\big) + b\big(|\boldsymbol{u}_j|^2 - |\boldsymbol{u}_{i,j}|^2\big) = \frac{ab}{a+b}|\boldsymbol{u}_i - \boldsymbol{u}_j|^2 \geq 0. \tag{8.31}$$

The first term in (8.24) is a positive-weight average of $\hat{\theta}_i$ and $\hat{\theta}_j$, while its second term is nonnegative by (8.31). The stated strictness conditions follow immediately. ■

8.4 Conservation Laws

Proposition 8.7 (Pairwise and Total Conservation)**.** *The operators* (8.20), *with parameters* (8.23)–(8.24), *conserve the mass of every species and conserve total momentum and total kinetic energy. More precisely,*

$$\int Q_{i,j}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} = 0, \tag{8.32a}$$
$$m_i\int \boldsymbol{v}Q_{i,j}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} + m_j\int \boldsymbol{v}Q_{j,i}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} = \boldsymbol{0}, \tag{8.32b}$$
$$m_i\int |\boldsymbol{v}|^2Q_{i,j}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} + m_j\int |\boldsymbol{v}|^2Q_{j,i}^{\mathrm{BGK}}\,\mathrm{d}\boldsymbol{v} = 0. \tag{8.32c}$$

Proof. Species-mass conservation is exactly (8.25). For $i \neq j$, the pairwise momentum and energy productions are the left-hand sides of (8.26) and (8.29); they vanish by the defining

balances (8.27) and (8.30). For $i = j$, the self-Maxwellian parameters in (8.22) match the mass, momentum, and energy moments of f_i, so the self-collision operator conserves all three quantities. Summing the cross-collision identities over unordered pairs and then adding the self-collision identities proves total momentum and energy conservation. ■

8.5 H-Theorem and Equilibria

Proposition 8.8 (H-Theorem for the Haack–Hauck–Murillo Model)**.** *Assume $f_i > 0$, all relevant moments are finite, and $\lambda_{i,j} > 0$. For the space-homogeneous system*

$$\partial_t f_i = \sum_j \lambda_{i,j}(M_{i,j} - f_i), \tag{8.33}$$

the information functional

$$H[\boldsymbol{f}] = \sum_i \int_{\mathbb{R}^d} f_i \ln f_i \, \mathrm{d}\boldsymbol{v} \tag{8.34}$$

satisfies $\mathrm{d}H/\mathrm{d}t \leq 0$. Equivalently, the physical entropy $S[\boldsymbol{f}] := -H[\boldsymbol{f}]$ satisfies $\mathrm{d}S/\mathrm{d}t \geq 0$. Equality holds if and only if every species is a Maxwellian with a common velocity and temperature:

$$f_i(\boldsymbol{v}) = n_i \left(\frac{m_i}{2\pi\theta^\infty}\right)^{d/2} \exp\left(-\frac{m_i|\boldsymbol{v} - \boldsymbol{u}^\infty|^2}{2\theta^\infty}\right). \tag{8.35}$$

Proof. Species-mass conservation removes the derivative of the linear part of $f_i \ln f_i$, so

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_{i,j} \lambda_{i,j} \int (M_{i,j} - f_i) \ln f_i \, \mathrm{d}\boldsymbol{v}. \tag{8.36}$$

Group the cross-collision contributions by unordered pairs. For $i \neq j$, add and subtract the logarithms of the pair Maxwellians to obtain

$$\begin{aligned} \mathcal{D}_{i,j} = & -\lambda_{i,j} \int (f_i - M_{i,j}) \ln \frac{f_i}{M_{i,j}} \, \mathrm{d}\boldsymbol{v} - \lambda_{j,i} \int (f_j - M_{j,i}) \ln \frac{f_j}{M_{j,i}} \, \mathrm{d}\boldsymbol{v} \\ & - \lambda_{i,j} \int (f_i - M_{i,j}) \ln M_{i,j} \, \mathrm{d}\boldsymbol{v} - \lambda_{j,i} \int (f_j - M_{j,i}) \ln M_{j,i} \, \mathrm{d}\boldsymbol{v}. \end{aligned} \tag{8.37}$$

The last line vanishes. Indeed,

$$\ln M_{i,j} = c_i + \frac{m_i}{\hat{\theta}_{i,j}} \boldsymbol{u}_{i,j} \cdot \boldsymbol{v} - \frac{m_i}{2\hat{\theta}_{i,j}} |\boldsymbol{v}|^2, \tag{8.38}$$

where c_i is independent of $\boldsymbol{v}$. The constant terms vanish by species-mass conservation;

because the pair targets share $\boldsymbol{u}_{i,j}$ and $\hat{\theta}_{i,j}$, the remaining terms vanish by pairwise momentum and energy conservation in Proposition 8.7.

The elementary inequality $(x-y)\ln(x/y) \geq 0$ for $x, y > 0$ now shows that $\mathcal{D}_{i,j} \leq 0$. The identical argument with $M_{i,i}$ handles self-collisions. Hence $\mathrm{d}H/\mathrm{d}t \leq 0$.

Equality requires $f_i = M_{i,i}$ and $f_i = M_{i,j}$ for every interacting pair. Thus each f_i is Maxwellian, and every pair has the same velocity and temperature. Since all $\lambda_{i,j}$ are positive, the interaction graph is connected, so these pairwise equalities propagate to all species. Conversely, distributions of the form (8.35) make every collision term vanish. ■

The common equilibrium parameters are fixed by the conserved total momentum and energy:

$$\boldsymbol{u}^{\infty} = \frac{\sum_i \rho_i \boldsymbol{u}_i(0)}{\sum_i \rho_i}, \qquad \theta^{\infty} = \frac{2}{dn}\left[\sum_i \left(\frac{1}{2}\rho_i |\boldsymbol{u}_i(0)|^2 + \frac{d}{2} n_i \hat{\theta}_i(0)\right) - \frac{1}{2}\rho |\boldsymbol{u}^{\infty}|^2\right]. \tag{8.39}$$

8.6 Fluid-Limit Connection and Scope

Haack, Hauck, and Murillo also apply a Chapman–Enskog expansion to the scaled equation

$$\partial_t f_i + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} f_i + \boldsymbol{a}_i \cdot \nabla_{\boldsymbol{v}} f_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}(M_{i,j} - f_i). \tag{8.40}$$

The H-theorem implies that the leading term as $\varepsilon \to 0$ lies in the common-equilibrium manifold:

$$f_i^{(0)} = n_i \left(\frac{m_i}{2\pi\theta}\right)^{d/2} \exp\left(-\frac{m_i|\boldsymbol{v}-\boldsymbol{u}|^2}{2\theta}\right). \tag{8.41}$$

Thus the model supplies the same common mixture velocity and temperature required by the multispecies Boltzmann equilibrium. The first correction generates diffusion, viscous stress, and heat flux, with transport coefficients that depend on the selected collision frequencies. Those calculations are not needed for the homogeneous moment analysis in Chapter 9; the general closure logic is developed in Chapters 3 and 4.

The structural tradeoff should be explicit. The construction guarantees positive pair temperatures, species-mass conservation, total momentum and energy conservation, and the H-theorem for arbitrary positive rates. A single rate per ordered pair generally cannot simultaneously match every momentum- and temperature-relaxation rate of the full Boltzmann operator. The model used in this report prioritizes its conservative and entropic structure; application-specific information enters through the positive functions $\lambda_{i,j}$.

Chapter 9

Asymptotic Relaxation and Numerics of Multispecies Moment Equations

This chapter is based on two recently published papers:

1. Reference [33]: **E. Habbershaw**, R.S. Glasby, J.R. Haack, C.D. Hauck, and **S.M. Wise**, *Asymptotic Relaxation of Moment Equations for a Multi-Species, Homogeneous BGK Model*, *SIAM Journal on Applied Mathematics* 85 (2025), 294–313. https://doi.org/10.1137/23M1596314

   This paper establishes long-time dynamics of the multispecies BGK moment equations. Understanding these model characteristics is vital for designing and implementing sophisticated numerical approximation methods.

2. Reference [34]: **E. Habbershaw**, C.D. Hauck, and **S.M. Wise**, *Implicit Update of the Moment Equations for a Multi-Species, Homogeneous BGK Model*, *SIAM Journal on Numerical Analysis* 63 (2025), 881–907. https://doi.org/10.1137/24M165421X

   This paper establishes stability, well-posedness, and convergence results for an iterative solver for the multispecies, spatially homogeneous BGK model. This solver provides a basis for future full multispecies BGK and ES-BGK algorithms.

We study the space-homogeneous, multispecies BGK equations from [32], which can be understood as a special, illustrative case of the equations in [8] or [45]. We do so for collision frequencies that are independent of phase-space velocity but depend on number density and

temperature. In particular, we analyze the ODE system for the momentum and energy moments; each number density is constant by species-mass conservation. This reduced system is important both for asymptotic analysis and for the development of numerical methods.

Indeed, many numerical simulations of the full BGK system (including phase-space advection) rely on implicit-explicit (IMEX) methods that treat the space-homogeneous component implicitly. In this setting, the result of solving the associated moments equations first is to effectively linearize and diagonalize (with respect to the phase-space velocity) the required implicit solve, yielding a significant reduction in computational cost. (See [17, 61] for applications to the single-species case and [63, Section 4] for an extension to the multispecies setting when the collision frequencies depend only on the number densities.)

In this chapter, we first establish existence and uniqueness of the momentum-energy ODE system, which follows from standard theory once a positive lower bound can be established for the species temperatures. We show monotonicity of the minimum temperature envelope as well as upper and lower bounds on the bulk velocity. We then prove exponential decay of the species momenta and energies to their steady-state values, thereby generalizing some of the results in [18] to a system with an arbitrary number of species.

9.1 Model Equations

Here we recap the multispecies BGK equation, specifically for the case that the collision frequencies are dependent upon species temperatures. We denote by f_i the kinetic distribution of particles of species $i \in \{1, \dots, N\}$ having mass m_i. More specifically, $f_i(\boldsymbol{x}, \boldsymbol{v}, t)$ is the density of species i particles at the point $\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$, with microscopic velocity $\boldsymbol{v} \in \mathbb{R}^d$, at time $t \geq 0$, with respect to the measure $\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{x}$. Associated to each f_i are the species number density n_i, mass density ρ_i, bulk velocity $\boldsymbol{u}_i$, and temperature $\hat{\theta}_i$, defined by

$$\rho_i = m_i n_i = m_i \int_{\mathbb{R}^d} f_i \,\mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u}_i = \frac{1}{\rho_i} \int_{\mathbb{R}^d} m_i \boldsymbol{v} f_i \,\mathrm{d}\boldsymbol{v}, \qquad \hat{\theta}_i = \frac{m_i}{n_i d} \int_{\mathbb{R}^d} |\boldsymbol{v} - \boldsymbol{u}_i|^2 f_i \,\mathrm{d}\boldsymbol{v}. \quad (9.1)$$

The species momentum density is $\rho_i \boldsymbol{u}_i$. The species total energy density is

$$\omega_i = \frac{1}{2}\rho_i |\boldsymbol{u}_i|^2 + \frac{d}{2} n_i \hat{\theta}_i. \quad (9.2)$$

BGK models are expressed in terms of Maxwellian distributions that depend on these moments.

Definition 9.1 (Maxwellian)**.** *Given $n > 0$, $\boldsymbol{u} \in \mathbb{R}^d$, and $\vartheta > 0$, the **Maxwellian** $M_{n,\boldsymbol{u},\vartheta}$ is the*

function

$$M_{n,\boldsymbol{u},\vartheta}(\boldsymbol{v}) = \frac{n}{(2\pi\vartheta)^{\frac{d}{2}}} \exp\left(-\frac{|\boldsymbol{v}-\boldsymbol{u}|^2}{2\vartheta}\right). \tag{9.3}$$

Direct Gaussian integration gives the Maxwellian moments

$$\int_{\mathbb{R}^d} M_{n,\boldsymbol{u},\vartheta}\,\mathrm{d}\boldsymbol{v} = n, \qquad \int_{\mathbb{R}^d} \boldsymbol{v} M_{n,\boldsymbol{u},\vartheta}\,\mathrm{d}\boldsymbol{v} = n\boldsymbol{u}, \qquad \int_{\mathbb{R}^d} |\boldsymbol{v}|^2 M_{n,\boldsymbol{u},\vartheta}\,\mathrm{d}\boldsymbol{v} = n|\boldsymbol{u}|^2 + dn\vartheta. \tag{9.4}$$

We consider in this chapter the multispecies BGK equation from [32], given by

$$\frac{\partial f_i}{\partial t} + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} f_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}(M_{i,j} - f_i), \quad \forall i \in \{1,\dots,N\}, \tag{9.5}$$

where $\varepsilon > 0$ is the Knudsen number, $\lambda_{i,j} > 0$ is the (nondimensional) frequency of collisions between species i and j (independent of $\boldsymbol{v}$), and

$$M_{i,j}(\boldsymbol{v}) = M_{n_i,\boldsymbol{u}_{i,j},\hat{\theta}_{i,j}/m_i}(\boldsymbol{v})$$

is a Maxwellian defined by (9.3), using the mixture velocities and temperatures

$$\boldsymbol{u}_{i,j} = \frac{\rho_i\lambda_{i,j}\boldsymbol{u}_i + \rho_j\lambda_{j,i}\boldsymbol{u}_j}{\rho_i\lambda_{i,j} + \rho_j\lambda_{j,i}}, \quad \hat{\theta}_{i,j} = \frac{n_i\lambda_{i,j}\hat{\theta}_i + n_j\lambda_{j,i}\hat{\theta}_j}{n_i\lambda_{i,j} + n_j\lambda_{j,i}} + \frac{1}{d}\frac{\rho_i\rho_j\lambda_{i,j}\lambda_{j,i}}{\rho_i\lambda_{i,j} + \rho_j\lambda_{j,i}}\frac{|\boldsymbol{u}_i - \boldsymbol{u}_j|^2}{n_i\lambda_{i,j} + n_j\lambda_{j,i}}. \tag{9.6}$$

The existence and uniqueness of non-negative mild solutions to (a more general) multispecies BGK equation was proved in [44] for periodic spatial domains and the collision frequencies that depend on the number densities.

9.1.1 Model Properties

The BGK collision operators in (9.5) satisfy the following invariance properties: For any i, j,

$$\int_{\mathbb{R}^d} \lambda_{i,j}(M_{i,j} - f_i)\,\mathrm{d}\boldsymbol{v} = 0 \tag{9.7}$$

$$m_i\int_{\mathbb{R}^d} \lambda_{i,j}(M_{i,j} - f_i)\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} + m_j\int_{\mathbb{R}^d} \lambda_{j,i}(M_{j,i} - f_j)\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} = \boldsymbol{0} \tag{9.8}$$

$$m_i\int_{\mathbb{R}^d} \lambda_{i,j}(M_{i,j} - f_i)|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v} + m_j\int_{\mathbb{R}^d} \lambda_{j,i}(M_{j,i} - f_j)|\boldsymbol{v}|^2\,\mathrm{d}\boldsymbol{v} = 0, \tag{9.9}$$

corresponding to conservation of species mass, total momentum, and total energy, respectively. Furthermore, the information functional $H = \sum_i \int_{\mathbb{R}^d} f_i \ln f_i\,\mathrm{d}\boldsymbol{v}$ is non-increasing, while

the physical entropy $S = -H$ is non-decreasing: $\mathrm{d}H/\mathrm{d}t \leq 0$ and $\mathrm{d}S/\mathrm{d}t \geq 0$. Equality holds if and only if $f_i = M_{n_i, \boldsymbol{u}_{\mathrm{eq}}, \hat{\theta}_{\mathrm{eq}}/m_i}(\boldsymbol{v})$ for all $i \in \{1, \ldots, N\}$, where $\boldsymbol{u}_{\mathrm{eq}}$ and $\hat{\theta}_{\mathrm{eq}}$ are the equilibrium bulk velocity and energy temperature, respectively, which are common to all species [32].

The collision frequencies $\lambda_{i,j}$ are typically expressed as functions of the species moments given in (9.1). To define these collision frequencies, [32] presents a recipe based on matching either momentum or energy relaxation rates of the Boltzmann equation, given either a differential cross section or a momentum transfer cross section. In this paper, we consider the case of hard sphere collisions, for which the momentum transfer cross section is independent of microscopic velocity and given by [7]

$$\sigma_{\mathrm{MT}} = \pi \left(\frac{d_i + d_j}{2} \right)^2 = \pi d_{i,j}^2, \tag{9.10}$$

where d_i and d_j are reference diameters for the particles of species i and j; reference diameters for several species are given in [7] and [15]. Following the recipe for matching energy relaxation rates from Section 4.3 of [32], we obtain the following collision frequencies for hard sphere interactions:

$$\lambda_{i,j}^{\mathrm{HS}} = \frac{32\pi^2}{3(2\pi)^{3/2}} \frac{m_i m_j}{(m_i + m_j)^2} (d_i + d_j)^2 n_j \sqrt{\frac{\hat{\theta}_i}{m_i} + \frac{\hat{\theta}_j}{m_j}}. \tag{9.11}$$

While the formula in (9.11) is specific to the physically relevant case $d = 3$, collision frequencies can be derived in arbitrary dimensions in a similar fashion.

The dependence of the collision frequencies on the temperature makes the BGK model considered here more complicated than many models considered in the literature which assume that $\lambda_{i,j}$ is a constant or depends only on n_j. This is largely due to the non-Lipschitz nature of the square root function near zero and the fact that the frequencies are time-dependent functions. While we focus on the hard spheres model in this paper for simplicity, the analysis provided here is readily applied to other models in the literature [7] .

9.1.2 Moment Equations

The focus of the present work is the space-homogeneous multispecies BGK equation, obtained by removing the advection terms from (9.5). In this case $f_i = f_i(\boldsymbol{v}, t)$ satisfies

$$\frac{\partial f_i}{\partial t} = \frac{1}{\varepsilon} \sum_j \lambda_{i,j} (M_{i,j} - f_i), \quad i \in \{1, \ldots, N\}. \tag{9.12}$$

Multiplication of (9.12) by m_i, $m_i\boldsymbol{v}$, and $m_i|\boldsymbol{v}|^2$, followed by integration with respect to $\boldsymbol{v}$, gives equations for the mass density, momentum density, and energy density of each species

$$\frac{\mathrm{d}\rho_i}{\mathrm{d}t} = 0, \tag{9.13a}$$

$$\frac{\mathrm{d}(\rho_i \boldsymbol{u}_i)}{\mathrm{d}t} = \frac{1}{\varepsilon}\sum_j A_{i,j}(\boldsymbol{u}_j - \boldsymbol{u}_i), \tag{9.13b}$$

$$\frac{\mathrm{d}\omega_i}{\mathrm{d}t} = \frac{1}{\varepsilon}\sum_j B_{i,j}\left(\frac{\omega_j}{n_j} - \frac{\omega_i}{n_i}\right) + \frac{1}{2\varepsilon}\sum_j B_{i,j}S_{i,j}(m_i - m_j), \tag{9.13c}$$

where the equations for $\rho_i\boldsymbol{u}_i$ and ω_i are derived using the definitions for $\boldsymbol{u}_{i,j}$ and $\hat{\theta}_{i,j}$ from (9.6) and we define $N \times N$ matrices A, B, C, D, F, G, and S by

$$[A]_{i,j} = \frac{\rho_i\rho_j\lambda_{i,j}\lambda_{j,i}}{\rho_i\lambda_{i,j} + \rho_j\lambda_{j,i}}, \quad [B]_{i,j} = \frac{n_i n_j\lambda_{i,j}\lambda_{j,i}}{n_i\lambda_{i,j} + n_j\lambda_{j,i}}, \quad [S]_{i,j} = |\boldsymbol{u}_{i,j}|^2, \quad [C]_{i,j} = B_{i,j}S_{i,j}, \tag{9.14a}$$

$$[D]_{i,j} = \Big(\sum_k A_{i,k}\Big)\delta_{i,j}, \quad [F]_{i,j} = \Big(\sum_k B_{i,k}\Big)\delta_{i,j}, \quad [G]_{i,j} = \Big(\sum_k C_{i,k}\Big)\delta_{i,j}. \tag{9.14b}$$

According to (9.13a), each individual species mass density ρ_i (or, equivalently, number density n_i) is conserved (i.e., constant in time). However, only the *total* momentum and energy are conserved. Indeed, summing (9.13b) and (9.13c) over all species yields:

$$\sum_{i=1}^{N}\frac{\mathrm{d}(\rho_i\boldsymbol{u}_i)}{\mathrm{d}t} = \boldsymbol{0} \qquad \text{and} \qquad \sum_{i=1}^{N}\frac{\mathrm{d}\omega_i}{\mathrm{d}t} = 0, \tag{9.15}$$

respectively.

In terms of the vectorized quantities

$$\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N]^\top \in \mathbb{R}^{N\times d} \quad \text{and} \quad \boldsymbol{\omega} = [\omega_1, \ldots, \omega_N]^\top \in \mathbb{R}^N, \tag{9.16}$$

the equations in (9.13) take the form

$$P\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = -\frac{1}{\varepsilon}(D - A)\boldsymbol{U}, \tag{9.17a}$$

$$\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t} = -\frac{1}{\varepsilon}(F - B)Q^{-1}\boldsymbol{\omega} + \frac{1}{2\varepsilon}(G - C)M\boldsymbol{1}, \tag{9.17b}$$

where $P = \mathrm{diag}\{\rho_k\} \in \mathbb{R}^{N\times N}$, $Q = \mathrm{diag}\{n_k\} \in \mathbb{R}^{N\times N}$, and $M = \mathrm{diag}\{m_k\} \in \mathbb{R}^{N\times N}$. To

simplify the analysis later, it will be convenient to introduce variables

$$\boldsymbol{W} = P^{\frac{1}{2}}\boldsymbol{U} \in \mathbb{R}^{N\times d} \quad \text{and} \quad \boldsymbol{\xi} = Q^{-\frac{1}{2}}\boldsymbol{\omega} \in \mathbb{R}^{N}. \tag{9.18}$$

In terms of $\boldsymbol{W}$ and $\boldsymbol{\xi}$, (9.17) takes the form

$$\frac{\mathrm{d}\boldsymbol{W}}{\mathrm{d}t} = -\frac{1}{\varepsilon}Z\boldsymbol{W}, \tag{9.19a}$$

$$\frac{\mathrm{d}\boldsymbol{\xi}}{\mathrm{d}t} = -\frac{1}{\varepsilon}\widehat{Z}\boldsymbol{\xi} + \frac{1}{2\varepsilon}Q^{-\frac{1}{2}}(G-C)M\mathbf{1}, \tag{9.19b}$$

where

$$Z = P^{-\frac{1}{2}}(D-A)P^{-\frac{1}{2}} \in \mathbb{R}^{N\times N} \qquad \text{and} \qquad \widehat{Z} = Q^{-\frac{1}{2}}(F-B)Q^{-\frac{1}{2}} \in \mathbb{R}^{N\times N}. \tag{9.20}$$

9.2 Properties of the ODE System

The main result of this section is to show that there is a unique solution to the ODE system (9.17) for all physically meaningful initial conditions. When the collision frequencies are constant in time, such a result follows directly from standard ODE theory for systems with Lipschitz dynamics. However, because the collision frequencies in (9.11) depend on $\sqrt{\hat{\theta}_i}$, only a local Lipschitz condition can be obtained for the right-hand side of (9.17). Thus the key step is to show that the species temperatures are bounded below away from zero. This bound will also be important for establishing exponential decay to steady state in Section 9.3. We also establish upper and lower bounds on the bulk velocity components that will be used in the decay estimates.

9.2.1 Positivity of the Temperature

To make the following proofs easier to manage, we introduce the parameters

$$\alpha_{i,j} = \frac{\rho_i\lambda_{i,j}}{\rho_i\lambda_{i,j} + \rho_j\lambda_{j,i}} \qquad \text{and} \qquad \beta_{i,j} = \frac{n_i\lambda_{i,j}}{n_i\lambda_{i,j} + n_j\lambda_{j,i}}, \tag{9.21}$$

which satisfy the conditions $\alpha_{i,j} + \alpha_{j,i} = 1 = \beta_{i,j} + \beta_{j,i}$. With these parameters, the mixture values in (9.6) can be expressed as

$$\boldsymbol{u}_{i,j} = \alpha_{i,j}\boldsymbol{u}_i + \alpha_{j,i}\boldsymbol{u}_j \qquad \text{and} \qquad \hat{\theta}_{i,j} = \beta_{i,j}\hat{\theta}_i + \beta_{j,i}\hat{\theta}_j + \frac{1}{d}m_i\alpha_{j,i}\beta_{i,j}\left|\boldsymbol{u}_i - \boldsymbol{u}_j\right|^2. \tag{9.22}$$

Lemma 9.2. *The temperatures $\hat{\theta}_i$, $i \in \{1, \dots, N\}$, satisfy the ODE*

$$\frac{\mathrm{d}\hat{\theta}_i}{\mathrm{d}t} = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\beta_{j,i}(\hat{\theta}_j - \hat{\theta}_i) + \frac{1}{\varepsilon d}\sum_j \lambda_{i,j} m_i \alpha_{j,i}(\alpha_{j,i} + \beta_{i,j})|\boldsymbol{u}_i - \boldsymbol{u}_j|^2. \tag{9.23}$$

Proof. The proof is a direct calculation. From (9.13b) and the fact that $A_{i,j} = \alpha_{j,i}\rho_i\lambda_{i,j}$, it follows that

$$\rho_i \frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{u}_i|^2 = \frac{1}{\varepsilon}\sum_j \rho_i\lambda_{i,j}(2\alpha_{j,i}\langle \boldsymbol{u}_i, \boldsymbol{u}_j\rangle - 2\alpha_{j,i}|\boldsymbol{u}_i|^2), \tag{9.24}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean inner product on $\mathbb{R}^d$. The equation in (9.23) can be obtained by (i) using (9.2) to express (9.13c) in terms of the temperatures and velocities; (ii) invoking (9.24) to eliminate the time derivative of $|\boldsymbol{u}_i|^2$; (iii) using the formula in (9.22) to replace $\boldsymbol{u}_{i,j}$ wherever it appears; and (iv) applying the elementary relations $\alpha_{j,i}^2 = \alpha_{j,i} - \alpha_{i,j}\alpha_{j,i} = 2\alpha_{j,i} + \alpha_{i,j}^2 - 1$. ∎

Theorem 9.3. *Suppose that, for some $t_\mathrm{f} > 0$, the system* (9.17) *has a local solution*

$$\boldsymbol{U} \in C^1\left([0, t_\mathrm{f}]; \mathbb{R}^{N\times d}\right), \qquad \boldsymbol{\omega} \in C^1\left([0, t_\mathrm{f}]; \mathbb{R}^N\right),$$

with initial conditions $\boldsymbol{U}(0)$ and $\boldsymbol{\omega}(0)$ such that $\hat{\theta}_i(0) > 0$ for every $i \in \{1, \dots, N\}$. Then $\hat{\theta}_i(t) \geq \hat{\theta}_\mathrm{min} \coloneqq \min_k\{\hat{\theta}_k(0)\}$ for all $t \in [0, t_\mathrm{f}]$ and all $i \in \{1, \dots, N\}$.

Proof. The proof relies on an integrating factor technique that yields a Grönwall-type estimate. Define a lower temperature envelope $\hat{\theta}_\star(t) = \min_k\{\hat{\theta}_k(t)\}$. From (9.23) it follows that, for $i \in \{1, \dots, N\}$,

$$\frac{\mathrm{d}\hat{\theta}_i}{\mathrm{d}t} + c_i(t)\hat{\theta}_i(t) \geq c_i(t)\hat{\theta}_\star(t), \qquad \text{where} \qquad c_i(t) = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}(t)\beta_{j,i}(t). \tag{9.25}$$

Let $a_i(t) = \int_0^t c_i(s)\,\mathrm{d}s$. Then $a_i'(t) = c_i(t)$ and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{a_i(t)}\hat{\theta}_i(t)\right) = e^{a_i(t)}\left[\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta}_i(t) + c_i(t)\hat{\theta}_i(t)\right] \overset{(9.25)}{\geq} e^{a_i(t)}c_i(t)\hat{\theta}_\star(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{a_i(t)}\right)\hat{\theta}_\star(t). \tag{9.26}$$

For each $\gamma > 0$, define the set

$$S_\gamma = \left\{ t \in [0, t_\mathrm{f}] \;\middle|\; \hat{\theta}_\star(t) < \frac{1}{1+\gamma}\hat{\theta}_\star(0) \right\}. \tag{9.27}$$

We now argue by contradiction. Assume that S_γ is nonempty and let $t_\star = \inf S_\gamma$. Since $\hat{\theta}_\star(0)$

lies strictly above the threshold, $t_\star > 0$. Continuity gives

$$\hat{\theta}_\star(t_\star) = \frac{1}{1+\gamma}\hat{\theta}_\star(0) < \hat{\theta}_\star(0). \tag{9.28}$$

However, because every $\hat{\theta}_i$ is continuous on $[0, t_\mathrm{f}]$, their pointwise minimum $\hat{\theta}_\star$ is also continuous. The definition of the first crossing time therefore gives $\hat{\theta}_\star(t) \geq \hat{\theta}_\star(t_\star)$ for every $t \in [0, t_\star]$ which, along with (9.26), implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{a_i(t)}\hat{\theta}_i(t)\right) \geq \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{a_i(t)}\right)\hat{\theta}_\star(t_\star), \qquad \forall t \in [0, t_\star]. \tag{9.29}$$

Integrating both sides of (9.29) from $[0, t_\star]$ and then multiplying the result by $e^{-a_i(t_\star)} \geq 0$ gives

$$\hat{\theta}_i(t_\star) \geq \hat{\theta}_i(0)e^{-a_i(t_\star)} + \hat{\theta}_\star(t_\star)\left(1 - e^{-a_i(t_\star)}\right) \geq \hat{\theta}_\star(0)e^{-a_i(t_\star)} + \hat{\theta}_\star(t_\star)\left(1 - e^{-a_i(t_\star)}\right). \tag{9.30}$$

Define $i_\star$ such that $\hat{\theta}_{i_\star}(t_\star) = \hat{\theta}_\star(t_\star)$. If $i = i_\star$, then (9.30) becomes

$$\hat{\theta}_\star(t_\star) \equiv \hat{\theta}_{i_\star}(t_\star) \geq \hat{\theta}_\star(0)e^{-a_{i_\star}(t_\star)} + \hat{\theta}_\star(t_\star)\left(1 - e^{-a_{i_\star}(t_\star)}\right), \tag{9.31}$$

which, after some simple algebra, implies that $\hat{\theta}_\star(t_\star) \geq \hat{\theta}_\star(0)$. However this results contradicts (9.28) which means that no such $t_\star$ exists and the set S_γ must be empty. As a consequence, $\hat{\theta}_\star(t) \geq \frac{1}{1+\gamma}\hat{\theta}_\star(0)$ for all $t \in [0, t_\mathrm{f}]$. Since $\gamma > 0$ is arbitrary, the limit $\gamma \to 0$ yields $\hat{\theta}_\star(t) \geq \hat{\theta}_\star(0) = \hat{\theta}_\mathrm{min}$ for all $t \in [0, t_\mathrm{f}]$. The proof is complete. ■

9.2.2 Velocity Bounds

For each component $\mathrm{j} \in \{1, \ldots, d\}$, define the lower and upper velocity envelopes

$$\underline{u}_k(t) = \min_j\{\boldsymbol{U}_{j,k}(t)\}, \qquad \overline{u}_k(t) = \max_j\{\boldsymbol{U}_{j,k}(t)\}, \tag{9.32}$$

and the components of the vector $\boldsymbol{u}_\mathrm{max} \in \mathbb{R}^d$

$$[\boldsymbol{u}_\mathrm{max}(t)]_k = \max\left\{|\underline{u}_k(t)|, |\overline{u}_k(t)|\right\}. \tag{9.33}$$

Theorem 9.4. *Suppose that the assumptions of Theorem 9.3 hold. Then for each* $\mathrm{j} \in$

$\{1, \ldots, d\}$, *the velocity components follow the inequalities*

$$\underline{u}_k(0) \leq \underline{u}_k(t) \leq \boldsymbol{U}_{i,k}(t) \leq \overline{u}_k(t) \leq \overline{u}_k(0) \tag{9.34}$$

for all $t \in [0, t_\mathrm{f}]$ *and all* $i \in \{1, \ldots, N\}$. *Further, for any* $t \in [0, t_\mathrm{f}]$, *and all* $i \in \{1, \ldots, N\}$,

$$\|\boldsymbol{u}_i(t)\|_2 \leq \|\boldsymbol{u}_{\max}(t)\|_2 \leq \|\boldsymbol{u}_{\max}(0)\|_2 =: u_{\max}. \tag{9.35}$$

Proof. The proof of the inequalities in (9.34) are similar to the proof of the temperature lower bound given in Theorem 9.3. From the velocity equation, (9.13b),

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}_{i,k} + \left(\frac{1}{\varepsilon}\sum_j \lambda_{i,j}\alpha_{j,i}\right)\boldsymbol{U}_{i,k} = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\alpha_{j,i}\boldsymbol{U}_{j,k}, \tag{9.36}$$

we obtain, using the definitions $\boldsymbol{U}_{j,k}(t) \leq \overline{u}_k(t)$, and $c_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\alpha_{j,i}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}_{i,k} + c_i(t)\boldsymbol{U}_{i,k} \leq c_i(t)\overline{u}_k(t). \tag{9.37}$$

From this point the proof of the upper bound in (9.34) follows that of Theorem 9.3, with (9.37) being the analog of (9.25). The proof of the lower bound is even more similar.

Finally, the bound in (9.35) follows from (9.34). Specifically, if a, b, and c are real-valued scalars such that $a \leq b \leq c$, then $|b| \leq \max\{|a|, |c|\}$. Hence

$$\|\boldsymbol{u}_i\|_2^2 = \sum_{k=1}^{d} \boldsymbol{U}_{i,k}^2 \leq \sum_{k=1}^{d} \max\left\{|\underline{u}_k(0)|, |\overline{u}_k(0)|\right\}^2 = \|\boldsymbol{u}_{\max}(0)\|_2^2 = u_{\max}^2. \tag{9.38}$$

■

9.2.3 Existence and Uniqueness

According to (9.13a), the number densities n_i and mass densities ρ_i are non-negative constants in time:

$$n_i(0) = n_i(t), \quad \rho_i(0) = \rho_i(t) = m_i n_i(t), \quad \forall\, i \in \{1, \ldots, N\}, \quad t \in [0, t_\mathrm{f}]. \tag{9.39}$$

To avoid degenerate cases, we assume

$$n_{\min} := \min_j n_j(0) > 0, \qquad \rho_{\min} := \min_j \rho_j(0) > 0.$$

Then, given any velocity-energy pair $(\boldsymbol{u}_i, \omega_i)$, the associated temperature $\hat{\theta}_i$ is given by (cf. (9.1))

$$\hat{\theta}_i = \vartheta_i(\boldsymbol{u}_i, \omega_i) := \frac{2}{dn_i}\omega_i - \frac{m_i}{d}|\boldsymbol{u}_i|^2, \tag{9.40}$$

and the set of all realizable velocity and energy states is given by

$$\mathcal{R} = \left\{(\boldsymbol{U}, \boldsymbol{\omega}) \in \mathbb{R}^{N\times d} \times \mathbb{R}^N \;\middle|\; \vartheta_i(\boldsymbol{u}_i, \omega_i) \geq 0,\ \forall\, i \in \{1, \dots, N\}\right\}, \tag{9.41}$$

where ϑ_i is defined in (9.40). Given positive scalars $\hat{\theta}_{\min} \in (0, \infty)$ and $\omega_{\text{tot}} \in (0, \infty)$, let

$$\mathcal{D}(\hat{\theta}_{\min}, \omega_{\text{tot}}) = \left\{(\boldsymbol{U}, \boldsymbol{\omega}) \in \mathcal{R} \;\middle|\; \sum_{i=1}^{N} \omega_i = \omega_{\text{tot}} \quad \text{and} \quad \vartheta_i(\boldsymbol{u}_i, \omega_i) \geq \hat{\theta}_{\min},\ \forall\, i \in \{1, \dots, N\}\right\}. \tag{9.42}$$

Recall that Theorem 9.4 gives

$$|\boldsymbol{u}_i| \leq u_{\max} := \|\boldsymbol{u}_{\max}(0)\|_2 \leq \sqrt{\frac{2\omega_{\text{tot}}}{\rho_{\min}}}$$

for all $(\boldsymbol{U}, \boldsymbol{\omega}) \in \mathcal{D}(\hat{\theta}_{\min}, \omega_{\text{tot}})$. Thus $\mathcal{D}(\hat{\theta}_{\min}, \omega_{\text{tot}})$ is a closed and bounded subset of $\mathbb{R}^{N\times d}\times\mathbb{R}^N$.

Lemma 9.5. *Let $\hat{\theta}_{\min} \in (0, \infty)$ and $\omega_{\text{tot}} \in (0, \infty)$ be given. Suppose that for some $t_{\text{f}} > 0$, there exists a local solution $\boldsymbol{U} \in C^1\left([0, t_{\text{f}}]; \mathbb{R}^{N\times d}\right)$ and $\boldsymbol{\omega} \in C^1\left([0, t_{\text{f}}]; \mathbb{R}^N\right)$ of the system in (9.17) with initial condition $(\boldsymbol{U}^0, \boldsymbol{\omega}^0) \in \mathcal{D}(\hat{\theta}_{\min}, \omega_{\text{tot}})$. Then $(\boldsymbol{U}(t), \boldsymbol{\omega}(t)) \in \mathcal{D}(\hat{\theta}_{\min}, \omega_{\text{tot}})$ for all $t \in [0, t_{\text{f}}]$. In particular, $\hat{\theta}_{\min} \leq \hat{\theta}_i(t) \leq \hat{\theta}_{\max} := \frac{2}{dn_{\min}}\omega_{\text{tot}}$ for all $t \in [0, t_{\text{f}}]$.*

Proof. To conclude that $(\boldsymbol{U}(t), \boldsymbol{\omega}(t)) \in \mathcal{D}$ for all $t \in [0, t_{\text{f}}]$, two conditions must be satisfied. The first condition:

$$\sum_{i=1}^{N} \omega_i(t) = \sum_{i=1}^{N} \omega_i(0) = \omega_{\text{tot}}, \quad \forall\, t \in [0, t_{\text{f}}], \tag{9.43}$$

follows immediately from (9.15), and the upper bound $\hat{\theta}_i(t) \leq \hat{\theta}_{\max}$ follows:

$$\hat{\theta}_i(t) = \frac{2}{dn_i}\omega_i(t) - \frac{m_i}{d}|\boldsymbol{u}_i(t)|^2 \leq \frac{2}{dn_i}\omega_i(t) \leq \frac{2}{dn_{\min}}\omega_{\text{tot}} = \hat{\theta}_{\max}. \tag{9.44}$$

The second condition:

$$\hat{\theta}_{\min} \leq \hat{\theta}_i(t) = \vartheta_i(\boldsymbol{u}_i(t), \omega_i(t)), \quad \forall\, t \in [0, t_{\text{f}}], \tag{9.45}$$

is a direct consequence of Theorem 9.3. ∎

Theorem 9.6. *Suppose that $m_i > 0$ and $n_i > 0$ for all $i \in \{1, \dots, N\}$. Then for any $(\boldsymbol{U}^0, \boldsymbol{\omega}^0) \in \operatorname{int} \mathcal{R}$, there exists a global, unique solution $(\boldsymbol{U}, \boldsymbol{\omega}) \in C^1([0,\infty); \mathcal{R})$ of the system (9.17) with the initial conditions $(\boldsymbol{U}(0), \boldsymbol{\omega}(0)) = (\boldsymbol{U}^0, \boldsymbol{\omega}^0)$. Moreover, the associated temperatures $\hat{\theta}_i(t) = \vartheta_i(\boldsymbol{u}_i(t), \omega_i(t))$ are bounded below by their initial values; that is*

$$\min_i \hat{\theta}_i(t) \geq \min_i \hat{\theta}_i(0), \qquad \forall t > 0. \tag{9.46}$$

Proof. The system (9.17) can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\boldsymbol{U}\\ \boldsymbol{\omega}\end{pmatrix} = \boldsymbol{f}(\boldsymbol{U}, \boldsymbol{\omega}) \quad \text{where} \quad \boldsymbol{f}(\boldsymbol{U}, \boldsymbol{\omega}) := \begin{pmatrix} -\frac{1}{\varepsilon} P^{-1}(D - A)\boldsymbol{U} \\ -\frac{1}{\varepsilon}(F - B)Q^{-1}\boldsymbol{\omega} + \frac{1}{2\varepsilon}(G - C)M\mathbf{1}\end{pmatrix}. \tag{9.47}$$

Let $\hat{\theta}_{\min} = \min_i \hat{\theta}_i(0)$ and $\omega_{\text{tot}} = \sum_i \omega_i(0)$. Since $(\boldsymbol{U}, \boldsymbol{\omega}) \in \operatorname{int} \mathcal{R}$, it follows that $\hat{\theta}_{\min} > 0$ and $\omega_{\text{tot}} > 0$. Moreover, there exists an $\eta > 0$ such that the closed η-neighborhood of $\mathcal{D}(\hat{\theta}_{\min}, \omega_{\text{tot}})$, denoted by $\mathcal{D}_\eta(\hat{\theta}_{\min}, \omega_{\text{tot}})$, is contained in $\operatorname{int} \mathcal{R}$; that is, $\mathcal{D}_\eta(\hat{\theta}_{\min}, \omega_{\text{tot}}) \subset \operatorname{int} \mathcal{R}$. In particular, $\min_i \hat{\theta}_i = \min_i \vartheta_i(\boldsymbol{u}_i, \omega_i)$ is bounded below on $\mathcal{D}_\eta(\hat{\theta}_{\min}, \omega_{\text{tot}})$. Hence, $\boldsymbol{f}$ is Lipschitz on $\mathcal{D}_\eta(\hat{\theta}_{\min}, \omega_{\text{tot}})$ with Lipschitz constant $L > 0$ and satisfies $\|\boldsymbol{f}(\boldsymbol{U}, \boldsymbol{\omega})\|_2 \leq \mathcal{M} < \infty$ there. Since $(\boldsymbol{U}^0, \boldsymbol{\omega}^0) \in \mathcal{D}$, there is a number $\beta = \beta(\eta) > 0$, independent of the point $(\boldsymbol{U}^0, \boldsymbol{\omega}^0)$, such that the closed ball $\overline{B}\left((\boldsymbol{U}^0, \boldsymbol{\omega}^0), \beta\right) \subset \mathcal{D}_\eta$. By the Picard–Lindelöf theorem, Theorem B.1, there exists a unique solution on the interval $[0, \delta]$, where $\delta = \min\left(\frac{1}{2L}, \frac{\beta}{\mathcal{M}}\right)$. By Lemma 9.5, $(\boldsymbol{U}(t), \boldsymbol{\omega}(t)) \in \mathcal{D}$ for all $t \in [0, \delta]$. Since $(\boldsymbol{U}(\delta), \boldsymbol{\omega}(\delta)) \in \mathcal{D}$, we can apply the Picard–Lindelöf theorem again with initial value $(\boldsymbol{U}(\delta), \boldsymbol{\omega}(\delta))$ and extend the solution to $[0, 2\delta]$. The same argument can be repeated indefinitely. Crucially, the uniform constants L, $\mathcal{M}$, and β prevent the continuation interval δ from shrinking. The result follows. ■

9.3 Long-time Behavior

Having proven the existence of unique, global, physically realizable solutions for (9.17), we now seek to characterize the asymptotic behavior of those solutions. In particular, we want to show that solutions converge to

$$\boldsymbol{U}^\infty(t) = \mathbf{1}\,(\boldsymbol{u}^\infty)^\top \qquad \text{and} \qquad \boldsymbol{\omega}^\infty(t) = \frac{|\boldsymbol{u}^\infty(t)|^2}{2} P\mathbf{1} + \frac{d\hat{\theta}^\infty(t)}{2} Q\mathbf{1}, \tag{9.48}$$

where

$$\boldsymbol{u}^\infty(t) = \frac{\sum_i \rho_i \boldsymbol{u}_i(t)}{\sum_i \rho_i} \qquad \text{and} \qquad \hat{\theta}^\infty(t) = \frac{\sum_i n_i \hat{\theta}_i(t)}{\sum_i n_i} + \frac{\sum_i \rho_i (|\boldsymbol{u}_i(t)|^2 - |\boldsymbol{u}^\infty(t)|^2)}{d \sum_i n_i}. \tag{9.49}$$

are constant functions of time.

Proposition 9.7. *Suppose that* $(\boldsymbol{U}, \boldsymbol{\omega}) \in C^1([0,\infty);\mathcal{R})$ *is a unique global-in-time solution to the system* (9.17a)-(9.17b) *with assumptions as given in Theorem 9.6. Further let* $\hat{\boldsymbol{\theta}} \in C^1\left([0,\infty);\mathbb{R}^N\right)$ *be the associated temperature vector, whose components satisfy* (9.23)*. Then* $\boldsymbol{u}^\infty$, $\hat{\theta}^\infty$, *and* $\boldsymbol{\omega}^\infty$ *are time invariant quantities, and* $\hat{\theta}^\infty > 0$.

Proof. By conservation of mass, (9.13a), and total momentum, (9.15), $\sum_i \rho_i$ and $\sum_i \rho_i \boldsymbol{u}_i$ are constant. Thus $\boldsymbol{u}^\infty$ is time invariant. To show that $\hat{\theta}^\infty$ is constant in time, multiply (9.49) by the expression $\frac{d}{2}\sum_i n_i$, which is constant by (9.13a), to obtain

$$\frac{d}{2}\left(\sum_i n_i\right)\hat{\theta}^\infty = \sum_i \left[\frac{d}{2} n_i \hat{\theta}_i + \frac{1}{2}\rho_i |\boldsymbol{u}_i|^2\right] - \frac{1}{2}\left(\sum_i \rho_i\right)|\boldsymbol{u}^\infty|^2 = \sum_i \omega_i - \frac{1}{2}\left(\sum_i \rho_i\right)|\boldsymbol{u}^\infty|^2. \tag{9.50}$$

The first term on the right-hand side of (9.50) is constant by conservation of total energy, (9.15); the second term is constant by conservation of species mass, (9.13a), and the fact that $\boldsymbol{u}^\infty$ is constant. Thus $\hat{\theta}^\infty$ is time invariant. It follows immediately from (9.48) that $\boldsymbol{\omega}^\infty$ is also time invariant.

By Theorem 9.6, $\hat{\theta}_i(t) \geq \hat{\theta}_{\min} > 0$ for every species and all $t \geq 0$. Therefore the first term in (9.49) is strictly positive. To conclude that $\hat{\theta}^\infty > 0$, it remains only to show that $\sum_i \rho_i(|\boldsymbol{u}_i|^2 - |\boldsymbol{u}^\infty|^2) \geq 0$. Let $r_i = \frac{\rho_i}{\sum_k \rho_k} > 0$. Then $\sum_i r_i = 1$ and $|\boldsymbol{u}^\infty|^2 = (\sum_i r_i \boldsymbol{u}_i)^\top (\sum_j r_j \boldsymbol{u}_j) = \sum_i \sum_j r_i r_j \boldsymbol{u}_i^\top \boldsymbol{u}_j$. Therefore

$$\sum_i \rho_i(|\boldsymbol{u}_i|^2 - |\boldsymbol{u}^\infty|^2) = \sum_k \rho_k \left[\sum_{i,j} r_i r_j |\boldsymbol{u}_i|^2 - \sum_{i,j} r_i r_j \boldsymbol{u}_i^\top \boldsymbol{u}_j\right] = \frac{1}{2}\sum_k \rho_k \sum_{i,j} r_i r_j |\boldsymbol{u}_i - \boldsymbol{u}_j|^2 \geq 0. \tag{9.51}$$

■

Remark 9.8. *The quantities* $\boldsymbol{u}^\infty$ *and* $\hat{\theta}^\infty$ *were referred to as the mixture mean velocity and mixture temperature, respectively, in [31]. To agree with the conventions of Chapter 2, we call them the equilibrium velocity and equilibrium species-frame temperature. The symbols* $\boldsymbol{u}_{i,j}$ *and* $\hat{\theta}_{i,j}$ *denote pairwise BGK target parameters; the unqualified terms mixture velocity and mixture temperature remain reserved for* $\boldsymbol{u}$ *and* θ.

The convergence of $\boldsymbol{u}_i$, $\hat{\theta}_i$, and $\boldsymbol{\omega}$ as $t \to \infty$ is established by the following result. To state the estimate without excluding coincident decay rates, we first introduce a continuous two-rate kernel.

Definition 9.9 (Two-Rate Decay Kernel)**.** *For $a, b > 0$ and $t \geq 0$, define the* ***two-rate decay kernel***

$$\mathcal{K}_{a,b}(t) = \begin{cases} \dfrac{e^{-at/\varepsilon} - e^{-bt/\varepsilon}}{b-a}, & a \neq b, \\ \dfrac{t}{\varepsilon} e^{-at/\varepsilon}, & a = b. \end{cases} \tag{9.52}$$

The second line is the continuous extension of the first as $b \to a$.

Theorem 9.10. *Under the assumptions of Proposition 9.7, for all $i \in \{1, \dots, N\}$,*

$$\lim_{t\to\infty} \boldsymbol{u}_i(t) = \boldsymbol{u}^\infty, \qquad \lim_{t\to\infty} \hat{\theta}_i(t) = \hat{\theta}^\infty, \qquad \text{and} \qquad \lim_{t\to\infty} \boldsymbol{\omega}(t) = \boldsymbol{\omega}^\infty, \tag{9.53}$$

and bounds for the decay rates of the bulk velocities and energies are given by

$$\|\boldsymbol{u}_i(t) - \boldsymbol{u}^\infty\|_2 \leq C_{\boldsymbol{U}} \exp\left(-\frac{z_{\min}}{\varepsilon} t\right), \quad i \in \{1, \dots, N\} \tag{9.54}$$

$$\|\boldsymbol{\omega}(t) - \boldsymbol{\omega}^\infty\|_2 \leq C_1 e^{-\frac{\widehat{z}_{\min}}{\varepsilon} t} + C_2 \mathcal{K}_{z_{\min}, \widehat{z}_{\min}}(t), \tag{9.55}$$

where $z_{\min} > 0$ and $\widehat{z}_{\min} > 0$ are lower bounds on the positive eigenvalues of Z and $\widehat{Z}$, respectively, and $C_{\boldsymbol{U}} > 0$, $C_1 > 0$, and $C_2 > 0$ are constants depending only on the initial conditions of the system.

9.3.1 Null Spaces of $D - A$, Z, $F - B$, and $\widehat{Z}$

To prove Theorem 9.10, we first characterize the null spaces of the matrices $D-A$, Z, $F-B$, and $\widehat{Z}$, which are defined in (9.14), and (9.20).

Lemma 9.11. *The matrices $D - A$ and Z are symmetric positive semi-definite (SPSD), each with a one dimensional null space. In particular, $\mathcal{N}(D - A) = \operatorname{span}(\{\mathbf{1}\})$ and $\mathcal{N}(Z) = \operatorname{span}\left(\left\{P^{\frac{1}{2}}\mathbf{1}\right\}\right)$. Moreover, the null space and range space of Z are invariant with respect to time.*

Proof. The matrix A is symmetric by its definition, and D is diagonal; hence $D - A$ is symmetric. The congruence defining Z then makes Z symmetric as well. For any $\boldsymbol{y} \in \mathbb{R}^N$,

$$\boldsymbol{y}^\top (D - A)\boldsymbol{y} = \frac{1}{2} \sum_{i,j} A_{i,j} (y_i - y_j)^2 \geq 0. \tag{9.56}$$

Moreover, since $A_{i,j} > 0$, $\frac{1}{2}\sum_{i,j} A_{i,j}(y_i - y_j)^2 = 0$ if and only if $y_i = y_j$ for all i and j, in which case $\boldsymbol{y} = c\mathbf{1}$ for some $c \in \mathbb{R}$. Thus $D - A$ is SPSD, with a one-dimensional null

space spanned by the eigenvector $\mathbf{1}$. Similarly $\boldsymbol{y}^\top Z\boldsymbol{y} = (P^{-\frac{1}{2}}\boldsymbol{y})^\top (D-A)(P^{-\frac{1}{2}}\boldsymbol{y}) \geq 0$, and $Z\boldsymbol{y} = \mathbf{0} \iff P^{-\frac{1}{2}}\boldsymbol{y} = c\mathbf{1} \iff \boldsymbol{y} = cP^{\frac{1}{2}}\mathbf{1}$, for some $c \in \mathbb{R}$. Thus Z is SPSD, with a one dimensional null space spanned by the eigenvector $P^{\frac{1}{2}}\mathbf{1}$. By conservation of mass, (9.13a), $P^{\frac{1}{2}}\mathbf{1}$ is independent of time. Thus the null space of Z is invariant with respect to time. The symmetry of Z implies that $\mathcal{R}(Z) = \mathcal{N}(Z)^\perp$ is also invariant. ■

Lemma 9.12. *The matrices $F-B$ and $\widehat{Z}$ are symmetric positive semi-definite (SPSD). In particular $\mathcal{N}(F-B) = \operatorname{span}(\{\mathbf{1}\})$ and $\mathcal{N}(\widehat{Z}) = \operatorname{span}\left(\left\{Q^{\frac{1}{2}}\mathbf{1}\right\}\right)$. Moreover, the null space and range space of $\widehat{Z}$ are invariant with respect to time.*

Proof. The matrices $F-B$ and $\widehat{Z}$ are symmetric. For every $\boldsymbol{y} \in \mathbb{R}^N$,

$$\boldsymbol{y}^\top (F-B)\boldsymbol{y} = \frac{1}{2}\sum_{i,j} B_{i,j}(y_i - y_j)^2 \geq 0. \tag{9.57}$$

Because $B_{i,j} > 0$ for $i \neq j$, equality holds exactly when all components of $\boldsymbol{y}$ are equal. Hence $F-B$ is SPSD and $\mathcal{N}(F-B) = \operatorname{span}\{\mathbf{1}\}$. The congruence relation

$$\boldsymbol{y}^\top \widehat{Z}\boldsymbol{y} = \left(Q^{-1/2}\boldsymbol{y}\right)^\top (F-B)\left(Q^{-1/2}\boldsymbol{y}\right) \tag{9.58}$$

then shows that $\widehat{Z}$ is SPSD and $\mathcal{N}(\widehat{Z}) = \operatorname{span}\{Q^{1/2}\mathbf{1}\}$. The matrix Q is constant in time by species-mass conservation, so this null space is time invariant. Finally, symmetry gives $\mathcal{R}(\widehat{Z}) = \mathcal{N}(\widehat{Z})^\perp$, and the range is therefore time invariant as well. ■

9.3.2 Velocity Relaxation Proof

Theorem 9.13. *Under the assumptions of Proposition 9.7, $\lim_{t\to\infty} \boldsymbol{u}_i(t) = \boldsymbol{u}^\infty$ and*

$$\|\boldsymbol{u}_i(t) - \boldsymbol{u}^\infty\|_2 \leq C_{\boldsymbol{U}} \exp\left(-\frac{z_{\min}}{\varepsilon}t\right), \quad i = 1, \ldots, N, \tag{9.59}$$

where $C_{\boldsymbol{U}} > 0$ is a constant that depends on the initial conditions of the system and $z_{\min} > 0$ is a lower bound on the positive eigenvalues of Z.

Proof. The proof is based on (9.19a) which is equivalent to (9.17a). Let $\boldsymbol{W}^\infty = P^{\frac{1}{2}}\boldsymbol{U}^\infty$, where $\boldsymbol{U}^\infty = \mathbf{1}\left(\boldsymbol{u}^\infty\right)^\top$ is given in (9.48). Then Lemma 9.11 implies that $Z\boldsymbol{W}^\infty = \mathbf{0} \in \mathbb{R}^{N\times d}$. Moreover, since P and $\boldsymbol{U}^\infty$ are independent of time, so too is $\boldsymbol{W}^\infty$. Thus in terms of $\widetilde{\boldsymbol{W}} := \boldsymbol{W} - \boldsymbol{W}^\infty$, (9.19a) takes the form

$$\frac{\mathrm{d}\widetilde{\boldsymbol{W}}}{\mathrm{d}t} = -\frac{1}{\varepsilon}Z\widetilde{\boldsymbol{W}}. \tag{9.60}$$

Write the element $\widetilde{\boldsymbol{W}}$ as a sum of the null and range space components with respect to Z: $\widetilde{\boldsymbol{W}} = \widetilde{\boldsymbol{W}}_{\mathcal{R}} + \widetilde{\boldsymbol{W}}_{\mathcal{N}}$. As shown in Appendix D, $\widetilde{\boldsymbol{W}}_{\mathcal{N}} = 0$. Hence $\widetilde{\boldsymbol{W}} = V\boldsymbol{A}$, where the columns of $V = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_{N-1}] \in \mathbb{R}^{N\times(N-1)}$ are a set of orthonormal eigenvectors of Z that span $\mathcal{R}(Z)$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_d] \in \mathbb{R}^{(N-1)\times d}$. Using this formulation, the ODE system (9.60) becomes

$$\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} = -\frac{1}{\varepsilon} V^\top Z V \boldsymbol{A}. \tag{9.61}$$

The eigenvalues of $V^\top Z V$ are strictly positive and bounded below by the quantity $z_{\min} \coloneqq N \max\{\rho_k\}^{-1} \min_{i,j} A_{i,j}$ (see Appendix C). Thus $\left(\boldsymbol{A}, V^\top Z V \boldsymbol{A}\right)_{\mathrm{F}} \geq z_{\min} \|\boldsymbol{A}\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Let $s(t) = \|\boldsymbol{A}\|_{\mathrm{F}}^2$. Then, using (9.61),

$$\frac{1}{2}\frac{\mathrm{d}s}{\mathrm{d}t} = \left(\boldsymbol{A}, \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t}\right)_{\mathrm{F}} = \left(\boldsymbol{A}, -\frac{1}{\varepsilon} V^\top Z V \boldsymbol{A}\right)_{\mathrm{F}} \leq -\frac{z_{\min}}{\varepsilon} \|\boldsymbol{A}\|_{\mathrm{F}}^2 = -\frac{z_{\min}}{\varepsilon} s \tag{9.62}$$

which implies that

$$\|\boldsymbol{A}(t)\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A}(0)\|_{\mathrm{F}}^2 \, e^{-\frac{2z_{\min}}{\varepsilon}t}. \tag{9.63}$$

Since $V^\top V = I_{N-1}$, it follows then $\|\widetilde{\boldsymbol{W}}\|_{\mathrm{F}}^2 = \|V\boldsymbol{A}\|_{\mathrm{F}}^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2$. Thus, (9.63) becomes

$$\|\boldsymbol{W}(t) - \boldsymbol{W}^\infty\|_{\mathrm{F}}^2 \leq \|\boldsymbol{W}(0) - \boldsymbol{W}^\infty\|_{\mathrm{F}}^2 \, e^{-\frac{2z_{\min}}{\varepsilon}t}. \tag{9.64}$$

Since $\boldsymbol{W} = P^{\frac{1}{2}}\boldsymbol{U}$, it follows that

$$\|\boldsymbol{W} - \boldsymbol{W}^\infty\|_{\mathrm{F}}^2 = \sum_{k=1}^{d}\sum_{j=1}^{N}\left[\rho_j \left((\boldsymbol{u}_j)_k - \boldsymbol{u}_k^\infty\right)^2\right] = \sum_{j=1}^{N} \rho_j \|\boldsymbol{u}_j - \boldsymbol{u}^\infty\|_2^2 \geq \min\{\rho_k\} \|\boldsymbol{u}_i - \boldsymbol{u}^\infty\|_2^2, \tag{9.65}$$

for any $i \in \{1, \dots, N\}$. Thus, (9.64) gives the decay bound

$$\|\boldsymbol{u}_i(t) - \boldsymbol{u}^\infty\|_2^2 \leq \frac{1}{\min\{\rho_k\}} \|\boldsymbol{W}(t) - \boldsymbol{W}^\infty\|_{\mathrm{F}}^2 \leq \frac{1}{\min\{\rho_k\}} \left\|\boldsymbol{W}^0 - \boldsymbol{W}^\infty\right\|_{\mathrm{F}}^2 e^{-\frac{2z_{\min}}{\varepsilon}t} =: C_{\boldsymbol{U}}^2 e^{-\frac{2z_{\min}}{\varepsilon}t}, \tag{9.66}$$

and $\lim_{t\to\infty} \boldsymbol{u}_i = \boldsymbol{u}^\infty$, as desired. ■

9.3.3 Energy Relaxation Proof

Theorem 9.14. *With the same assumptions as in Proposition 9.7,* $\lim_{t\to\infty} \boldsymbol{\omega}(t) = \boldsymbol{\omega}^\infty$ *and*

$$\|\boldsymbol{\omega}(t) - \boldsymbol{\omega}^\infty\|_2 \leq C_1 e^{-\frac{\widehat{z}_{\min}}{\varepsilon}t} + C_2 \mathcal{K}_{z_{\min}, \widehat{z}_{\min}}(t), \tag{9.67}$$

where $C_1 > 0$ and $C_2 > 0$ are constants that depend on the initial conditions of the system (9.17) *and $\widehat{z}_{\min}$ is a lower bound on the positive eigenvalues of $\widehat{Z}$.*

Proof. The proof is based on (9.19), which is equivalent to (9.17). Set $\boldsymbol{\xi}^\infty := Q^{-\frac{1}{2}}\boldsymbol{\omega}^\infty$, and note that

$$\widehat{Z}\boldsymbol{\xi}^\infty = \frac{|\boldsymbol{u}^\infty|^2}{2} Q^{-\frac{1}{2}}(F-B)M\mathbf{1}. \tag{9.68}$$

Since Q and $\boldsymbol{\omega}^\infty$ are time invariant, so too is $\boldsymbol{\xi}^\infty$. Hence $\widetilde{\boldsymbol{\xi}} := \boldsymbol{\xi} - \boldsymbol{\xi}^\infty$ satisfies the ODE

$$\frac{\mathrm{d}\widetilde{\boldsymbol{\xi}}}{\mathrm{d}t} = -\frac{1}{\varepsilon}\widehat{Z}\widetilde{\boldsymbol{\xi}} + \frac{1}{2\varepsilon}Q^{-\frac{1}{2}}\left[(G-C) - |\boldsymbol{u}^\infty|^2(F-B)\right]M\mathbf{1}. \tag{9.69}$$

Consider the null and range space components of $\widetilde{\boldsymbol{\xi}}$ with respect to $\widehat{Z}$: $\widetilde{\boldsymbol{\xi}} = \widetilde{\boldsymbol{\xi}}_{\mathcal{R}} + \widetilde{\boldsymbol{\xi}}_{\mathcal{N}}$. It can be shown that $\widetilde{\boldsymbol{\xi}}_{\mathcal{N}} \equiv \mathbf{0}$. (See Appendix D.) Therefore $\widetilde{\boldsymbol{\xi}} = \widetilde{\boldsymbol{\xi}}_{\mathcal{R}} = \widehat{V}\boldsymbol{b}$, where $\widehat{V} = [\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_{N-1}] \in \mathbb{R}^{N\times(N-1)}$ is a matrix whose columns form an orthonormal basis for the range of $\widehat{Z}$ and $\boldsymbol{b} = \widehat{V}^\top\widetilde{\boldsymbol{\xi}} \in \mathbb{R}^{N-1}$. Multiplication of (9.69) by $\boldsymbol{b}^\top\widehat{V}^\top$ gives

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{b}\|_2^2 = &-\frac{1}{\varepsilon}\boldsymbol{b}^\top\widehat{V}^\top\widehat{Z}\widehat{V}\boldsymbol{b} \\ &+ \frac{1}{2\varepsilon}\boldsymbol{b}^\top\widehat{V}^\top Q^{-\frac{1}{2}}\left[(G-C) - |\boldsymbol{u}^\infty|^2(F-B)\right]M\mathbf{1}.\end{aligned} \tag{9.70}$$

The eigenvalues of $\widehat{V}^\top\widehat{Z}\widehat{V}$ are strictly positive and bounded below by

$$\widehat{z}_{\min} := N\max\{n_k\}^{-1}\min_{i,j} B_{i,j};$$

see Appendix C. Consequently,

$$\boldsymbol{b}^\top\widehat{V}^\top\widehat{Z}\widehat{V}\boldsymbol{b} \geq \widehat{z}_{\min}\|\boldsymbol{b}\|_2^2 \qquad \text{for every } \boldsymbol{b}\in\mathbb{R}^{N-1}.$$

With this bound and an application of the Cauchy–Schwarz inequality, (9.70) becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{b}\|_2^2 \leq &-\frac{2\widehat{z}_{\min}}{\varepsilon}\|\boldsymbol{b}\|_2^2 \\ &+ \frac{\|\boldsymbol{b}\|_2}{\varepsilon\min\{n_k\}^{1/2}}\left\|\left[(G-C) - |\boldsymbol{u}^\infty|^2(F-B)\right]M\mathbf{1}\right\|_2.\end{aligned} \tag{9.71}$$

The next step is to bound the source term in (9.71). Standard norm inequalities and the

definitions in (9.14), and Appendix C, give

$$\left\| \left[(G-C) - |\boldsymbol{u}^\infty|^2 (F-B) \right] M\mathbf{1} \right\|_2 \leq 2 \max\{m_k\} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} B_{\max} \left| |\boldsymbol{u}_{i,j}|^2 - |\boldsymbol{u}^\infty|^2 \right|. \tag{9.72}$$

Meanwhile, to bound $|S_{i,j} - |\boldsymbol{u}^\infty|^2|$, we use (i) the decay bound in (9.59), (ii) the definitions of $\boldsymbol{u}_{i,j}$ and $\boldsymbol{u}^\infty$, (iii) the triangle inequality and Cauchy Schwarz inequality, (iv) the fact that $\alpha_{i,j} + \alpha_{j,i} = 1$, and (v) the fact that $|\boldsymbol{u}_{i,j}| \leq \max_{i,j}\{|\boldsymbol{u}_i|, |\boldsymbol{u}_j|\} \leq u_{\max}$ to obtain

$$\begin{aligned} \left| |\boldsymbol{u}_{i,j}|^2 - |\boldsymbol{u}^\infty|^2 \right| &= \left| (\boldsymbol{u}_{i,j} + \boldsymbol{u}^\infty)^\top (\boldsymbol{u}_{i,j} - \boldsymbol{u}^\infty) \right| \leq 2u_{\max} \left\| \boldsymbol{u}_{i,j} - \boldsymbol{u}^\infty \right\|_2 \\ &\leq 2u_{\max} \left(\alpha_{i,j} \|\boldsymbol{u}_i - \boldsymbol{u}^\infty\|_2 + \alpha_{j,i} \|\boldsymbol{u}_j - \boldsymbol{u}^\infty\|_2 \right) \leq 2C_{\boldsymbol{U}} u_{\max} \exp\left(-\frac{z_{\min}}{\varepsilon} t \right). \end{aligned} \tag{9.73}$$

Together (9.72) and (9.73) give

$$\left\| \left[(G-C) - |\boldsymbol{u}^\infty|^2 (F-B) \right] M\mathbf{1} \right\|_2 \leq 4N(N-1) C_{\boldsymbol{U}} u_{\max} B_{\max} \max\{m_k\} e^{-\frac{z_{\min}}{\varepsilon} t}, \tag{9.74}$$

and (9.71) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\boldsymbol{b}\|_2 \leq -\frac{\widehat{z}_{\min}}{\varepsilon} \|\boldsymbol{b}\|_2 + \frac{1}{\varepsilon} C_0 e^{-\frac{z_{\min}}{\varepsilon} t}, \tag{9.75}$$

where $C_0 = \frac{1}{2} \frac{4N(N-1) C_{\boldsymbol{U}} u_{\max} B_{\max} \max\{m_k\}}{\min\{n_k\}^{\frac{1}{2}}}$. At times when $\|\boldsymbol{b}\|_2 > 0$, this follows by dividing by $\|\boldsymbol{b}\|_2$; at a zero, the same inequality follows by applying the calculation to $\sqrt{\|\boldsymbol{b}\|_2^2 + \delta^2}$ and letting $\delta \downarrow 0$. Equivalently,

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\|\boldsymbol{b}\|_2 e^{\frac{\widehat{z}_{\min}}{\varepsilon} t} \right) \leq \frac{C_0}{\varepsilon} e^{\frac{\widehat{z}_{\min} - z_{\min}}{\varepsilon} t}. \tag{9.76}$$

Integrating both sides above in t and using (9.52) gives

$$\|\boldsymbol{b}\|_2 \leq \|\boldsymbol{b}^0\|_2 e^{-\frac{\widehat{z}_{\min}}{\varepsilon} t} + C_0 \mathcal{K}_{z_{\min}, \widehat{z}_{\min}}(t). \tag{9.77}$$

Since $\widehat{V}^\top \widehat{V} = I_{N-1}$, it follows that $\|\widetilde{\boldsymbol{\xi}}\|_2^2 = \|\widehat{V}\boldsymbol{b}\|_2^2 = \|\boldsymbol{b}\|_2^2$. Thus, (9.77) becomes

$$\|\boldsymbol{\xi} - \boldsymbol{\xi}^\infty\|_2 \leq \left\| \boldsymbol{\xi}^0 - \boldsymbol{\xi}^\infty \right\|_2 e^{-\frac{\widehat{z}_{\min}}{\varepsilon} t} + C_0 \mathcal{K}_{z_{\min}, \widehat{z}_{\min}}(t). \tag{9.78}$$

Further, since

$$\|\boldsymbol{\xi}-\boldsymbol{\xi}^{\infty}\|_2^2=\|Q^{-\frac{1}{2}}(\boldsymbol{\omega}-\boldsymbol{\omega}^{\infty})\|_2^2=\sum_{i=1}^{N}n_i^{-1}|\omega_i-\omega_i^{\infty}|^2\geq\frac{1}{\max\{n_k\}}\|\boldsymbol{\omega}-\boldsymbol{\omega}^{\infty}\|_2^2, \tag{9.79}$$

it follows that

$$\|\boldsymbol{\omega}(t)-\boldsymbol{\omega}^{\infty}\|_2\leq\max\{n_k\}^{\frac{1}{2}}\left(\left\|\boldsymbol{\xi}^0-\boldsymbol{\xi}^{\infty}\right\|_2e^{-\frac{\widehat{z}_{\min}}{\varepsilon}t}+C_0\mathcal{K}_{z_{\min},\widehat{z}_{\min}}(t)\right), \tag{9.80}$$

which recovers (9.67) with $C_1:=\max\{n_k\}^{\frac{1}{2}}\left\|\boldsymbol{\xi}^0-\boldsymbol{\xi}^{\infty}\right\|_2$ and $C_2:=C_0\max\{n_k\}^{\frac{1}{2}}$. Thus $\lim_{t\to\infty}\boldsymbol{\omega}(t)=\boldsymbol{\omega}^{\infty}$, as desired. ■

9.3.4 Temperature Relaxation

Corollary 9.15. *With the same assumptions as in Proposition 9.7, for each species $i\in\{1,\dots,N\}$, $\lim_{t\to\infty}\hat{\theta}_i(t)=\hat{\theta}^{\infty}$, and*

$$\left|\hat{\theta}_i-\hat{\theta}^{\infty}\right|\leq\frac{2}{d\min\{n_k\}}\left[C_1e^{-\frac{\widehat{z}_{\min}}{\varepsilon}t}+C_2\mathcal{K}_{z_{\min},\widehat{z}_{\min}}(t)\right]+\frac{2\max\{m_k\}u_{\max}C_{\boldsymbol{U}}}{d}e^{-\frac{z_{\min}}{\varepsilon}t}. \tag{9.81}$$

Proof. For each i,

$$\hat{\theta}_i=\frac{2}{dn_i}\omega_i-\frac{m_i}{d}|\boldsymbol{u}_i|^2\qquad\text{and}\qquad\hat{\theta}^{\infty}=\frac{2}{dn_i}\omega_i^{\infty}-\frac{m_i}{d}|\boldsymbol{u}^{\infty}|^2. \tag{9.82}$$

Therefore

$$\left|\hat{\theta}_i-\hat{\theta}^{\infty}\right|\leq\frac{2}{d\min\{n_k\}}|\omega_i-\omega_i^{\infty}|+\frac{\max\{m_k\}}{d}\left||\boldsymbol{u}^{\infty}|^2-|\boldsymbol{u}_i|^2\right|. \tag{9.83}$$

The velocity decay bound (9.59), gives

$$\left||\boldsymbol{u}^{\infty}|^2-|\boldsymbol{u}_i|^2\right|\leq2u_{\max}\|\boldsymbol{u}^{\infty}-\boldsymbol{u}_i\|_2\leq2u_{\max}C_{\boldsymbol{U}}e^{-\frac{z_{\min}}{\varepsilon}t}. \tag{9.84}$$

Using this and the energy decay bound (9.67), the result follows. ■

Remark 9.16. *The result in Corollary 9.15 can also be verified directly using the temperature equation in Equation* (9.23) *and an approach similar to the proof of Theorem 9.14. The difference in the proofs comes from the fact that the nullspace component analogous to $\widetilde{\boldsymbol{\xi}}_{\mathcal{N}}$ is no longer zero when using Equation* (9.23).

9.4 Numerical Confirmation of Theory

In this section, we compute temperature and velocity profiles using a fully implicit (backward) Euler time stepping scheme.[1] We assume a slab geometry in the x_1 direction. Thus while $d = 3$, the quantities of interest depend only on x_1. Moreover the bulk velocities in the x_2 and x_3 direction are identically zero, i.e., $U_{i,j} = 0$ for all $j \in \{2, 3\}$.

We consider systems with three different gas species chosen from a collection of four possible elements: Helium (He), Argon (Ar), Krypton (Kr), and Xenon (Xe). The masses and diameters of these elements are taken from [15] and are given below in SI units:

$$\begin{pmatrix} m_{\mathrm{He}} \\ m_{\mathrm{Ar}} \\ m_{\mathrm{Kr}} \\ m_{\mathrm{Xe}} \end{pmatrix} = \begin{pmatrix} 6.6464731 \\ 66.335209 \\ 139.14984 \\ 218.01714 \end{pmatrix} \times 10^{-27} \text{ kg}, \qquad \begin{pmatrix} d_{\mathrm{He}} \\ d_{\mathrm{Ar}} \\ d_{\mathrm{Kr}} \\ d_{\mathrm{Xe}} \end{pmatrix} = \begin{pmatrix} 2.193 \\ 3.659 \\ 4.199 \\ 4.939 \end{pmatrix} \times 10^{-10} \text{ m}. \tag{9.85}$$

In all of the examples below, the monotonicity of the minimum temperature (Theorem 9.3) and the upper and lower bounds on the bulk velocities (Theorem 9.4) are respected.

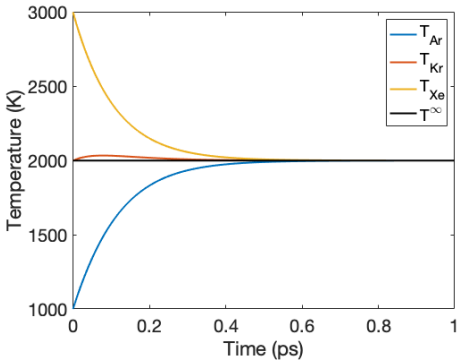
Example 1: Temperature Decay, Ar–Kr–Xe Mixture. The purpose of this example is to demonstrate temperature relaxation when the bulk velocities are zero. The initial number densities, velocities, and temperatures are given by

$$\begin{pmatrix} n^0_{\mathrm{Ar}} \\ n^0_{\mathrm{Kr}} \\ n^0_{\mathrm{Xe}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \times 10^{28} \text{ m}^{-3}, \qquad \boldsymbol{U}^0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ m/s}, \qquad \hat{\boldsymbol{\theta}}^0 = \begin{pmatrix} 1000 \\ 2000 \\ 3000 \end{pmatrix} \text{ K}. \tag{9.86}$$

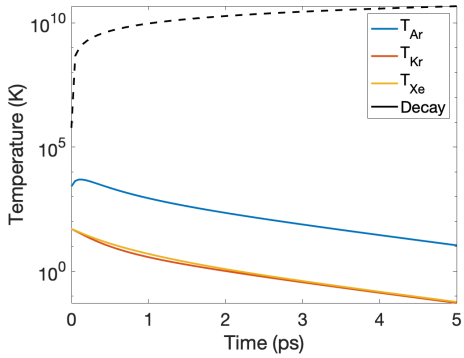
Figure 9.1 shows the relaxation of the temperatures to their steady state value. In this case the decay rate established in (9.81) is fairly sharp, even though the temperatures may not approach steady state monotonically.

Example 2: Velocity Decay, Ar–Kr–Xe Mixture. The purpose of this example is to demonstrate the velocity relaxation, with minimal effects from the temperature. A positive velocity is given to one particle type, and each temperature is set to the same constant. The

[1] The backward Euler method preserves the monotonicity properties of the temperature and bulk velocity established in Section 9.2. This fact will be proved in future work.



(a) Species and steady state temperatures.



(b) Deviation from steady state temperature and analytical decay estimate from (9.81).

Figure 9.1: Temperature Decay of Ar–Kr–Xe Mixture in Example 1.

initial number densities, velocities, and temperatures are given by

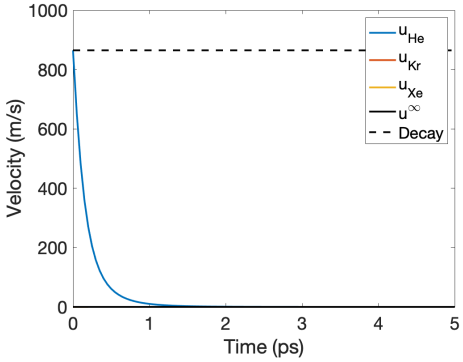
$$\begin{pmatrix} n^0_{\mathrm{Ar}} \\ n^0_{\mathrm{Kr}} \\ n^0_{\mathrm{Xe}} \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \times 10^{28}\ \mathrm{m}^{-3}, \qquad \boldsymbol{U}^0 = \begin{pmatrix} 100 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\ \mathrm{m/s}, \qquad \hat{\boldsymbol{\theta}}^0 = \begin{pmatrix} 1000 \\ 1000 \\ 1000 \end{pmatrix}\ \mathrm{K}. \quad (9.87)$$

Figure 9.2 shows the relaxation of the velocities and temperatures to the steady state values. The decay rate established for the velocities and temperatures in (9.59) and (9.81) underestimate the true decay rates in the example. The results also demonstrate that, unlike the two-species case, particle velocities may not converge to the steady-state value monotonically.

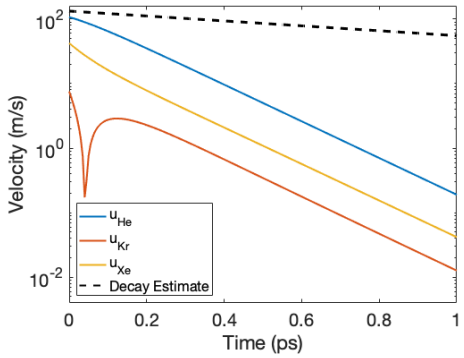
Example 3: Velocity-Temperature Relaxation, He–Kr–Xe Mixture This test case exercises the model due to the large kinetic energy differences between the species. The initial number densities, velocities, and temperatures are given by

$$\begin{pmatrix} n^0_{\mathrm{He}} \\ n^0_{\mathrm{Kr}} \\ n^0_{\mathrm{Xe}} \end{pmatrix} = \begin{pmatrix} 0.01 \\ 1 \\ 1 \end{pmatrix} \times 10^{28}\ \mathrm{m}^{-3}, \qquad \boldsymbol{U}^0 = \begin{pmatrix} 864.8 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\ \mathrm{m/s}, \qquad \hat{\boldsymbol{\theta}}^0 = \begin{pmatrix} 3000 \\ 300 \\ 300 \end{pmatrix}\ \mathrm{K}. \quad (9.88)$$

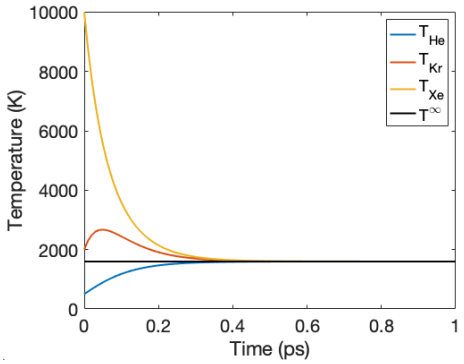
The momentum equilibration has a corresponding kinetic energy redistribution which causes non-monotonic changes in the temperature profile, as seen in Figure 9.3c. Because of the large differences in the particle masses and number densities, the estimates for the decay rates established in (9.59) and (9.81) are very weak.



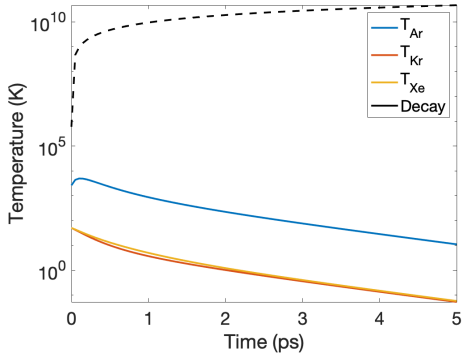
(a) Species and steady state velocities.



(b) Deviation from steady state velocity and analytical decay estimate from (9.59).

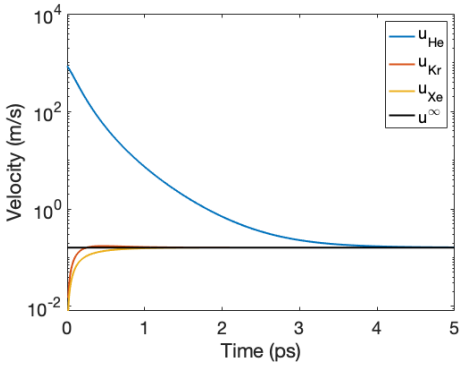


(c) Species and steady state temperatures.

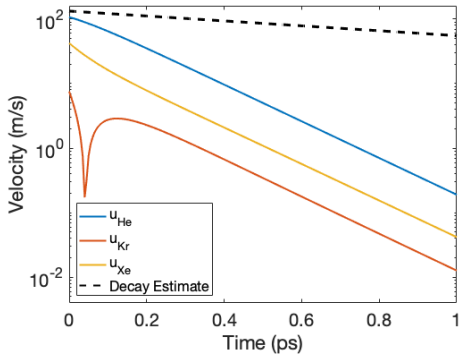


(d) Deviation from steady state velocity and analytical decay estimate from (9.81).

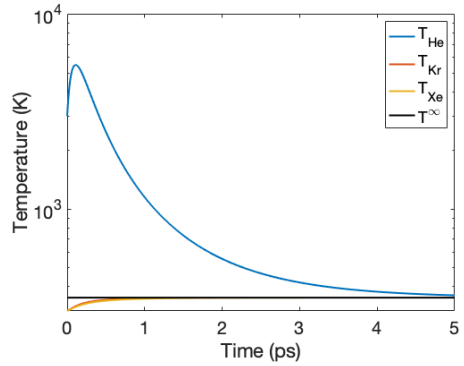
Figure 9.2: Velocity and Temperature Decay of Ar–Kr–Xe Mixture in Example 2.



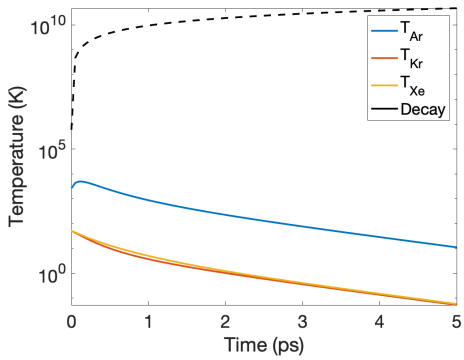
(a) Species and steady state velocities.



(b) Deviation from steady state velocity and analytical decay estimate from (9.59).



(c) Species and steady state temperatures.



(d) Deviation from steady state velocity and analytical decay estimate from (9.81).

Figure 9.3: Velocity and Temperature Decay of He–Kr–Xe Mixture in Example 3.

9.5 A Novel Nonlinear GST Iterative Solver

This section is based on our recent paper [34]:

E. Habbershaw, C.D. Hauck, and S.M. Wise, *Implicit Update of the Moment Equations for a Multi-Species, Homogeneous BGK Model*, SINUM (SIAM J. Numer. Anal.) 63 (2025) 881–907.
https://doi.org/10.1137/24M165421X

There we derived and analyzed a new nonlinear Gauss–Seidel-type (GST) solver for the space-homogeneous, multispecies BGK model. We will not give the tedious details of the method or the analyses in this document. Rather, we will content ourselves with some numerical simulations using the GST iterative method.

We first consider a slab geometry, in which the distribution functions f_i are invariant in the x_2 and x_3 direction. In this setting, the multispecies BGK equation takes the form

$$\frac{\partial f_i}{\partial t} + v_1 \partial_x f_i = \frac{1}{\varepsilon} \sum_j \lambda_{i,j}(M_{i,j} - f_i) \quad \forall i \in \{1, \dots, N\}, \tag{9.89}$$

where $[\boldsymbol{u}_{i,j}]_2 = [\boldsymbol{u}_{i,j}]_3 = 0$, and to simplify the notation, we let $x_1 = x$, $[\boldsymbol{u}_i]_1 = u_i$, and $[\boldsymbol{u}_{i,j}]_1 = u_{i,j}$. Rather than solve (9.89) in the full three-dimensional velocity space, we apply a multispecies extension of Chu reduction [16]. For each $i \in \{1, \dots, N\}$, let

$$g_i(x, v_1, t) = \int_{\mathbb{R}^2} f_i(x, \boldsymbol{v}, t) \, \mathrm{d}v_2 \, \mathrm{d}v_3, \tag{9.90a}$$

$$h_i(x, v_1, t) = \int_{\mathbb{R}^2} (v_2^2 + v_3^2) f_i(x, \boldsymbol{v}, t) \, \mathrm{d}v_2 \, \mathrm{d}v_3. \tag{9.90b}$$

Then (9.89) can be reduced to a system of equations for g_i and h_i:

$$\frac{\partial g_i}{\partial t} + v_1 \frac{\partial g_i}{\partial x} = \frac{1}{\varepsilon} \sum_j \lambda_{i,j}(G_{i,j} - g_i), \tag{9.91a}$$

$$\frac{\partial h_i}{\partial t} + v_1 \frac{\partial h_i}{\partial x} = \frac{1}{\varepsilon} \sum_j \lambda_{i,j}(H_{i,j} - h_i), \tag{9.91b}$$

where $G_{ij}(v_1) = M^{(1)}_{n_i, u_{i,j}, \hat{\theta}_{i,j}/m_i}(v_1)$ is a Maxwellian with $d = 1$, $H_{i,j} = 2\frac{\hat{\theta}_{i,j}}{m_i} G_{i,j}$, and

$$n_i = \int_{\mathbb{R}} g_i \, \mathrm{d}v_1, \quad n_i u_i = \int_{\mathbb{R}} v_1 g_i \, \mathrm{d}v_1, \quad \frac{3n_i}{m_i}\hat{\theta}_i = \int_{\mathbb{R}} (v_1 - u_i)^2 g_i \, \mathrm{d}v_1 + \int_{\mathbb{R}} h_i \, \mathrm{d}v_1. \tag{9.92}$$

We use the IMEX approach outlined in [34] to solve (9.91). The Butcher tableaux for

the IMEX method is taken from Table II of [58]. The implicit step takes the form of a backward Euler method. The nonlinear GST method still forms the key component of the implicit step. The spatial discretization of the advection operator uses a second-order finite volume scheme with minmod-type slope limiters, as described in earlier chapters. The velocity is discretized using standard discrete velocity models based on evenly spaced points and a midpoint quadrature rule. Details of these discretizations can be found in Chapter 6. The strategy for selecting the time step is to use the natural hyperbolic restriction of the advection operators in the equations. In practice, we observe that the GST method converges without onerous restrictions. However, for general problems, these restrictions provide a rigorous fall-back whenever the GST method fails to converge at the hyperbolic time step [34].

Anderson Acceleration (AA) [80] is a commonly used method to speed up fixed point iterative methods. We wrap the GST method with an Anderson Acceleration loop, and find that using this method decreases the iteration counts required for convergence of the method.

9.5.1 Sod problem with two identical species

To verify the multispecies code, we perform a simulation of the Sod shock tube problem [73], a Riemann problem for the compressible Euler equations of a perfect gas. The initial data are $(n_{\mathrm{L}}, \boldsymbol{u}_{\mathrm{L}}, \theta_{\mathrm{L}}) = (1, \mathbf{0}, 1)$ for $x \leq 0$, and $(n_{\mathrm{R}}, \boldsymbol{u}_{\mathrm{R}}, \theta_{\mathrm{R}}) = (0.125, \mathbf{0}, 0.8)$ for $x \geq 0$. An analytical solution for the Riemann problem can be found in [77]. For small values of ε, both the single-species BGK equation [17] and multispecies BGK (M-BGK) equations [32] with properly initialized data formally recover the Sod solution. Because the slab problem (9.89) is formulated in three velocity dimensions, the adiabatic index for the limiting Euler equations is $\gamma = 5/3$.

In the current test, we treat the left- and right-hand side of the domain as different particle species, each having mass $m_i = 1$. The kinetic initial data is

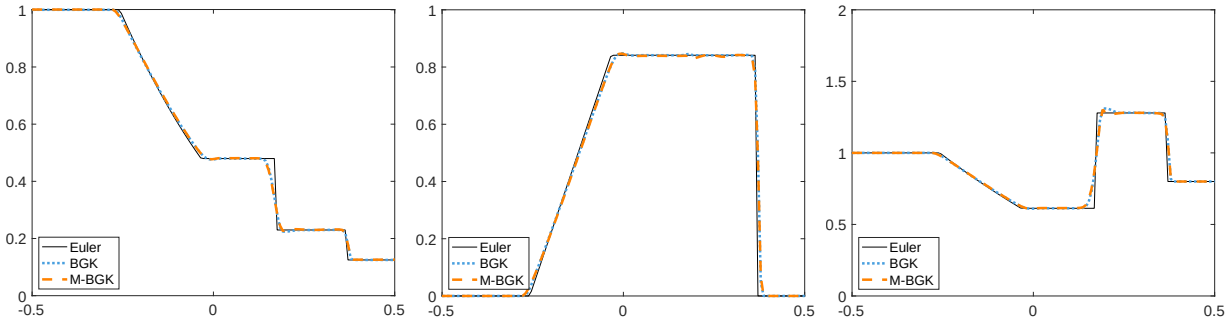
$$f_1(x, \boldsymbol{v}, 0) = \begin{cases} M^{(3)}_{n_{\mathrm{L}}, \boldsymbol{u}_{\mathrm{L}}, \theta_{\mathrm{L}}}, & x \leq 0, \\ M^{(3)}_{0.001, \boldsymbol{u}_{\mathrm{R}}, \theta_{\mathrm{R}}}, & x > 0, \end{cases} \quad \text{and} \quad f_2(x, \boldsymbol{v}, 0) = \begin{cases} M^{(3)}_{0.001, \boldsymbol{u}_{\mathrm{L}}, \theta_{\mathrm{L}}}, & x \leq 0, \\ M^{(3)}_{n_{\mathrm{R}}, \boldsymbol{u}_{\mathrm{R}}, \theta_{\mathrm{R}}}, & x > 0, \end{cases} \tag{9.93}$$

where $M^{(d)}_{n, \boldsymbol{u}, \theta}$ denotes the d-dimensional Maxwellian, defined in (9.3). Trace densities are used in (9.93) instead of voids in order to avoid negativity in the numerical solutions.

The numerical solution is computed on a discretization of the phase space $X \times V = [-1, 1] \times [-10, 10]$, using periodic boundary conditions and a mesh with $N_x = 256$ by $N_v = 192$ cells of uniform size. The time step selection is driven by the advection in (9.91):

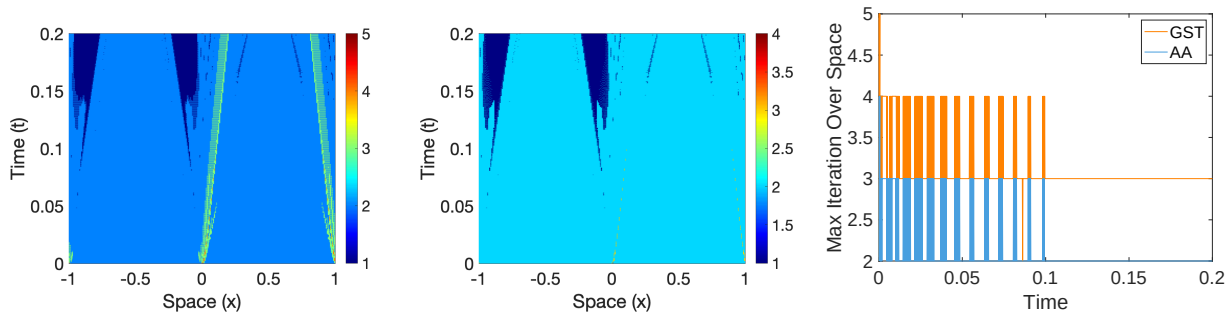
$\Delta t = 0.9\frac{2/Nx}{2\max\{|v_l|\}}$. Figure 9.4 shows single- and multispecies simulations of the model with $\varepsilon = 10^{-4}$, run to a final time of $t_f = 0.2$, compared with the Sod solution for the associated Euler equations. The results show that both BGK and M-BGK models perform well as numerical solvers for the Sod problem in this highly collisional regime.

Figure 9.5 contains a comparison of the iteration counts required for convergence at a tolerance of 10^{-8}. The iteration count for the GST method ranges from one to five over the simulation, with an average of approximately 1.982 iterations, and the Anderson Acceleration method reduces the number of iterations to a range of one to four, with an average of approximately 1.943 iterations.



(a) Total number density. **(b)** Total velocity. **(c)** Total temperature.

Figure 9.4: The total density, total bulk velocity, and total temperature moments are compared to the solution of the Euler equations at time $t = 0.2$, with $\varepsilon = 10^{-4}$.



(a) Iterations required for convergence of GST. **(b)** Iterations required for convergence of AA. **(c)** Maximum iteration count over time.

Figure 9.5: The number of iterations required for convergence is decreased by the application of Anderson Acceleration (AA) to the GST algorithm. The maximum iteration number over the spatial grid is strictly greater for the GST algorithm, compared to the Anderson Acceleration approach.

9.5.2 Three Species Mixture: Ar–Kr–Xe

The goal of this test is to demonstrate the benefits of the IMEX method when the problem becomes stiff (i.e., as $\varepsilon \to 0$). The following CFL-type conditions on the time step are required for the explicit and IMEX methods (which use a minmod-type slope limiter in the flux computation) to ensure stability and positivity of the numerical solution:

$$\Delta t^{\mathfrak{n}} \leq \begin{cases} \frac{\varepsilon \Delta x}{2\varepsilon \max\{|v_l|\} + \Delta x \Lambda^{\mathfrak{n}}} =: \Delta t^{\mathfrak{n}}_{\mathsf{EX}}, \\ \frac{\Delta x}{2\max\{|v_l|\}} =: \Delta t_{\mathsf{IMEX}}, \end{cases} \quad \text{where} \quad \Lambda^{\mathfrak{n}} = \max_{i=1}^{N} \left\{ \max_{\kappa} \left\{ \sum_{j=1}^{N} \lambda^{\mathfrak{n}}_{i,j}(x_\kappa) \right\} \right\}. \tag{9.94}$$

As $\varepsilon \to 0$, the condition on $\Delta t^{\mathfrak{n}}_{\mathsf{EX}}$ becomes restrictive.

To compare the methods, we simulate an interface problem, with a mixture of Argon, Krypton, and Xenon; the masses and diameters of these particle species are taken from [15] and are given below in SI units:

$$\begin{pmatrix} m_{\mathsf{Ar}} \\ m_{\mathsf{Kr}} \\ m_{\mathsf{Xe}} \end{pmatrix} = \begin{pmatrix} 6.6335209 \\ 13.914984 \\ 21.801714 \end{pmatrix} \times 10^{-26} \text{ kg}, \qquad \begin{pmatrix} d_{\mathsf{Ar}} \\ d_{\mathsf{Kr}} \\ d_{\mathsf{Xe}} \end{pmatrix} = \begin{pmatrix} 3.659 \\ 4.199 \\ 4.939 \end{pmatrix} \times 10^{-10} \text{ m}. \tag{9.95}$$

The setup for the test is similar to that of Section 9.5.1: a mixture of gases at rest is initialized with a high temperature on the left, and a low temperature on the right of the domain. Specifically, the mixture of Ar, Kr, and Xe is initialized on the left half of the one-dimensional spatial domain with number densities $(n^0_{\mathsf{Ar}}, n^0_{\mathsf{Kr}}, n^0_{\mathsf{Xe}}) = (5, 5, 0.5) \times 10^{25}$ m^{-3}, and common energy temperature $\hat{\theta}_{\mathsf{L}} = 10$; on the right half, the mixture is initialized with number densities $(n^0_{\mathsf{Ar}}, n^0_{\mathsf{Kr}}, n^0_{\mathsf{Xe}}) = (0.5, 0.5, 5) \times 10^{25}$ m^{-3}, and common energy temperature $\hat{\theta}_{\mathsf{R}} = 1$. The initial kinetic distributions are given by

$$f_{\mathsf{Ar}}(x, \boldsymbol{v}, 0) = \chi_{[-1,0]}(x) M^{(3)}_{5,\mathbf{0},10/m_{\mathsf{Ar}}}(\boldsymbol{v}) + \chi_{(0,1]}(x) M^{(3)}_{0.5,\mathbf{0},1/m_{\mathsf{Ar}}}(\boldsymbol{v}), \tag{9.96a}$$

$$f_{\mathsf{Kr}}(x, \boldsymbol{v}, 0) = \chi_{[-1,0]}(x) M^{(3)}_{5,\mathbf{0},10/m_{\mathsf{Kr}}}(\boldsymbol{v}) + \chi_{(0,1]}(x) M^{(3)}_{0.5,\mathbf{0},1/m_{\mathsf{Kr}}}(\boldsymbol{v}), \tag{9.96b}$$

$$f_{\mathsf{Xe}}(x, \boldsymbol{v}, 0) = \chi_{[-1,0]}(x) M^{(3)}_{0.5,\mathbf{0},10/m_{\mathsf{Xe}}}(\boldsymbol{v}) + \chi_{(0,1]}(x) M^{(3)}_{5,\mathbf{0},1/m_{\mathsf{Xe}}}(\boldsymbol{v}), \tag{9.96c}$$

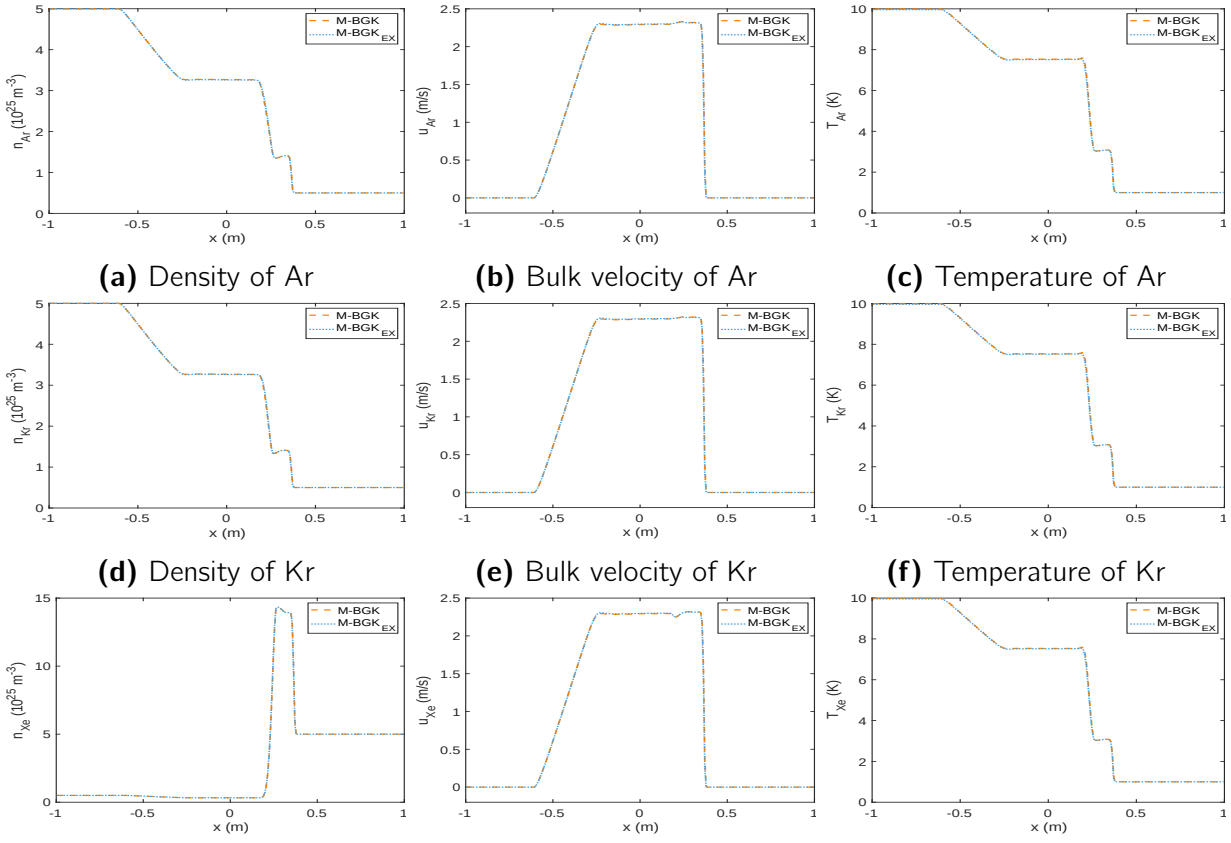
where $\chi_{\mathcal{S}}$ is the indicator function on the set $\mathcal{S}$, and $M^{(d)}_{n,\boldsymbol{u},\theta}$ denotes the d-dimensional Maxwellian, defined in (9.3).

The numerical solution is computed on a discretization of the phase space $X \times V = [-1, 1] \times [-30, 30]$, using reflective boundaries; the mesh consists of $N_x = 256 \times N_v = 192$ cells of uniform size. The IMEX method (using the GST solver) uses a time step of size

$\Delta t = 0.9\Delta t_{\text{IMEX}}$, and the fully explicit method uses a time step of size $\Delta t^{\text{n}} = 0.9\Delta t^{\text{n}}_{\text{EX}}$; the simulation is run to a final time of $t_{\text{f}} = 1$ μs.

Moment plots for numerical solutions computed with the IMEX and fully explicit method are compared in Figure 9.6 for $\varepsilon \approx 0.001$. The computed solutions align closely. However, the explicit method requires 7089 time steps, while the IMEX method only requires 854 steps. In practice, as ε decreases, the number of time steps for the explicit method grows, slowing the simulation, while the IMEX method experiences no slow down in computation time.

Figure 9.7 contains a comparison of the iteration counts required for convergence at a tolerance of 10^{-8}. The iteration count for the GST method ranges from two to five during the simulation, with an average of approximately 2.0989 iterations, and the Anderson Acceleration method reduces the number of iterations to a range of two to four, with an average of approximately 2.0001 iterations.

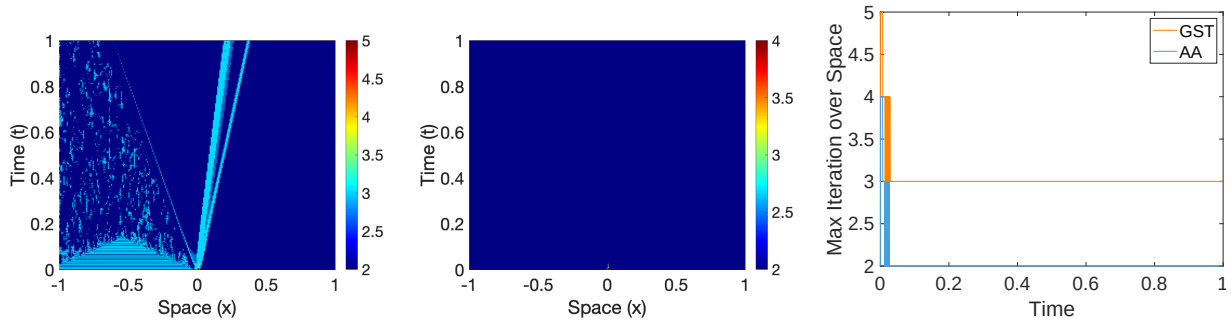


(a) Density of Ar **(b)** Bulk velocity of Ar **(c)** Temperature of Ar

(d) Density of Kr **(e)** Bulk velocity of Kr **(f)** Temperature of Kr

(g) Density of Xe **(h)** Bulk velocity of Xe **(i)** Temperature of Xe

Figure 9.6: Noble gas mixture of Ar, Kr, and Xe: Moment plots of species number density, bulk velocity, and temperature, are given at $t_{\text{f}} = 1$ μs, for a mixture with $\varepsilon \approx 0.00103$. M-BGK uses the GST algorithm for the moment solve. M-BGK$_{\text{EX}}$ uses a fully explicit method.



(a) Iterations required for convergence of GST.

(b) Iterations required for convergence of AA.

(c) Maximum iteration count over time.

Figure 9.7: The number of iterations required for convergence is decreased by the application of Anderson Acceleration (AA) to the GST algorithm. The maximum iteration number over the spatial grid is strictly greater for the GST algorithm, compared to the Anderson Acceleration approach.

Test	Tolerance	GST iterations	AA iterations	Time steps
Two-species Sod	10^{-8}	1–5 (mean 1.982)	1–4 (mean 1.943)	—
Ar–Kr–Xe mixture	10^{-8}	2–5 (mean 2.0989)	2–4 (mean 2.0001)	IMEX: 854; explicit: 7089

Table 9.1: Archived nonlinear-solver and time-step counts. For the Ar–Kr–Xe test, the IMEX method uses approximately 8.30 times fewer time steps than the fully explicit method. AA denotes Anderson acceleration applied to the GST iteration.

9.6 Conclusions

In this chapter, we have studied the moment equations associated to a recently developed multispecies BGK kinetic model for rarefied gas dynamics in the spatially homogeneous setting. The model includes collision frequencies that, unlike most models in the literature, depend on the species temperatures. This fact complicates the analysis of the moment equations and subsequent numerical tools. We have proven that all species temperatures are bounded below by a positive, non-decreasing temperature envelope, thus establishing that species temperatures always remain positive. We have also established upper and lower bounds on the components of the bulk velocity that are in terms of the initial data.

Using the lower bound on the species temperatures, we have shown that the moments always stay within a bounded, time-invariant set of physically realizable states, which, in turn, leads to the existence of unique global solutions. Finally, we have proved that unique equilibria exist for the moment equations, and solutions converge to these equilibria exponentially in time. In addition, we have established bounds on the convergence rates to equilibria. We concluded the examination of the space-homogeneous problem with some basic numerical simulations that demonstrate some of the established theoretical behavior.

The numerical results in this chapter are computed with a backward Euler method, using an iterative scheme to solve the relevant nonlinear algebraic equations at each time-step. In future work, we will describe the scheme more fully and prove convergence under a suitable time step restriction that does not require resolution of the parameter ε. We will use the backward Euler method as a component in an IMEX scheme for simulating multispecies BGK models that include phase-space advection.

In our paper [34], we presented a simple iterative approach for solving a set of implicit moment equations as part of a larger IMEX-based strategy for computing solutions of a multispecies BGK model. The main challenge is that the collision frequencies in the moment equations themselves depend on the moments in a complicated way. The key contribution is a rigorous contraction mapping proof, which shows that the iterative method converges to a unique solution under reasonable time steps that do not degenerate to zero when the collision parameter $\varepsilon \to 0$. This implicit solve is a key component in the IMEX discretization of the multispecies BGK model, the implementation of which is performed in slab geometry but closed in three-dimensional velocity space using a Chu reduction technique [16].

In future work, we will examine similar problems for multispecies ellipsoidal statistical models (ES-BGK) [39, 2] that, unlike standard BGK models, recover in highly collisional regimes, the correct Prandtl number for the compressible Navier–Stokes equations. We will

also extend the current code structure used here to generate results for more general problems in multi-dimensional domains. The ES-BGK model is the subject of the next chapter.

Chapter 10

ES-BGK Models and Numerics

10.1 Purpose and Organization

The standard BGK model fixes the relaxation rate of the stress tensor and heat flux in a way that produces a unit Prandtl number. The ellipsoidal-statistical BGK (ES-BGK) model replaces the isotropic Maxwellian target by an anisotropic Gaussian and thereby separates these two relaxation scales. The model was introduced by Holway [39]; its conservative and entropic formulation was established by Andries, Le Tallec, Perlat, and Perthame [2]. Modern entropy-production estimates cover the full monatomic parameter range, including the critical value that yields the physical Prandtl number [43].

This chapter uses the notation and Chapman–Enskog conventions of Chapters 2–4. Section 10.2 defines the single-species model and proves its basic structural properties. Section 10.3 derives its Euler and Navier–Stokes–Fourier limits by the normal-solution formalism of Chapter 3. Section 10.4 gives a closed implicit collision update, and Section 10.5 restores the numerical demonstration from the Year Four Report. Section 10.6 develops pair Gaussians, their moment equations, and both restricted and general multispecies tensor closures.

10.2 The Single-Species ES-BGK Model

Let $f = f(\boldsymbol{x}, \boldsymbol{v}, t) > 0$ describe particles of mass $m > 0$. The **number density**, **mass density**, **bulk velocity**, and **temperature in energy units** are

$$n = \int_{\mathbb{R}^d} f \, \mathrm{d}\boldsymbol{v}, \qquad \rho = mn, \qquad \boldsymbol{u} = \frac{1}{n} \int_{\mathbb{R}^d} \boldsymbol{v} f \, \mathrm{d}\boldsymbol{v}, \qquad \theta = \frac{m}{dn} \int_{\mathbb{R}^d} |\boldsymbol{v} - \boldsymbol{u}|^2 f \, \mathrm{d}\boldsymbol{v}. \tag{10.1}$$

The **thermal variance** is $\vartheta = \theta/m$, the **scalar pressure** is $p = n\theta = \rho\vartheta$, and the **pressure tensor** is

$$\mathsf{P} = m\int_{\mathbb{R}^d} (\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u}) f \,\mathrm{d}\boldsymbol{v}, \qquad \operatorname{tr}\mathsf{P} = dp. \tag{10.2}$$

The **deviatoric pressure tensor** and **heat flux** are

$$\Pi = \mathsf{P} - p\mathsf{I}, \qquad \boldsymbol{q} = \frac{m}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}-\boldsymbol{u}|^2(\boldsymbol{v}-\boldsymbol{u}) f \,\mathrm{d}\boldsymbol{v}. \tag{10.3}$$

Thus $\operatorname{tr}\Pi = 0$. These definitions agree with Chapter 3; in particular, θ is not a variance, while $\vartheta = \theta/m$ is.

Definition 10.1 (Temperature Tensor and ES-BGK Target)**.** *For a parameter*

$$-\frac{1}{d-1} \le \nu < 1, \tag{10.4}$$

*the **temperature tensor** is*

$$\mathsf{T}_\nu = (1-\nu)\vartheta\mathsf{I} + \nu\frac{\mathsf{P}}{\rho} = \vartheta\mathsf{I} + \frac{\nu}{\rho}\Pi. \tag{10.5}$$

*The **ellipsoidal Gaussian target** is*

$$\mathcal{G}_\nu[f](\boldsymbol{v}) = \frac{n}{\sqrt{\det(2\pi\mathsf{T}_\nu)}}\exp\left[-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{u})\cdot\mathsf{T}_\nu^{-1}(\boldsymbol{v}-\boldsymbol{u})\right]. \tag{10.6}$$

The ES-BGK equation is

$$\partial_t f + \boldsymbol{v}\cdot\nabla_{\boldsymbol{x}} f = \frac{\lambda}{\varepsilon}\big(\mathcal{G}_\nu[f] - f\big), \qquad \lambda > 0, \quad \varepsilon > 0. \tag{10.7}$$

The endpoint $\nu = -1/(d-1)$ gives $\mathrm{Pr} = (d-1)/d$; for $d = 3$, this is the monatomic value $2/3$. The choice $\nu = 0$ gives the ordinary BGK model. The value $\nu = 1$ is excluded because the stress relaxation rate vanishes.

Lemma 10.2 (Positivity and Moments of the Target)**.** *Assume* P *is positive definite. Then* T_ν *is positive definite for the parameter range in* (10.4)*, and, therefore, the ellipsoidal Gaussian target is well defined. Furthermore,*

$$\int_{\mathbb{R}^d} \mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} = n, \tag{10.8a}$$

$$\int_{\mathbb{R}^d} \boldsymbol{v}\mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} = n\boldsymbol{u}, \tag{10.8b}$$

$$m\int_{\mathbb{R}^d}(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})\mathcal{G}_\nu\,\mathrm{d}\boldsymbol{v}=\rho\mathsf{T}_\nu, \tag{10.8c}$$

$$m\int_{\mathbb{R}^d}\boldsymbol{v}\otimes\boldsymbol{v}\mathcal{G}_\nu\,\mathrm{d}\boldsymbol{v}=\rho\boldsymbol{u}\otimes\boldsymbol{u}+\rho\mathsf{T}_\nu=\rho\boldsymbol{u}\otimes\boldsymbol{u}+(1-\nu)p\mathsf{I}+\nu\mathsf{P}, \tag{10.8d}$$

$$\frac{m}{2}\int_{\mathbb{R}^d}|\boldsymbol{v}|^2\mathcal{G}_\nu\,\mathrm{d}\boldsymbol{v}=\frac{1}{2}\rho|\boldsymbol{u}|^2+\frac{d}{2}n\theta. \tag{10.8e}$$

Proof. Let $s_1,\dots,s_d>0$ be the eigenvalues of P/ρ. Since $\sum_k s_k=d\vartheta$, the eigenvalues of T_ν are $t_k=(1-\nu)\vartheta+\nu s_k$. For $0\le\nu<1$, we have $t_k\ge(1-\nu)\vartheta>0$. If $-1/(d-1)\le\nu<0$, then $s_k<d\vartheta$ and

$$t_k>(1-\nu)\vartheta+\nu d\vartheta=[1+(d-1)\nu]\vartheta\ge 0.$$

Thus every t_k is strictly positive throughout the admissible parameter range. Consequently, T_ν is positive definite, $\det\mathsf{T}_\nu>0$, and both T_ν^{-1} and $\mathsf{T}_\nu^{-1/2}$ exist.

We now calculate the target moments. Let

$$\boldsymbol{z}=\mathsf{T}_\nu^{-1/2}(\boldsymbol{v}-\boldsymbol{u}),\qquad \boldsymbol{v}=\boldsymbol{u}+\mathsf{T}_\nu^{1/2}\boldsymbol{z},\qquad \mathrm{d}\boldsymbol{v}=\sqrt{\det\mathsf{T}_\nu}\,\mathrm{d}\boldsymbol{z},$$

and introduce the standard centered Gaussian

$$\gamma(\boldsymbol{z})=(2\pi)^{-d/2}\exp\left(-\frac{1}{2}|\boldsymbol{z}|^2\right).$$

The determinant in the target normalization cancels the Jacobian, so

$$\mathcal{G}_\nu(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}=n\gamma(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{10.9}$$

Using

$$\int_{\mathbb{R}^d}\gamma\,\mathrm{d}\boldsymbol{z}=1,\qquad \int_{\mathbb{R}^d}\boldsymbol{z}\gamma\,\mathrm{d}\boldsymbol{z}=\boldsymbol{0},\qquad \int_{\mathbb{R}^d}\boldsymbol{z}\otimes\boldsymbol{z}\,\gamma\,\mathrm{d}\boldsymbol{z}=\mathsf{I},$$

we obtain

$$\int_{\mathbb{R}^d}\mathcal{G}_\nu\,\mathrm{d}\boldsymbol{v}=n, \tag{10.10}$$

$$\int_{\mathbb{R}^d}\boldsymbol{v}\mathcal{G}_\nu\,\mathrm{d}\boldsymbol{v}=n\int_{\mathbb{R}^d}\left(\boldsymbol{u}+\mathsf{T}_\nu^{1/2}\boldsymbol{z}\right)\gamma(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}=n\boldsymbol{u}, \tag{10.11}$$

$$m\int_{\mathbb{R}^d}(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})\mathcal{G}_\nu\,\mathrm{d}\boldsymbol{v}=mn\mathsf{T}_\nu^{1/2}\left(\int_{\mathbb{R}^d}\boldsymbol{z}\otimes\boldsymbol{z}\,\gamma(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\right)\mathsf{T}_\nu^{1/2}=\rho\mathsf{T}_\nu. \tag{10.12}$$

The centered first moment also gives the raw second moment explicitly:

$$m\int_{\mathbb{R}^d} \boldsymbol{v}\otimes\boldsymbol{v}\mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} = m\int_{\mathbb{R}^d} \big[(\boldsymbol{v}-\boldsymbol{u})+\boldsymbol{u}\big]\otimes\big[(\boldsymbol{v}-\boldsymbol{u})+\boldsymbol{u}\big]\mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} \tag{10.13}$$

$$= \rho\boldsymbol{u}\otimes\boldsymbol{u} + \rho\mathsf{T}_\nu \tag{10.14}$$

$$= \rho\boldsymbol{u}\otimes\boldsymbol{u} + (1-\nu)p\mathsf{I} + \nu\mathsf{P}. \tag{10.15}$$

Here the two cross terms vanish, and the final equality follows from $\rho\mathsf{T}_\nu = (1-\nu)p\mathsf{I}+\nu\mathsf{P}$.

Finally, taking one half of the trace of the raw second-moment identity gives

$$\frac{m}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}|^2\mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} = \frac{1}{2}\rho|\boldsymbol{u}|^2 + \frac{m}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}-\boldsymbol{u}|^2\mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} \tag{10.16}$$

$$= \frac{1}{2}\rho|\boldsymbol{u}|^2 + \frac{\rho}{2}\mathrm{tr}\,\mathsf{T}_\nu. \tag{10.17}$$

Because

$$\mathrm{tr}\,\mathsf{T}_\nu = (1-\nu)d\vartheta + \nu\mathrm{tr}(\mathsf{P}/\rho) = d\vartheta = \frac{d\theta}{m},$$

and $\rho = mn$, this becomes

$$\frac{m}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}|^2\mathcal{G}_\nu \,\mathrm{d}\boldsymbol{v} = \frac{1}{2}\rho|\boldsymbol{u}|^2 + \frac{d}{2}n\theta,$$

which proves the final moment identity. ■

Proposition 10.3 (Conservation Laws)**.** *Smooth, rapidly decaying solutions of* (10.7) *satisfy*

$$\partial_t\rho + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}) = 0, \tag{10.18a}$$

$$\partial_t(\rho\boldsymbol{u}) + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}\otimes\boldsymbol{u}+\mathsf{P}) = \boldsymbol{0}, \tag{10.18b}$$

$$\partial_t\omega + \nabla_{\boldsymbol{x}}\cdot(\omega\boldsymbol{u}+\mathsf{P}\boldsymbol{u}+\boldsymbol{q}) = 0, \tag{10.18c}$$

where the ***total energy density*** *is* $\omega = \frac{1}{2}\rho|\boldsymbol{u}|^2 + \frac{d}{2}n\theta$.

Proof. Taking the mass, momentum, and energy moments of (10.7), Lemma 10.2 shows that every collision contribution vanishes. Decomposing the transport moments into centered and bulk parts gives (10.18) exactly as in Proposition 3.5. ■

The energy balance gives only the trace of the second moment dynamics. An anisotropic model also requires the full tensor balance. We first make the third-order tensor notation explicit.

Remark 10.4 (Third-Order Tensor Notation)**.** *For vectors* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^d$ *and a second-order tensor* $\mathsf{B} \in \mathbb{R}^{d\times d}$*, define*

$$(\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c})_{ijk} = a_i b_j c_k, \tag{10.19}$$

$$(\boldsymbol{a} \otimes \mathsf{B})_{ijk} = a_i B_{jk}, \qquad i, j, k \in \{1, \dots, d\}. \tag{10.20}$$

Both products are third-order tensors in $\mathbb{R}^{d\times d\times d}$*. If* B *is symmetric, we use the unnormalized symmetrization*

$$\big[\mathrm{Sym}(\boldsymbol{a} \otimes \mathsf{B})\big]_{ijk} = a_i B_{jk} + a_j B_{ik} + a_k B_{ij}. \tag{10.21}$$

For a third-order tensor C*, its spatial divergence is the second-order tensor defined by*

$$(\nabla_{\boldsymbol{x}} \cdot \mathsf{C})_{ij} = \sum_{k=1}^{d} \partial_{x_k} C_{ijk}. \tag{10.22}$$

With these conventions, the **third centered moment** is the fully symmetric third-order tensor $\mathsf{R} \in \mathbb{R}^{d\times d\times d}$ defined componentwise by

$$R_{ijk} = m \int_{\mathbb{R}^d} (v_i - u_i)(v_j - u_j)(v_k - u_k) f \, \mathrm{d}\boldsymbol{v}, \qquad i, j, k \in \{1, \dots, d\}. \tag{10.23}$$

Equivalently,

$$\mathsf{R} = m \int_{\mathbb{R}^d} (\boldsymbol{v} - \boldsymbol{u}) \otimes (\boldsymbol{v} - \boldsymbol{u}) \otimes (\boldsymbol{v} - \boldsymbol{u}) f \, \mathrm{d}\boldsymbol{v}. \tag{10.24}$$

The component formula shows directly that R is unchanged by every permutation of its three indices.

Proposition 10.5 (Pressure-tensor balance)**.** *Smooth, rapidly decaying solutions of* (10.7) *satisfy*

$$\begin{aligned} &\partial_t\big(\rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}\big) + \nabla_{\boldsymbol{x}} \cdot \Big(\rho \boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{u} + \mathrm{Sym}(\boldsymbol{u} \otimes \mathsf{P}) + \mathsf{R}\Big) \\ &\qquad = -\frac{\lambda(1-\nu)}{\varepsilon} \Pi. \end{aligned} \tag{10.25}$$

Taking one half of its trace recovers the total-energy balance in (10.18)*.*

Proof. Multiply (10.7) by $m\boldsymbol{v} \otimes \boldsymbol{v}$ and integrate in velocity. The raw second moment is

$$m \int_{\mathbb{R}^d} \boldsymbol{v} \otimes \boldsymbol{v} f \, \mathrm{d}\boldsymbol{v} = \rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}, \tag{10.26}$$

while expansion about $\boldsymbol{u}$ gives

$$m\int_{\mathbb{R}^d} \boldsymbol{v}\otimes\boldsymbol{v}\otimes\boldsymbol{v} f\,\mathrm{d}\boldsymbol{v} = \rho\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} + \mathrm{Sym}(\boldsymbol{u}\otimes\mathsf{P}) + \mathsf{R}. \tag{10.27}$$

By Lemma 10.2, the target has raw second moment $\rho\boldsymbol{u}\otimes\boldsymbol{u} + \rho\mathsf{T}_\nu$. Hence the collision moment is

$$\frac{\lambda}{\varepsilon}(\rho\mathsf{T}_\nu - \mathsf{P}) = -\frac{\lambda(1-\nu)}{\varepsilon}(\mathsf{P} - p\mathsf{I}), \tag{10.28}$$

which proves (10.25). Contraction of the first two tensor indices gives the energy flux decomposition and therefore the stated trace relation. ■

Lemma 10.6 (Determinant Inequality)**.** *Assuming* (10.4) *holds,*

$$\det\mathsf{T}_\nu \geq \det(\mathsf{P}/\rho). \tag{10.29}$$

Proof. Let $s_1,\dots,s_d$ be the eigenvalues of P/ρ, and let $t_1,\dots,t_d$ be the eigenvalues of T_ν. Because $\mathsf{T}_\nu = (1-\nu)\vartheta\mathsf{I} + \nu\mathsf{P}/\rho$, the two tensors have the same eigenvectors and

$$t_k = (1-\nu)\vartheta + \nu s_k, \qquad k = 1,\dots,d.$$

To make the interpolation argument precise, introduce a new scalar variable α and set

$$g_k(\alpha) \coloneqq (1-\alpha)\vartheta + \alpha s_k, \qquad k = 1,\dots,d.$$

We first verify that all these quantities are positive on the admissible interval. For $0 \leq \alpha \leq 1$, $g_k(\alpha)$ is a convex combination of the positive numbers ϑ and s_k, and is therefore positive. The identity $\sum_{\ell=1}^d s_\ell = d\vartheta$ gives

$$g_k\left(-\frac{1}{d-1}\right) = \frac{d\vartheta - s_k}{d-1} = \frac{1}{d-1}\sum_{\ell\neq k} s_\ell > 0.$$

Since g_k is affine and $g_k(0) = \vartheta > 0$, it is positive throughout $[-1/(d-1), 0]$ as well. Consequently, every g_k is strictly positive on the full interval $[-1/(d-1), 1]$. The common positivity set

$$\mathcal{D} \coloneqq \{\alpha\in\mathbb{R} \mid g_k(\alpha) > 0 \text{ for every } k\}$$

is open and contains this entire closed interval.

We may therefore define

$$F(\alpha) \coloneqq \sum_{k=1}^{d} \ln\left(g_k(\alpha)\right)$$

on $\mathcal{D}$. Direct differentiation gives

$$F''(\alpha) = -\sum_{k=1}^{d} \frac{(s_k - \vartheta)^2}{\left((1-\alpha)\vartheta + \alpha s_k\right)^2} \leq 0.$$

Thus F is concave on the admissible interval. At the endpoints, $F(1) = \sum_k \ln s_k$, while the arithmetic–geometric mean inequality gives $F(0) = d \ln \vartheta \geq \sum_k \ln s_k$. Concavity proves the claim for $0 \leq \nu \leq 1$.

At $\nu = -1/(d-1)$,

$$t_k = \frac{\sum_{\ell \neq k} s_\ell}{d-1} \geq \left(\prod_{\ell \neq k} s_\ell\right)^{1/(d-1)}.$$

Multiplication over k gives $\prod_k t_k \geq \prod_k s_k$. Concavity between $-1/(d-1)$ and 0 completes the proof. ■

Theorem 10.7 (H-Theorem for the Homogeneous ES–BGK Equation)**.** *Let $f = f(\boldsymbol{v}, t) > 0$ be a sufficiently regular, rapidly decaying solution on a time interval $[0, T]$ of the spatially homogeneous ES–BGK initial-value problem*

$$\partial_t f(\boldsymbol{v}, t) = \frac{\lambda}{\varepsilon}\left(\mathcal{G}_\nu[f](\boldsymbol{v}, t) - f(\boldsymbol{v}, t)\right), \qquad f(\boldsymbol{v}, 0) = f_0(\boldsymbol{v}), \tag{10.30}$$

where $\lambda > 0$, $\varepsilon > 0$, ν lies in the range (10.4)*, and the pressure tensor remains positive definite. Assume also that the integrations and differentiations below are justified and that the information functional*

$$H[f(t)] \coloneqq \int_{\mathbb{R}^d} f(\boldsymbol{v}, t) \ln f(\boldsymbol{v}, t)\, \mathrm{d}\boldsymbol{v} \tag{10.31}$$

is finite. Then

$$\frac{\mathrm{d}}{\mathrm{d}t} H[f(t)] = \frac{\lambda}{\varepsilon} \int_{\mathbb{R}^d} (\mathcal{G}_\nu[f] - f) \ln f \, \mathrm{d}\boldsymbol{v} \leq 0. \tag{10.32}$$

Equivalently, the physical entropy $S[f] \coloneqq -H[f]$ satisfies $\mathrm{d}S/\mathrm{d}t \geq 0$. At any time t, equality in (10.32) *holds if and only if $f(\cdot, t)$ is the isotropic Maxwellian with moments $(n, \boldsymbol{u}, \theta)$.*

Proof. Integrating (10.30) and using the mass identity in Lemma 10.2 gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} f \, \mathrm{d}\boldsymbol{v} = \frac{\lambda}{\varepsilon} \int_{\mathbb{R}^d} (\mathcal{G}_\nu[f] - f)\, \mathrm{d}\boldsymbol{v} = 0. \tag{10.33}$$

Consequently, differentiation of H and the chain rule yield

$$\frac{\mathrm{d}}{\mathrm{d}t}H[f(t)] = \int_{\mathbb{R}^d} (\ln f + 1)\partial_t f \,\mathrm{d}\boldsymbol{v} \tag{10.34}$$

$$= \frac{\lambda}{\varepsilon}\int_{\mathbb{R}^d} (\mathcal{G}_\nu[f] - f)(\ln f + 1)\,\mathrm{d}\boldsymbol{v} \tag{10.35}$$

$$= \frac{\lambda}{\varepsilon}\int_{\mathbb{R}^d} (\mathcal{G}_\nu[f] - f)\ln f \,\mathrm{d}\boldsymbol{v}, \tag{10.36}$$

where the term containing 1 vanishes by (10.33).

Let $\mathcal{G}_1$ denote the Gaussian with the same mass, mean, and full covariance P/ρ as f. Because $\ln \mathcal{G}_1$ is a linear combination of 1, the components of $\boldsymbol{v}$, and the entries of $\boldsymbol{v}\otimes\boldsymbol{v}$, moment matching gives

$$H[f] - H[\mathcal{G}_1] = \int_{\mathbb{R}^d} f \ln\left(\frac{f}{\mathcal{G}_1}\right)\,\mathrm{d}\boldsymbol{v} \geq 0. \tag{10.37}$$

The inequality is the nonnegativity of relative information after normalizing both distributions by their common mass n; equality holds only when $f = \mathcal{G}_1$ almost everywhere. Thus $H[\mathcal{G}_1] \leq H[f]$. Direct evaluation of the two Gaussian information functionals and Lemma 10.6 give

$$H[\mathcal{G}_\nu] - H[\mathcal{G}_1] = -\frac{n}{2}\ln\frac{\det \mathsf{T}_\nu}{\det(\mathsf{P}/\rho)} \leq 0.$$

It follows that $H[\mathcal{G}_\nu] \leq H[f]$.

Now let $\Phi(z) = z\ln z$, which is strictly convex on $(0,\infty)$. The tangent-line inequality

$$\Phi(y) \geq \Phi(x) + \Phi'(x)(y - x) \tag{10.38}$$

with $x = f$ and $y = \mathcal{G}_\nu$ gives, after integration,

$$H[\mathcal{G}_\nu] - H[f] \geq \int_{\mathbb{R}^d} (\mathcal{G}_\nu - f)(\ln f + 1)\,\mathrm{d}\boldsymbol{v} \tag{10.39}$$

$$= \int_{\mathbb{R}^d} (\mathcal{G}_\nu - f)\ln f \,\mathrm{d}\boldsymbol{v}, \tag{10.40}$$

where equal mass again removes the linear term. Combining this inequality with $H[\mathcal{G}_\nu] \leq H[f]$ and (10.36) proves (10.32).

It remains to characterize equality. If $\mathrm{d}H/\mathrm{d}t = 0$, then equality must hold in the integrated strict-convexity inequality, and therefore $f = \mathcal{G}_\nu[f]$ almost everywhere. Taking the centered

second moment of this identity and using Lemma 10.2 gives

$$\mathsf{P} = \rho \mathsf{T}_\nu = (1-\nu) p \mathsf{I} + \nu \mathsf{P}. \tag{10.41}$$

Since $\nu < 1$, this implies $\mathsf{P} = p\mathsf{I}$. Hence $\mathsf{T}_\nu = \vartheta \mathsf{I}$ and f is the isotropic Maxwellian with moments $(n, \boldsymbol{u}, \theta)$. Conversely, for that Maxwellian one has $\mathcal{G}_\nu[f] = f$, so equality holds in (10.32). ■

This proof follows the structural argument of [2]; sharper quantitative entropy-production estimates are available in [43].

10.3 Chapman–Enskog Expansion

We now apply exactly the normal-solution convention of Chapter 3. Before specializing to first order, we record the order-by-order organization used in the Year Four Report. Formally, write

$$f \sim \sum_{k\geq 0} \varepsilon^k f^{(k)}, \qquad \mathcal{G}_\nu[f] \sim \sum_{k\geq 0} \varepsilon^k \mathcal{G}^{(k)}, \qquad \partial_t \sim \sum_{k\geq 0} \varepsilon^k \partial_t^{(k)}. \tag{10.42}$$

The macroscopic fields are assigned entirely to the leading term. Thus

$$m \int_{\mathbb{R}^d} f^{(0)} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2/2 \end{pmatrix} \mathrm{d}\boldsymbol{v} = \begin{pmatrix} \rho \\ \rho \boldsymbol{u} \\ \omega \end{pmatrix}, \tag{10.43a}$$

$$m \int_{\mathbb{R}^d} f^{(k)} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2/2 \end{pmatrix} \mathrm{d}\boldsymbol{v} = \boldsymbol{0}, \qquad k \geq 1. \tag{10.43b}$$

Consequently, the nonequilibrium moments have expansions

$$\Pi \sim \sum_{k\geq 1} \varepsilon^k \Pi^{(k)}, \qquad \boldsymbol{q} \sim \sum_{k\geq 1} \varepsilon^k \boldsymbol{q}^{(k)}, \tag{10.44}$$

where

$$\Pi^{(k)} = m \int_{\mathbb{R}^d} \mathsf{A}(\boldsymbol{c}) f^{(k)} \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{q}^{(k)} = \frac{m}{2} \int_{\mathbb{R}^d} |\boldsymbol{c}|^2 \boldsymbol{c} f^{(k)} \, \mathrm{d}\boldsymbol{v}. \tag{10.45}$$

Here and below $\boldsymbol{c} = \boldsymbol{v} - \boldsymbol{u}$, and A is the traceless quadratic mode defined in (10.54).

For a normal solution, $f^{(k)}$ depends on $\boldsymbol{x}$ and t through the macroscopic fields and their

spatial derivatives through order k:

$$f^{(k)} = f^{(k)}[U, \nabla U, \dots, \nabla^k U](\boldsymbol{v}), \qquad U = (n, \boldsymbol{u}, \theta). \tag{10.46}$$

Time derivatives of these functionals are interpreted by the chain rule. For example,

$$\partial_t^{(a)} f^{(k)} = \sum_{|\alpha| \le k} \frac{\partial f^{(k)}}{\partial(\partial_x^\alpha U)} \partial_x^\alpha \big(\partial_t^{(a)} U\big). \tag{10.47}$$

Substitution of (10.42) into the kinetic equation first gives

$$\mathcal{G}^{(0)} = f^{(0)} = M[U]. \tag{10.48}$$

At every subsequent order $r \ge 0$, it gives the recursion

$$\sum_{a+b=r} \partial_t^{(a)} f^{(b)} + \boldsymbol{v} \cdot \nabla_x f^{(r)} = \lambda\big(\mathcal{G}^{(r+1)} - f^{(r+1)}\big). \tag{10.49}$$

The corresponding coefficients of the exact balance laws are

$$\partial_t^{(0)} \rho + \nabla \cdot (\rho \boldsymbol{u}) = 0, \qquad \partial_t^{(k)} \rho = 0, \quad k \ge 1, \tag{10.50a}$$

$$\partial_t^{(0)} (\rho \boldsymbol{u}) + \nabla \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + p \mathsf{I}) = \boldsymbol{0}, \qquad \partial_t^{(k)} (\rho \boldsymbol{u}) + \nabla \cdot \Pi^{(k)} = \boldsymbol{0}, \quad k \ge 1, \tag{10.50b}$$

$$\partial_t^{(0)} \omega + \nabla \cdot \big((\omega + p)\boldsymbol{u}\big) = 0, \quad \partial_t^{(k)} \omega + \nabla \cdot \big(\Pi^{(k)} \boldsymbol{u} + \boldsymbol{q}^{(k)}\big) = 0, \quad k \ge 1. \tag{10.50c}$$

Equations (10.47)–(10.50) provide the route to Burnett and higher closures; see [83, 74]. We use only the case $r = 0$ below.

Equivalently, set $f^{(1)} = M\phi^{(1)}$ and write

$$f = M\big(1 + \varepsilon\phi^{(1)} + O(\varepsilon^2)\big), \qquad M = n(2\pi\vartheta)^{-d/2} \exp\left(-\frac{|\boldsymbol{c}|^2}{2\vartheta}\right), \qquad \boldsymbol{c} = \boldsymbol{v} - \boldsymbol{u}, \tag{10.51}$$

with the moment normalization

$$\int_{\mathbb{R}^d} M\phi^{(1)} \begin{pmatrix} 1 \\ \boldsymbol{v} \\ |\boldsymbol{v}|^2 \end{pmatrix} \mathrm{d}\boldsymbol{v} = \boldsymbol{0}. \tag{10.52}$$

We use the Chapman–Enskog time-derivative convention introduced in Chapter 3:

$$\partial_t U \sim \partial_t^{(0)} U + \varepsilon \partial_t^{(1)} U + \cdots, \qquad U = (n, \boldsymbol{u}, \theta). \tag{10.53}$$

Here $\partial_t^{(0)} U$ is the leading evolution coefficient defined by (3.66). The notation is extended to any function of U by the chain rule; thus $\partial_t^{(0)}$ acts by evolving the macroscopic fields with the compressible Euler equations. It is not differentiation with respect to a separate time variable. At leading order, $\mathsf{P} = p\mathsf{I}$, hence $\mathsf{T}_\nu = \vartheta\mathsf{I}$ and $\mathcal{G}_\nu = M$; the order-zero solvability conditions are precisely the Euler equations of Proposition 3.7.

Define the **traceless quadratic mode** and **heat-flux mode**

$$\mathsf{A}(\boldsymbol{c}) \coloneqq \boldsymbol{c}\otimes\boldsymbol{c} - \frac{|\boldsymbol{c}|^2}{d}\mathsf{I}, \qquad \boldsymbol{B}(\boldsymbol{c}) \coloneqq \left(\frac{|\boldsymbol{c}|^2}{2\vartheta} - \frac{d+2}{2}\right)\boldsymbol{c}. \tag{10.54}$$

They are orthogonal to the collision invariants in the Maxwellian inner product used in (3.32).

Proposition 10.8 (First-Order ES-BGK Closure)**.** *Under the formal Chapman–Enskog assumptions of Chapter 3, the first-order constitutive laws are*

$$\Pi = -\varepsilon\mu_\nu\left(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^\top - \frac{2}{d}(\nabla\cdot\boldsymbol{u})\mathsf{I}\right) + O(\varepsilon^2), \qquad \boldsymbol{q} = -\varepsilon\kappa\nabla\theta + O(\varepsilon^2), \tag{10.55}$$

where

$$\mu_\nu = \frac{p}{\lambda(1-\nu)}, \qquad \kappa = \frac{d+2}{2m}\frac{p}{\lambda}, \qquad \mathrm{Pr} = \frac{c_p\mu_\nu}{\kappa} = \frac{1}{1-\nu}, \qquad c_p = \frac{d+2}{2m}. \tag{10.56}$$

Proof. The purpose of the calculation is to determine the part of $\phi^{(1)}$ that contributes to the nonequilibrium stress and heat flux.

Step 1: Compute the transport derivative of the Maxwellian. Set

$$\mathsf{D}^\circ(\boldsymbol{u}) \coloneqq \frac{1}{2}\left(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^\top\right) - \frac{1}{d}(\nabla\cdot\boldsymbol{u})\mathsf{I}. \tag{10.57}$$

At order zero, the number density n, bulk velocity $\boldsymbol{u}$, and temperature θ satisfy the compressible Euler equations. In the notation of (10.53), $\partial_t^{(0)}$ means that these fields are differentiated using that Euler evolution. Rewriting the Euler mass, momentum, and temperature equations in primitive variables, and defining the associated material derivative $D_t^{(0)} = \partial_t^{(0)} + \boldsymbol{u}\cdot\nabla$, gives

$$D_t^{(0)}\log n = -\nabla\cdot\boldsymbol{u}, \qquad D_t^{(0)}\boldsymbol{u} = -\frac{1}{\rho}\nabla p, \qquad D_t^{(0)}\log\theta = -\frac{2}{d}\nabla\cdot\boldsymbol{u}. \tag{10.58}$$

Because $\rho = mn$ and $p = n\theta$, the momentum equation in (10.58) may also be written as

$$D_t^{(0)}\boldsymbol{u} = -\frac{1}{mn}\nabla(n\theta) = -\vartheta\nabla\log n - \frac{1}{m}\nabla\theta. \tag{10.59}$$

We now compute the left-hand side of the first-order kinetic equation. Since $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{c}$, the free-transport derivative decomposes as

$$\mathcal{D}^{(0)} := \partial_t^{(0)} + \boldsymbol{v} \cdot \nabla = D_t^{(0)} + \boldsymbol{c} \cdot \nabla. \tag{10.60}$$

Differentiate

$$\log M = \log n - \frac{d}{2} \log(2\pi\vartheta) - \frac{|\boldsymbol{c}|^2}{2\vartheta}, \qquad \boldsymbol{c} = \boldsymbol{v} - \boldsymbol{u}, \tag{10.61}$$

at fixed $\boldsymbol{v}$. The identities

$$\mathcal{D}^{(0)}\boldsymbol{c} = -\mathcal{D}^{(0)}\boldsymbol{u}, \qquad \mathcal{D}^{(0)}|\boldsymbol{c}|^2 = -2\boldsymbol{c} \cdot \mathcal{D}^{(0)}\boldsymbol{u} \tag{10.62}$$

give

$$\begin{aligned} \frac{\mathcal{D}^{(0)}M}{M} &= \mathcal{D}^{(0)} \log M \\ &= D_t^{(0)} \log n + \boldsymbol{c} \cdot \nabla \log n - \frac{d}{2} \left(D_t^{(0)} \log \theta + \boldsymbol{c} \cdot \nabla \log \theta \right) \\ &\quad + \frac{1}{\vartheta} \boldsymbol{c} \cdot \left(D_t^{(0)} \boldsymbol{u} + (\boldsymbol{c} \cdot \nabla)\boldsymbol{u} \right) + \frac{|\boldsymbol{c}|^2}{2\vartheta} \left(D_t^{(0)} \log \theta + \boldsymbol{c} \cdot \nabla \log \theta \right). \end{aligned} \tag{10.63}$$

This formula makes the cancellations transparent. Substitute (10.58) and (10.59). The term $\boldsymbol{c} \cdot \nabla \log n$ cancels the density-gradient part of $\vartheta^{-1}\boldsymbol{c} \cdot D_t^{(0)}\boldsymbol{u}$. Since $\vartheta = \theta/m$, the remaining temperature-gradient terms combine as

$$\left(\frac{|\boldsymbol{c}|^2}{2\vartheta} - \frac{d+2}{2} \right) \boldsymbol{c} \cdot \nabla \log \theta = \boldsymbol{B}(\boldsymbol{c}) \cdot \nabla \log \theta. \tag{10.64}$$

The terms containing velocity gradients are

$$\frac{1}{\vartheta}(\boldsymbol{c} \otimes \boldsymbol{c}) : \nabla \boldsymbol{u} - \frac{|\boldsymbol{c}|^2}{d\vartheta} \nabla \cdot \boldsymbol{u}. \tag{10.65}$$

Only the symmetric part of $\nabla \boldsymbol{u}$ contributes because $\boldsymbol{c} \otimes \boldsymbol{c}$ is symmetric. Removing its trace therefore gives

$$\frac{1}{\vartheta} \mathsf{A}(\boldsymbol{c}) : \mathsf{D}^{\circ}(\boldsymbol{u}) = \frac{m}{\theta} \mathsf{A}(\boldsymbol{c}) : \mathsf{D}^{\circ}(\boldsymbol{u}). \tag{10.66}$$

Consequently,

$$\frac{\mathcal{D}^{(0)}M}{M} = \frac{m}{\theta} \mathsf{A}(\boldsymbol{c}) : \mathsf{D}^{\circ}(\boldsymbol{u}) + \boldsymbol{B}(\boldsymbol{c}) \cdot \nabla \log \theta. \tag{10.67}$$

Thus the transport source has exactly two non-hydrodynamic parts: a symmetric trace-free quadratic mode and an odd cubic heat-flux mode.

Step 2: Linearize the ES–BGK target and solve for $\phi^{(1)}$. Write

$$\mathsf{P} = p\mathsf{I} + \varepsilon\Pi^{(1)} + \mathcal{O}(\varepsilon^2), \qquad \Pi^{(1)} = m\int_{\mathbb{R}^d} M\mathsf{A}(\boldsymbol{c})\phi^{(1)}\,\mathrm{d}\boldsymbol{v}. \tag{10.68}$$

To see the second identity, recall that $\Pi = \mathsf{P} - p\mathsf{I}$ and use the normalization conditions: the first-order changes of n, $\boldsymbol{u}$, and the scalar temperature all vanish. In particular, $\operatorname{tr}\Pi^{(1)} = 0$. It follows from (10.5) that

$$\mathsf{T}_\nu = \vartheta\mathsf{I} + \varepsilon\frac{\nu}{\rho}\Pi^{(1)} + \mathcal{O}(\varepsilon^2). \tag{10.69}$$

We next expand the Gaussian target with respect to its covariance. For a symmetric perturbation $\delta\mathsf{T}$ of a covariance tensor, Jacobi's formula and the derivative of a matrix inverse give

$$\delta\log\mathcal{G} = -\frac{1}{2}\operatorname{tr}(\mathsf{T}^{-1}\delta\mathsf{T}) + \frac{1}{2}\boldsymbol{c}\cdot\mathsf{T}^{-1}\delta\mathsf{T}\mathsf{T}^{-1}\boldsymbol{c}. \tag{10.70}$$

Indeed, the first term comes from differentiating $(\det\mathsf{T})^{-1/2}$, and the second comes from differentiating the quadratic form in the exponential. Set $\delta\mathsf{T} = (\nu/\rho)\Pi^{(1)}$ and evaluate (10.70) at $\mathsf{T} = \vartheta\mathsf{I}$. The trace term vanishes, while the quadratic term becomes

$$\begin{aligned}\delta\log\mathcal{G} &= \frac{\nu}{2\rho\vartheta^2}(\boldsymbol{c}\otimes\boldsymbol{c}) : \Pi^{(1)} \\ &= \frac{\nu}{2\rho\vartheta^2}\mathsf{A}(\boldsymbol{c}) : \Pi^{(1)} = \nu\frac{m}{2n\theta^2}\Pi^{(1)} : \mathsf{A}(\boldsymbol{c}).\end{aligned} \tag{10.71}$$

In the second line we used $\operatorname{tr}\Pi^{(1)} = 0$; in the last equality we used $\rho = mn$ and $\vartheta = \theta/m$. Hence

$$\mathcal{G}_\nu[f] = M\left[1 + \varepsilon\nu\mathcal{P}_2\phi^{(1)}\right] + \mathcal{O}(\varepsilon^2), \tag{10.72}$$

where we have defined the traceless-quadratic projection

$$\mathcal{P}_2\phi^{(1)} \coloneqq \frac{m}{2n\theta^2}\Pi^{(1)} : \mathsf{A}(\boldsymbol{c}). \tag{10.73}$$

The terminology "projection" can be checked directly. If $\psi(\boldsymbol{c}) = \mathsf{A}(\boldsymbol{c}) : \mathsf{C}$ for a symmetric trace-free tensor C, the centered Gaussian fourth-moment identity gives $\mathcal{P}_2\psi = \psi$. If ψ is odd in $\boldsymbol{c}$, its contribution to $\Pi^{(1)}$ vanishes by parity and $\mathcal{P}_2\psi = 0$. Thus the ES–BGK target retains the fraction ν of a stress perturbation and no part of a heat-flux perturbation.

We now insert the expansions into the kinetic equation. Its left-hand side is $\mathcal{D}f = \mathcal{D}^{(0)}M + O(\varepsilon)$, whereas its right-hand side is

$$\frac{\lambda}{\varepsilon}(\mathcal{G}_\nu - f) = -\lambda M(\mathsf{Id} - \nu\mathcal{P}_2)\phi^{(1)} + O(\varepsilon). \tag{10.74}$$

Equating the coefficients of ε^0 and dividing by M gives

$$\frac{\mathcal{D}^{(0)}M}{M} = -\lambda\left(\mathrm{Id} - \nu\mathcal{P}_2\right)\phi^{(1)}. \tag{10.75}$$

Equation (10.67) is a sum of a stress mode and a heat-flux mode. On the former, $\mathrm{Id} - \nu\mathcal{P}_2$ acts by multiplication by $1-\nu$; on the latter, it acts by multiplication by 1. We may therefore invert the operator separately on the two modes and obtain

$$\phi^{(1)} = -\frac{m}{\lambda(1-\nu)\theta}\mathsf{A}(\boldsymbol{c}) : \mathsf{D}^{\circ}(\boldsymbol{u}) - \frac{1}{\lambda}\boldsymbol{B}(\boldsymbol{c})\cdot\nabla\log\theta. \tag{10.76}$$

The first summand is even and trace-free, and the second is odd with its coefficient chosen to have zero momentum moment. Consequently both are orthogonal, in the Maxwellian inner product, to 1, $\boldsymbol{v}$, and $|\boldsymbol{v}|^2$. Thus (10.76) satisfies the normalization (10.52) and is the unique normalized first-order correction.

Step 3: Take the stress moment. For a centered Maxwellian with covariance $\vartheta\mathsf{I}$,

$$\int_{\mathbb{R}^d} M c_i c_j c_k c_\ell \,\mathrm{d}\boldsymbol{v} = n\vartheta^2(\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}). \tag{10.77}$$

Contract this formula with a symmetric trace-free tensor D. The term containing $\delta_{ij}\delta_{k\ell}$ vanishes because $\operatorname{tr}\mathsf{D} = 0$, and the other two terms each produce D. Therefore

$$\int_{\mathbb{R}^d} M\mathsf{A}(\boldsymbol{c})\big(\mathsf{A}(\boldsymbol{c}) : \mathsf{D}\big)\,\mathrm{d}\boldsymbol{v} = 2n\vartheta^2\mathsf{D}. \tag{10.78}$$

The heat-flux term in (10.76) is odd, whereas A is even, so its stress moment is zero. Using (10.78) for the remaining term gives

$$\Pi = \varepsilon m\int_{\mathbb{R}^d} M\mathsf{A}\phi^{(1)}\,\mathrm{d}\boldsymbol{v} + \mathcal{O}(\varepsilon^2) \tag{10.79}$$

$$= -\frac{2\varepsilon p}{\lambda(1-\nu)}\mathsf{D}^{\circ}(\boldsymbol{u}) + \mathcal{O}(\varepsilon^2) \tag{10.80}$$

$$= -\varepsilon\mu_\nu\left(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\top} - \frac{2}{d}(\nabla\cdot\boldsymbol{u})\mathsf{I}\right) + \mathcal{O}(\varepsilon^2), \tag{10.81}$$

where $p = n\theta$ and

$$\mu_\nu = \frac{p}{\lambda(1-\nu)}. \tag{10.82}$$

Step 4: Take the heat-flux moment. The heat flux is $\boldsymbol{q} = (m/2)\int \boldsymbol{c}|\boldsymbol{c}|^2 f\,\mathrm{d}\boldsymbol{v}$. The equilibrium Maxwellian has zero heat flux because its integrand is odd. The stress term in (10.76) also

gives zero: it is even, so its product with $\boldsymbol{c}|\boldsymbol{c}|^2$ is odd. Only the heat-flux mode remains.

For completeness, isotropy implies that the tensor $\int M\boldsymbol{c}|\boldsymbol{c}|^2 \otimes \boldsymbol{B}\,\mathrm{d}\boldsymbol{v}$ is a scalar multiple of I. Its trace, and hence that scalar, can be evaluated from

$$\int_{\mathbb{R}^d} M|\boldsymbol{c}|^4\,\mathrm{d}\boldsymbol{v} = nd(d+2)\vartheta^2, \qquad \int_{\mathbb{R}^d} M|\boldsymbol{c}|^6\,\mathrm{d}\boldsymbol{v} = nd(d+2)(d+4)\vartheta^3. \tag{10.83}$$

Indeed, using the definition of $\boldsymbol{B}$ in (10.54),

$$\operatorname{tr}\int_{\mathbb{R}^d} M\boldsymbol{c}|\boldsymbol{c}|^2 \otimes \boldsymbol{B}\,\mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} M|\boldsymbol{c}|^4\left(\frac{|\boldsymbol{c}|^2}{2\vartheta} - \frac{d+2}{2}\right)\mathrm{d}\boldsymbol{v} \tag{10.84}$$

$$= nd(d+2)\vartheta^2. \tag{10.85}$$

Dividing by d gives

$$\int_{\mathbb{R}^d} M\,\boldsymbol{c}|\boldsymbol{c}|^2 \otimes \boldsymbol{B}(\boldsymbol{c})\,\mathrm{d}\boldsymbol{v} = n(d+2)\vartheta^2\mathsf{I}. \tag{10.86}$$

Substituting the heat-flux part of (10.76) into the definition of $\boldsymbol{q}$ therefore gives

$$\boldsymbol{q} = \varepsilon\frac{m}{2}\int_{\mathbb{R}^d} M\boldsymbol{c}|\boldsymbol{c}|^2\phi^{(1)}\,\mathrm{d}\boldsymbol{v} + \mathcal{O}(\varepsilon^2) \tag{10.87}$$

$$= -\varepsilon\frac{m}{2\lambda}n(d+2)\vartheta^2\nabla\log\theta + \mathcal{O}(\varepsilon^2) \tag{10.88}$$

$$= -\varepsilon\frac{d+2}{2m}\frac{p}{\lambda}\nabla\theta + \mathcal{O}(\varepsilon^2). \tag{10.89}$$

In the last line we used $\vartheta = \theta/m$, $p = n\theta$, and $\nabla\log\theta = (\nabla\theta)/\theta$. Hence Fourier's law holds with

$$\kappa = \frac{d+2}{2m}\frac{p}{\lambda}. \tag{10.90}$$

Finally, the specific heat at constant pressure for a monatomic gas is $c_p = (d+2)/(2m)$. The Prandtl number associated with the two coefficients just obtained is therefore

$$\mathrm{Pr} = \frac{c_p\mu_\nu}{\kappa} = \frac{1}{1-\nu}. \tag{10.91}$$

This proves Equations (10.55) to (10.56). Substitution into the exact balance laws in Equation (10.18) gives the compressible Navier–Stokes–Fourier system. As in Chapter 3, the expansion is formal: it assumes a smooth normal solution and neglects terms of order ε^2. ■

Thus $\nu = 1 - 1/\mathrm{Pr}$. In three velocity dimensions, $\nu = -1/2$ yields $\mathrm{Pr} = 2/3$. This is the principal advantage of ES-BGK over BGK. The derivation agrees with the classical analysis in [39, 2, 74].

10.4 Implicit Collision Update

To derive the implicit collision update, we temporarily suppress the transport term in the full ES–BGK equation and fix a spatial position $\boldsymbol{x}$. The distribution then satisfies the velocity-space ordinary differential equation

$$\partial_t f = \frac{\lambda}{\varepsilon}\big(\mathcal{G}_\nu[f] - f\big), \tag{10.92}$$

where the frozen dependence on $\boldsymbol{x}$ has been suppressed. In an IMEX or operator-split discretization, this collision-only evolution is the **collision stage**, or **collision substep**, of the full update. During this stage, distinct spatial points evolve independently.

The target $\mathcal{G}_\nu[f]$ has the same mass, momentum, and total energy as f. Therefore, taking the three collision-invariant moments of (10.92) gives

$$\partial_t n = 0, \qquad \partial_t(n\boldsymbol{u}) = \boldsymbol{0}, \qquad \partial_t\left(\frac{1}{2}\rho|\boldsymbol{u}|^2 + \frac{d}{2}n\theta\right) = 0. \tag{10.93}$$

Since $n > 0$ and $\rho = mn$, it follows that n, $\boldsymbol{u}$, and the scalar temperature θ remain fixed throughout the collision stage. The anisotropic pressure tensor P may nevertheless evolve while its trace $dp = dn\theta$ remains fixed. Taking the centered second moment of (10.92) gives the closed equation

$$\partial_t \mathsf{P} = \frac{\lambda}{\varepsilon}(\rho\mathsf{T}_\nu - \mathsf{P}) = -\frac{\lambda(1-\nu)}{\varepsilon}\Pi. \tag{10.94}$$

Proposition 10.9 (Backward-Euler ES-BGK Collision Step)**.** *Fix a spatial point and let $\varepsilon > 0$, $\Delta t > 0$, $\lambda > 0$, and $-1/(d-1) \le \nu < 1$. Here λ is held fixed during the collision stage; in particular, it may be evaluated from the collision-invariant moments of f^n. Suppose that $f^n \ge 0$ has a finite second moment, positive number density n^n, and positive-definite pressure tensor P^n. Compute $\rho^n = mn^n$, $\boldsymbol{u}^n$, θ^n, $\vartheta^n = \theta^n/m$, and $p^n = n^n\theta^n$, and define*

$$\mathsf{P}^{n+1} = \frac{\varepsilon}{\varepsilon + \Delta t\lambda(1-\nu)}\mathsf{P}^n + \frac{\Delta t\lambda(1-\nu)}{\varepsilon + \Delta t\lambda(1-\nu)}p^n\mathsf{I}, \tag{10.95}$$

$$\mathsf{T}_\nu^{n+1} = (1-\nu)\vartheta^n\mathsf{I} + \nu\frac{\mathsf{P}^{n+1}}{\rho^n}, \qquad \mathcal{G}_\nu^{n+1} = \mathcal{G}(n^n, \boldsymbol{u}^n, \mathsf{T}_\nu^{n+1}), \tag{10.96}$$

$$f^{n+1} = \frac{\varepsilon}{\varepsilon + \Delta t\lambda}f^n + \frac{\Delta t\lambda}{\varepsilon + \Delta t\lambda}\mathcal{G}_\nu^{n+1}. \tag{10.97}$$

Then T_ν^{n+1} is positive definite, so $\mathcal{G}_\nu^{n+1}$ is well defined. Moreover, the distribution in (10.97)

has moments

$$n^{n+1} = n^n, \qquad \boldsymbol{u}^{n+1} = \boldsymbol{u}^n, \qquad \theta^{n+1} = \theta^n, \qquad \mathsf{P}[f^{n+1}] = \mathsf{P}^{n+1}, \tag{10.98}$$

and is the unique solution of the fully implicit backward-Euler equation

$$\frac{f^{n+1} - f^n}{\Delta t} = \frac{\lambda}{\varepsilon}\big(\mathcal{G}_\nu[f^{n+1}] - f^{n+1}\big). \tag{10.99}$$

Consequently, the collision step preserves mass, momentum, total energy, and nonnegativity.

Proof. Because the collision stage preserves n, $\boldsymbol{u}$, and θ, it also preserves ρ, ϑ, and p. Backward Euler applied to (10.94) therefore gives

$$\frac{\mathsf{P}^{n+1} - \mathsf{P}^n}{\Delta t} = -\frac{\lambda(1-\nu)}{\varepsilon}\big(\mathsf{P}^{n+1} - p^n \mathsf{I}\big). \tag{10.100}$$

Solving this linear tensor equation for P^{n+1} gives (10.95). Its two coefficients are positive and sum to one. Since P^n is positive definite and $p^n > 0$, the tensor P^{n+1} is positive definite. Taking its trace in (10.95) yields

$$\operatorname{tr}\mathsf{P}^{n+1} = \operatorname{tr}\mathsf{P}^n = dp^n. \tag{10.101}$$

Thus the updated pressure tensor has the same scalar temperature as f^n. The admissible range of ν and Lemma 10.2 now imply that T_ν^{n+1} is positive definite and that the Gaussian in (10.96) is well defined.

It remains to verify that this explicitly constructed Gaussian is the target of the updated distribution, rather than merely a target formed from the old moments. Set

$$a = \frac{\varepsilon}{\varepsilon + \Delta t\lambda}, \qquad b = \frac{\Delta t\lambda}{\varepsilon + \Delta t\lambda}, \qquad a + b = 1. \tag{10.102}$$

By the target-moment identities in Equation (10.8), equation (10.97) gives

$$\int_{\mathbb{R}^d} f^{n+1}\,\mathrm{d}\boldsymbol{v} = an^n + bn^n = n^n, \tag{10.103}$$

$$\int_{\mathbb{R}^d} \boldsymbol{v} f^{n+1}\,\mathrm{d}\boldsymbol{v} = an^n\boldsymbol{u}^n + bn^n\boldsymbol{u}^n = n^n\boldsymbol{u}^n. \tag{10.104}$$

Hence the mean velocity of f^{n+1} is $\boldsymbol{u}^n$. Its centered second moment is consequently

$$\begin{aligned}\mathsf{P}[f^{n+1}] &= a\mathsf{P}^n + b\rho^n \mathsf{T}_\nu^{n+1} \\ &= a\mathsf{P}^n + b\big((1-\nu)p^n\mathsf{I} + \nu\mathsf{P}^{n+1}\big) = \mathsf{P}^{n+1}.\end{aligned} \tag{10.105}$$

The last equality is exactly (10.100), after multiplication by Δt and rearrangement. Together with (10.101), this proves all the identities in (10.98). In particular,

$$\mathcal{G}_\nu^{n+1} = \mathcal{G}_\nu[f^{n+1}]. \tag{10.106}$$

Finally, multiplying (10.97) by $\varepsilon + \Delta t\lambda$ and using (10.106) gives (10.99). Conversely, taking the invariant and centered second moments of any solution of (10.99) forces (10.98) and (10.95); the kinetic equation then forces (10.97). This proves uniqueness.

The weights a and b in (10.102) are nonnegative. Since $f^n \geq 0$ and $\mathcal{G}_\nu^{n+1} > 0$, (10.97) also gives $f^{n+1} \geq 0$. The mass, momentum, and energy identities already proved give the claimed conservation properties. ■

This moment-first inversion is the key observation behind asymptotic-preserving IMEX schemes for ES-BGK [25]. It fits the transport/collision splitting and Runge–Kutta framework of Chapter 6; no nonlinear iteration over the velocity grid is required for the single-species collision step.

10.5 Numerical Demonstration of the IMEX Method

We now restore the numerical experiment reported in the Year Four Report. Its purpose is modest but useful: it verifies that the moment-first collision update of Proposition 10.9 can be used inside an IMEX method without solving a nonlinear system on the velocity grid.

10.5.1 Slab problem and initial data

We solve the ES–BGK equation in a one-dimensional physical domain with two velocity coordinates,

$$\partial_t f + v_x \partial_x f = \frac{\lambda}{\varepsilon}\big(\mathcal{G}_\nu[f] - f\big), \qquad x \in [-1, 2], \qquad (v_x, v_y) \in [-15, 15]^2. \tag{10.107}$$

This is a $1x2v$ slab reduction: the distribution is independent of the second spatial coordinate but retains both velocity components. It can therefore represent transverse momentum and

an anisotropic pressure tensor.

The initial distribution is piecewise Maxwellian,

$$f(x, \boldsymbol{v}, 0) = \begin{cases} M_{n_L, \boldsymbol{u}_L, \theta_L}(\boldsymbol{v}), & -1 \leq x < 0, \\ M_{n_R, \boldsymbol{u}_R, \theta_R}(\boldsymbol{v}), & 0 \leq x \leq 2, \end{cases} \tag{10.108}$$

with

$$n_L = 1, \qquad \boldsymbol{u}_L = (2.5\sqrt{2}, 0)^\top, \qquad \theta_L = 1, \tag{10.109}$$

$$n_R = 1, \qquad \boldsymbol{u}_R = \boldsymbol{0}, \qquad \theta_R = 1.05. \tag{10.110}$$

Far-field inflow data are imposed by filling the left and right ghost cells with the corresponding initial Maxwellians.

10.5.2 Discretization and parameters

The physical interval is divided into $N_x = 300$ uniform finite volumes, so $\Delta x = 1/100$. Transport is discretized by the second-order upwind finite-volume method of Chapter 6, using piecewise linear reconstruction and the θ-minmod limiter; the interface state is selected from the left or right reconstruction according to the sign of v_x. The velocity square is represented by an 80×80 Cartesian grid with

$$\Delta v_x = \Delta v_y = \frac{30}{80}, \tag{10.111}$$

and velocity moments are computed with midpoint quadrature.

The explicit transport term and implicit collision term are coupled with a two-stage, second-order IMEX Runge–Kutta method. At each implicit stage we first update the invariant moments and pressure tensor, then form the target Gaussian, and finally update the distribution. Thus the stage uses the same moment-first idea as Proposition 10.9. The time step is chosen from the transport CFL condition,

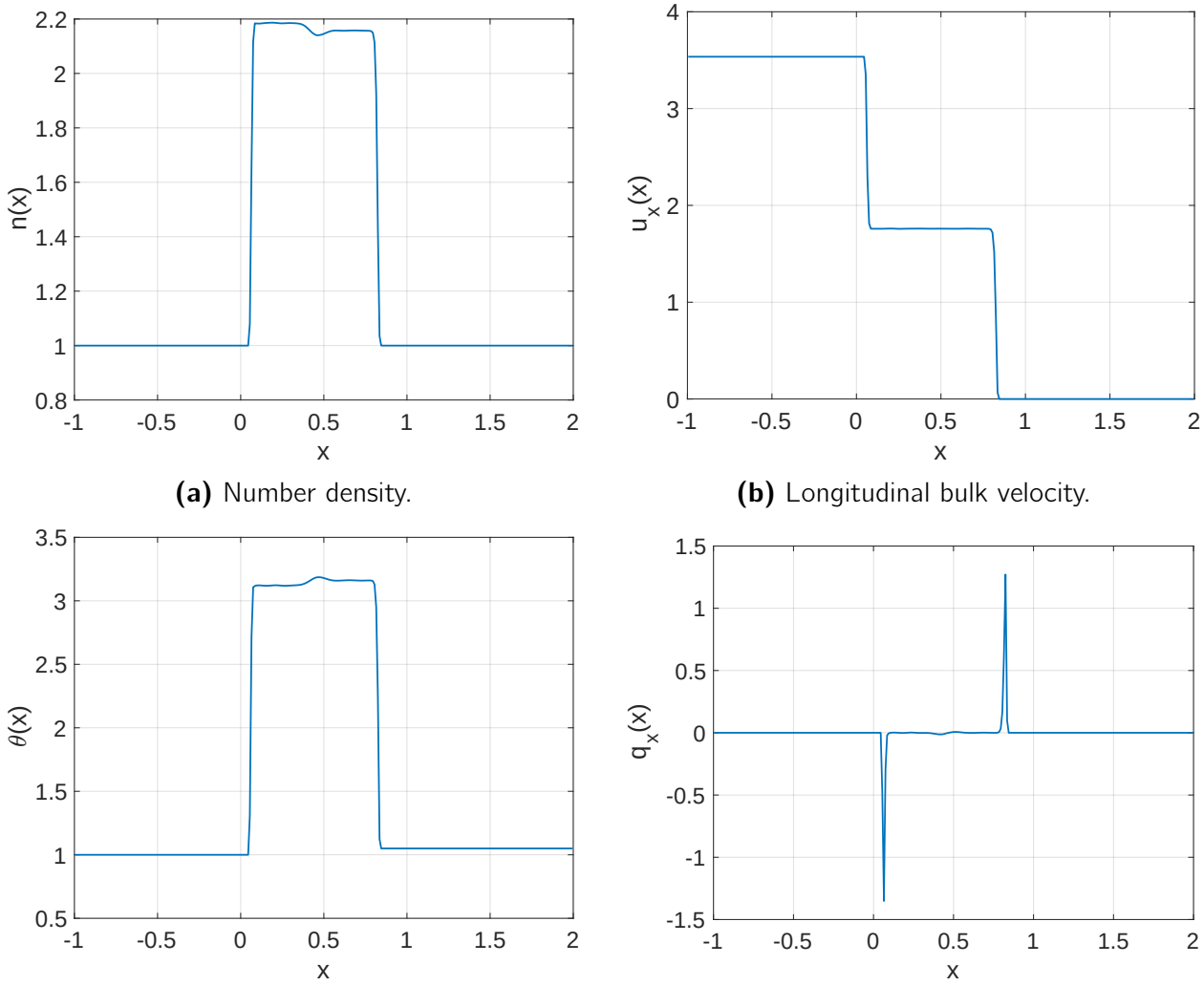
$$\Delta t = 0.9 \frac{\Delta x}{2 \max_\ell |v_{x,\ell}|} = \frac{0.9(1/100)}{2(15)}. \tag{10.112}$$

The factor 2 is the transport restriction used with the piecewise-linear minmod reconstruction

in this implementation. Following the benchmark in [25], we take

$$\nu = \frac{1}{2}, \qquad \lambda(x,t) = \frac{3\pi}{\sqrt{2}} \frac{0.436}{1-\nu}\, n(x,t), \qquad \varepsilon = 10^{-3}, \qquad t_F = 0.25. \tag{10.113}$$

The factor 0.436 is the benchmark value used to match the prescribed viscosity coefficient. These choices specify the numerical benchmark; the relation between ν and the Chapman–Enskog Prandtl number is the one derived in Proposition 10.8.



(a) Number density. **(b)** Longitudinal bulk velocity.

(c) Temperature. **(d)** Longitudinal heat flux.

Figure 10.1: Moment profiles for the $1x2v$ ES–BGK slab problem at $t_F = 0.25$ with $\varepsilon = 10^{-3}$.

Figure 10.1 shows the number density, longitudinal bulk velocity, temperature, and longitudinal heat flux at $t = t_F$. The profiles remain sharply resolved in the stiff regime $\varepsilon = 10^{-3}$, while the heat flux is localized near the two transition layers. This experiment is a numerical demonstration, not a convergence study; quantitative accuracy would require mesh refine-

ment and comparison with a reference solution.

10.6 Multispecies ES-BGK Models

10.6.1 Pair Gaussians and their moments

The conservative multispecies BGK model in Chapter 8 uses a pair Maxwellian $M_{i,j}$ in the collision of species i with species j. Its number density is n_i, while its pair velocity $\boldsymbol{u}_{i,j}$ and species-frame pair temperature $\hat{\theta}_{i,j}$ are chosen to exchange momentum and energy conservatively. A multispecies ES–BGK model keeps those three scalar and vector moments but replaces $M_{i,j}$ by the anisotropic pair Gaussian

$$\mathcal{G}_{i,j}(\boldsymbol{v}) = \frac{n_i}{\sqrt{\det(2\pi \mathsf{T}_{i,j})}} \exp\left[-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{u}_{i,j})\cdot \mathsf{T}_{i,j}^{-1}(\boldsymbol{v}-\boldsymbol{u}_{i,j})\right], \tag{10.114}$$

where $\mathsf{T}_{i,j}$ is a symmetric positive-definite covariance tensor. Because $\mathsf{T}_{i,j}$ is a velocity covariance, whereas $\hat{\theta}_{i,j}$ is a temperature in energy units, conservation of the pair energy requires

$$\operatorname{tr} \mathsf{T}_{i,j} = d\frac{\hat{\theta}_{i,j}}{m_i}. \tag{10.115}$$

The kinetic equations are

$$\partial_t f_i + \boldsymbol{v}\cdot\nabla_x f_i = \frac{1}{\varepsilon}\sum_{j=1}^{N} Q_{i,j}^{\mathrm{ES}}, \qquad Q_{i,j}^{\mathrm{ES}} := \lambda_{i,j}(\mathcal{G}_{i,j} - f_i). \tag{10.116}$$

The conservative HHM pair velocity and temperature from Chapter 8 are the default choice in this report. The Year Four Report also used a broader ordered-pair parameterization, which is useful when momentum and temperature relaxation rates are fitted separately. We record it here so that the two model classes can be compared.

Definition 10.10 (Ordered-pair scalar parameters)**.** *The **ordered-pair scalar parameters** are specified as follows. For $i \neq j$, choose*

$$\boldsymbol{u}_{i,j} = \alpha_{i,j}\boldsymbol{u}_i + (1-\alpha_{i,j})\boldsymbol{u}_j, \qquad 0 \le \alpha_{i,j} \le 1, \tag{10.117}$$

and

$$\hat{\theta}_{i,j} = \beta_{i,j}\hat{\theta}_i + (1-\beta_{i,j})\hat{\theta}_j + \frac{\gamma_{i,j}}{d}|\boldsymbol{u}_i - \boldsymbol{u}_j|^2, \qquad 0 \le \beta_{i,j} \le 1, \quad \gamma_{i,j} \ge 0. \tag{10.118}$$

For self-collisions, set $\boldsymbol{u}_{i,i} = \boldsymbol{u}_i$ *and* $\hat{\theta}_{i,i} = \hat{\theta}_i$.

Proposition 10.11 (Reverse-pair parameters and admissibility)**.** *For positive collision frequencies, the ordered-pair targets in Definition 10.10 exchange equal and opposite momentum and energy if the reverse parameters satisfy*

$$\alpha_{j,i} = 1 - \frac{\rho_i \lambda_{i,j}}{\rho_j \lambda_{j,i}}(1 - \alpha_{i,j}), \tag{10.119a}$$

$$\beta_{j,i} = 1 - \frac{n_i \lambda_{i,j}}{n_j \lambda_{j,i}}(1 - \beta_{i,j}), \tag{10.119b}$$

$$\gamma_{j,i} = \frac{\rho_i \lambda_{i,j}(1 - \alpha_{i,j})(\alpha_{i,j} + \alpha_{j,i}) - n_i \lambda_{i,j} \gamma_{i,j}}{n_j \lambda_{j,i}}. \tag{10.119c}$$

A sufficient set of scalar admissibility conditions is

$$\max\left\{0, 1 - \frac{\rho_j \lambda_{j,i}}{\rho_i \lambda_{i,j}}\right\} \le \alpha_{i,j} \le 1, \tag{10.120a}$$

$$\max\left\{0, 1 - \frac{n_j \lambda_{j,i}}{n_i \lambda_{i,j}}\right\} \le \beta_{i,j} \le 1, \tag{10.120b}$$

$$0 \le \gamma_{i,j} \le m_i(1 - \alpha_{i,j})(\alpha_{i,j} + \alpha_{j,i}). \tag{10.120c}$$

Under these bounds, both ordered pair temperatures are positive whenever $\hat{\theta}_i$ *and* $\hat{\theta}_j$ *are positive.*

Proof. The momentum source produced by the i–j collision is $\lambda_{i,j}\rho_i(\boldsymbol{u}_{i,j} - \boldsymbol{u}_i)$. Adding the reverse source and using (10.117) gives

$$\rho_i \lambda_{i,j}(1 - \alpha_{i,j}) = \rho_j \lambda_{j,i}(1 - \alpha_{j,i}), \tag{10.121}$$

which is equivalent to (10.119a). The coefficients of $\hat{\theta}_i - \hat{\theta}_j$ in the pair energy balance give (10.119b); the remaining coefficient of $|\boldsymbol{u}_i - \boldsymbol{u}_j|^2$ gives (10.119c).

The lower bounds in (10.120a)–(10.120b) are precisely the conditions $\alpha_{j,i} \ge 0$ and $\beta_{j,i} \ge 0$; their upper bounds imply the corresponding reverse parameters do not exceed one. Finally, (10.120c) makes the numerator in (10.119c) nonnegative. Thus $\gamma_{j,i} \ge 0$, and both pair temperatures are nonnegative convex combinations of positive species temperatures plus a nonnegative drift term. ■

The HHM construction in Proposition 8.4 instead chooses common pair values $\boldsymbol{u}_{i,j} = \boldsymbol{u}_{j,i}$ and $\hat{\theta}_{i,j} = \hat{\theta}_{j,i}$ directly from the conservation laws. It is more restrictive but removes the

free ordered-pair parameters. The tensor constructions below apply to either scalar model, provided its pair momentum and energy sources cancel.

Lemma 10.12 (Moments of a Pair Gaussian)**.** *If* $\mathsf{T}_{i,j}$ *is positive definite, then*

$$\int_{\mathbb{R}^d} \mathcal{G}_{i,j}\,\mathrm{d}\boldsymbol{v} = n_i, \qquad m_i \int_{\mathbb{R}^d} \boldsymbol{v}\mathcal{G}_{i,j}\,\mathrm{d}\boldsymbol{v} = \rho_i \boldsymbol{u}_{i,j}, \tag{10.122}$$

$$m_i \int_{\mathbb{R}^d} \boldsymbol{v}\otimes\boldsymbol{v}\mathcal{G}_{i,j}\,\mathrm{d}\boldsymbol{v} = \rho_i \boldsymbol{u}_{i,j}\otimes\boldsymbol{u}_{i,j} + \rho_i \mathsf{T}_{i,j}, \tag{10.123}$$

$$\frac{m_i}{2}\int_{\mathbb{R}^d} |\boldsymbol{v}|^2 \mathcal{G}_{i,j}\,\mathrm{d}\boldsymbol{v} = \frac{\rho_i}{2}|\boldsymbol{u}_{i,j}|^2 + \frac{m_i n_i}{2}\mathrm{tr}\,\mathsf{T}_{i,j}. \tag{10.124}$$

Proof. Set $\boldsymbol{z} = \mathsf{T}_{i,j}^{-1/2}(\boldsymbol{v}-\boldsymbol{u}_{i,j})$. The affine change of variables used in Lemma 10.2 reduces all four integrals to moments of the standard centered Gaussian. Its mean is zero and its covariance is I, which gives (10.122)–(10.124). ■

Proposition 10.13 (Conservation criterion for pair Gaussians)**.** *Suppose the scalar pair velocities and temperatures have pairwise canceling momentum and energy sources. This holds, in particular, for the HHM values in* (8.23)–(8.24) *and for the ordered-pair family in Proposition 10.11. If* (10.115) *also holds, then every species mass is conserved, and the i–j and j–i collision terms exchange equal and opposite amounts of total momentum and total energy.*

Proof. By Lemma 10.12, the zeroth moment of $Q_{i,j}^{\mathrm{ES}}$ vanishes. Its momentum and energy moments are identical to those of the pair Maxwellian $M_{i,j}$ because (10.115) turns (10.124) into

$$\frac{\rho_i}{2}|\boldsymbol{u}_{i,j}|^2 + \frac{d}{2}n_i\hat{\theta}_{i,j}. \tag{10.125}$$

The assumed scalar pair balance therefore makes the corresponding *i–j* and *j–i* momentum and energy sources cancel. For the HHM choice this is Proposition 8.7; for the ordered-pair choice it is Proposition 10.11. ■

10.6.2 The homogeneous moment system

The tensor equations are easiest to see during a homogeneous collision step. In this subsection n_i is constant, but $\boldsymbol{u}_i$, $\hat{\theta}_i$, and the species-frame pressure tensor

$$\widehat{\mathsf{P}}_i = m_i \int_{\mathbb{R}^d} (\boldsymbol{v}-\boldsymbol{u}_i)\otimes(\boldsymbol{v}-\boldsymbol{u}_i) f_i\,\mathrm{d}\boldsymbol{v} \tag{10.126}$$

may evolve. We also define its deviatoric part by

$$\widehat{\Pi}_i := \widehat{\mathsf{P}}_i - n_i\hat{\theta}_i\mathsf{I}, \qquad \operatorname{tr}\widehat{\Pi}_i = 0. \tag{10.127}$$

The hat is essential: both tensors are measured relative to the species velocity $\boldsymbol{u}_i$, in agreement with Appendix G.

Proposition 10.14 (Homogeneous multispecies moment equations)**.** *A sufficiently integrable solution of the homogeneous form of* (10.116) *satisfies*

$$\partial_t n_i = 0, \tag{10.128a}$$

$$\partial_t(\rho_i\boldsymbol{u}_i) = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\rho_i(\boldsymbol{u}_{i,j} - \boldsymbol{u}_i), \tag{10.128b}$$

$$\partial_t\omega_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\left[\frac{\rho_i}{2}(|\boldsymbol{u}_{i,j}|^2 - |\boldsymbol{u}_i|^2) + \frac{d}{2}n_i(\hat{\theta}_{i,j} - \hat{\theta}_i)\right], \tag{10.128c}$$

$$\partial_t\widehat{\mathsf{P}}_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\left[\rho_i\mathsf{T}_{i,j} + \rho_i(\boldsymbol{u}_{i,j} - \boldsymbol{u}_i)\otimes(\boldsymbol{u}_{i,j} - \boldsymbol{u}_i) - \widehat{\mathsf{P}}_i\right]. \tag{10.128d}$$

Here $\omega_i = \rho_i|\boldsymbol{u}_i|^2/2 + dn_i\hat{\theta}_i/2$ is the species total kinetic-energy density.

Proof. Multiply the homogeneous kinetic equation successively by 1, $m_i\boldsymbol{v}$, $m_i|\boldsymbol{v}|^2/2$, and $m_i(\boldsymbol{v}-\boldsymbol{u}_i)\otimes(\boldsymbol{v}-\boldsymbol{u}_i)$, and integrate in velocity. The first three equations follow immediately from Lemma 10.12. When differentiating the centered second moment, the derivative of $\boldsymbol{u}_i$ creates cross terms proportional to $\int(\boldsymbol{v}-\boldsymbol{u}_i)f_i\,\mathrm{d}\boldsymbol{v}$, which vanish. Finally,

$$\begin{aligned} m_i\int_{\mathbb{R}^d}(\boldsymbol{v}-\boldsymbol{u}_i)\otimes(\boldsymbol{v}-\boldsymbol{u}_i)\mathcal{G}_{i,j}\,\mathrm{d}\boldsymbol{v} \\ = \rho_i\mathsf{T}_{i,j} + \rho_i(\boldsymbol{u}_{i,j}-\boldsymbol{u}_i)\otimes(\boldsymbol{u}_{i,j}-\boldsymbol{u}_i), \end{aligned} \tag{10.129}$$

which proves (10.128d). ∎

Taking one half of the trace of (10.128d) reproduces the internal-energy equation obtained from (10.128b)–(10.128c). Thus the trace of the pressure system is not an additional independent evolution law; it is a consistency condition on the tensor model.

10.6.3 Restricted tensor model

The simplest anisotropic closure uses only the pressure of the species being updated:

$$\mathsf{T}_{i,j} = \frac{\hat{\theta}_{i,j}}{m_i}\mathsf{I} + \frac{\nu_{i,j}}{\rho_i}\widehat{\Pi}_i. \tag{10.130}$$

This is the covariance form of the restricted tensor used in the Year Four Report and by Klingenberg, Pirner, and Puppo [46]. In the ordered-pair notation of Definition 10.10, their specialization is obtained by identifying $\nu_{i,j} = \beta_{i,j}$:

$$\frac{1}{m_i}\left[\hat{\theta}_{i,j}\mathsf{I} + \beta_{i,j}\left(\frac{\widehat{\mathsf{P}}_i}{n_i} - \hat{\theta}_i\mathsf{I}\right)\right] = \frac{\hat{\theta}_{i,j}}{m_i}\mathsf{I} + \frac{\beta_{i,j}}{\rho_i}\widehat{\Pi}_i. \tag{10.131}$$

We retain the symbol $\nu_{i,j}$ in the restricted tensor ansatz so that scalar interspecies energy exchange and anisotropic stress relaxation remain visibly distinct modeling choices. The physical motivation is the "persistence of velocity" discussed by Jeans: collisions need not erase directional information at the same rate at which they relax scalar temperature differences [41, 42]. The anisotropic correction is trace-free, so the trace constraint is automatic. Substitution into Proposition 10.14 gives the closed equation

$$\partial_t\widehat{\mathsf{P}}_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\left[n_i\hat{\theta}_{i,j}\mathsf{I} + \nu_{i,j}\widehat{\Pi}_i + \rho_i\delta\boldsymbol{u}_{i,j}\otimes\delta\boldsymbol{u}_{i,j} - \widehat{\mathsf{P}}_i\right], \tag{10.132}$$

$$\delta\boldsymbol{u}_{i,j} \coloneqq \boldsymbol{u}_{i,j} - \boldsymbol{u}_i.$$

Proposition 10.15 (Backward-Euler update for the restricted tensor)**.** *Assume that the backward-Euler updates of $\boldsymbol{u}_i$, ω_i, and hence $\hat{\theta}_i$ have already been computed. If the coefficients are evaluated at the new time, then the pressure update is determined explicitly by*

$$\widehat{\mathsf{P}}_i^{n+1} = \frac{1}{a_i}\left[\widehat{\mathsf{P}}_i^{n} + \frac{\Delta t}{\varepsilon}\sum_j \lambda_{i,j}^{n+1}\left(n_i\hat{\theta}_{i,j}^{n+1}\mathsf{I} - \nu_{i,j}^{n+1}n_i\hat{\theta}_i^{n+1}\mathsf{I} + \rho_i\delta\boldsymbol{u}_{i,j}^{n+1}\otimes\delta\boldsymbol{u}_{i,j}^{n+1}\right)\right], \tag{10.133}$$

$$a_i = 1 + \frac{\Delta t}{\varepsilon}\sum_j \lambda_{i,j}^{n+1}(1-\nu_{i,j}^{n+1}). \tag{10.134}$$

Provided $a_i > 0$, this is the unique solution of the backward-Euler pressure equation.

Proof. In (10.132), replace $\widehat{\Pi}_i^{n+1}$ by $\widehat{\mathsf{P}}_i^{n+1} - n_i\hat{\theta}_i^{n+1}\mathsf{I}$. Move every multiple of $\widehat{\mathsf{P}}_i^{n+1}$ to the left. Its scalar coefficient is exactly a_i; division by this coefficient gives (10.133). ∎

This algebra proves solvability of the tensor update, but not positivity of the resulting covariance. Admissible bounds on $\nu_{i,j}$ must be imposed for the particular model. Likewise, trace preservation alone does not prove an H-theorem.

10.6.4 General tensor model

To exchange anisotropy between species, one may use the trace-preserving ansatz

$$\mathsf{T}_{i,j} = \frac{\hat{\theta}_{i,j}}{m_i}\mathsf{I} + \mu_{i,j}\left[\beta_{i,j}\frac{\widehat{\Pi}_i}{\rho_i} + (1-\beta_{i,j})\frac{\widehat{\Pi}_j}{\rho_j}\right]. \tag{10.135}$$

Both corrections are trace-free, so (10.115) still holds. Notice the plus sign between the two species contributions: each pressure anisotropy contributes to the covariance according to its weight. Formula (10.135) reduces to an isotropic pair Maxwellian when $\mu_{i,j} = 0$. Writing the ansatz directly in covariance variables makes the mass scaling explicit: both $\widehat{\Pi}_i/\rho_i$ and $\widehat{\Pi}_j/\rho_j$ have units of velocity squared. The plus sign is essential because the two weighted pressure anisotropies contribute to, rather than subtract from, the target covariance.

Proposition 10.16 (Structure of the general pressure system)**.** *For the ansatz* (10.135)*, the homogeneous pressure equations are*

$$\partial_t \widehat{\mathsf{P}}_i = \frac{1}{\varepsilon}\sum_j \lambda_{i,j}\Big[n_i\hat{\theta}_{i,j}\mathsf{I} + \mu_{i,j}\beta_{i,j}\widehat{\Pi}_i + \mu_{i,j}(1-\beta_{i,j})\frac{\rho_i}{\rho_j}\widehat{\Pi}_j + \rho_i\delta\boldsymbol{u}_{i,j}\otimes\delta\boldsymbol{u}_{i,j} - \widehat{\mathsf{P}}_i\Big]. \tag{10.136}$$

Once the velocities and scalar temperatures are known, backward Euler applied to (10.136) *is a finite-dimensional linear system for the symmetric tensors* $\widehat{\mathsf{P}}_1^{n+1},\dots,\widehat{\mathsf{P}}_N^{n+1}$.

Proof. Insert (10.135) into (10.128d) and use $\rho_i = m_i n_i$. This gives (10.136). Every occurrence of an unknown pressure tensor is linear because $\widehat{\Pi}_i = \widehat{\mathsf{P}}_i - n_i\hat{\theta}_i\mathsf{I}$, and the scalar temperatures have already been determined by the energy update. A symmetric $d\times d$ tensor has $d(d+1)/2$ independent entries; vectorizing those entries converts the backward-Euler equations into an ordinary block linear system of size $Nd(d+1)/2$. ∎

The coupled part becomes especially transparent after taking the deviatoric part. For any square tensor A, write

$$\operatorname{dev}\mathsf{A} \coloneqq \mathsf{A} - \frac{1}{d}(\operatorname{tr}\mathsf{A})\mathsf{I}. \tag{10.137}$$

Proposition 10.17 (Linear system for the general backward-Euler step)**.** *Let $h = \Delta t/\varepsilon$, and suppose the new velocities and scalar temperatures have already been determined. Backward Euler applied to the deviatoric part of* (10.136) *gives*

$$a_i \widehat{\Pi}_i^{n+1} - \sum_{j \neq i} b_{i,j} \widehat{\Pi}_j^{n+1} = \mathsf{R}_i, \tag{10.138}$$

$$a_i \coloneqq 1 + h\lambda_{i,i}^{n+1}(1 - \mu_{i,i}^{n+1}) + h \sum_{j \neq i} \lambda_{i,j}^{n+1}(1 - \mu_{i,j}^{n+1}\beta_{i,j}^{n+1}), \tag{10.139}$$

$$b_{i,j} \coloneqq h\lambda_{i,j}^{n+1}\mu_{i,j}^{n+1}(1 - \beta_{i,j}^{n+1})\frac{\rho_i}{\rho_j}, \tag{10.140}$$

$$\mathsf{R}_i \coloneqq \widehat{\Pi}_i^{n} + h \sum_{j} \lambda_{i,j}^{n+1}\rho_i \operatorname{dev}\left(\delta \boldsymbol{u}_{i,j}^{n+1} \otimes \delta \boldsymbol{u}_{i,j}^{n+1}\right). \tag{10.141}$$

Consequently, each independent component of the deviatoric tensors satisfies the same $N \times N$ scalar linear system. If

$$a_i > \sum_{j \neq i} |b_{i,j}| \qquad \text{for every } i, \tag{10.142}$$

and $b_{i,j} \geq 0$ for $j \neq i$, then the system has a unique solution, and the standard block Gauss–Seidel iteration converges. If the model contains no self-collision term, set $\lambda_{i,i} = 0$.

Proof. Apply dev to (10.136). All multiples of I vanish, $\operatorname{dev} \widehat{\mathsf{P}}_i = \widehat{\Pi}_i$, and the remaining equation is

$$\partial_t \widehat{\Pi}_i = \frac{1}{\varepsilon} \sum_j \lambda_{i,j} \Bigg[-(1 - \mu_{i,j}\beta_{i,j})\widehat{\Pi}_i + \mu_{i,j}(1 - \beta_{i,j})\frac{\rho_i}{\rho_j}\widehat{\Pi}_j + \rho_i \operatorname{dev}(\delta \boldsymbol{u}_{i,j} \otimes \delta \boldsymbol{u}_{i,j}) \Bigg]. \tag{10.143}$$

Discretize this equation by backward Euler and collect the new-time tensor terms. When $j = i$, the two tensor terms combine because $\rho_i/\rho_i = 1$:

$$-(1 - \mu_{i,i}\beta_{i,i})\widehat{\Pi}_i + \mu_{i,i}(1 - \beta_{i,i})\widehat{\Pi}_i = -(1 - \mu_{i,i})\widehat{\Pi}_i.$$

The terms with $j \neq i$ remain coupled. This gives (10.138)–(10.141). Vectorizing the independent tensor entries shows that the coefficient matrix is the same for each entry. Strict row diagonal dominance implies nonsingularity by the Levy–Desplanques theorem. Moreover, when $b_{i,j} \geq 0$, the matrix has positive diagonal entries and nonpositive off-diagonal entries. Strict diagonal dominance then makes it a nonsingular M-matrix, for which the Gauss–Seidel iteration converges. ∎

The restricted model decouples the tensor solve by species. The general model couples it through the terms containing $\widehat{\Pi}_j$. A block Gauss–Seidel iteration can therefore be implemented by solving (10.138) for one species tensor at a time, using the newest available tensors for the other species. Condition (10.142) is sufficient, though not necessary, for convergence. The construction and admissible parameter restrictions are discussed in [46].

For implementation, let $q = d(d+1)/2$ and let svec store the independent entries of a symmetric tensor. If C is the $N \times N$ matrix with diagonal entries a_i and off-diagonal entries $-b_{i,j}$, then (10.138) can be written as

$$(\mathsf{C} \otimes \mathsf{I}_q) \begin{bmatrix} \operatorname{svec} \widehat{\Pi}_1^{n+1} \\ \vdots \\ \operatorname{svec} \widehat{\Pi}_N^{n+1} \end{bmatrix} = \begin{bmatrix} \operatorname{svec} \mathsf{R}_1 \\ \vdots \\ \operatorname{svec} \mathsf{R}_N \end{bmatrix}. \tag{10.144}$$

Thus one factorization of C serves every tensor component. It is usually preferable, however, to exploit the much smaller componentwise systems rather than to assemble the Kronecker product.

10.6.5 Nonlinear moment-first GST solve

The preceding backward-Euler formulas become nonlinear when the collision frequencies depend on the new moments. As in the multispecies BGK solver of Chapter 8, this nonlinearity can be confined to moment space: freeze the rates, update the moments in stages, and then refresh the rates. No nonlinear solve on the velocity grid is required.

We first display the two scalar stages for the ordered-pair family in Definition 10.10. Set $h = \Delta t/\varepsilon$, and introduce

$$\mathsf{M}_\rho = \operatorname{diag}(\rho_1, \dots, \rho_N), \qquad \mathsf{M}_n = \operatorname{diag}(n_1, \dots, n_N). \tag{10.145}$$

At nonlinear iterate ℓ, evaluate the rates $\lambda_{i,j}^{\ell}$ from the current moment iterate and define

$$A_{i,j}^{\ell} = \rho_i \lambda_{i,j}^{\ell} (1 - \alpha_{i,j}), \qquad D_{i,j}^{\ell} = \delta_{i,j} \sum_k A_{i,k}^{\ell}, \tag{10.146}$$

$$B_{i,j}^{\ell} = n_i \lambda_{i,j}^{\ell} (1 - \beta_{i,j}), \qquad F_{i,j}^{\ell} = \delta_{i,j} \sum_k B_{i,k}^{\ell}. \tag{10.147}$$

Let $\boldsymbol{U} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_N)^\top$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_N)^\top$. The frozen-rate velocity stage is the $N \times N$ system

$$\left[\mathsf{M}_\rho + h(\mathsf{D}^{\ell} - \mathsf{A}^{\ell})\right] \boldsymbol{U}^{\ell+1} = \mathsf{M}_\rho \boldsymbol{U}^{n}. \tag{10.148}$$

The same coefficient matrix is used for every velocity component.

For a velocity array $\boldsymbol{U}$, define

$$[\boldsymbol{s}(\boldsymbol{U})]_i \coloneqq |\boldsymbol{u}_i|^2, \tag{10.149}$$

$$[\boldsymbol{r}^\ell(\boldsymbol{U})]_i \coloneqq \sum_j \lambda_{i,j}^\ell \left[\rho_i\big(|\boldsymbol{u}_{i,j}|^2 - |\boldsymbol{u}_i|^2\big) + n_i\gamma_{i,j}|\boldsymbol{u}_i - \boldsymbol{u}_j|^2\right], \tag{10.150}$$

where each $\boldsymbol{u}_{i,j}$ is formed from the velocities in its argument. The temperature stage is then

$$\left[\mathsf{M}_n + h(\mathsf{F}^\ell - \mathsf{B}^\ell)\right]\hat{\boldsymbol{\theta}}^{\ell+1} = \mathsf{M}_n\hat{\boldsymbol{\theta}}^n - \frac{1}{d}\mathsf{M}_\rho\left[\boldsymbol{s}(\boldsymbol{U}^{\ell+1}) - \boldsymbol{s}(\boldsymbol{U}^n)\right] + \frac{h}{d}\boldsymbol{r}^\ell(\boldsymbol{U}^{\ell+1}). \tag{10.151}$$

Equations (10.148) and (10.151) are simply the backward-Euler momentum and total-energy equations with the rates frozen. For the HHM pair targets, the corresponding two stages are the conservative velocity and energy solves developed in Chapter 8; the tensor stage below is unchanged.

Definition 10.18 (Nonlinear moment-first GST iteration)**.** *The **nonlinear moment-first GST iteration** is the following procedure. Initialize $\boldsymbol{U}^0 = \boldsymbol{U}^n$, $\hat{\boldsymbol{\theta}}^0 = \hat{\boldsymbol{\theta}}^n$, and $\widehat{\mathsf{P}}_i^0 = \widehat{\mathsf{P}}_i^n$. Given iterate ℓ, evaluate every moment-dependent collision frequency $\lambda_{i,j}^\ell$. With those rates fixed, solve* (10.148) *and then* (10.151)*; for HHM targets, use the analogous Chapter 8 scalar stages. Next update the pressure tensors with Proposition 10.15 for the restricted ansatz, or solve* (10.138) *for the general ansatz, always using the newest velocities and temperatures and the frozen rates $\lambda_{i,j}^\ell$. Finally recompute the rates from $(\boldsymbol{U}^{\ell+1}, \hat{\boldsymbol{\theta}}^{\ell+1}, \{\widehat{\mathsf{P}}_i^{\ell+1}\})$ and repeat until both the relative moment change and the backward-Euler residual are below the prescribed tolerance.*

If the collision frequencies depend only on density and scalar temperature, they may be refreshed immediately after the temperature stage. If they also depend on anisotropy, the pressure stage must precede the refresh. Under-relaxation can be used to preserve positive scalar temperatures and positive-definite covariances during early iterates. Anderson acceleration may also be applied to the complete moment map, as in the scalar multispecies solver [34].

Proposition 10.19 (Consistency and conservation of the GST fixed point)**.** *Suppose the iteration in Definition 10.18 converges, its linear systems are solved exactly, and the collision-rate map is continuous at the limit. Then the limiting moments satisfy the fully implicit backward-Euler discretization of* (10.128)*. If the scalar pair targets satisfy the pairwise*

balance in Proposition 10.13, the limiting update preserves every species mass and the total momentum and total energy.

Proof. At a fixed point, the rates used in the three stages equal the rates evaluated from their output moments. Consequently, the frozen-rate velocity, energy, and pressure equations become the corresponding fully implicit backward-Euler equations. Summing the species momentum and energy equations cancels each ordered pair with its reverse by Proposition 10.13; the zeroth-moment equations leave each n_i unchanged. ■

After convergence, reconstruct the pair Gaussians from the limiting moments and apply the distribution update below. If a maximum iteration count is reached first, the step should be rejected or repeated with a smaller time step; silently accepting an unconverged moment iterate forfeits the stated implicit accuracy and conservation tolerance.

10.6.6 Distribution update and admissibility

After the moment solve, form every $\mathcal{G}_{i,j}^{n+1}$ and set

$$f_i^{n+1} = \frac{\varepsilon f_i^n + \Delta t \sum_j \lambda_{i,j}^{n+1} \mathcal{G}_{i,j}^{n+1}}{\varepsilon + \Delta t \sum_j \lambda_{i,j}^{n+1}}. \tag{10.152}$$

This is a convex combination of f_i^n and the pair Gaussians whenever the collision frequencies are nonnegative and all covariance tensors are positive definite. It then preserves nonnegativity. Conservation follows from Proposition 10.13 after summing the species updates.

Three logically different requirements must not be conflated:

1. the trace constraint gives the correct pair energy;
2. positive definiteness makes each pair Gaussian well defined and preserves nonnegativity in (10.152);
3. an H-theorem requires an additional entropy comparison for the complete collection of pair targets.

The first property does not imply either of the other two. For this reason, the restricted and general formulas above describe model families; a particular implementation must state and verify its admissible parameter range.

10.7 Conclusions

The single-species ES–BGK model retains the conservation and information dissipation structure of BGK while correcting its Prandtl number. The Chapman–Enskog calculation differs from the BGK calculation in Chapter 3 only in the relaxation of the trace-free quadratic mode: deviatoric pressure relaxes at rate $\lambda(1-\nu)$, whereas heat flux relaxes at rate λ. This gives $\mathrm{Pr} = 1/(1-\nu)$.

For mixtures, replacing pair Maxwellians by pair Gaussians preserves species mass and the total momentum and energy whenever the pair velocities, temperatures, and covariance traces agree with Chapter 8. The resulting pressure equations close once a restricted or general tensor ansatz has been selected. The restricted ansatz produces separate linear tensor solves; the general ansatz produces a coupled block system. For moment-dependent collision frequencies, the nonlinear GST iteration combines either tensor update with the ordered velocity and temperature solves while retaining the fully implicit conservation laws at convergence. Positive definiteness and an H-theorem require additional parameter restrictions and cannot be inferred from trace preservation or conservation alone. Keeping these three requirements distinct is essential both in the analysis and in an implicit implementation.

Chapter 11

Dynamical Low-Rank Approximation Methods

This chapter contains our preliminary investigation of dynamical low-rank approximation (DLRA) and integrates its matrix-manifold foundations, projector-splitting algorithms, and finite-volume applications with the kinetic models developed earlier in this report. The fixed-rank matrix formulation and Dirac–Frenkel projection follow the foundational treatment of Koch and Lubich [47], while the robust projector-splitting algorithm is due to Lubich and Oseledets [53]. The continuous phase-space formulation and the kinetic applications build on the Vlasov and Boltzmann–BGK developments in [24, 23, 40, 19].

11.1 Foundations of Dynamical Low-Rank Approximation

11.1.1 Matrix Differential Equations and Low-Rank Factorizations

We begin by fixing the matrix norms used below. Throughout this chapter, $\|\cdot\|_2$ denotes the matrix 2-norm, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, and $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^n$. For a matrix $A \in \mathbb{R}^{m\times n}$, these norms are defined by

$$\|A\|_{\mathrm{F}} = \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\Big)^{\frac{1}{2}}$$
$$\|A\|_2 = \sup_{\substack{x\in\mathbb{R}^n \\ \|x\|=1}} \|Ax\|. \tag{11.1}$$

The Frobenius norm is generated by the inner product

$$\langle A, B\rangle_{\mathrm{F}} = \operatorname{tr}(A^{\top}B) = \operatorname{tr}(AB^{\top}).$$

The goal of this section is to investigate dynamical low-rank approximations (DLRA) to matrix differential equations. When $A \in \mathbb{R}^{m\times n}$ and both dimensions are large, storing A requires mn numbers. We instead approximate it by a matrix of prescribed rank $r \leq \min\{m, n\}$ written as

$$Y = USV^{\top}, \qquad U^{\top}U = V^{\top}V = I_r, \qquad S \in \mathrm{GL}_r(\mathbb{R}).$$

This representation requires only $mr + r^2 + nr$ numbers. The core S need not be diagonal; its invertibility is precisely what makes the rank of Y equal to r.

We consider the nonautonomous matrix initial-value problem

$$\begin{aligned} \dot{A}(t) &= \mathcal{R}(A(t), t), \\ A(t_0) &= A_0. \end{aligned} \tag{11.2}$$

Here $\mathcal{R} : \mathbb{R}^{m\times n} \times [t_0, T] \to \mathbb{R}^{m\times n}$ is assumed continuous in t and locally Lipschitz in its matrix argument, uniformly on compact time intervals. These standard hypotheses ensure local existence and uniqueness. The autonomous case is recovered when $\mathcal{R}(A, t)$ is independent of t.

The DLRA construction replaces $A(t)$ by a curve $Y(t)$ in the fixed-rank set

$$\mathcal{M}_r = \left\{ Y \in \mathbb{R}^{m\times n} \;\middle|\; \operatorname{rank} Y = r \right\}.$$

Appendix E gives the abstract definition of a topological manifold and collects related results. In this chapter, every matrix manifold is a subset of a finite-dimensional Euclidean space, so the following concrete definition is sufficient.

Definition 11.1 (Smooth embedded manifold)**.** *Let $\mathcal{M} \subseteq \mathbb{R}^N$ and let $0 \leq d \leq N$. The set $\mathcal{M}$ is a d-**dimensional smooth embedded manifold** if, for every $p \in \mathcal{M}$, there are an open set $\Omega \subseteq \mathbb{R}^d$, an open neighborhood $\mathcal{O} \subseteq \mathbb{R}^N$ of p, and a smooth map $\phi : \Omega \to \mathbb{R}^N$ satisfying the following conditions.*

(i) $\phi(\Omega) = \mathcal{M} \cap \mathcal{O}$.

(ii) *The map $\phi : \Omega \to \mathcal{M} \cap \mathcal{O}$ is a homeomorphism, where $\mathcal{M} \cap \mathcal{O}$ has the topology inherited from $\mathbb{R}^N$.*

(iii) *The derivative $D\phi[u] : \mathbb{R}^d \to \mathbb{R}^N$ has rank d for every $u \in \Omega$.*

Here smooth means that derivatives of all orders exist and are continuous. The map ϕ is called a ***local parameterization****, and its inverse is called a* ***coordinate chart****.*

Thus, near each of its points, a d-dimensional smooth embedded manifold looks like a smoothly curved copy of an open subset of $\mathbb{R}^d$ with d independent coordinate directions. We identify $\mathbb{R}^{m\times n}$ with $\mathbb{R}^{mn}$ when applying this definition to matrices.

Example 11.2 (Linear subspaces)**.** Let $V \subseteq \mathbb{R}^N$ be a d-dimensional linear subspace with basis $v_1, \dots, v_d$. The map

$$\phi : \mathbb{R}^d \to V, \qquad \phi(a_1, \dots, a_d) = \sum_{j=1}^{d} a_j v_j,$$

is a global parameterization. It is a homeomorphism onto V, and its derivative has the linearly independent columns $v_1, \dots, v_d$. Therefore every finite-dimensional linear subspace is a smooth embedded manifold, with dimension equal to its dimension as a vector space.

The next result justifies treating this set as a smooth manifold.

Proposition 11.3 (Fixed-rank matrix manifold)**.** *Let $1 \le r \le \min\{m, n\}$. Then $\mathcal{M}_r$ is a smooth embedded submanifold of $\mathbb{R}^{m\times n}$ of dimension*

$$r(m + n - r).$$

Proof. First consider matrices whose upper-left $r \times r$ block is invertible. Write

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}, \qquad \begin{aligned} &B \in \mathbb{R}^{r\times r}, && C \in \mathbb{R}^{r\times(n-r)}, \\ &D \in \mathbb{R}^{(m-r)\times r}, && E \in \mathbb{R}^{(m-r)\times(n-r)}, \end{aligned}$$

where blocks with zero rows or columns are allowed when $r = m$ or $r = n$. Define

$$\mathcal{U} = \left\{ \begin{bmatrix} B & C \\ D & E \end{bmatrix} \in \mathbb{R}^{m\times n} \;\middle|\; \det B \neq 0 \right\}.$$

Thus $\mathcal{U}$ consists precisely of the matrices whose upper-left $r \times r$ block is invertible. The operation that extracts this block is linear, and the determinant is continuous, so $\mathcal{U}$ is open in $\mathbb{R}^{m\times n}$. For $A \in \mathcal{U}$, the two block matrices

$$G = \begin{bmatrix} I_r & 0 \\ -DB^{-1} & I_{m-r} \end{bmatrix}, \qquad H = \begin{bmatrix} I_r & -B^{-1}C \\ 0 & I_{n-r} \end{bmatrix}$$

are invertible, and direct multiplication gives

$$GAH = \begin{bmatrix} B & 0 \\ 0 & E - DB^{-1}C \end{bmatrix}.$$

Multiplication by invertible matrices does not change rank. Since B has rank r,

$$\operatorname{rank}(A) = r + \operatorname{rank}(E - DB^{-1}C).$$

Consequently, A has rank r if and only if $E = DB^{-1}C$.

Now define the open set

$$L_0 = \left\{ (B, C, D) \in \mathbb{R}^{r\times r} \times \mathbb{R}^{r\times(n-r)} \times \mathbb{R}^{(m-r)\times r} \;\middle|\; \det B \neq 0 \right\}.$$

The portion of $\mathcal{M}_r$ contained in $\mathcal{U}$ is the graph of the smooth map

$$(B, C, D) \longmapsto DB^{-1}C, \qquad (B, C, D) \in L_0.$$

Equivalently, it is parameterized by

$$\phi(B, C, D) = \begin{bmatrix} B & C \\ D & DB^{-1}C \end{bmatrix}.$$

A graph of a smooth map over an open set is a smooth embedded submanifold. Its dimension is the dimension of L_0, namely

$$r^2 + r(n - r) + (m - r)r = r(m + n - r).$$

Finally, every rank-r matrix has a nonzero $r \times r$ minor. Row and column permutations can move that minor to the upper-left corner. Since multiplication by permutation matrices is a smooth linear diffeomorphism, the corresponding permuted copies of the patch above are smooth embedded submanifolds. These patches cover $\mathcal{M}_r$, which proves the result. ■

A smooth embedded manifold may be viewed, locally, as a smooth surface inside a Euclidean space. Its allowable instantaneous directions are described by its tangent space.

Definition 11.4 (Tangent space)**.** *Let $\mathcal{M} \subseteq \mathbb{R}^{m\times n}$ be a smooth embedded manifold and let*

$Y \in \mathcal{M}$. Its ***tangent space at*** *Y is*

$$T_Y\mathcal{M} = \left\{ \dot{\gamma}(0) \;\middle|\; \gamma : (-\delta, \delta) \to \mathcal{M} \text{ is a } C^1 \text{ curve for some } \delta > 0 \text{ and } \gamma(0) = Y \right\}. \quad (11.3)$$

Thus a tangent vector is the velocity, at $t = 0$, of a differentiable curve that remains on the manifold.

$T_Y\mathcal{M}$ is a vector space. To see why, let $\phi : U \subseteq \mathbb{R}^d \to \mathcal{M}$ be a local parameterization with $\phi(u_0) = Y$. The chain rule shows that the velocity of every curve through Y has the form $D\phi[u_0](c)$ for some $c \in \mathbb{R}^d$. Conversely, the curve $\gamma(t) = \phi(u_0 + tc)$ has velocity $D\phi[u_0](c)$. Therefore,

$$T_Y\mathcal{M} = \operatorname{range} D\phi[u_0].$$

The derivative $D\phi[u_0]$ is a linear map, so its range is a vector space.

For an embedded matrix manifold, $T_Y\mathcal{M}$ is therefore a linear subspace of the ambient vector space $\mathbb{R}^{m\times n}$; in particular, it contains the zero matrix and is closed under addition and scalar multiplication. If $\mathcal{M}$ has dimension d, then $T_Y\mathcal{M}$ also has dimension d. It inherits the Frobenius inner product and norm. Because it is finite-dimensional, it is closed, so every ambient matrix Z has a unique orthogonal projection $P_{T_Y\mathcal{M}}Z$ onto it. The tangent space depends on the base point Y and should not be confused with the affine tangent plane $Y + T_Y\mathcal{M}$; see [50, 1] for further background.

We seek a curve $Y(t) \in \mathcal{M}_r$ that approximates the solution $A(t)$ of the full matrix equation:

$$A(t) \approx Y(t).$$

Choose an initial approximation $Y_0 = U_0 S_0 V_0^\top \in \mathcal{M}_r$ to A_0. When A_0 has rank at least r, a common choice is its rank-r truncated singular value decomposition; if $A_0 \in \mathcal{M}_r$, we may simply take $Y_0 = A_0$.

Because $Y(t)$ is constrained to remain in $\mathcal{M}_r$, its velocity must satisfy

$$\dot{Y}(t) \in T_{Y(t)}\mathcal{M}_r.$$

The full equation suggests the velocity $\mathcal{R}(Y(t), t)$, but this matrix need not be tangent to $\mathcal{M}_r$. We therefore choose, among all admissible tangent vectors, the one closest to $\mathcal{R}(Y(t), t)$ in the Frobenius norm:

$$\dot{Y}(t) = \operatorname*{arg\,min}_{W \in T_{Y(t)}\mathcal{M}_r} \frac{1}{2}\|W - \mathcal{R}(Y(t), t)\|_{\mathrm{F}}^2, \qquad Y(t_0) = Y_0. \quad (11.4)$$

This is the Dirac–Frenkel variational principle in the matrix setting [47].

Proposition 11.5 (Equivalent forms of the DLRA equation)**.** *For fixed $Y \in \mathcal{M}_r$ and $Z \in \mathbb{R}^{m\times n}$, the following statements are equivalent:*

1. *W minimizes $\frac{1}{2}\|W - Z\|_{\mathsf{F}}^2$ over $T_Y\mathcal{M}_r$, that is,*

$$W = \operatorname*{arg\,min}_{X\in T_Y\mathcal{M}_r} \frac{1}{2}\|X - Z\|_{\mathsf{F}}^2;$$

2. *$W \in T_Y\mathcal{M}_r$ satisfies the Galerkin Condition, that is,*

$$\langle W - Z, X\rangle_{\mathsf{F}} = 0 \quad \text{for every } X \in T_Y\mathcal{M}_r; \tag{11.5}$$

3. *W is the projection onto $T_Y\mathcal{M}_r$, that is, $W = P_{T_Y\mathcal{M}_r}Z$.*

Proof. The tangent space $T_Y\mathcal{M}_r$ is finite-dimensional and hence closed. The orthogonal decomposition theorem gives unique matrices $W \in T_Y\mathcal{M}_r$ and $Z_\perp \in (T_Y\mathcal{M}_r)^\perp$ such that $Z = W + Z_\perp$. For any $H \in T_Y\mathcal{M}_r$,

$$\|H - Z\|_{\mathsf{F}}^2 = \|H - W\|_{\mathsf{F}}^2 + \|Z_\perp\|_{\mathsf{F}}^2.$$

Thus W is the unique minimizer. Moreover, $Z - W = Z_\perp$ is orthogonal to the tangent space, which is exactly the Galerkin condition. Conversely, the Galerkin condition identifies W as the orthogonal projection of Z onto $T_Y\mathcal{M}_r$. ∎

Applying the proposition with $Z = \mathcal{R}(Y(t), t)$ gives

$$\dot{Y}(t) = P_{T_{Y(t)}\mathcal{M}_r}\mathcal{R}(Y(t), t), \qquad Y(t_0) = Y_0. \tag{11.6}$$

Proposition 11.6 (Local existence and preservation of rank)**.** *Suppose that $\mathcal{R}$ is continuous in t and locally Lipschitz in its matrix argument, as assumed above. For every $Y_0 \in \mathcal{M}_r$, there is a time $\delta > 0$ for which Equation* (11.6) *has a unique solution satisfying*

$$Y(t) \in \mathcal{M}_r \quad \text{and hence} \quad \operatorname{rank} Y(t) = r \qquad (t_0 \le t < t_0 + \delta).$$

The rank remains r throughout the maximal interval on which the solution exists as a solution on $\mathcal{M}_r$.

Proof. For $Y \in \mathcal{M}_r$, the map $Z \mapsto P_{T_Y\mathcal{M}_r} Z$ is linear, and the projector depends smoothly on the base point Y; an explicit formula is derived later in Equation (11.31). Consequently,

$$G(Y, t) = P_{T_Y\mathcal{M}_r} \mathcal{R}(Y, t)$$

is continuous in t and locally Lipschitz in Y on $\mathcal{M}_r$. Moreover, $G(Y, t) \in T_Y\mathcal{M}_r$ by construction, so G is a tangent vector field on the manifold. In a local parameterization of $\mathcal{M}_r$, the projected equation becomes an ordinary differential equation in a Euclidean space. The Picard–Lindelöf theorem gives a unique local solution in those coordinates. Mapping that solution back through the parameterization gives a curve in $\mathcal{M}_r$, and therefore its rank remains r. ■

Example 11.7 (Finite-time loss of rank)**.** The preceding conclusion is local in time. Existence of the original full-matrix problem does not imply that the projected solution remains in $\mathcal{M}_r$ for the same length of time. Take $m = n = r = 1$, so $\mathcal{M}_1 = \mathbb{R} \setminus \{0\}$, and let $\mathcal{R}(A, t) = -1$. The tangent space is all of $\mathbb{R}$ at every $Y \neq 0$, and hence the projected problem with $Y(0) = 1$ is

$$\dot{Y} = -1, \qquad Y(0) = 1.$$

Its solution is $Y(t) = 1 - t$. It belongs to $\mathcal{M}_1$ for $0 \leq t < 1$, but at $t = 1$ its rank drops to zero. The full equation still has a solution beyond this time, whereas the projected equation on $\mathcal{M}_1$ has reached the boundary of its manifold. For a general rank-r matrix, the analogous warning sign is that its smallest nonzero singular value approaches zero. If the solution remains bounded and that singular value stays bounded away from zero, the usual continuation theorem permits the local solution to be extended.

The remainder of the chapter proceeds in four stages. We first develop the geometry of fixed-rank matrices and derive the projected equations. We then introduce robust splitting integrators, extend the construction to separable functions, and finally apply it to finite-volume discretizations of kinetic equations. Recent applications include kinetic and fluid equations [22, 23, 40, 19] and machine learning [71, 66, 69, 70, 65, 67, 68, 82, 49].

11.1.2 The Fixed-Rank Matrix Manifold

This subsection introduces the matrix manifolds used in the fixed-rank representation: the manifold of fixed-rank matrices, the Stiefel manifold, and the general linear group. We establish their manifold structure and compute the corresponding tangent spaces.

It is important to distinguish exact rank from bounded rank. By Proposition 11.3, the exact-rank set $\mathcal{M}_r$ is a smooth manifold, whereas

$$\mathcal{M}_{\leq r} = \{Y \mid \operatorname{rank} Y \leq r\}$$

is a stratified set whose lower-rank strata are singular relative to the rank-r stratum. A factorization $Y = USV^\top$ with $S \in \mathrm{GL}_r(\mathbb{R})$ represents an element of $\mathcal{M}_r$. If S is singular, the same product has rank strictly below r and belongs only to $\mathcal{M}_{\leq r}$. We begin with a standard linear-algebraic fact about rank-r matrices for $r < \min\{m, n\}$.

Proposition 11.8 (Compact SVD Factorization)**.** *Let $Y \in \mathbb{R}^{m\times n}$ have rank $r \leq \min\{n, m\}$. Then there exist $S \in \mathbb{R}^{r\times r}$, $U \in \mathbb{R}^{m\times r}$, and $V \in \mathbb{R}^{n\times r}$ satisfying*

$$Y = USV^\top \tag{11.7}$$

where S is invertible and both U and V have orthonormal columns, that is, $U^\top U = V^\top V = I_r$.

Proof. Write a singular value decomposition $Y = U'S'V'^\top$, where $U' \in \mathbb{R}^{m\times m}$ and $V' \in \mathbb{R}^{n\times n}$ are orthogonal and $S' \in \mathbb{R}^{m\times n}$ is diagonal, with singular values $\sigma_1 \geq \cdots \geq \sigma_p$ and $p = \min\{m, n\}$. Since Y has rank r, $\sigma_r > 0$ and all remaining singular values vanish. Partition U' and V' as

$$U' = \begin{bmatrix} U'_1 & U'_2 \end{bmatrix},$$

and

$$V' = \begin{bmatrix} V'_1 & V'_2 \end{bmatrix},$$

where $U'_1 \in \mathbb{R}^{m\times r}$, $U'_2 \in \mathbb{R}^{m\times(m-r)}$, $V'_1 \in \mathbb{R}^{n\times r}$, and $V'_2 \in \mathbb{R}^{n\times(n-r)}$. Partition S' conformably:

$$S' = \begin{bmatrix} S'_1 & 0 \\ 0 & 0 \end{bmatrix},$$

where $S'_1 \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing the positive singular values of Y and is thus an invertible matrix. So,

$$\begin{aligned} Y &= U'S'V'^\top \\ &= \begin{bmatrix} U'_1 & U'_2 \end{bmatrix} \begin{bmatrix} S'_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V'^\top_1 \\ V'^\top_2 \end{bmatrix} \\ &= U'_1 S'_1 V'^\top_1. \end{aligned} \tag{11.8}$$

The orthogonality of U' and V' now shows directly that U_1' and V_1' have orthonormal columns:

$$\begin{aligned} I_m &= U'^{\top} U' \\ &= \begin{bmatrix} U_1'^{\top} \\ U_2'^{\top} \end{bmatrix} \begin{bmatrix} U_1' & U_2' \end{bmatrix} \\ &= \begin{bmatrix} U_1'^{\top} U_1' & U_1'^{\top} U_2' \\ U_2'^{\top} U_1' & U_2'^{\top} U_2' \end{bmatrix}, \end{aligned} \tag{11.9}$$

and

$$\begin{aligned} I_n &= V'^{\top} V' \\ &= \begin{bmatrix} V_1'^{\top} \\ V_2'^{\top} \end{bmatrix} \begin{bmatrix} V_1' & V_2' \end{bmatrix} \\ &= \begin{bmatrix} V_1'^{\top} V_1' & V_1'^{\top} V_2' \\ V_2'^{\top} V_1' & V_2'^{\top} V_2' \end{bmatrix}. \end{aligned} \tag{11.10}$$

Take $U = U_1'$, $S = S_1'$, and $V = V_1'$. Then, $Y = USV^{\top}$. ■

The decomposition in the result is often called a **compact SVD** or **reduced SVD**.

Definition 11.9 (Orthogonal Group)**.** *The **orthogonal group of order** r is*

$$\mathrm{O}(r) = \left\{ Q \in \mathbb{R}^{r \times r} \;\middle|\; Q^{\top} Q = I_r \right\}.$$

*It is a group under matrix multiplication. For every $Q \in \mathrm{O}(r)$, $Q^{-1} = Q^{\top}$, and hence also $QQ^{\top} = I_r$. Its elements are called **orthogonal matrices**.*

Remark 11.10. *The decomposition in Proposition 11.8 is not unique. Suppose $Y = USV^{\top}$ satisfies the conditions of Proposition 11.8. Take two arbitrary orthogonal matrices $P, Q \in \mathrm{O}(r)$. Then,*

$$Y = USV^{\top} = UQQ^{\top} SPP^{\top} V^{\top}. \tag{11.11}$$

Set $\widetilde{U} = UQ$, $\widetilde{V} = VP$, and $\widetilde{S} = Q^{\top} SP$. The matrix $\widetilde{S}$ is invertible because it is a product of invertible matrices. The orthonormality of the transformed factors follows from

$$\begin{aligned} \widetilde{U}^{\top} \widetilde{U} &= Q^{\top} U^{\top} UQ \\ &= Q^{\top} Q \end{aligned}$$

$$
\begin{aligned}
&= I_r, \\
\widetilde{V}^\top \widetilde{V} &= P^\top V^\top V P \\
&= P^\top P \\
&= I_r.
\end{aligned} \tag{11.12}
$$

Thus, $\widetilde{U}$, $\widetilde{S}$, and $\widetilde{V}$ satisfy the conditions of Proposition 11.8 and $Y = USV^\top = \widetilde{U}\widetilde{S}\widetilde{V}^\top$, thus yielding infinitely many representations for any such element in $\mathcal{M}_r$.

We next introduce the remaining spaces needed to describe the fixed-rank factors and their tangent geometry: the symmetric matrices, the Stiefel manifold, and the general linear group. We begin with the symmetric matrices.

Definition 11.11 (Symmetric Matrices)**.** *Let $n \in \mathbb{N}$. The **space of symmetric matrices** in $\mathbb{R}^{n\times n}$ is the set of all matrices A in $\mathbb{R}^{n\times n}$ equal to their own transpose, $A^\top = A$. We denote this space by*

$$
\mathrm{Sym}(n) = \left\{ A \in \mathbb{R}^{n\times n} \;\middle|\; A^\top = A \right\}. \tag{11.13}
$$

Theorem 11.12. *Let $n \in \mathbb{N}$. Then, $\mathrm{Sym}(n) \subset \mathbb{R}^{n\times n}$ is a smooth submanifold of $\mathbb{R}^{n\times n}$.*

Proof. This follows from the fact that $\mathrm{Sym}(n)$ is a linear subspace of $\mathbb{R}^{n\times n}$, which is finite-dimensional. ■

The factors U and V in Proposition 11.8 have orthonormal columns and therefore lie on Stiefel manifolds, which we introduce next.

Definition 11.13 (Stiefel Manifold)**.** *Let $m, n \in \mathbb{N}$ where $m \geq n$. Then,*

$$
\mathrm{St}(m, n) = \left\{ X \in \mathbb{R}^{m\times n} \;\middle|\; X^\top X = I_n \right\},
$$

*the set of all $m \times n$ matrices with orthonormal columns, is called the **Stiefel manifold**.*

Proposition 11.14. *The Stiefel manifold forms a smooth submanifold of $\mathbb{R}^{m\times n}$.*

Proof. Define $\Phi(X) = X^\top X - I_n$ for $X \in \mathbb{R}^{m\times n}$. Then $\Phi^{-1}(0) = \mathrm{St}(m, n)$, and Φ maps into $\mathrm{Sym}(n)$. The map Φ is smooth, and its derivative at Y is the linear map

$$
D\Phi[Y] : \mathbb{R}^{m\times n} \to \mathrm{Sym}(n), \qquad D\Phi[Y](H) = Y^\top H + H^\top Y.
$$

Let $H = \frac{1}{2}YA$ for arbitrary $A \in \mathrm{Sym}(n)$. Then

$$\begin{aligned} D\Phi[Y]\left(\frac{1}{2}YA\right) &= Y^\top\left(\frac{1}{2}YA\right) + \left(\frac{1}{2}YA\right)^\top Y \\ &= \frac{1}{2}Y^\top YA + \frac{1}{2}A^\top Y^\top Y \\ &= \frac{1}{2}A + \frac{1}{2}A^\top \\ &= \frac{1}{2}A + \frac{1}{2}A \\ &= A. \end{aligned} \tag{11.14}$$

Thus $D\Phi[Y]$ is surjective for every $Y \in \Phi^{-1}(0) = \mathrm{St}(m,n)$, so 0 is a regular value of Φ. Applying Theorem E.8 now shows that $\mathrm{St}(m,n)$ is a smooth embedded submanifold of $\mathbb{R}^{m\times n}$. ■

Definition 11.15 ($\mathrm{GL}_r(\mathbb{R})$, general linear group over $\mathbb{R}$)**.** *Let $r \in \mathbb{N}$. Then,*

$$\mathrm{GL}_r(\mathbb{R}) = \left\{X \in \mathbb{R}^{r\times r} \;\middle|\; \det(X) \neq 0\right\},$$

the set of all invertible $r \times r$ matrices, is called the ***general linear group****.*

Proposition 11.16. *The general linear group, $\mathrm{GL}_r(\mathbb{R}) \subset \mathbb{R}^{r\times r}$, forms a smooth matrix submanifold.*

Proof. The determinant is a polynomial in the entries of a matrix and is therefore smooth; in particular, it is continuous. Since $\mathbb{R}\setminus\{0\}$ is open,

$$\mathrm{GL}_r(\mathbb{R}) = (\det)^{-1}(\mathbb{R}\setminus\{0\})$$

is open in $\mathbb{R}^{r\times r}$. Every open subset of a Euclidean space is a smooth embedded submanifold of full dimension: the identity map on the open set supplies a smooth coordinate chart. Hence $\mathrm{GL}_r(\mathbb{R})$ is a smooth embedded submanifold of $\mathbb{R}^{r\times r}$ of dimension r^2. ■

As noted in Remark 11.10, a matrix $Y \in \mathcal{M}_r$ does not determine a unique triple (U,S,V) through the factorization $Y = USV^\top$. The following definitions and propositions describe this redundancy precisely.

Definition 11.17 (Factor parameter space and reconstruction map)**.** *Define*

$$\mathcal{P}_r = \mathrm{St}(m,r) \times \mathrm{GL}_r(\mathbb{R}) \times \mathrm{St}(n,r).$$

*The **reconstruction map** is*

$$\pi : \mathcal{P}_r \to \mathcal{M}_r, \qquad \pi(U, S, V) \coloneqq USV^\top.$$

Thus a point of $\mathcal{P}_r$ is a choice of factors, whereas its image under π is the matrix represented by those factors.

Proposition 11.18 (Orthogonal change-of-basis action)**.** *The formula*

$$(U, S, V) \cdot (Q, P) = (UQ, Q^\top SP, VP)$$

defines a smooth right action of $\mathrm{O}(r) \times \mathrm{O}(r)$ on $\mathcal{P}_r$. Moreover, the reconstruction map is unchanged by this action:

$$\pi\big((U, S, V) \cdot (Q, P)\big) = \pi(U, S, V).$$

Proof. The identity (I_r, I_r) leaves every triple unchanged. If $g_1 = (Q_1, P_1)$ and $g_2 = (Q_2, P_2)$, then

$$\big((U, S, V) \cdot g_1\big) \cdot g_2 = (UQ_1Q_2, Q_2^\top Q_1^\top SP_1P_2, VP_1P_2) = (U, S, V) \cdot (g_1 g_2), \tag{11.15}$$

where

$$(g_1 g_2) = (Q_1Q_2, P_1P_2).$$

Hence the formula defines a right group action, and it is smooth because it uses only matrix multiplication and transposition. Finally,

$$\pi\big((U, S, V) \cdot (Q, P)\big) = (UQ)(Q^\top SP)(VP)^\top = USV^\top,$$

which proves the invariance of π. ∎

The matrices Q and P change the orthonormal bases used for the column and row spaces of Y, while the transformation of S compensates for those changes.

Definition 11.19 (Factor quotient)**.** *Two triples (U, S, V) and $(\widetilde{U}, \widetilde{S}, \widetilde{V})$ in $\mathcal{P}_r$ are **equivalent** if there is a pair $(Q, P) \in \mathrm{O}(r) \times \mathrm{O}(r)$ such that*

$$(\widetilde{U}, \widetilde{S}, \widetilde{V}) = (U, S, V) \cdot (Q, P).$$

*The equivalence class of (U, S, V) is its **orbit**. The set of all such equivalence classes is the*

quotient

$$\mathcal{P}_r/\big(\mathrm{O}(r)\times\mathrm{O}(r)\big).$$

Proposition 11.20 (Quotient representation of the fixed-rank manifold)**.** *Two triples in $\mathcal{P}_r$ represent the same matrix if and only if they are equivalent. Consequently, the reconstruction map induces a bijection*

$$\overline{\pi}:\mathcal{P}_r/\big(\mathrm{O}(r)\times\mathrm{O}(r)\big)\longrightarrow\mathcal{M}_r,\qquad \overline{\pi}\big([(U,S,V)]\big)=USV^\top.$$

The quotient is a smooth manifold, and under this smooth structure $\overline{\pi}$ is a diffeomorphism.

Proof. Proposition 11.18 shows that equivalent triples represent the same matrix. Conversely, suppose

$$USV^\top=\widetilde{U}\widetilde{S}\widetilde{V}^\top.$$

Because S and $\widetilde{S}$ are invertible, the two sides have the same column and row spaces. Therefore there are $Q,P\in\mathrm{O}(r)$ such that $\widetilde{U}=UQ$ and $\widetilde{V}=VP$. Multiplication of the displayed identity on the left by $U^\top$ and on the right by V gives $\widetilde{S}=Q^\top SP$. Thus the triples are equivalent, so the fibers of π are exactly the group orbits. Surjectivity follows from Proposition 11.8; hence $\overline{\pi}$ is a well-defined bijection.

The action is free: if $(U,S,V)\cdot(Q,P)=(U,S,V)$, then the full column rank of U and V forces $Q=P=I_r$. It is proper because $\mathrm{O}(r)\times\mathrm{O}(r)$ is compact. Thus Theorem E.13 gives the quotient a smooth manifold structure. In the fixed-rank coordinate patches used in the proof of Proposition 11.3, the factors can be chosen smoothly; these local choices show that $\overline{\pi}$ and its inverse are smooth. This is the standard quotient description of the fixed-rank manifold; see [1, 21] for complete treatments. ■

We will not need to work with the quotient explicitly. For a particular $Y\in\mathcal{M}_r$, we choose any one factorization $Y=USV^\top$. The tangent space obtained below depends only on Y, and the later gauge conditions remove the corresponding infinitesimal freedom in the factors.

We now derive the tangent space of the fixed-rank manifold. Recall from Definition 11.4 that its elements are velocities of curves that remain on the manifold. The reconstruction map introduced in Definition 11.17 provides the connection between tangent directions of the factors and tangent vectors of the represented matrix.

Proposition 11.21 (Differential of the reconstruction map)**.** *Let*

$$\pi:\mathcal{P}_r\to\mathcal{M}_r,\qquad \pi(U,S,V)=USV^\top,$$

and let $Y = \pi(U, S, V)$. Then

$$D\pi_{(U,S,V)}(H_U, H_S, H_V) = H_U S V^\top + U H_S V^\top + U S H_V^\top, \tag{11.16}$$

where

$$H_U \in T_U \mathrm{St}(m, r), \qquad H_S \in T_S \mathrm{GL}_r(\mathbb{R}), \qquad H_V \in T_V \mathrm{St}(n, r).$$

Moreover, $D\pi_{(U,S,V)}$ maps onto $T_Y \mathcal{M}_r$. Consequently, every $Z \in T_Y \mathcal{M}_r$ admits a representation

$$Z = H_U S V^\top + U H_S V^\top + U S H_V^\top. \tag{11.17}$$

This representation is generally not unique.

Proof. Let $(U(t), S(t), V(t))$ be a C^1 curve in $\mathcal{P}_r$ through (U, S, V). Differentiating

$$Y(t) = \pi(U(t), S(t), V(t)) = U(t) S(t) V(t)^\top$$

gives

$$\dot{Y}(t) = \dot{U}(t) S(t) V(t)^\top + U(t) \dot{S}(t) V(t)^\top + U(t) S(t) \dot{V}(t)^\top. \tag{11.18}$$

Evaluating at $t = 0$ proves Equation (11.16).

It remains to prove surjectivity. Let $Z \in T_Y \mathcal{M}_r$ and choose a C^1 curve $Y(t) \in \mathcal{M}_r$ such that $Y(0) = Y$ and $\dot{Y}(0) = Z$. In the local fixed-rank coordinate patches used in the proof of Proposition 11.3, the factors of $Y(t)$ can be chosen smoothly. If the resulting factors at $t = 0$ differ from the prescribed triple (U, S, V), the equivalence established in Proposition 11.20 supplies constant orthogonal changes of basis that make the initial factors equal to (U, S, V). Thus we may write

$$Y(t) = U(t) S(t) V(t)^\top, \qquad (U(0), S(0), V(0)) = (U, S, V).$$

Setting

$$H_U = \dot{U}(0), \qquad H_S = \dot{S}(0), \qquad H_V = \dot{V}(0)$$

in Equation (11.18) yields Equation (11.17). Hence $D\pi_{(U,S,V)}$ maps onto $T_Y \mathcal{M}_r$. Nonuniqueness follows from the orthogonal change-of-basis freedom described in Proposition 11.18. ∎

We next characterize the three factor tangent spaces. The symbols H_U, H_S, and H_V denote generic tangent directions, whereas $\dot{U}$, $\dot{S}$, and $\dot{V}$ below are reserved for time derivatives along a solution curve.

Lemma 11.22. *Let $m, r \in \mathbb{N}$ and $U \in \mathrm{St}(m, r) \subset \mathbb{R}^{m\times r}$ be an element of the Stiefel manifold. Then the tangent space of the Stiefel manifold is given by the set*

$$T_U\mathrm{St}(m, r) = \left\{ H_U \in \mathbb{R}^{m\times r} \;\middle|\; H_U^\top U = -U^\top H_U \right\}. \tag{11.19}$$

Equivalently, the tangent space is the set of all matrices $H_U \in \mathbb{R}^{m\times r}$ such that the matrix $H_U^\top U$ is skew-symmetric.

Proof. Let $U(t) \in \mathrm{St}(m, r)$ be a C^1 curve for $t \in \mathbb{R}$. Then $U(t)^\top U(t) = I_r$ for all t. Taking the time derivative of both sides yields

$$O = \dot{U}(t)^\top U(t) + U(t)^\top \dot{U}(t), \tag{11.20}$$

where O is the $r \times r$ zero matrix. Setting $t = 0$ and replacing the derivatives with the tangent vector notation yields

$$H_U^\top U = -U^\top H_U. \tag{11.21}$$

Conversely, suppose that $H_U \in \mathbb{R}^{m\times r}$ satisfies

$$H_U^\top U + U^\top H_U = 0. \tag{11.22}$$

For t sufficiently close to zero, define

$$\gamma(t) = (U + tH_U)\left((U + tH_U)^\top (U + tH_U)\right)^{-1/2}. \tag{11.23}$$

The matrix

$$(U + tH_U)^\top (U + tH_U)$$

is symmetric positive definite for all sufficiently small t, so its inverse square root is well defined. Moreover,

$$\begin{aligned}\gamma(t)^\top \gamma(t) &= \left((U + tH_U)^\top (U + tH_U)\right)^{-1/2} \\ &\quad \cdot (U + tH_U)^\top (U + tH_U)\left((U + tH_U)^\top (U + tH_U)\right)^{-1/2} \\ &= I_r. \end{aligned} \tag{11.24}$$

Thus, $\gamma(t) \in \mathrm{St}(m, r)$ and $\gamma(0) = U$. Using $U^\top U = I_r$ and the assumed condition on H_U,

we have

$$\begin{aligned}(U+tH_U)^\top(U+tH_U) &= U^\top U + t\left(H_U^\top U + U^\top H_U\right) + t^2 H_U^\top H_U \\ &= I_r + t^2 H_U^\top H_U. \end{aligned} \tag{11.25}$$

Consequently,

$$\left(I_r + t^2 H_U^\top H_U\right)^{-1/2} = I_r + \mathcal{O}(t^2), \tag{11.26}$$

and hence

$$\gamma(t) = (U+tH_U)\left(I_r + \mathcal{O}(t^2)\right) = U + tH_U + \mathcal{O}(t^2). \tag{11.27}$$

It follows from taking a derivative that

$$\dot\gamma(0) = H_U. \tag{11.28}$$

Therefore, H_U is the velocity of a curve in $\mathrm{St}(m,r)$ passing through U, and hence $H_U \in T_U\mathrm{St}(m,r)$. This proves the reverse inclusion and completes the proof. ■

Lemma 11.23. *Let $r \in \mathbb{N}$ and $S \in \mathrm{GL}_r(\mathbb{R})$ be an element of the general linear group. Then, $T_S\mathrm{GL}_r(\mathbb{R}) \cong \mathbb{R}^{r\times r}$.*

Proof. Recall from the proof of Proposition 11.16 that $\mathrm{GL}_r(\mathbb{R}) \subset \mathbb{R}^{r\times r}$ is an open subset. By Proposition 3.10 in [50], if $U \subseteq \mathcal{M}$ is an open subset of a manifold and $p \in U$, then there exists an isomorphism $T_pU \cong T_p\mathcal{M}$. Hence, $T_S\mathrm{GL}_r(\mathbb{R}) \cong \mathbb{R}^{r^2} \cong \mathbb{R}^{r\times r}$. ■

Theorem 11.24 (Gauge-fixed tangent space)**.** *Let $Y = USV^\top \in \mathcal{M}_r$, where $U^\top U = V^\top V = I_r$ and $S \in \mathrm{GL}_r(\mathbb{R})$. Then*

$$T_Y\mathcal{M}_r = \left\{ H_USV^\top + UH_SV^\top + USH_V^\top \;\middle|\; \begin{array}{c} H_U \in \mathbb{R}^{m\times r}, \quad H_S \in \mathbb{R}^{r\times r}, \quad H_V \in \mathbb{R}^{n\times r}, \\ U^\top H_U = 0, \quad V^\top H_V = 0 \end{array} \right\}. \tag{11.29}$$

For the fixed factorization $Y = USV^\top$, the three tangent components H_U, H_S, and H_V in this representation are unique.

Proof. By Proposition 11.21, every tangent vector has a representation in terms of factor tangent directions. That representation is not unique: time-dependent orthogonal changes of basis can transfer skew-symmetric components between the factors. The gauge conditions remove exactly this freedom. Moreover, the Stiefel tangent-space characterization above shows that every triple satisfying the gauge conditions in Equation (11.29) consists of valid factor tangent directions. Hence the space on the right is a linear subspace of $T_Y\mathcal{M}_r$.

To prove uniqueness, suppose a gauge-fixed triple represents the zero tangent vector. Multiplication on the left by $U^\top$ and on the right by V gives $H_S = 0$. Multiplication of the remaining identity on the right by V gives $H_U S = 0$, hence $H_U = 0$ because S is invertible. Multiplication on the left by $U^\top$ then gives $S H_V^\top = 0$, so $H_V = 0$.

The space on the right of Equation (11.29) has dimension

$$(m-r)r + r^2 + (n-r)r = r(m+n-r),$$

which is the dimension of $\mathcal{M}_r$. Since it is a linear subspace of $T_Y\mathcal{M}_r$ with the same dimension, the two spaces agree. ■

Writing $P_U = UU^\top$ and $P_V = VV^\top$, the tangent space is the orthogonal direct sum

$$T_Y\mathcal{M}_r = P_U^\perp \mathbb{R}^{m\times n} P_V \oplus P_U \mathbb{R}^{m\times n} P_V \oplus P_U \mathbb{R}^{m\times n} P_V^\perp. \tag{11.30}$$

Consequently, for any $Z \in \mathbb{R}^{m\times n}$,

$$P_{T_Y\mathcal{M}_r}(Z) = P_U Z + Z P_V - P_U Z P_V. \tag{11.31}$$

The gauge is a coordinate choice, not an additional condition on the matrix curve $Y(t)$.

11.1.3 Tangent-Space Projection and Factor Evolution Equations

We now derive the three factor equations for $\dot{U}$, $\dot{S}$, and $\dot{V}$. The gauge-fixed tangent-space representation makes the derivation unique and avoids the ambiguity caused by time-dependent changes of basis.

Lemma 11.25. *Let $P_U = UU^\top$ and $P_V = VV^\top$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ have orthonormal columns. Then, P_U is an orthogonal projection onto the column space of U (denoted $\overline{U}$) and P_V is an orthogonal projection onto the column space of V (denoted $\overline{V}$).*

Proof. We will show P_U is an orthogonal projection onto the column space of U. First, we show P_U is a projection.

$$P_U^2 = (UU^\top)(UU^\top) = U(U^\top U)U^\top = UU^\top = P_U. \tag{11.32}$$

A standard fact in linear algebra is that a Hilbert space can be written uniquely as the direct sum of a closed subspace and its orthogonal complement. In this case, we have

$\mathbb{R}^m = \overline{U} \bigoplus \overline{U}^\perp$. If $x \in \mathbb{R}^m$, we can rewrite it as

$$x = x + UU^\top x - UU^\top x = UU^\top x + (I - UU^\top)x \tag{11.33}$$

We show that $(I - UU^\top)x \in U^\perp$. An element of $\overline{U}$ has the form Uy for some $y \in \mathbb{R}^r$. So,

$$\begin{aligned}\langle (I - UU^\top)x, Uy\rangle &= \langle x, Uy\rangle - \langle UU^\top x, Uy\rangle \\ &= \langle x, Uy\rangle - \langle U^\top x, U^\top Uy\rangle \\ &= \langle x, Uy\rangle - \langle U^\top x, y\rangle \\ &= \langle x, Uy\rangle - \langle x, Uy\rangle \\ &= 0. \end{aligned} \tag{11.34}$$

Thus, Equation (11.33) is the unique decomposition of x in $\overline{U} \bigoplus \overline{U}^\perp$. By the definition of orthogonal projection, $P_U v = UU^\top v$. The proof that P_V is an orthogonal projection onto $\overline{V}$ follows the same argument. ■

From here onward, we use $P_U^\perp = I - P_U$ and $P_V^\perp = I - P_V$ to denote the projections onto the orthogonal complements of $\overline{U}$ and $\overline{V}$.

We can now state the main result of this discussion.

Theorem 11.26. *Let $Y = USV^\top \in \mathcal{M}_r$, where $S \in \mathrm{GL}_r(\mathbb{R})$ and U and V have orthonormal columns. Set $R = \mathcal{R}(Y(t), t)$. Then the Galerkin and gauge conditions imply*

$$\begin{aligned}\dot{S} &= U^\top RV, \\ \dot{U} &= P_U^\perp RVS^{-1}, \\ \dot{V} &= P_V^\perp R^\top US^{-\top}. \end{aligned} \tag{11.35}$$

Proof. Write

$$\dot{Y} = \dot{U}SV^\top + U\dot{S}V^\top + USV^\top, \qquad U^\top \dot{U} = V^\top \dot{V} = 0.$$

Choose the tangent test matrix $Z = UAV^\top$, where $A \in \mathbb{R}^{r\times r}$ is arbitrary. The gauge conditions eliminate the $\dot{U}$ and $\dot{V}$ terms, so the Galerkin condition gives

$$\langle \dot{S} - U^\top RV, A\rangle_{\mathsf{F}} = 0 \quad \text{for every } A.$$

Therefore $\dot{S} = U^\top RV$.

Next choose $Z = KSV^\top$, where $U^\top K = 0$. Using the gauge conditions and the formula

for $\dot{S}$ gives

$$\langle \dot{U}S - P_U^{\perp} RV, K \rangle_{\mathrm{F}} = 0 \quad \text{for every } K \text{ with } U^{\top} K = 0.$$

The matrix $\dot{U}S - P_U^{\perp} RV$ itself has columns orthogonal to U. It is therefore an admissible choice of K, and its Frobenius norm must vanish. Since S is invertible,

$$\dot{U} = P_U^{\perp} RVS^{-1}.$$

Finally choose $Z = USL^{\top}$, where $V^{\top} L = 0$. The same calculation yields

$$\langle \dot{V}S^{\top} - P_V^{\perp} R^{\top} U, L \rangle_{\mathrm{F}} = 0 \quad \text{for every } L \text{ with } V^{\top} L = 0.$$

The difference in the first slot belongs to this test space, so it is zero and

$$\dot{V} = P_V^{\perp} R^{\top} US^{-\top}.$$

This proves all three equations. The subspace argument in the last two steps is essential: orthogonality to a subspace forces a vector to vanish only after one knows that the vector lies in that subspace. ■

Substituting Equation (11.35) into the derivative of $Y = USV^{\top}$ recovers the tangent-space projector:

$$\begin{aligned} \dot{Y}(t) &= P_{T_{Y(t)}\mathcal{M}_r} \mathcal{R}(Y(t), t) \\ &= U(t)U(t)^{\top} \mathcal{R}(Y(t), t) - U(t)U(t)^{\top} \mathcal{R}(Y(t), t) V(t)V(t)^{\top} \\ &\qquad + \mathcal{R}(Y(t), t) V(t)V(t)^{\top}. \end{aligned} \tag{11.36}$$

Here $Y(t) = U(t)S(t)V(t)^{\top}$.

11.1.4 Projector-Splitting and BUG Integrators

In this section, we discuss the primary methods for evolving the low-rank approximation in time. We first motivate the need for robust low-rank time integrators.

The projected equation is $\dot{Y} = P_{T_Y \mathcal{M}_r} \mathcal{R}(Y, t)$. Directly time-stepping the factor equations is unattractive because the equations for $\dot{U}$ and $\dot{V}$ contain S^{-1} and become ill-conditioned when the smallest singular value is small. Moreover, an ambient-space Runge–Kutta update need not have rank at most r. The projector-splitting integrator avoids both difficulties by evolving factorized subproblems without forming S^{-1} [53].

We begin with the projector-splitting integrator of [53]. Recall that for $\dot{A}(t) = \mathcal{R}(A(t), t)$, the fixed-rank approximation solves

$$\begin{aligned} \dot{Y}(t) &= P_{T_Y \mathcal{M}_r} \mathcal{R}(Y(t), t) \\ &= \mathcal{R}(Y(t), t) P_V - P_U \mathcal{R}(Y(t), t) P_V + P_U \mathcal{R}(Y(t), t), \\ Y(t_0) &= U_0 S_0 V_0^\top. \end{aligned} \tag{11.37}$$

Let $t_{n+1} = t_n + \tau$ and $Y_n = U_n S_n V_n^\top$. During the first two subflows, $P_{V_n} = V_n V_n^\top$ is frozen. After the first subflow is factored, $P_{U_{n+1}} = U_{n+1} U_{n+1}^\top$ is frozen during the second and third subflows. The exact subproblems are

$$\dot{Y}_{\mathrm{I}}(t) = \mathcal{R}(Y_{\mathrm{I}}(t), t) P_{V_n}, \qquad Y_{\mathrm{I}}(t_n) = Y_n, \tag{11.38a}$$

$$\dot{Y}_{\mathrm{II}}(t) = -P_{U_{n+1}} \mathcal{R}(Y_{\mathrm{II}}(t), t) P_{V_n}, \qquad Y_{\mathrm{II}}(t_n) = Y_{\mathrm{I}}(t_{n+1}), \tag{11.38b}$$

$$\dot{Y}_{\mathrm{III}}(t) = P_{U_{n+1}} \mathcal{R}(Y_{\mathrm{III}}(t), t), \qquad Y_{\mathrm{III}}(t_n) = Y_{\mathrm{II}}(t_{n+1}). \tag{11.38c}$$

If Φ_{I}, Φ_{II}, and Φ_{III} denote the exact solution maps of these subproblems over $[t_n, t_{n+1}]$, the Lie–Trotter step is

$$Y_{n+1} = \Phi_{\mathrm{III}} \circ \Phi_{\mathrm{II}} \circ \Phi_{\mathrm{I}}(Y_n).$$

This formula defines the splitting method independently of how the subproblems are solved numerically. The experiments below use Forward Euler for each subproblem, so their total temporal error contains both Lie–Trotter splitting error and subproblem discretization error. The following result is Lemma 3.1 of [53].

Lemma 11.27 (Factorized solutions of the splitting subproblems)**.** *The solutions of the three subproblems in* (11.38) *admit the following low-rank factorizations. More precisely, the frozen bases give*

$$Y_{\mathrm{I}}(t) = K(t) V_n^\top, \qquad Y_{\mathrm{II}}(t) = U_{n+1} S(t) V_n^\top, \qquad Y_{\mathrm{III}}(t) = U_{n+1} L(t)^\top.$$

Consequently, every subflow has rank at most r. Its rank is exactly r only while $K(t)$, $S(t)$, or $L(t)$, respectively, has full rank.

(i) *The solution of* (11.38a) *can be written as*

$$Y_{\mathrm{I}}(t) = U_{\mathrm{I}}(t) S_{\mathrm{I}}(t) V_{\mathrm{I}}(t)^\top,$$

where

$$\frac{d}{dt}\big(U_{\mathrm{I}}(t)S_{\mathrm{I}}(t)\big) = \mathcal{R}(Y_{\mathrm{I}}(t),t)V_{\mathrm{I}}(t), \qquad \dot{V}_{\mathrm{I}}(t) = 0.$$

(ii) *The solution of* (11.38b) *can be written as*

$$Y_{\mathrm{II}}(t) = U_{\mathrm{II}}(t)S_{\mathrm{II}}(t)V_{\mathrm{II}}(t)^\top,$$

where

$$\dot{S}_{\mathrm{II}}(t) = -U_{\mathrm{II}}(t)^\top \mathcal{R}(Y_{\mathrm{II}}(t),t)V_{\mathrm{II}}(t),$$

and

$$\dot{U}_{\mathrm{II}}(t) = 0, \qquad \dot{V}_{\mathrm{II}}(t) = 0.$$

(iii) *The solution of* (11.38c) *can be written as*

$$Y_{\mathrm{III}}(t) = U_{\mathrm{III}}(t)S_{\mathrm{III}}(t)V_{\mathrm{III}}(t)^\top,$$

where

$$\frac{d}{dt}\big(V_{\mathrm{III}}(t)S_{\mathrm{III}}(t)^\top\big) = \mathcal{R}(Y_{\mathrm{III}}(t),t)^\top U_{\mathrm{III}}(t), \qquad \dot{U}_{\mathrm{III}}(t) = 0.$$

Proof. First note that, for $Z \in \mathbb{R}^{m\times n}$, the terms $ZVV^\top$, $-UU^\top ZVV^\top$, and $UU^\top Z$ lie in the tangent space $T_Y\mathcal{M}_r$, as the following calculations show.

$$\begin{aligned} P_{T_Y\mathcal{M}_r}(ZVV^\top) &= ZVV^\top VV^\top - UU^\top ZVV^\top VV^\top + UU^\top ZVV^\top \\ &= ZVV^\top - UU^\top ZVV^\top + UU^\top ZVV^\top \\ &= ZVV^\top. \end{aligned} \tag{11.39}$$

Similarly,

$$\begin{aligned} P_{T_Y\mathcal{M}_r}(-UU^\top ZVV^\top) &= -UU^\top ZVV^\top VV^\top + UU^\top UU^\top ZVV^\top VV^\top - UU^\top UU^\top ZVV^\top \\ &= -UU^\top ZVV^\top VV^\top \\ &= -UU^\top ZVV^\top. \end{aligned} \tag{11.40}$$

Finally,

$$\begin{aligned} P_{T_Y\mathcal{M}_r}(UU^\top Z) &= UU^\top ZVV^\top - UU^\top UU^\top ZVV^\top + UU^\top UU^\top Z \\ &= UU^\top ZVV^\top - UU^\top ZVV^\top + UU^\top Z \end{aligned}$$

$$= UU^\top Z. \tag{11.41}$$

Hence the three vector fields have the required factorized forms. Their solutions remain in $\mathcal{M}_{\leq r}$; they remain in the smooth manifold $\mathcal{M}_r$ only as long as the evolving factor has full rank. On any interval on which this condition holds, write $Y_{\mathrm{I}}(t) = U_{\mathrm{I}}(t)S_{\mathrm{I}}(t)V_{\mathrm{I}}(t)^\top$, $Y_{\mathrm{II}}(t) = U_{\mathrm{II}}(t)S_{\mathrm{II}}(t)V_{\mathrm{II}}(t)^\top$, and $Y_{\mathrm{III}}(t) = U_{\mathrm{III}}(t)S_{\mathrm{III}}(t)V_{\mathrm{III}}(t)^\top$.

(i) First subflow. Set

$$Y_{\mathrm{I}}(t) = U_{\mathrm{I}}(t)S_{\mathrm{I}}(t)V_{\mathrm{I}}(t)^\top$$

and define

$$K_{\mathrm{I}}(t) = U_{\mathrm{I}}(t)S_{\mathrm{I}}(t).$$

It follows that

$$Y_{\mathrm{I}}(t) = K_{\mathrm{I}}(t)V_{\mathrm{I}}(t)^\top.$$

Differentiating this expression gives

$$\dot{Y}_{\mathrm{I}}(t) = \dot{K}_{\mathrm{I}}(t)V_{\mathrm{I}}(t)^\top + K_{\mathrm{I}}(t)\dot{V}_{\mathrm{I}}(t)^\top.$$

We choose a representation for which

$$\dot{V}_{\mathrm{I}}(t) = 0.$$

Substitution into (11.38a) then yields

$$\dot{K}_{\mathrm{I}}(t)V_{\mathrm{I}}(t)^\top = \mathcal{R}(Y_{\mathrm{I}}(t), t)V_{\mathrm{I}}(t)V_{\mathrm{I}}(t)^\top.$$

Right multiplication by $V_{\mathrm{I}}(t)$, together with $V_{\mathrm{I}}(t)^\top V_{\mathrm{I}}(t) = I_r$, gives

$$\dot{K}_{\mathrm{I}}(t) = \mathcal{R}(Y_{\mathrm{I}}(t), t)V_{\mathrm{I}}(t).$$

Since $\dot{V}_{\mathrm{I}}(t) = 0$, the matrix $V_{\mathrm{I}}(t)$ remains constant throughout this substep. (ii) Second subflow. Set

$$Y_{\mathrm{II}}(t) = U_{\mathrm{II}}(t)S_{\mathrm{II}}(t)V_{\mathrm{II}}(t)^\top.$$

Differentiating gives

$$\begin{aligned}\dot{Y}_{\mathrm{II}}(t) = \dot{U}_{\mathrm{II}}(t)S_{\mathrm{II}}(t)V_{\mathrm{II}}(t)^\top + U_{\mathrm{II}}(t)\dot{S}_{\mathrm{II}}(t)V_{\mathrm{II}}(t)^\top \\ + U_{\mathrm{II}}(t)S_{\mathrm{II}}(t)\dot{V}_{\mathrm{II}}(t)^\top.\end{aligned}$$

We choose a representation for which

$$\dot{U}_{\rm II}(t) = 0, \qquad \dot{V}_{\rm II}(t) = 0.$$

Substitution into (11.38b) therefore gives

$$U_{\rm II}(t)\dot{S}_{\rm II}(t)V_{\rm II}(t)^\top = -U_{\rm II}(t)U_{\rm II}(t)^\top \mathcal{R}(Y_{\rm II}(t), t)V_{\rm II}(t)V_{\rm II}(t)^\top.$$

Multiplying from the left by $U_{\rm II}(t)^\top$ and from the right by $V_{\rm II}(t)$ yields

$$\dot{S}_{\rm II}(t) = -U_{\rm II}(t)^\top \mathcal{R}(Y_{\rm II}(t), t)V_{\rm II}(t).$$

(iii) Third subflow. Set

$$Y_{\rm III}(t) = U_{\rm III}(t)S_{\rm III}(t)V_{\rm III}(t)^\top$$

and define

$$L_{\rm III}(t) = V_{\rm III}(t)S_{\rm III}(t)^\top.$$

Thus,

$$Y_{\rm III}(t) = U_{\rm III}(t)L_{\rm III}(t)^\top.$$

Differentiating gives

$$\dot{Y}_{\rm III}(t) = \dot{U}_{\rm III}(t)L_{\rm III}(t)^\top + U_{\rm III}(t)\dot{L}_{\rm III}(t)^\top.$$

We choose a representation for which

$$\dot{U}_{\rm III}(t) = 0.$$

Substitution into (11.38c) then yields

$$U_{\rm III}(t)\dot{L}_{\rm III}(t)^\top = U_{\rm III}(t)U_{\rm III}(t)^\top \mathcal{R}(Y_{\rm III}(t), t).$$

Multiplying from the left by $U_{\rm III}(t)^\top$ gives

$$\dot{L}_{\rm III}(t)^\top = U_{\rm III}(t)^\top \mathcal{R}(Y_{\rm III}(t), t).$$

Taking the transpose therefore gives

$$\dot{L}_{\rm III}(t) = \mathcal{R}(Y_{\rm III}(t), t)^\top U_{\rm III}(t).$$

This completes the proof. ■

11.1.4.1 Projector-Splitting K-, S-, and L-Steps

The factorized subflows in Lemma 11.27 lead to the projector-splitting integrator of [53]. One time step consists of a K-step, an S-step, and an L-step. By evolving the factors directly, the method retains the low-rank representation and avoids forming S^{-1}. For simplicity, we discretize each subproblem below with Forward Euler.

Definition 11.28 (Projector-splitting integrator)**.** *Let $\tau > 0$, let $t_{n+1} = t_n + \tau$, and suppose $Y_n = U_n S_n V_n^\top$ is the rank-r approximation at time t_n. One step of the projector-splitting integrator advances Y_n to Y_{n+1} through the following three substeps.*

*K-**step.** Set*

$$K(t_n) = U_n S_n$$

and solve the matrix ODE

$$\dot{K}(t) = \mathcal{R}\big(K(t)V_n^\top, t\big)V_n, \qquad t \in [t_n, t_{n+1}].$$

After computing $K(t_{n+1})$, perform a QR factorization

$$K(t_{n+1}) = U_{n+1}\widehat{S}_{n+1},$$

where the columns of U_{n+1} are orthonormal.

*S-**step.** Set*

$$S(t_n) = \widehat{S}_{n+1}$$

and solve

$$\dot{S}(t) = -U_{n+1}^\top \mathcal{R}\big(U_{n+1}S(t)V_n^\top, t\big)V_n, \qquad t \in [t_n, t_{n+1}].$$

Denote the value obtained at the end of this step by

$$\widetilde{S}_{n+1} = S(t_{n+1}).$$

*L-**step.** Set*

$$L(t_n) = V_n \widetilde{S}_{n+1}^\top$$

and solve

$$\dot{L}(t) = \mathcal{R}\big(U_{n+1}L(t)^\top, t\big)^\top U_{n+1}, \qquad t \in [t_n, t_{n+1}].$$

Finally, perform a QR factorization

$$L(t_{n+1}) = V_{n+1} S_{n+1}^\top,$$

where the columns of V_{n+1} are orthonormal.

The approximation at the new time is then

$$Y_{n+1} = U_{n+1} S_{n+1} V_{n+1}^\top.$$

This method is condensed into the algorithm below.

Algorithm 1 Projector-splitting update for $Y_n = U_n S_n V_n^\top$ using Forward Euler time stepping, as presented in [53]

Require: $Y_n = U_n S_n V_n^\top$ and time step $\tau > 0$

Ensure: $Y_{n+1} = U_{n+1} S_{n+1} V_{n+1}^\top$

K-step

$K_n \leftarrow U_n S_n$

$K_{n+1} \leftarrow K_n + \tau \mathcal{R}\left(K_n V_n^\top, t_n\right) V_n$

$\left[U_{n+1}, \widehat{S}_{n+1}\right] \leftarrow \mathrm{QR}(K_{n+1})$

S-step

$\widetilde{S}_{n+1} \leftarrow \widehat{S}_{n+1} - \tau U_{n+1}^\top \mathcal{R}\left(U_{n+1} \widehat{S}_{n+1} V_n^\top, t_n\right) V_n$

L-step

$L_n \leftarrow V_n \widetilde{S}_{n+1}^\top$

$L_{n+1} \leftarrow L_n + \tau \mathcal{R}\left(U_{n+1} L_n^\top, t_n\right)^\top U_{n+1}$

$\left[V_{n+1}, S_{n+1}^\top\right] \leftarrow \mathrm{QR}(L_{n+1})$

return $U_{n+1}, S_{n+1}, V_{n+1}$

The basis-update and Galerkin (BUG) integrator, originally called the unconventional robust integrator, was introduced in [14]. Its K- and L-basis updates, together with their QR factorizations, can be performed in parallel. Like the projector-splitting method, it remains robust to small singular values by updating the K and L factors rather than the derivatives $\dot{U}$, $\dot{S}$, and $\dot{V}$ directly. It also preserves symmetry or skew-symmetry when the differential equation preserves the corresponding matrix structure. A further motivation for this integrator

concerns the negative sign in the S-step of the projector-splitting integrator. For example, consider the ODE

$$\mathcal{R}(Y(t)) = -\lambda Y(t), \tag{11.42}$$

or

$$\dot{Y}(t) = -\lambda Y(t) \tag{11.43}$$

for $\lambda > 0$, $t \geq 0$, and $Y(0) = Y_0$. The exact solution is given by

$$Y(t) = e^{-\lambda t} Y_0. \tag{11.44}$$

Factor the initial value as

$$Y_0 = U_0 S_0 V_0^\top. \tag{11.45}$$

The K-step yields

$$\begin{aligned} \dot{K}(t) &= \mathcal{R}\big(K(t)V_0^\top\big)V_0 \\ &= -\lambda K(t) V_0^\top V_0 \\ &= -\lambda K(t). \end{aligned} \tag{11.46}$$

The S-step yields

$$\begin{aligned} \dot{S}(t) &= -U_1^\top \mathcal{R}\big(U_1 S(t) V_0^\top\big) V_0 \\ &= -U_1^\top \big(-\lambda U_1 S(t) V_0^\top\big) V_0 \\ &= \lambda S(t). \end{aligned} \tag{11.47}$$

Thus,

$$S(t) = e^{\lambda(t-t_0)} S(t_0). \tag{11.48}$$

Thus, the S-subproblem is exponentially increasing. When the three subproblems are solved exactly, this intermediate growth is canceled by the dissipative K- and L-steps, and the complete projector-splitting step reproduces the expected decay. Numerically, however, the anti-dissipative S-subproblem can amplify perturbations and splitting errors, particularly for strongly dissipative or stiff equations. The unconventional integrator avoids this artificial sign reversal by replacing the middle splitting step with a Galerkin evolution in the updated bases. The algorithm proceeds as follows.

Definition 11.29 (Fixed-rank BUG integrator)**.** *Let $\tau > 0$, let $t_{n+1} = t_n + \tau$, and suppose*

the rank-r approximation at time t_n is

$$Y_n = U_n S_n V_n^\top. \tag{11.49}$$

One step of the fixed-rank basis-update and Galerkin integrator advances Y_n to Y_{n+1} as follows.

Basis update. *Update $U_n \longrightarrow U_{n+1}$ and $V_n \longrightarrow V_{n+1}$ in parallel.*

*K-**step:** Integrate from $t = t_n$ to t_{n+1} the $m \times r$ matrix ODE*

$$\dot{K}(t) = \mathcal{R}\big(K(t)V_n^\top, t\big)V_n, \qquad K(t_n) = U_n S_n, \tag{11.50}$$

to obtain

$$K(t_{n+1}) = K_{n+1}, \tag{11.51}$$

and then perform a QR decomposition

$$U_{n+1}R_{n+1} = \mathrm{QR}(K_{n+1}). \tag{11.52}$$

Now set

$$M = U_{n+1}^\top U_n \in \mathbb{R}^{r\times r}. \tag{11.53}$$

*L-**step:** Integrate from $t = t_n$ to t_{n+1} the $n \times r$ matrix ODE*

$$\dot{L}(t) = \mathcal{R}\big(U_n L(t)^\top, t\big)^\top U_n, \qquad L(t_n) = V_n S_n^\top, \tag{11.54}$$

and set

$$V_{n+1}\widetilde{R}_{n+1} = \mathrm{QR}\big(L(t_{n+1})\big). \tag{11.55}$$

Define the $r \times r$ matrix

$$N = V_{n+1}^\top V_n. \tag{11.56}$$

Galerkin core update. *Update $S_n \longrightarrow S_{n+1}$.*

*S-**step:** Integrate from $t = t_n$ to t_{n+1} the $r \times r$ matrix ODE*

$$\dot{S}(t) = U_{n+1}^\top \mathcal{R}\big(U_{n+1}S(t)V_{n+1}^\top, t\big)V_{n+1}, \qquad S(t_n) = MS_nN^\top, \tag{11.57}$$

to get $S(t_{n+1})$. Set

$$S_{n+1} = S(t_{n+1}). \tag{11.58}$$

Each of the differential equations in the substeps can be solved using a standard integrator such as an explicit or implicit Runge–Kutta method. If $\mathcal{R}(Y, t)$ is linear, an exponential

integrator can be used as well. The S-step can be viewed as a Galerkin condition on the subspace

$$\mathcal{W} = \left\{ U_{n+1} Z V_{n+1}^\top \;\middle|\; Z \in \mathbb{R}^{r\times r} \right\} \subseteq \mathbb{R}^{m\times n}. \tag{11.59}$$

This is an r^2-dimensional subspace. Since U and V are held constant, we are effectively computing an approximation of the form

$$Y_S(t) = U_{n+1} S(t) V_{n+1}^\top. \tag{11.60}$$

Thus,

$$\begin{aligned} \dot{Y}_S(t) &= U_{n+1} \dot{S}(t) V_{n+1}^\top \\ &= U_{n+1} U_{n+1}^\top \mathcal{R}\big(U_{n+1} S(t) V_{n+1}^\top, t\big) V_{n+1} V_{n+1}^\top. \end{aligned} \tag{11.61}$$

Defining the residual

$$R_{\text{res}}(t) = \dot{Y}_S(t) - \mathcal{R}(Y_S(t), t), \tag{11.62}$$

the Galerkin condition on $\mathcal{W}$ is posed as finding $Y_S(t)$ such that

$$\left\langle \dot{Y}_S(t) - \mathcal{R}(Y_S(t), t), U_{n+1} Z V_{n+1}^\top \right\rangle_{\mathsf{F}} = 0 \tag{11.63}$$

over all $Z \in \mathbb{R}^{r\times r}$. Equivalently,

$$\left\langle U_{n+1}^\top \left(U_{n+1} \dot{S}(t) V_{n+1}^\top - \mathcal{R}(Y_S(t), t) \right) V_{n+1}, Z \right\rangle_{\mathsf{F}} = 0 \tag{11.64}$$

over all $Z \in \mathbb{R}^{r\times r}$. Hence,

$$U_{n+1}^\top U_{n+1} \dot{S}(t) V_{n+1}^\top V_{n+1} - U_{n+1}^\top \mathcal{R}\big(U_{n+1} S(t) V_{n+1}^\top, t\big) V_{n+1} = 0, \tag{11.65}$$

or equivalently,

$$\dot{S}(t) - U_{n+1}^\top \mathcal{R}\big(U_{n+1} S(t) V_{n+1}^\top, t\big) V_{n+1} = 0. \tag{11.66}$$

This yields the Galerkin part of the BUG method. The first step updates the bases U_{n+1} and V_{n+1}. The second step implements a Galerkin method whose S-step has no artificial minus sign. The resulting method is summarized below.

Algorithm 2 Fixed-rank BUG update for $Y_n = U_n S_n V_n^\top$ using Forward Euler time stepping, as presented in [14]

Require: $Y_n = U_n S_n V_n^\top$ and time step $\tau > 0$

Ensure: $Y_{n+1} = U_{n+1} S_{n+1} V_{n+1}^\top$

K-step

$K_n \leftarrow U_n S_n$

$K_{n+1} \leftarrow K_n + \tau \mathcal{R}\big(K_n V_n^\top, t_n\big) V_n$

$[U_{n+1}, R_{n+1}] \leftarrow \mathrm{QR}(K_{n+1})$

$M \leftarrow U_{n+1}^\top U_n$

L-step

$L_n \leftarrow V_n S_n^\top$

$L_{n+1} \leftarrow L_n + \tau \mathcal{R}\big(U_n L_n^\top, t_n\big)^\top U_n$

$\left[V_{n+1}, \widetilde{R}_{n+1}\right] \leftarrow \mathrm{QR}(L_{n+1})$

$N \leftarrow V_{n+1}^\top V_n$

The K- and L-steps may be performed in parallel.

S-step

$\widetilde{S}_n \leftarrow M S_n N^\top$

$S_{n+1} \leftarrow \widetilde{S}_n + \tau U_{n+1}^\top \mathcal{R}\Big(U_{n+1} \widetilde{S}_n V_{n+1}^\top, t_n\Big) V_{n+1}$

return $U_{n+1}, S_{n+1}, V_{n+1}$

The projector-splitting and BUG integrators also satisfy exactness and robust error properties under appropriate regularity, fixed-rank, and basis-overlap assumptions. We do not discuss these analytical results here, since the purpose of this section is to describe the construction of the integrators.

The rank-adaptive robust integrator of [13] extends the BUG method described above. A fixed-rank method may become inaccurate when the rank required to approximate the solution changes over time. The adaptive method addresses this limitation by constructing augmented basis spaces, evolving the Galerkin approximation in those enlarged spaces, and then truncating the result at a prescribed tolerance.

Definition 11.30 (Rank-adaptive BUG integrator)**.** *Let $\tau > 0$, let $t_{n+1} = t_n + \tau$, let $\vartheta > 0$ be a truncation tolerance, and suppose $Y_n = U_n S_n V_n^\top$ is the rank-r_n approximation at time*

t_n. One step of the rank-adaptive BUG integrator is defined as follows.

Augmented basis construction. *Construct the augmented basis matrices $\widehat{U}_{n+1}$ and $\widehat{V}_{n+1}$ in parallel.*

*K-**step:** Integrate from $t = t_n$ to t_{n+1} the $m \times r_n$ matrix ODE*

$$\dot{K}(t) = \mathcal{R}\big(K(t)V_n^\top, t\big)V_n, \qquad K(t_n) = U_n S_n, \tag{11.67}$$

to obtain

$$K(t_{n+1}) = K_{n+1}. \tag{11.68}$$

Rather than replacing U_n by an orthonormal basis of K_{n+1}, as in the fixed-rank BUG method, augment the new directions with the previous basis. Form the matrix

$$\begin{bmatrix} K_{n+1} & U_n \end{bmatrix} \in \mathbb{R}^{m \times 2r_n} \tag{11.69}$$

and compute

$$\widehat{U}_{n+1} R^U_{n+1} = \mathrm{QR}\left(\begin{bmatrix} K_{n+1} & U_n \end{bmatrix}\right). \tag{11.70}$$

Thus, the columns of

$$\widehat{U}_{n+1} \in \mathbb{R}^{m \times 2r_n} \tag{11.71}$$

form an orthonormal basis for

$$\mathrm{range}\left(\begin{bmatrix} K_{n+1} & U_n \end{bmatrix}\right). \tag{11.72}$$

Define

$$\widehat{M} = \widehat{U}_{n+1}^\top U_n \in \mathbb{R}^{2r_n \times r_n}. \tag{11.73}$$

*L-**step:** Integrate from $t = t_n$ to t_{n+1} the $n \times r_n$ matrix ODE*

$$\dot{L}(t) = \mathcal{R}\big(U_n L(t)^\top, t\big)^\top U_n, \qquad L(t_n) = V_n S_n^\top, \tag{11.74}$$

to obtain

$$L(t_{n+1}) = L_{n+1}. \tag{11.75}$$

Augment the new directions with the previous basis by forming

$$\begin{bmatrix} L_{n+1} & V_n \end{bmatrix} \in \mathbb{R}^{n \times 2r_n}, \tag{11.76}$$

and compute

$$\widehat{V}_{n+1} R^V_{n+1} = \mathrm{QR}\left(\begin{bmatrix} L_{n+1} & V_n \end{bmatrix}\right). \tag{11.77}$$

The columns of

$$\widehat{V}_{n+1} \in \mathbb{R}^{n\times 2r_n} \tag{11.78}$$

therefore form an orthonormal basis for

$$\operatorname{range}\left(\begin{bmatrix} L_{n+1} & V_n \end{bmatrix}\right). \tag{11.79}$$

Define

$$\widehat{N} = \widehat{V}_{n+1}^\top V_n \in \mathbb{R}^{2r_n\times r_n}. \tag{11.80}$$

Since the columns of U_n and V_n lie in the corresponding augmented spaces, we have

$$U_n = \widehat{U}_{n+1}\widehat{M}, \qquad V_n = \widehat{V}_{n+1}\widehat{N}. \tag{11.81}$$

Consequently, the previous approximation can be represented exactly in the augmented basis spaces:

$$Y_n = \widehat{U}_{n+1}\big(\widehat{M}S_n\widehat{N}^\top\big)\widehat{V}_{n+1}^\top. \tag{11.82}$$

This identity motivates the initial condition for the augmented S-step.

Augmented Galerkin core update. *Update the core matrix in the augmented basis spaces.*

***Augmented** S**-step:** Integrate from $t = t_n$ to t_{n+1} the $2r_n \times 2r_n$ matrix ODE*

$$\dot{\widehat{S}}(t) = \widehat{U}_{n+1}^\top \mathcal{R}\left(\widehat{U}_{n+1}\widehat{S}(t)\widehat{V}_{n+1}^\top, t\right)\widehat{V}_{n+1}, \tag{11.83}$$

subject to the initial condition

$$\widehat{S}(t_n) = \widehat{M}S_n\widehat{N}^\top. \tag{11.84}$$

This produces the augmented core matrix

$$\widehat{S}_{n+1} = \widehat{S}(t_{n+1}) \in \mathbb{R}^{2r_n\times 2r_n}. \tag{11.85}$$

Rank truncation. *Truncate the augmented approximation to the prescribed tolerance ϑ.*

Compute the singular value decomposition

$$\widehat{S}_{n+1} = \widehat{P}\widehat{\Sigma}\widehat{Q}^\top, \tag{11.86}$$

where

$$\widehat{\Sigma} = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_{2r_n}\right), \qquad \sigma_1 \geq \cdots \geq \sigma_{2r_n} \geq 0. \tag{11.87}$$

Choose $r_{n+1} \leq 2r_n$ to be the smallest integer satisfying

$$\left(\sum_{j=r_{n+1}+1}^{2r_n} \sigma_j^2 \right)^{1/2} \leq \vartheta. \tag{11.88}$$

Let

$$P_{n+1} = \widehat{P}(:, 1 : r_{n+1}), \qquad Q_{n+1} = \widehat{Q}(:, 1 : r_{n+1}), \tag{11.89}$$

and define

$$S_{n+1} = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_{r_{n+1}}\right). \tag{11.90}$$

The updated basis matrices are then

$$U_{n+1} = \widehat{U}_{n+1} P_{n+1} \in \mathbb{R}^{m \times r_{n+1}}, \tag{11.91}$$

and

$$V_{n+1} = \widehat{V}_{n+1} Q_{n+1} \in \mathbb{R}^{n \times r_{n+1}}. \tag{11.92}$$

Therefore, the approximation at time t_{n+1} is

$$Y_{n+1} = U_{n+1} S_{n+1} V_{n+1}^{\top}. \tag{11.93}$$

In the fixed-rank BUG method, the new bases are constructed only from K_{n+1} and L_{n+1}. In the rank-adaptive method, the old bases U_n and V_n are retained and combined with the new directions. Consequently, the temporary approximation space can grow from dimension r_n to at most $2r_n$. The subsequent SVD determines which of these old and new directions should be retained. Depending on the discarded singular values and the tolerance ϑ, the new rank r_{n+1} may be smaller than, equal to, or larger than r_n.

Algorithm 3 Rank-adaptive BUG update for $Y_n = U_n S_n V_n^\top$ using Forward Euler time stepping, as presented in [13]

Require: $Y_n = U_n S_n V_n^\top$ of rank r_n, time step $\tau > 0$, and truncation tolerance $\vartheta > 0$

Ensure: $Y_{n+1} = U_{n+1} S_{n+1} V_{n+1}^\top$ of rank $r_{n+1} \leq 2r_n$

K-step and augmented column basis

$K_n \leftarrow U_n S_n$

$K_{n+1} \leftarrow K_n + \tau \mathcal{R}\left(K_n V_n^\top, t_n\right) V_n$

$\left[\widehat{U}, R_U\right] \leftarrow \mathrm{QR}([K_{n+1}, U_n])$

$\widehat{M} \leftarrow \widehat{U}^\top U_n$

L-step and augmented row basis

$L_n \leftarrow V_n S_n^\top$

$L_{n+1} \leftarrow L_n + \tau \mathcal{R}\left(U_n L_n^\top, t_n\right)^\top U_n$

$\left[\widehat{V}, R_V\right] \leftarrow \mathrm{QR}([L_{n+1}, V_n])$

$\widehat{N} \leftarrow \widehat{V}^\top V_n$

The K- and L-steps may be performed in parallel.

Augmented S-step

$\widehat{S}_n \leftarrow \widehat{M} S_n \widehat{N}^\top$

$\widehat{S}_{n+1} \leftarrow \widehat{S}_n + \tau \widehat{U}^\top \mathcal{R}\left(\widehat{U} \widehat{S}_n \widehat{V}^\top, t_n\right) \widehat{V}$

Truncation step

$\left[\widehat{P}, \widehat{\Sigma}, \widehat{Q}\right] \leftarrow \mathrm{SVD}\left(\widehat{S}_{n+1}\right)$

Let

$$\widehat{\Sigma} = \mathrm{diag}\left(\sigma_1, \ldots, \sigma_{2r_n}\right), \qquad \sigma_1 \geq \cdots \geq \sigma_{2r_n} \geq 0$$

Choose the smallest $r_{n+1} \leq 2r_n$ such that

$$\left(\sum_{j=r_{n+1}+1}^{2r_n} \sigma_j^2 \right)^{1/2} \leq \vartheta$$

Let $\widehat{P}_{r_{n+1}}$ and $\widehat{Q}_{r_{n+1}}$ contain the first r_{n+1} columns of $\widehat{P}$ and $\widehat{Q}$, respectively

Let $\widehat{\Sigma}_{r_{n+1}}$ be the leading $r_{n+1} \times r_{n+1}$ principal block of $\widehat{\Sigma}$

$U_{n+1} \leftarrow \widehat{U} \widehat{P}_{r_{n+1}}$

$S_{n+1} \leftarrow \widehat{\Sigma}_{r_{n+1}}$

$V_{n+1} \leftarrow \widehat{V} \widehat{Q}_{r_{n+1}}$

return $U_{n+1}, S_{n+1}, V_{n+1}, r_{n+1}$

11.1.5 Continuous Dynamical Low-Rank Approximation

In this section, we provide a brief overview of continuous dynamical low-rank approximation, which replaces matrix spaces with function spaces. For partial differential equations (PDEs), the goal is to approximate the solution by a sum of separable functions, called a low-rank function. A matrix DLRA method first discretizes the PDE and then converts it into a time-dependent matrix ODE; the continuous formulation instead produces projected evolution equations before a particular spatial discretization is chosen. The resulting dynamics mirror the matrix-ODE formulation. Because kinetic equations are a primary application, we assume that the PDE domain has a phase-space structure.

Let $H_x = L^2(\Omega_x)$ and $H_v = L^2(\Omega_v)$, and assume these Hilbert spaces are separable. We identify $L^2(\Omega_x \times \Omega_v)$ with the Hilbert tensor product $H_x \otimes H_v$. Consider

$$\begin{aligned} \partial_t f(x, v, t) &= \mathcal{G}(f, t), \\ f(x, v, t_0) &= f_0(x, v). \end{aligned} \tag{11.94}$$

The operator $\mathcal{G}$ may contain explicit dependence on x and v in addition to the displayed dependence on f and t. Continuous DLRA seeks an approximation of the form

$$f_r(x, v, t) = \sum_{i=1}^{r} \sum_{j=1}^{r} X_i(x, t) S_{ij}(t) V_j(v, t), \tag{11.95}$$

where $X_i(\cdot, t) \in H_x$, $V_j(\cdot, t) \in H_v$, $S(t) \in \mathrm{GL}_r(\mathbb{R})$, and

$$\langle X_i(\cdot, t), X_j(\cdot, t) \rangle_{L^2(\Omega_x)} = \langle V_i(\cdot, t), V_j(\cdot, t) \rangle_{L^2(\Omega_v)} = \delta_{ij} \tag{11.96}$$

for each $t \in [t_0, T]$. The invertibility of $S(t)$ guarantees exact separated rank r.

Definition 11.31 (Separated rank)**.** *For $g \in H_x \otimes H_v$, its **separated rank** is*

$$\mathrm{rank}_{\mathrm{sep}}(g) = \min \left\{ q \;\middle|\; g(x, v) = \sum_{k=1}^{q} a_k(x) b_k(v),\ a_k \in H_x,\ b_k \in H_v \right\},$$

with rank $+\infty$ if no finite representation exists. Equivalently, this is the rank of the Hilbert–Schmidt operator from H_v to H_x whose kernel is g.

We use

$$\mathcal{M}_r = \{ g \in H_x \otimes H_v \mid \mathrm{rank}_{\mathrm{sep}}(g) = r \}.$$

The set of separated rank at most r also contains lower-rank strata and is not the same smooth manifold. At each time, the DLRA velocity minimizes

$$\frac{1}{2}\|\delta f_r - \mathcal{G}(f_r, t)\|^2_{L^2(\Omega_x \times \Omega_v)} \tag{11.97}$$

over $\delta f_r \in T_{f_r}\mathcal{M}_r$ at every $t \geq 0$. Equivalently, we compute

$$\partial_t f_r(x, v, t) = \underset{\delta f_r \in T_{f_r}\mathcal{M}_r}{\arg\min} \frac{1}{2}\|\delta f_r - \mathcal{G}(f_r, t)\|^2_{L^2(\Omega_x \times \Omega_v)}. \tag{11.98}$$

Recall the $L^2(\Omega_x \times \Omega_v)$ norm is generated by the inner product

$$\langle f, g\rangle_{L^2(\Omega_x \times \Omega_v)} = \int_{\Omega_x \times \Omega_v} f(x, v) g(x, v) \mathrm{d}x\mathrm{d}v, \tag{11.99}$$

Here the function spaces are real-valued. Then, Equation (11.97) can be written as a Galerkin condition: find $\partial_t f_r \in T_{f_r}\mathcal{M}_r$ such that

$$\langle \partial_t f_r - \mathcal{G}(f_r, t), \delta f_r\rangle_{L^2(\Omega_x \times \Omega_v)} = 0 \tag{11.100}$$

for all $\delta f_r \in T_{f_r}\mathcal{M}_r$. As in Proposition 11.5, this is equivalent to orthogonal projection because each tangent space is closed.

Definition 11.32 (Fixed-Rank Function Manifold)**.** *The **fixed-rank function manifold** has the equivalent factorized description*

$$\mathcal{M}_r = \left\{ f(x, v) = \sum_{i=1}^{r}\sum_{j=1}^{r} X_i(x) S_{ij} V_j(v) \;\middle|\; \langle X_i, X_j\rangle_{L^2(\Omega_x)} = \langle V_i, V_j\rangle_{L^2(\Omega_v)} = \delta_{ij}, S \in \mathrm{GL}_r(\mathbb{R}) \right\} \tag{11.101}$$

Any function in this set can be written compactly as $f(x, v) = X(x)SV(v)^\top$, where $X = (X_1, \ldots, X_r)$ and $V = (V_1, \ldots, V_r)$.

Theorem 11.33 (Fixed-rank function manifold)**.** *Under the separability assumptions above, $\mathcal{M}_r$ is a smooth embedded Hilbert submanifold of $H_x \otimes H_v$.*

Proof reference. This is the standard fixed-rank Hilbert-manifold theorem; see Appendix A of [4]. Under the stated separability assumptions, local charts are obtained by splitting each factor space into the span of the current orthonormal factors and its orthogonal complement. The invertibility of S keeps the representation on the exact-rank stratum, and these charts give the asserted smooth embedded Hilbert-manifold structure. ■

As in standard theory on smooth manifolds, this implies the existence of a tangent space at every point in $\mathcal{M}_r$. The definition of the tangent space in the context of Hilbert manifolds is analogous to that in the finite dimensional case.

We next construct the relevant orthogonal projections.

Let

$$\mathcal{X} = \text{span}\{X_1, \dots, X_r\} \subseteq L^2(\Omega_x)$$

and

$$\mathcal{V} = \text{span}\{V_1, \dots, V_r\} \subseteq L^2(\Omega_v),$$

where

$$\langle X_i, X_j \rangle_{L^2(\Omega_x)} = \delta_{ij} \qquad \text{and} \qquad \langle V_i, V_j \rangle_{L^2(\Omega_v)} = \delta_{ij}.$$

Proposition 11.34. *Define $P_X : L^2(\Omega_x) \to \mathcal{X}$ by*

$$P_X g = \sum_{i=1}^{r} X_i \langle X_i, g \rangle_{L^2(\Omega_x)}.$$

Then P_X is the orthogonal projection of $L^2(\Omega_x)$ onto $\mathcal{X}$.

Proof. First, $P_X g \in \mathcal{X}$ for every $g \in L^2(\Omega_x)$. Moreover, using the orthonormality of $\{X_i\}_{i=1}^r$, we obtain

$$\begin{aligned} P_X^2 g &= P_X \left(\sum_{k=1}^{r} X_k \langle X_k, g \rangle_{L^2(\Omega_x)} \right) \\ &= \sum_{i=1}^{r} \sum_{k=1}^{r} X_i \langle X_i, X_k \rangle_{L^2(\Omega_x)} \langle X_k, g \rangle_{L^2(\Omega_x)} \\ &= \sum_{i=1}^{r} X_i \langle X_i, g \rangle_{L^2(\Omega_x)} \\ &= P_X g. \end{aligned} \tag{11.102}$$

Thus, P_X is a projection.

It remains to show that it is an orthogonal projection. Let $h \in \mathcal{X}$. Then

$$h = \sum_{j=1}^{r} \alpha_j X_j$$

for some $\alpha_1, \ldots, \alpha_r \in \mathbb{R}$. Therefore,

$$\begin{aligned}\langle g - P_X g, h\rangle_{L^2(\Omega_x)} &= \sum_{j=1}^{r} \alpha_j \langle g, X_j\rangle_{L^2(\Omega_x)} \\ &\quad - \sum_{i=1}^{r}\sum_{j=1}^{r} \alpha_j \langle X_i, g\rangle_{L^2(\Omega_x)} \langle X_i, X_j\rangle_{L^2(\Omega_x)} \\ &= \sum_{j=1}^{r} \alpha_j \langle g, X_j\rangle_{L^2(\Omega_x)} - \sum_{j=1}^{r} \alpha_j \langle X_j, g\rangle_{L^2(\Omega_x)} \\ &= 0. \end{aligned} \tag{11.103}$$

Hence, $g - P_X g$ is orthogonal to $\mathcal{X}$, so P_X is the orthogonal projection onto $\mathcal{X}$. ■

Proposition 11.35. *Define $P_V : L^2(\Omega_v) \to \mathcal{V}$ by*

$$P_V g = \sum_{j=1}^{r} V_j \langle V_j, g\rangle_{L^2(\Omega_v)}.$$

Then P_V is the orthogonal projection of $L^2(\Omega_v)$ onto $\mathcal{V}$.

Proof. The proof is identical to the proof for P_X. Indeed, orthonormality gives

$$\begin{aligned} P_V^2 g &= \sum_{i=1}^{r}\sum_{j=1}^{r} V_i \langle V_i, V_j\rangle_{L^2(\Omega_v)} \langle V_j, g\rangle_{L^2(\Omega_v)} \\ &= \sum_{j=1}^{r} V_j \langle V_j, g\rangle_{L^2(\Omega_v)} \\ &= P_V g. \end{aligned} \tag{11.104}$$

Furthermore, if $h \in \mathcal{V}$, then the orthonormality of $\{V_j\}_{j=1}^{r}$ implies

$$\langle g - P_V g, h\rangle_{L^2(\Omega_v)} = 0.$$

Therefore, P_V is the orthogonal projection onto $\mathcal{V}$. ■

When these projectors act on a phase-space function, they must be lifted to the product space. We use the same symbols for the lifted operators and define

$$(P_X g)(x, v) = \sum_{i=1}^{r} X_i(x) \int_{\Omega_x} X_i(x') g(x', v) \mathrm{d}x', \tag{11.105}$$

$$(P_V g)(x,v) = \sum_{j=1}^{r} \left(\int_{\Omega_v} g(x,v') V_j(v') \mathrm{d}v' \right) V_j(v). \tag{11.106}$$

Thus P_X projects only the x-dependence and P_V projects only the v-dependence. Fubini's theorem and the orthonormality of the bases show that

$$P_X P_V = P_V P_X.$$

The continuous tangent-space projector is therefore

$$P_{T_{f_r}\mathcal{M}_r} = P_X + P_V - P_X P_V, \tag{11.107}$$

which is the function-space analogue of Equation (11.31).

Proposition 11.36 (Continuous equations of motion)**.** *Let $G = \mathcal{G}(f_r, t)$ and impose*

$$\langle X_i, \partial_t X_j \rangle_x = 0, \qquad \langle V_i, \partial_t V_j \rangle_v = 0.$$

Define

$$G_{V,j}(x) = \int_{\Omega_v} G(x,v) V_j(v) \mathrm{d}v, \qquad G_{X,i}(v) = \int_{\Omega_x} X_i(x) G(x,v) \mathrm{d}x.$$

Then the Galerkin condition is equivalent to

$$\dot{S}_{ij} = \langle X_i, G_{V,j} \rangle_x, \tag{11.108}$$

$$\sum_{i=1}^{r} \partial_t X_i \, S_{ij} = (I - P_X) G_{V,j}, \tag{11.109}$$

$$\sum_{j=1}^{r} S_{ij} \, \partial_t V_j = (I - P_V) G_{X,i}. \tag{11.110}$$

Because S is invertible, these equations uniquely determine the gauge-fixed factor derivatives.

Proof. Differentiate the separated representation of f_r. Testing the Galerkin condition first with $\delta f = \sum_{i,j} X_i H_{ij} V_j$ for an arbitrary matrix H gives Equation (11.108). Next, vary only the spatial factors, with each $\delta X_i \in \mathcal{X}^{\perp}$. The core and velocity-derivative terms vanish by the gauge conditions, leaving

$$\left\langle \sum_i \partial_t X_i S_{ij} - (I - P_X) G_{V,j}, \delta X \right\rangle_x = 0 \quad \text{for every } \delta X \in \mathcal{X}^{\perp}.$$

The expression in the first slot also belongs to $\mathcal{X}^\perp$, so it vanishes. This proves Equation (11.109). Testing with velocity variations in $\mathcal{V}^\perp$ proves Equation (11.110) in the same way. ■

The solution of Equation (11.97) can now be written as

$$\begin{aligned}\partial_t f_r &= P_{T_{f_r}\mathcal{M}_r}\mathcal{G}(f_r, t) \\ &= P_V\mathcal{G}(f_r, t) - P_X P_V\mathcal{G}(f_r, t) + P_X\mathcal{G}(f_r, t). \end{aligned} \tag{11.111}$$

The projector-splitting, BUG, and adaptive BUG integrators from Definitions 11.28 to 11.30 can be applied to this projected evolution while retaining a factorized representation.

To relate this continuous projection to a matrix discretization, let w_i^x and w_j^v be the spatial and velocity quadrature weights. The discrete approximation to the L^2 inner product is

$$\langle A, B\rangle_h = \sum_{i=1}^{N_x}\sum_{j=1}^{N_v} w_i^x w_j^v A_{ij}B_{ij} = \left\langle W_x^{1/2}AW_v^{1/2}, W_x^{1/2}BW_v^{1/2}\right\rangle_{\mathsf{F}}, \tag{11.112}$$

where $W_x = \mathrm{diag}(w_i^x)$ and $W_v = \mathrm{diag}(w_j^v)$. Thus a direct discretization uses weighted orthogonality,

$$X^\top W_x X = I_r, \qquad V^\top W_v V = I_r,$$

or, equivalently, applies ordinary Frobenius geometry to the scaled matrix $W_x^{1/2}AW_v^{1/2}$. On the uniform midpoint grids used below, both weight matrices are scalar multiples of the identity, so ordinary Frobenius projection gives the same tangent direction. This is the precise sense in which the discretize-then-project method represents the continuous L^2 projection.

11.1.6 Computational Benefits and Limitations

The benefits of DLRA are most pronounced for high-dimensional matrix ODEs. If $m, n \in \mathbb{N}$ are both very large and $Y(t) \in \mathbb{R}^{m\times n}$ is approximated by $Y_r(t) = U(t)S(t)V(t)^\top$, then the storage cost is reduced from $\mathcal{O}(mn)$ to $\mathcal{O}(rm + r^2 + rn)$. This reduction leads to a corresponding computational advantage when two conditions are true. First, the solution must remain accurately approximable for moderate rank r. Second, the quantities required by the low-rank integrator must be computable from U, S, V without forming the full approximation Y_r or a full-dimensional intermediate quantity. As was discussed in [38], operators possessing

a separable structure are especially suitable. For example, a discretized operator of the form

$$\mathcal{R}(Y,t) = \sum_{\ell=1}^{N} A_\ell(t) Y B_\ell(t)^\top \tag{11.113}$$

for $A_\ell(t) \in \mathbb{R}^{m\times m}$, $B_\ell(t) \in \mathbb{R}^{n\times n}$, and N small enough can be evaluated directly on low-rank factors:

$$\mathcal{R}(USV^\top, t) = \sum_{\ell=1}^{N} (A_\ell(t)U) S (B_\ell(t)V)^\top. \tag{11.114}$$

In the context of PDEs, many linear transport operators possess such a structure upon discretization, although this depends on the PDE itself, boundary conditions, choice of discretization, and other factors. Certain bilinear operators can still be expressed by a manageable number of separable terms. By contrast, general nonlinear entrywise operations, nonlinear reconstructions, and nonseparable terms may require storage on the order of $\mathcal{O}(mn)$, which can effectively eliminate the benefits of DLRA. In these cases, additional techniques such as interpolation, collocation, hyper-reduction, or problem-specific techniques are needed to overcome these obstacles.

11.2 Projector-Splitting DLRA for the $1x1v$ BGK Equation

11.2.1 The $1x1v$ BGK Model and Its Moments

In this section, we develop a dynamical low-rank approximation for the $1x1v$ Boltzmann–BGK equation. Its collision term uses the local Maxwellian M_f, defined from the moments below:

$$\partial_t f(x,v,t) + v\partial_x f(x,v,t) = \frac{1}{\varepsilon}(M_f(x,v,t) - f(x,v,t)), \tag{11.115}$$

Here $x \in \Omega_x \subseteq \mathbb{R}$, $v \in \mathbb{R}$, and $\varepsilon > 0$ is the Knudsen number. This nondimensional $1x1v$ model uses unit particle mass, so number and mass density coincide numerically; we use n to match the report-wide notation. The continuous BGK model is posed on the full velocity line. A bounded velocity interval is introduced only when the equation is discretized below.

The moments are velocity-weighted integrals of f that represent macroscopic properties of the gas. We first define the number density, bulk velocity, and temperature.

$$n(x,t) = \int_{\mathbb{R}} f(x,v,t)\mathrm{d}v, \tag{11.116}$$

$$u(x,t) = \frac{1}{n(x,t)} \int_{\mathbb{R}} v f(x,v,t) \mathrm{d}v, \tag{11.117}$$

$$\theta(x,t) = \frac{1}{n(x,t)} \int_{\mathbb{R}} (v - u(x,t))^2 f(x,v,t) \, \mathrm{d}v. \tag{11.118}$$

These moments determine the local Maxwellian appearing in Equation (11.115), which represents the local equilibrium distribution:

$$M_f(x,v,t) = \frac{n(x,t)}{\sqrt{2\pi\theta(x,t)}} \exp\left(-\frac{(v-u(x,t))^2}{2\theta(x,t)}\right). \tag{11.119}$$

By construction on $\mathbb{R}$,

$$\int_{\mathbb{R}} (M_f - f)\phi(v) \mathrm{d}v = 0, \qquad \phi(v) \in \mathrm{span}\{1, v, v^2/2\}. \tag{11.120}$$

The same identity is generally false after velocity truncation and quadrature if one merely samples this continuous Maxwellian.

The following is an adaptation of the discretization in Chapter 4 of [35]. Choose $L > 0$ and $V_{\max} > 0$, set $\Omega_x = [-L, L]$, and truncate velocity to $\Omega_v = [-V_{\max}, V_{\max}]$. Divide these intervals into N_x and N_v uniform cells, respectively, with

$$h_x = \frac{2L}{N_x} \text{ and } h_v = \frac{2V_{\max}}{N_v}. \tag{11.121}$$

For $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_v$, define the cell centers and interfaces by

$$x_i = -L + \left(i - \frac{1}{2}\right) h_x, \quad x_{i\pm\frac{1}{2}} = x_i \pm \frac{h_x}{2}, \qquad v_j = -V_{\max} + \left(j - \frac{1}{2}\right) h_v, \quad v_{j\pm\frac{1}{2}} = v_j \pm \frac{h_v}{2}.$$

The phase-space cells are

$$C_{ij} = C_i^x \times C_j^v = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [v_{j-\frac{1}{2}}, v_{j+\frac{1}{2}}], \qquad |C_{ij}| = h_x h_v.$$

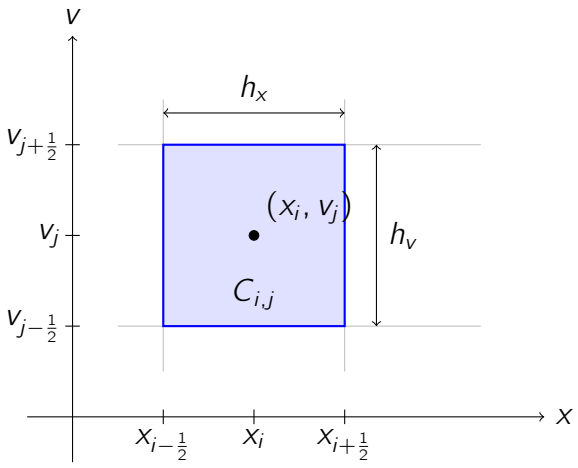


Figure 11.1: A finite-volume cell $C_{i,j}$ in truncated phase space $\Omega_x \times \Omega_v$.

Next, we define the average of the solution f on the cell $C_{i,j}$ by

$$\overline{f}_{i,j}(t) = \frac{1}{|C_{i,j}|}\int_{C_{i,j}} f(x,v,t)\mathrm{d}x\mathrm{d}v = \frac{1}{h_x h_v}\int_{C_{i,j}} f(x,v,t)\mathrm{d}x\mathrm{d}v. \tag{11.122}$$

Our goal is to approximate the averages at each cell. Note that $|C_{i,j}| = h_x h_v$. Computing the time derivative of both sides yields

$$\begin{aligned}\frac{d}{dt}\overline{f}_{i,j}(t) &= \frac{1}{h_x h_v}\int_{C_{i,j}} \partial_t f(x,v,t)\,\mathrm{d}x\mathrm{d}v \\ &= -\frac{1}{h_x h_v}\int_{C_{i,j}} \partial_x(vf(x,v,t))\,\mathrm{d}x\mathrm{d}v + \frac{1}{\varepsilon h_x h_v}\int_{C_{i,j}} (M_f - f)(x,v,t)\,\mathrm{d}x\mathrm{d}v. \end{aligned}\tag{11.123}$$

By the fundamental theorem of calculus, we have that

$$\begin{aligned}\int_{C_{i,j}} \partial_x(vf)\,\mathrm{d}x\mathrm{d}v &= \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \partial_x(vf)\,\mathrm{d}x\mathrm{d}v \\ &= \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} v\left[f(x_{i+\frac{1}{2}},v,t) - f(x_{i-\frac{1}{2}},v,t)\right]\mathrm{d}v. \end{aligned}\tag{11.124}$$

Define the physical transport flux by $q(f;v) = vf$. Then Equation (11.123) becomes

$$\begin{aligned}\frac{d}{dt}\overline{f}_{i,j}(t) = -\frac{1}{h_x h_v}\int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} \big(q(f(x_{i+\frac{1}{2}},v,t);v) - q(f(x_{i-\frac{1}{2}},v,t);v)\big)\mathrm{d}v \\ + \frac{1}{\varepsilon}\big(\overline{M}_{f,i,j}(t) - \overline{f}_{i,j}(t)\big). \end{aligned}\tag{11.125}$$

Here $\overline{M}_{f,i,j}(t)$ denotes the exact cell average of $M_f = M[f]$ on $C_{i,j}$. This identity is exact. To obtain a numerical method, we approximate the velocity integral by the midpoint rule. Set

$$q_{i\pm\frac{1}{2}}(v,t) = q(f(x_{i\pm\frac{1}{2}},v,t);v).$$

$$\int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} \left[q_{i+\frac{1}{2}}(v,t) - q_{i-\frac{1}{2}}(v,t)\right] \mathrm{d}v \approx h_v\left[q_{i+\frac{1}{2}}(v_j,t) - q_{i-\frac{1}{2}}(v_j,t)\right]. \tag{11.126}$$

Hence,

$$\frac{1}{h_x h_v}\int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} \left[q_{i+\frac{1}{2}}(v,t) - q_{i-\frac{1}{2}}(v,t)\right] \mathrm{d}v \approx \frac{q_{i+\frac{1}{2}}(v_j,t) - q_{i-\frac{1}{2}}(v_j,t)}{h_x}. \tag{11.127}$$

The physical flux depends on interface traces of f, while the finite-volume method stores only the cell averages $\overline{f}_{i,j}(t)$. We therefore reconstruct the traces from the piecewise-linear function $p_{i,j}(x,v,t) = \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x - x_i)$ on C^x_i.

$$\begin{aligned} f(x_{i+\frac{1}{2}},v,t) &\approx p_{i,j}(x_{i+\frac{1}{2}},v,t) \\ &= \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x_{i+\frac{1}{2}} - x_i) \\ f(x_{i-\frac{1}{2}},v,t) &\approx p_{i,j}(x_{i-\frac{1}{2}},v,t) \\ &= \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x_{i-\frac{1}{2}} - x_i). \end{aligned} \tag{11.128}$$

where the slope $\sigma^x_{i,j}$ is defined by

$$\sigma^x_{i,j} = \mathrm{minmod}\left(\left(\frac{\overline{f}_{i+1,j} - \overline{f}_{i-1,j}}{2h_x}\right), \beta_{\mathrm{lim}}\left(\frac{\overline{f}_{i,j} - \overline{f}_{i-1,j}}{h_x}\right), \beta_{\mathrm{lim}}\left(\frac{\overline{f}_{i+1,j} - \overline{f}_{i,j}}{h_x}\right)\right), \tag{11.129}$$

where $\beta_{\mathrm{lim}} \in [1,2]$ is the limiter parameter and the minmod function is defined by

$$\mathrm{minmod}(a,b,c) = \begin{cases} \min(a,b,c), & a,b,c \geq 0, \\ \max(a,b,c), & a,b,c \leq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11.130}$$

Typically, we set $\beta_{\mathrm{lim}} = 2$ because this gives more accurate results in the Sod problem. In general, adjacent reconstructed traces are unequal:

$$p_{i,j}(x_{i+\frac{1}{2}},v_j,t) \neq p_{i+1,j}(x_{i+\frac{1}{2}},v_j,t).$$

We define

$$f^-_{i+\frac{1}{2},j}(t) = p_{i,j}(x_{i+\frac{1}{2}}, v_j, t)$$
$$f^+_{i+\frac{1}{2},j}(t) = p_{i+1,j}(x_{i+\frac{1}{2}}, v_j, t). \tag{11.131}$$

The physical interface flux is replaced by a numerical flux:

$$\begin{aligned} q(f(x_{i+\frac{1}{2}}, v_j, t); v_j) &\approx \widehat{q}\big(f^-_{i+\frac{1}{2},j}(t), f^+_{i+\frac{1}{2},j}(t); v_j\big) \\ &= \widehat{q}_{i+\frac{1}{2},j}(t). \end{aligned} \tag{11.132}$$

For $q(f; v) = vf$, the upwind choice is

$$\begin{aligned} q(f(x_{i+\frac{1}{2}}, v_j, t); v_j) &\approx \widehat{q}(f^-_{i+\frac{1}{2},j}(t), f^+_{i+\frac{1}{2},j}(t); v_j) \\ &:= \begin{cases} v_j f^-_{i+\frac{1}{2},j}(t) \text{ if } v_j \geq 0 \\ v_j f^+_{i+\frac{1}{2},j}(t) \text{ if } v_j < 0. \end{cases} \end{aligned} \tag{11.133}$$

We denote this upwind numerical flux by $\widehat{q}_{i+\frac{1}{2},j}(t)$. Lastly, the exact cell-average Maxwellian $\overline{M}_{f,i,j}(t)$ is replaced by the sampled approximation $\widehat{M}_{f,i,j}(t)$, which is defined as follows. First, the discrete moments are defined by

$$n_i(t) = \sum_{j=1}^{N_v} \overline{f}_{i,j}(t) h_v, \tag{11.134}$$

$$u_i(t) = \frac{1}{n_i(t)} \sum_{j=1}^{N_v} v_j \overline{f}_{i,j}(t) h_v, \tag{11.135}$$

and

$$\theta_i(t) = \frac{1}{n_i(t)} \sum_{j=1}^{N_v} (v_j - u_i(t))^2 \overline{f}_{i,j}(t) h_v. \tag{11.136}$$

The sampled Maxwellian is therefore defined by

$$\widehat{M}_{f,i,j}(t) = \frac{n_i(t)}{\sqrt{2\pi\theta_i(t)}} \exp\Big(-\frac{(v_j - u_i(t))^2}{2\theta_i(t)}\Big). \tag{11.137}$$

This is the continuous Maxwellian evaluated at the velocity nodes; it is not, in general, a discrete moment-matching Maxwellian. Therefore

$$h_v \sum_{j=1}^{N_v} (\widehat{M}_{f,i,j} - \overline{f}_{i,j})\phi(v_j) \neq 0$$

for at least one $\phi \in \{1, v, v^2/2\}$. One may instead construct a discrete Maxwellian by solving a nonlinear moment-matching problem. The computations reported here use the sampled Maxwellian and treat the mismatch as velocity-truncation and quadrature error. The sampled Maxwellian $\widehat{M}_{f,i,j}(t)$ then replaces the exact cell average $\overline{M}_{f,i,j}(t)$, giving the numerical scheme

$$\frac{d}{dt}\overline{f}_{i,j}(t) = -\frac{1}{h_x}(\widehat{q}_{i+\frac{1}{2},j}(t) - \widehat{q}_{i-\frac{1}{2},j}(t)) + \frac{1}{\varepsilon}(\widehat{M}_{f,i,j}(t) - \overline{f}_{i,j}(t)) \tag{11.138}$$

subject to an initial condition.

We advance this spatial discretization with an implicit-explicit (IMEX) scheme, treating transport explicitly and collision implicitly. We first write the BGK equation as

$$\frac{\partial}{\partial t} f(x, v, t) = \mathcal{T}[f](x, v, t) + Q[f](x, v, t), \tag{11.139}$$

where

$$\mathcal{T}[f](x, v, t) = -v\frac{\partial}{\partial x} f(x, v, t), \tag{11.140}$$

and

$$Q[f](x, v, t) = \frac{1}{\varepsilon}(M_f(x, v, t) - f(x, v, t)). \tag{11.141}$$

As in Remark 6.1, $Q[f]$ is shorthand for the unary BGK collision approximation $Q^{\mathrm{BGK}}[f]$; it does not denote the bilinear Boltzmann operator $Q[f, f]$.

Let N_{st} be the number of IMEX stages. With k denoting the current stage and ℓ a previous stage, the complete stage equation is

$$f^{(k)} = f^n + \Delta t \sum_{\ell<k} \widetilde{a}_{k\ell}\mathcal{T}[f^{(\ell)}] + \Delta t \sum_{\ell\leq k} a_{k\ell} Q[f^{(\ell)}], \qquad k = 1, \ldots, N_{\mathrm{st}}. \tag{11.142}$$

Collect all known terms into

$$g^{(k)} = f^n + \Delta t \sum_{\ell<k} \widetilde{a}_{k\ell} \mathcal{T}[f^{(\ell)}] + \frac{\Delta t}{\varepsilon} \sum_{\ell<k} a_{k\ell} \big(M[f^{(\ell)}] - f^{(\ell)}\big). \tag{11.143}$$

Then

$$f^{(k)} = g^{(k)} + \lambda_k \big(M[f^{(k)}] - f^{(k)}\big), \qquad \lambda_k = \frac{\Delta t\, a_{kk}}{\varepsilon}. \tag{11.144}$$

If the collision operator exactly matches the three collision-invariant moments, taking moments of Equation (11.144) gives

$$U(f^{(k)}) = U(g^{(k)}), \qquad M[f^{(k)}] = M[g^{(k)}],$$

and the implicit stage has the explicit representation

$$f^{(k)} = \frac{g^{(k)} + \lambda_k M[g^{(k)}]}{1 + \lambda_k}. \tag{11.145}$$

This derivation is exact for the continuous full-velocity BGK operator and for a discrete Maxwellian constructed to match the discrete moments. It is only approximate for the sampled Maxwellian used here, because truncation and midpoint quadrature introduce the mismatch described above.

After all stages, the IMEX update is

$$f^{n+1} = f^n + \Delta t \sum_{k=1}^{N_{\mathrm{st}}} \widetilde{b}_k \mathcal{T}[f^{(k)}] + \Delta t \sum_{k=1}^{N_{\mathrm{st}}} b_k Q[f^{(k)}]. \tag{11.146}$$

Here $\widetilde{A} = (\widetilde{a}_{k\ell})$ is strictly lower triangular and $A = (a_{k\ell})$ is lower triangular. We reserve r for low rank and ν for the ES–BGK parameter used elsewhere in the report.

Following [35], we replace the implicit and explicit terms by their discrete counterparts. Let $f^n_{i,j}$ approximate the cell average of the distribution over $C_{i,j}$ at time $t_n = t_{n-1} + \Delta t$, and let $f^{(\ell)}_{i,j}$ denote its ℓth IMEX Runge–Kutta stage. Define

$$\begin{aligned} \mathcal{T}^\star_{i,j} &= \frac{\widehat{q}^\star_{i-\frac{1}{2},j} - \widehat{q}^\star_{i+\frac{1}{2},j}}{h_x} \\ Q^\star_{i,j} &= \frac{1}{\varepsilon}(M^\star_{f,i,j} - f^\star_{i,j}), \end{aligned} \tag{11.147}$$

where $\widehat{q}_{i\pm\frac{1}{2},j}$ denotes the numerical flux computed from the cell averages $f^\star_{i,j}$. Here, $\star$ stands for any specified stage or time level. The discrete moments and Maxwellian are computed

via

$$
\begin{aligned}
n_i^\star &= h_v \sum_{j=1}^{N_v} f_{i,j}^\star, \\
u_i^\star &= \frac{h_v}{n_i^\star} \sum_{j=1}^{N_v} v_j f_{i,j}^\star, \\
\theta_i^\star &= \frac{h_v}{n_i^\star} \sum_{j=1}^{N_v} |v_j|^2 f_{i,j}^\star - |u_i^\star|^2,
\end{aligned} \tag{11.148}
$$

and the Maxwellian by

$$
M_{f,i,j}^\star = \frac{n_i^\star}{(2\pi\theta_i^\star)^{1/2}} \exp\left(\frac{-(v_j - u_i^\star)^2}{2\theta_i^\star}\right). \tag{11.149}
$$

Then, the algorithm is run by computing the terms

$$
\begin{aligned}
g_{i,j}^{(k)} &= f_{i,j}^n + \Delta t \sum_{\ell=1}^{k-1} \widetilde{a}_{k\ell} \mathcal{T}_{i,j}^{(\ell)} + \Delta t \sum_{\ell=1}^{k-1} a_{k\ell} Q_{i,j}^{(\ell)}, \\
f_{i,j}^{(k)} &\approx \frac{g_{i,j}^{(k)} + \lambda_k \widehat{M}[g^{(k)}]_{i,j}}{1 + \lambda_k}, \qquad \lambda_k = \frac{\Delta t \, a_{kk}}{\varepsilon}, \\
f^{n+1} &= f^n + \Delta t \sum_{k=1}^{N_{\mathrm{st}}} \widetilde{b}_k \mathcal{T}[f^{(k)}] + \Delta t \sum_{k=1}^{N_{\mathrm{st}}} b_k Q[f^{(k)}].
\end{aligned} \tag{11.150}
$$

Here $\widehat{M}[g^{(k)}]$ is the continuous Maxwellian sampled at the velocity nodes after its parameters are computed from the midpoint moments of $g^{(k)}$. The approximation sign is intentional: without an exactly moment-matching discrete Maxwellian, this is not the exact solution of the discrete implicit stage. The experiments track the resulting mass, momentum, and energy drift.

The implementation uses constant-extrapolation, or zero-gradient, boundary conditions. The ghost-cell values satisfy

$$
\begin{aligned}
\overline{f}_{0,j} &= \overline{f}_{1,j} \\
\overline{f}_{N_x+1,j} &= \overline{f}_{N_x,j}, \qquad j = 1, \ldots, N_v.
\end{aligned} \tag{11.151}
$$

11.2.2 Finite-Volume Matrix Formulation and DLRA Update

The fixed-rank manifold, tangent-space projection, and projector-splitting integrator were developed above. For the kinetic discretization, velocity is the row index and space is the column index:

$$[\mathbf{f}(t)]_{j,i} = \overline{f}_{i,j}(t), \qquad \mathbf{f}(t) \in \mathbb{R}^{N_v \times N_x}.$$

We reserve $\mathcal{R}_h$ for the semidiscrete kinetic right-hand side.

We next derive the slope-limiter reconstruction as a matrix ODE. The physical transport flux is $q(f; v) = vf$, and the slope-limited upwind flux at a spatial interface is

$$\widehat{q}_{i+\frac{1}{2},j}(t) = \begin{cases} v_j f^-_{i+\frac{1}{2},j}(t) \text{ if } v_j \geq 0 \\ v_j f^+_{i+\frac{1}{2},j}(t) \text{ if } v_j < 0, \end{cases} \tag{11.152}$$

and

$$\widehat{q}_{i-\frac{1}{2},j}(t) = \begin{cases} v_j f^-_{i-\frac{1}{2},j}(t) \text{ if } v_j \geq 0 \\ v_j f^+_{i-\frac{1}{2},j}(t) \text{ if } v_j < 0. \end{cases} \tag{11.153}$$

Both of these upwind approximations can be rewritten as

$$\widehat{q}_{i+\frac{1}{2},j}(t) = v_j^+ f^-_{i+\frac{1}{2},j}(t) - v_j^- f^+_{i+\frac{1}{2},j}(t). \tag{11.154}$$

and

$$\widehat{q}_{i-\frac{1}{2},j}(t) = v_j^+ f^-_{i-\frac{1}{2},j}(t) - v_j^- f^+_{i-\frac{1}{2},j}(t). \tag{11.155}$$

Here, $v_j^+ = \max(v_j, 0)$ and $v_j^- = \max(-v_j, 0)$. The approximations $f^{\pm}_{i+\frac{1}{2},j}(t)$ and $f^{\pm}_{i-\frac{1}{2},j}(t)$ are defined in Equation (11.128). Specifically,

$$\begin{aligned} p_{i,j}(x, v, t) &= \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x - x_i) \\ p_{i-1,j}(x, v, t) &= \overline{f}_{i-1,j}(t) + \sigma^x_{i-1,j}(x - x_{i-1}) \\ p_{i+1,j}(x, v, t) &= \overline{f}_{i+1,j}(t) + \sigma^x_{i+1,j}(x - x_{i+1}). \end{aligned} \tag{11.156a}$$

The terms that appear in the upwind formulation are given below and simplified.

$$
\begin{aligned}
f^{-}_{i+\frac{1}{2},j}(t) &\approx p_{i,j}(x_{i+\frac{1}{2}}, v, t) \\
&= \overline{f}_{i,j}(t) + \frac{h_x}{2}\sigma^x_{i,j} \\
f^{+}_{i-\frac{1}{2},j}(t) &\approx p_{i,j}(x_{i-\frac{1}{2}}, v, t) \\
&= \overline{f}_{i,j}(t) - \frac{h_x}{2}\sigma^x_{i,j} \\
f^{+}_{i+\frac{1}{2},j}(t) &\approx p_{i+1,j}(x_{i+\frac{1}{2}}, v, t) \\
&= \overline{f}_{i+1,j}(t) - \frac{h_x}{2}\sigma^x_{i+1,j} \\
f^{-}_{i-\frac{1}{2},j}(t) &\approx p_{i-1,j}(x_{i-\frac{1}{2}}, v, t) \\
&= \overline{f}_{i-1,j}(t) + \frac{h_x}{2}\sigma^x_{i-1,j}. \qquad (11.157a)
\end{aligned}
$$

These equations yield a matrix formulation for the numerical upwind flux. Define $\Lambda[\mathbf{f}] \in \mathbb{R}^{N_v \times N_x}$ by $[\Lambda[\mathbf{f}]]_{j,i} = \sigma^x_{i,j}$. Define $B^+ = \mathrm{diag}(v_j^+)$ and $B^- = \mathrm{diag}(v_j^-)$ in $\mathbb{R}^{N_v \times N_v}$. For any state matrix Y, $M[Y]$ is the sampled continuous Maxwellian computed from the discrete moments of Y.

For the constant-extrapolation boundary condition, let $L_{-1}, L_{+1} \in \mathbb{R}^{N_x \times N_x}$ be the shift matrices defined by

$$
(YL_{-1})_{j,i} = Y_{j,\max\{i-1,1\}}, \qquad (YL_{+1})_{j,i} = Y_{j,\min\{i+1,N_x\}}.
$$

Equivalently, these matrices have the form

$$
L_{-1} = \begin{pmatrix}
1 & 1 & 0 & 0 & 0 & \dots \\
0 & 0 & 1 & 0 & 0 & \dots \\
0 & 0 & 0 & 1 & 0 & \dots \\
0 & 0 & 0 & 0 & 1 & \dots \\
0 & 0 & 0 & 0 & 0 & \dots \\
. & . & . & . & . & \dots
\end{pmatrix}, \qquad
L_{+1} = \begin{pmatrix}
\dots & . & . & . & . & \\
\dots & 0 & 0 & 0 & 0 & 0 \\
\dots & 1 & 0 & 0 & 0 & 0 \\
\dots & 0 & 1 & 0 & 0 & 0 \\
\dots & 0 & 0 & 1 & 0 & 0 \\
\dots & 0 & 0 & 0 & 1 & 1
\end{pmatrix}.
$$

Then the transport term $[\widehat{q}_{i-1/2,j}(t) - \widehat{q}_{i+1/2,j}(t)]/h_x$ can be written in matrix form as

$$
\frac{1}{h_x}\left[B^+\mathbf{f}(t)L_{-1} - (B^- + B^+)\mathbf{f}(t) + B^-\mathbf{f}(t)L_{+1}\right]
$$

$$+\frac{1}{2}\left[B^{+}\Lambda[\mathbf{f}(t)]L_{-1}+(B^{-}-B^{+})\Lambda[\mathbf{f}(t)]-B^{-}\Lambda[\mathbf{f}(t)]L_{+1}\right]. \tag{11.158}$$

To simplify the notation, we introduce the following definitions.

Definition 11.37 ($T_{\text{lin}}(Y)$ and $T_{\text{slope}}(Y)$)**.** *Let $Y \in \mathbb{R}^{N_v \times N_x}$ and matrices $B^{\pm}$, $\Lambda[Y]$, and $L_{\pm}$ be as defined in Equation* (11.158)*. The* ***linear transport operator*** *and* ***slope-correction operator*** *are*

$$T_{\text{lin}}(Y)=\frac{1}{h_x}[B^{+}YL_{-1}-(B^{-}+B^{+})Y+B^{-}YL_{+1}], \tag{11.159}$$

and

$$T_{\text{slope}}(Y)=\frac{1}{2}[B^{+}\Lambda[Y]L_{-1}+(B^{-}-B^{+})\Lambda[Y]-B^{-}\Lambda[Y]L_{+1}]. \tag{11.160}$$

Then, the full matrix ODE can be written as

$$\dot{\mathbf{f}}(t)=\frac{1}{\varepsilon}M[\mathbf{f}(t)]-\frac{1}{\varepsilon}\mathbf{f}(t)+T_{\text{lin}}(\mathbf{f}(t))+T_{\text{slope}}(\mathbf{f}(t)). \tag{11.161}$$

More compactly,

$$\dot{\mathbf{f}}(t)=\mathcal{R}_h(\mathbf{f}(t),t), \tag{11.162}$$

where

$$\mathcal{R}_h(Y,t)=\frac{1}{\varepsilon}M[Y]-\frac{1}{\varepsilon}Y+T_{\text{lin}}(Y)+T_{\text{slope}}(Y). \tag{11.163}$$

The storage advantage of DLRA is realized only if $\mathcal{R}_h(USV^{\top},t)$ can be evaluated without forming the full product $USV^{\top}$ [38]. The linear transport terms have this property; the sampled Maxwellian and nonlinear slope limiter generally do not.

$$T_{\text{slope}}(USV^{\top})=\frac{1}{2}[B^{+}\Lambda[USV^{\top}]L_{-1}+(B^{-}-B^{+})\Lambda[USV^{\top}]-B^{-}\Lambda[USV^{\top}]L_{+1}]. \tag{11.164}$$

The minmod map is nonlinear and acts entrywise, so $\Lambda[USV^{\top}]$ generally has rank much larger than r. In the implementation used for the experiments, the full matrix $USV^{\top}$ is reconstructed, the limiter is applied entrywise, and the resulting full limiter matrix is used in the projected factor update. This costs $\mathcal{O}(N_xN_v)$ storage and work for that part of the method; no fully low-rank evaluation of the limiter is claimed. A future hyper-reduced implementation would need to specify its approximation procedure and tolerance. If the

limiter were independently recompressed, that truncation would introduce another error and could disturb discrete moments.

Remark 11.38 (Positivity and conservation)**.** *Neither orthogonal tangent projection nor low-rank truncation automatically preserves nonnegativity of the phase density. Likewise, even if the underlying finite-volume operator is conservative, its orthogonal projection need not preserve every discrete moment unless the corresponding invariant directions lie in the relevant trial and test spaces. The sampled Maxwellian introduces an additional moment defect. For these reasons the experiments report positivity and conservation diagnostics rather than asserting these properties a priori.*

For the slope-limiter system in (11.163), we use a Forward Euler realization of the projector-splitting integrator with time step $\tau > 0$. Let $Y_n = U_n S_n V_n^\top$ be the low-rank approximation at time t_n. The update $Y_n \to Y_{n+1} = U_{n+1} S_{n+1} V_{n+1}^\top$ proceeds as follows.

K-step

Set $K_n = U_n S_n$. Then,

$$\begin{aligned} K_{n+1} = K_n + \tau\Big[&\frac{1}{\varepsilon} M[K_n V_n^\top] V_n - \frac{1}{\varepsilon} K_n \\ &+ \frac{1}{h_x}[B^+ K_n (V_n^\top L_{-1} V_n) - (B^- + B^+) K_n + B^- K_n (V_n^\top L_{+1} V_n)] \\ &+ \frac{1}{2}[B^+ \Lambda[K_n V_n^\top] L_{-1} V_n + (B^- - B^+) \Lambda[K_n V_n^\top] V_n - B^- \Lambda[K_n V_n^\top] L_{+1} V_n]\Big]. \end{aligned} \tag{11.165}$$

We then take $[U_{n+1}, \widehat{S}_{n+1}] = \mathrm{QR}(K_{n+1})$. This yields the update for U_{n+1}.

S-step

The standard projector-splitting S-step is

$$\widetilde{S}_{n+1} = \widehat{S}_{n+1} - \tau U_{n+1}^\top \mathcal{R}_h(U_{n+1} \widehat{S}_{n+1} V_n^\top, t_n) V_n. \tag{11.166}$$

Expanding the right-hand side gives

$$\begin{aligned} \widetilde{S}_{n+1} = \widehat{S}_{n+1} - \tau \Bigg[&\frac{1}{\varepsilon} U_{n+1}^\top M[U_{n+1} \widehat{S}_{n+1} V_n^\top] V_n - \frac{1}{\varepsilon} \widehat{S}_{n+1} \\ &+ \frac{1}{h_x}\Big[U_{n+1}^\top B^+ U_{n+1} \widehat{S}_{n+1} \big(V_n^\top L_{-1} V_n\big) \\ &\quad - U_{n+1}^\top (B^- + B^+) U_{n+1} \widehat{S}_{n+1} \\ &\quad + U_{n+1}^\top B^- U_{n+1} \widehat{S}_{n+1} \big(V_n^\top L_{+1} V_n\big)\Big] \end{aligned}$$

$$+\frac{1}{2}\Big[U_{n+1}^{\top}B^{+}\Lambda[U_{n+1}\widehat{S}_{n+1}V_n^{\top}]L_{-1}V_n + U_{n+1}^{\top}(B^{-}-B^{+})\Lambda[U_{n+1}\widehat{S}_{n+1}V_n^{\top}]V_n - U_{n+1}^{\top}B^{-}\Lambda[U_{n+1}\widehat{S}_{n+1}V_n^{\top}]L_{+1}V_n\Big]\Big]. \tag{11.167}$$

This produces the intermediate core matrix $\widetilde{S}_{n+1} \in \mathbb{R}^{r\times r}$, which is then used to initialize the L-step.

L-step

Set

$$L_n = V_n\widetilde{S}_{n+1}^{\top}.$$

The standard projector-splitting L-step is

$$L_{n+1} = L_n + \tau\mathcal{R}_h(U_{n+1}L_n^{\top}, t_n)^{\top}U_{n+1}. \tag{11.168}$$

Because $U_{n+1}^{\top}U_{n+1} = I_r$, we have

$$(U_{n+1}L_n^{\top})^{\top}U_{n+1} = L_nU_{n+1}^{\top}U_{n+1} = L_n.$$

Expanding the right-hand side therefore gives

$$L_{n+1} = L_n + \tau\Bigg[\frac{1}{\varepsilon}M[U_{n+1}L_n^{\top}]^{\top}U_{n+1} - \frac{1}{\varepsilon}L_n + \frac{1}{h_x}\Big[L_{-1}^{\top}L_nU_{n+1}^{\top}(B^{+})^{\top}U_{n+1} - L_nU_{n+1}^{\top}(B^{-}+B^{+})^{\top}U_{n+1} + L_{+1}^{\top}L_nU_{n+1}^{\top}(B^{-})^{\top}U_{n+1}\Big] + \frac{1}{2}\Big[L_{-1}^{\top}\Lambda[U_{n+1}L_n^{\top}]^{\top}(B^{+})^{\top}U_{n+1} + \Lambda[U_{n+1}L_n^{\top}]^{\top}(B^{-}-B^{+})^{\top}U_{n+1} - L_{+1}^{\top}\Lambda[U_{n+1}L_n^{\top}]^{\top}(B^{-})^{\top}U_{n+1}\Big]\Bigg]. \tag{11.169}$$

We then compute

$$[V_{n+1}, S_{n+1}^{\top}] = \mathrm{QR}(L_{n+1}). \tag{11.170}$$

This yields the updated spatial factor V_{n+1} and core matrix S_{n+1}, and hence the full update

$Y_{n+1} = U_{n+1}S_{n+1}V_{n+1}^{\top}$. The following subsection presents numerical experiments for a Sod shock-tube initial condition.

11.2.3 Sod Shock-Tube Experiments

For these experiments, the $1x1v$ finite-volume solver derived in Chapter 6 serves as the reference solver. The classical Sod shock-tube test begins with gases in two different thermodynamic states separated by a diaphragm. After the diaphragm is removed, the kinetic evolution is governed by the BGK equation with piecewise-Maxwellian initial data:

$$\begin{aligned} &\partial_t f(x,v,t) + v\partial_x f(x,v,t) = \frac{1}{\varepsilon}(M_f(x,v,t) - f(x,v,t)), \\ &f(x,v,0) = f_0(x,v), \end{aligned} \tag{11.171}$$

where

$$f_0(x,v) = \begin{cases} \frac{n_L}{\sqrt{2\pi\theta_L}} \exp\left(-\frac{|v-u_L|^2}{2\theta_L}\right), & x \in [-L,0],\, v \in [-V_{\max}, V_{\max}] \\ \frac{n_R}{\sqrt{2\pi\theta_R}} \exp\left(-\frac{|v-u_R|^2}{2\theta_R}\right), & x \in (0,L],\, v \in [-V_{\max}, V_{\max}]. \end{cases} \tag{11.172}$$

For our test, the left and right states are

$$(n_L, u_L, \theta_L) = (1.0, 0.0, 1.0), \qquad (n_R, u_R, \theta_R) = (0.125, 0.0, 0.8).$$

The corresponding Euler Riemann problem has an exact solution that provides a useful qualitative comparison. The DLRA experiments below use ranks 8, 16, 20, and 24, with

$$L = 0.5, \qquad V_{\max} = 10, \qquad \varepsilon \in \{10^{-3}, 10^{-4}\}, \qquad t_{\text{fin}} = 0.2.$$

The plots show the Sod moments and the phase-density error for the highest rank, $r = 24$. We choose a time step corresponding to the CFL constant $C = 0.12$, with $N_x = N_v = 64$. The reference is the full finite-volume solver on this same grid and timestep. Thus it isolates the low-rank error but is not asserted to be a grid-converged solution. We report

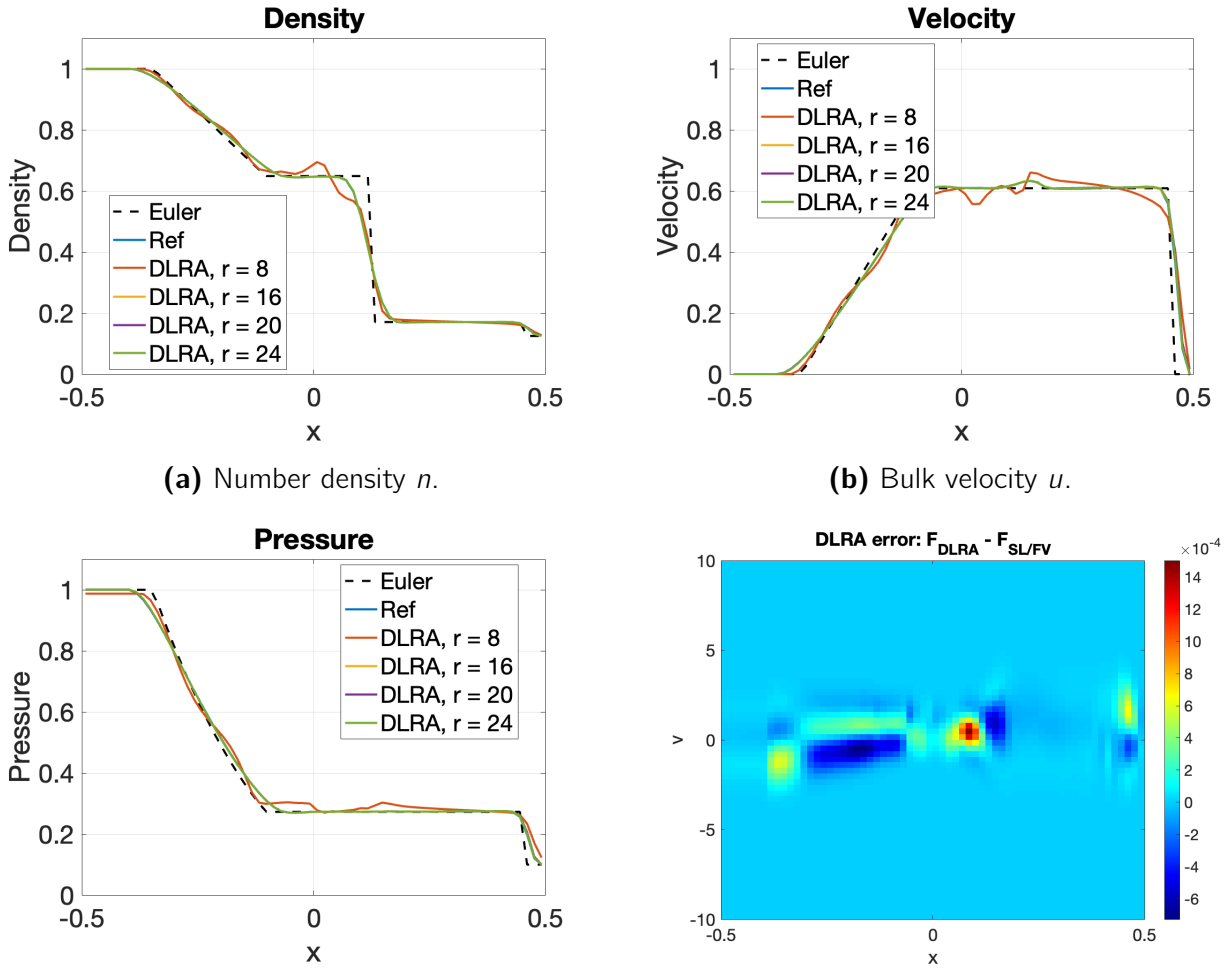
$$E_f(t) = \frac{\|Y_r(t) - Y_{\text{ref}}(t)\|_{\text{F}}}{\|Y_{\text{ref}}(t)\|_{\text{F}}}$$

and use the analogous relative discrete ℓ^2 norm in x for each macroscopic moment. For a matrix Z with singular values σ_k, the numerical rank at tolerance 10^{-10} is

$$\text{rank}_{10^{-10}}(Z) = \#\left\{k \;\middle|\; \sigma_k/\sigma_1 > 10^{-10}\right\}.$$

We apply this definition to the low-rank state Y, sampled Maxwellian $M[Y]$, and slope-limiter matrix $\Lambda[Y]$.

We begin with the moderately stiff case $\varepsilon = 10^{-3}$.



(a) Number density n.

(b) Bulk velocity u.

(c) Pressure.

(d) Distribution error for the highest prescribed rank.

Figure 11.2: Macroscopic quantities and distribution error for the one-dimensional Sod experiment with $\varepsilon = 10^{-3}$. The DLRA results are compared with the slope-limited full finite-volume reference solution.

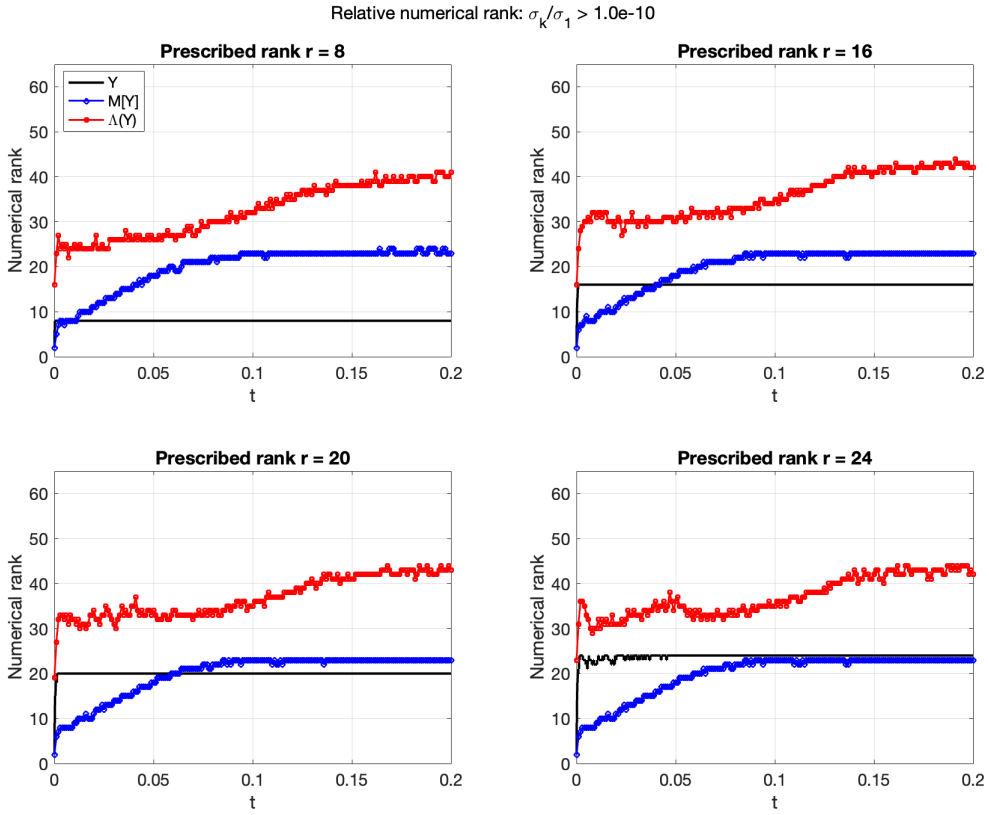
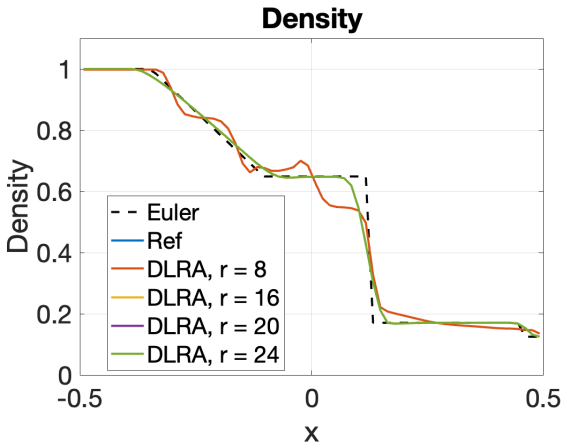
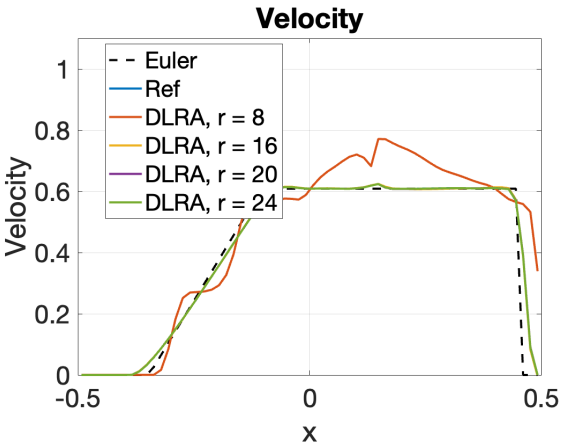


Figure 11.3: Numerical ranks of the DLRA approximation Y, the sampled Maxwellian $M[Y]$, and the slope-limiter matrix $\Lambda[Y]$ for the Sod experiment with $\varepsilon = 10^{-3}$. Numerical rank is measured using the relative threshold $\sigma_k/\sigma_1 > 10^{-10}$.

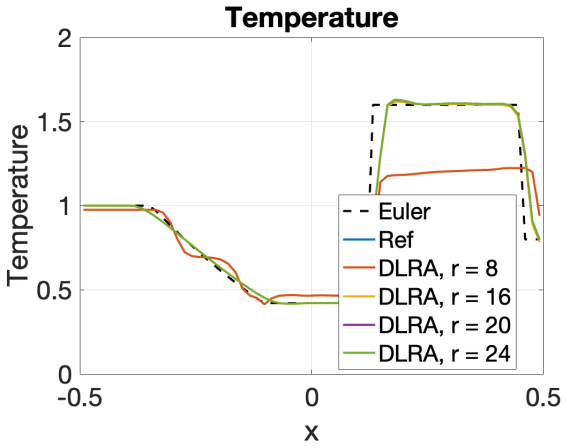
We next consider the more strongly collisional case $\varepsilon = 10^{-4}$.



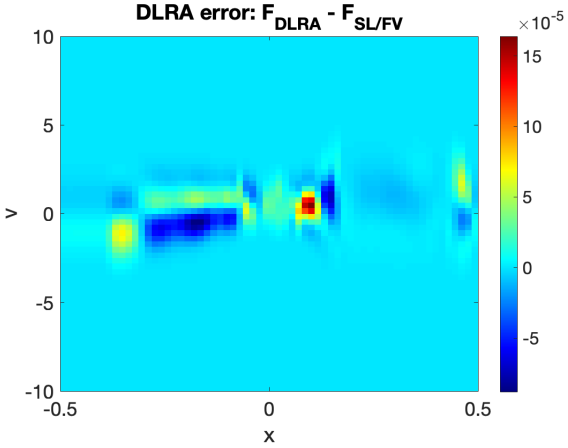
(a) Number density n.



(b) Bulk velocity u.



(c) Temperature θ.



(d) Distribution error for the highest prescribed rank.

Figure 11.4: Macroscopic quantities and distribution error for the one-dimensional Sod experiment with $\varepsilon = 10^{-4}$. The DLRA results are compared with the slope-limited full finite-volume reference solution.

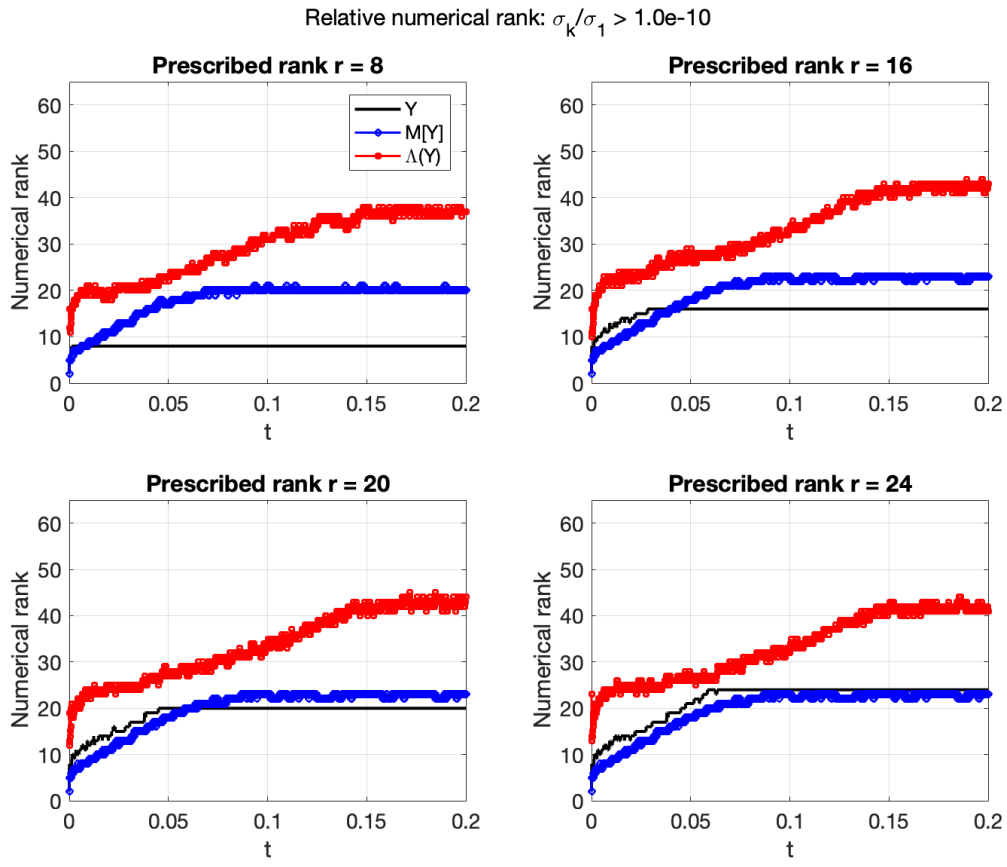


Figure 11.5: Numerical ranks of the DLRA approximation Y, the sampled Maxwellian $M[Y]$, and the slope-limiter matrix $\Lambda[Y]$ for the Sod experiment with $\varepsilon = 10^{-4}$. Numerical rank is measured using the relative threshold $\sigma_k/\sigma_1 > 10^{-10}$.

11.3 DLRA for Slope-Limited Collisionless Transport

11.3.1 Slope-Limited Finite-Volume Discretization

To isolate the effect of the minmod limiter on the DLRA approximation, we remove the BGK collision term and study the resulting $1x1v$ collisionless transport equation. We then track the numerical rank of the slope-limiter matrix in a high-resolution finite-volume discretization. Let $L > 0$ and $V_{\max} > 0$, and take $\Omega_x \times \Omega_v = [-L, L] \times [-V_{\max}, V_{\max}]$. The model is

$$\partial_t f(x, v, t) + v \partial_x f(x, v, t) = 0 \tag{11.173}$$

for $(x, v) \in \Omega_x \times \Omega_v$. We assume periodic spatial boundary conditions. Divide Ω_x into N_x cells and Ω_v into N_v cells, and set

$$h_x = \frac{2L}{N_x} \text{ and } h_v = \frac{2V_{\max}}{N_v}. \tag{11.174}$$

For $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_v$, the cell centers are $x_i = -L + (i - \frac{1}{2})h_x$ and $v_j = -V_{\max} + (j - \frac{1}{2})h_v$. Hence

$$\Omega_x \times \Omega_v = \bigcup_{i=1}^{N_x} \bigcup_{j=1}^{N_v} C_{ij} = \bigcup_{i=1}^{N_x} \bigcup_{j=1}^{N_v} [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [v_{j-\frac{1}{2}}, v_{j+\frac{1}{2}}]. \tag{11.175}$$

We approximate the solution via an average on each cell, defined by

$$\overline{f}_{i,j}(t) = \frac{1}{h_x h_v} \int_{C_{i,j}} f(x, v, t) \mathrm{d}x\mathrm{d}v = \frac{1}{h_x h_v} \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x, v, t) \mathrm{d}x\mathrm{d}v. \tag{11.176}$$

Differentiating under the integral sign gives

$$\begin{aligned}
\frac{d}{dt}\overline{f}_{i,j}(t) &= \frac{1}{h_x h_v} \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \partial_t f(x, v, t) \, \mathrm{d}x\mathrm{d}v \\
&= -\frac{1}{h_x h_v} \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} v \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \partial_x f(x, v, t) \, \mathrm{d}x\mathrm{d}v \\
&= -\frac{1}{h_x h_v} \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} v \big[f(x_{i+\frac{1}{2}}, v, t) - f(x_{i-\frac{1}{2}}, v, t) \big] \, \mathrm{d}v \\
&= \frac{1}{h_x h_v} \int_{v_{j-\frac{1}{2}}}^{v_{j+\frac{1}{2}}} v \big[f(x_{i-\frac{1}{2}}, v, t) - f(x_{i+\frac{1}{2}}, v, t) \big] \, \mathrm{d}v.
\end{aligned} \tag{11.177}$$

We introduce a midpoint rule approximation to the above integral to obtain

$$\begin{aligned}
\frac{d}{dt}\overline{f}_{i,j}(t) &\approx \frac{1}{h_x h_v} v_j \big[f(x_{i-\frac{1}{2}}, v_j, t) - f(x_{i+\frac{1}{2}}, v_j, t) \big] h_v \\
&= \frac{v_j f(x_{i-\frac{1}{2}}, v_j, t) - v_j f(x_{i+\frac{1}{2}}, v_j, t)}{h_x}.
\end{aligned} \tag{11.178}$$

Using the same physical flux $q(f; v) = vf$ as above, the previous equation becomes

$$\frac{q(f(x_{i-\frac{1}{2}}, v_j, t); v_j) - q(f(x_{i+\frac{1}{2}}, v_j, t); v_j)}{h_x}. \tag{11.179}$$

Hence, the dynamics can be written as

$$\frac{d}{dt}\overline{f}_{i,j}(t) = \frac{q(f(x_{i-\frac{1}{2}}, v_j, t); v_j) - q(f(x_{i+\frac{1}{2}}, v_j, t); v_j)}{h_x}. \tag{11.180}$$

The interface values are approximated with the same slope reconstruction and numerical flux $\widehat{q}$ used in the BGK discretization.

We now introduce the slope-limiter reconstruction. Because the finite-volume method stores only the cell averages $\overline{f}_{i,j}(t)$, we reconstruct the interface traces with $p_{i,j}(x, v, t) = \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x - x_i)$.

$$\begin{aligned} f(x_{i+\frac{1}{2}}, v, t) &\approx p_{i,j}(x_{i+\frac{1}{2}}, v, t) \\ &= \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x_{i+\frac{1}{2}} - x_i) \\ f(x_{i-\frac{1}{2}}, v, t) &\approx p_{i,j}(x_{i-\frac{1}{2}}, v, t) \\ &= \overline{f}_{i,j}(t) + \sigma^x_{i,j}(x_{i-\frac{1}{2}} - x_i). \end{aligned} \tag{11.181}$$

where the slope, $\sigma^x_{i,j}$ is defined by

$$\sigma^x_{i,j} = \text{minmod}\left(\left(\frac{\overline{f}_{i+1,j} - \overline{f}_{i-1,j}}{2h_x}\right), \beta_{\text{lim}}\left(\frac{\overline{f}_{i,j} - \overline{f}_{i-1,j}}{h_x}\right), \beta_{\text{lim}}\left(\frac{\overline{f}_{i+1,j} - \overline{f}_{i,j}}{h_x}\right)\right), \tag{11.182}$$

where $\beta_{\text{lim}} \in [1, 2]$ is the limiter parameter and the minmod function is defined by

$$\text{minmod}(a, b, c) = \begin{cases} \min(a, b, c), & a, b, c \geq 0, \\ \max(a, b, c), & a, b, c \leq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11.183}$$

Typically, we set $\beta_{\text{lim}} = 2$ because this gives more accurate results in the Sod problem. In general, adjacent reconstructed traces are unequal:

$$p_{i,j}(x_{i+\frac{1}{2}}, v_j, t) \neq p_{i+1,j}(x_{i+\frac{1}{2}}, v_j, t).$$

We define

$$\begin{aligned} f^-_{i+\frac{1}{2},j}(t) &= p_{i,j}(x_{i+\frac{1}{2}}, v_j, t) \\ f^+_{i+\frac{1}{2},j}(t) &= p_{i+1,j}(x_{i+\frac{1}{2}}, v_j, t) \\ f^-_{i-\frac{1}{2},j}(t) &= p_{i-1,j}(x_{i-\frac{1}{2}}, v_j, t) \end{aligned}$$

$$f^{+}_{i-\frac{1}{2},j}(t) = p_{i,j}(x_{i-\frac{1}{2}}, v_j, t). \tag{11.184}$$

We then take the numerical flux approximation

$$\begin{aligned} q(f(x_{i+\frac{1}{2}}, v_j, t); v_j) &\approx \widehat{q}\big(f^{-}_{i+\frac{1}{2},j}(t), f^{+}_{i+\frac{1}{2},j}(t); v_j\big) \\ &= \widehat{q}_{i+\frac{1}{2},j}(t), \\ q(f(x_{i-\frac{1}{2}}, v_j, t); v_j) &\approx \widehat{q}\big(f^{-}_{i-\frac{1}{2},j}(t), f^{+}_{i-\frac{1}{2},j}(t); v_j\big) \\ &= \widehat{q}_{i-\frac{1}{2},j}(t). \end{aligned} \tag{11.185}$$

For $q(f; v) = vf$, the upwind numerical flux is

$$\begin{aligned} q(f(x_{i+\frac{1}{2}}, v_j, t); v_j) &\approx \widehat{q}(f^{-}_{i+\frac{1}{2},j}(t), f^{+}_{i+\frac{1}{2},j}(t); v_j) \\ &= \begin{cases} v_j f^{-}_{i+\frac{1}{2},j}(t) \text{ if } v_j \geq 0 \\ v_j f^{+}_{i+\frac{1}{2},j}(t) \text{ if } v_j < 0. \end{cases} \\ q(f(x_{i-\frac{1}{2}}, v_j, t); v_j) &\approx \widehat{q}(f^{-}_{i-\frac{1}{2},j}(t), f^{+}_{i-\frac{1}{2},j}(t); v_j) \\ &= \begin{cases} v_j f^{-}_{i-\frac{1}{2},j}(t) \text{ if } v_j \geq 0 \\ v_j f^{+}_{i-\frac{1}{2},j}(t) \text{ if } v_j < 0. \end{cases} \end{aligned} \tag{11.186}$$

Both numerical fluxes can be written as

$$\begin{aligned} \widehat{q}(f^{-}_{i+\frac{1}{2},j}(t), f^{+}_{i+\frac{1}{2},j}(t); v_j) &= v_j^{+} f^{-}_{i+\frac{1}{2},j}(t) - v_j^{-} f^{+}_{i+\frac{1}{2},j}(t) \\ &= v_j^{+} p_{i,j}(x_{i+\frac{1}{2}}, v_j, t) - v_j^{-} p_{i+1,j}(x_{i+\frac{1}{2}}, v_j, t) \\ \widehat{q}(f^{-}_{i-\frac{1}{2},j}(t), f^{+}_{i-\frac{1}{2},j}(t); v_j) &= v_j^{+} f^{-}_{i-\frac{1}{2},j}(t) - v_j^{-} f^{+}_{i-\frac{1}{2},j}(t) \\ &= v_j^{+} p_{i-1,j}(x_{i-\frac{1}{2}}, v_j, t) - v_j^{-} p_{i,j}(x_{i-\frac{1}{2}}, v_j, t). \end{aligned} \tag{11.187}$$

Using the definition of $p_{i,j}(x, v, t)$ in Equation (11.181), the first flux becomes

$$\begin{aligned} v_j^{+} p_{i-1,j}(x_{i-\frac{1}{2}}, v_j, t) - v_j^{-} p_{i,j}(x_{i-\frac{1}{2}}, v_j, t) &= v_j^{+}\left(\overline{f}_{i-1,j} + \frac{h_x}{2}\sigma^{x}_{i-1,j}\right) - v_j^{-}\left(\overline{f}_{i,j} - \frac{h_x}{2}\sigma^{x}_{i,j}\right) \\ &= v_j^{+}\overline{f}_{i-1,j} + \frac{h_x}{2} v_j^{+}\sigma^{x}_{i-1,j} - v_j^{-}\overline{f}_{i,j} + \frac{h_x}{2} v_j^{-}\sigma^{x}_{i,j}. \end{aligned} \tag{11.188}$$

Similarly,

$$\begin{aligned}
&v_j^+ p_{i,j}(x_{i+\frac{1}{2}}, v_j, t) - v_j^- p_{i+1,j}(x_{i+\frac{1}{2}}, v_j, t) \\
&= v_j^+ \left(\bar{f}_{i,j}(t) + \sigma_{i,j}^x \left(x_{i+\frac{1}{2}} - x_i\right)\right) - v_j^- (\bar{f}_{i+1,j}(t) + \sigma_{i+1,j}^x \left(x_{i+\frac{1}{2}} - x_{i+1}\right)) \\
&= v_j^+ (\bar{f}_{i,j}(t) + \frac{h_x}{2}\sigma_{i,j}^x) - v_j^- (\bar{f}_{i+1,j}(t) - \frac{h_x}{2}\sigma_{i+1,j}^x) \\
&= v_j^+ \bar{f}_{i,j}(t) + \frac{h_x}{2} v_j^+ \sigma_{i,j}^x - v_j^- \bar{f}_{i+1,j}(t) + \frac{h_x}{2} v_j^- \sigma_{i+1,j}^x. \qquad (11.189)
\end{aligned}$$

Thus, the finite-volume scheme can be written as

$$\begin{aligned}
\frac{d}{dt}\overline{f}_{i,j}(t) &= \frac{\widehat{q}_{i-\frac{1}{2},j}(t) - \widehat{q}_{i+\frac{1}{2},j}(t)}{h_x} \\
&= \frac{1}{h_x}\Big[v_j^+ \overline{f}_{i-1,j} - v_j^- \overline{f}_{i,j} - v_j^+ \overline{f}_{i,j} + v_j^- \overline{f}_{i+1,j} \\
&\qquad + \frac{h_x}{2}\big(v_j^+ \sigma_{i-1,j}^x + v_j^- \sigma_{i,j}^x - v_j^+ \sigma_{i,j}^x - v_j^- \sigma_{i+1,j}^x\big)\Big]. \qquad (11.190)
\end{aligned}$$

This expression can then be condensed into the form below.

We next reformulate the semidiscrete system as a matrix ODE. We collect the finite-volume cell averages into the matrix

$$Y(t) = \left[\overline{f}_{i,j}(t)\right]_{j,i} \in \mathbb{R}^{N_v \times N_x}, \qquad (11.191)$$

where the row index j corresponds to velocity and the column index i corresponds to position. Similarly, define the limited-slope matrix

$$\Sigma(Y) = \left[\sigma_{i,j}^x\right]_{j,i} \in \mathbb{R}^{N_v \times N_x}. \qquad (11.192)$$

For periodic spatial boundary conditions, let $P_-, P_+ \in \mathbb{R}^{N_x \times N_x}$ denote the backward and forward cyclic-shift matrices, respectively, defined by

$$(YP_-)_{j,i} = Y_{j,i-1}, \qquad (YP_+)_{j,i} = Y_{j,i+1}, \qquad (11.193)$$

where the spatial indices are interpreted periodically. In particular,

$$P_+ P_- = P_- P_+ = I_{N_x}. \qquad (11.194)$$

The backward, centered, and forward slope-candidate matrices are

$$A(Y) = \frac{\beta_{\text{lim}}}{h_x} Y \left(I_{N_x} - P_- \right),$$
$$B(Y) = \frac{1}{2h_x} Y \left(P_+ - P_- \right),$$
$$C(Y) = \frac{\beta_{\text{lim}}}{h_x} Y \left(P_+ - I_{N_x} \right), \tag{11.195}$$

respectively, where $\beta_{\text{lim}} \in [1, 2]$. The limited-slope matrix is therefore

$$\Sigma(Y) = \text{minmod}\left(A(Y), B(Y), C(Y) \right), \tag{11.196}$$

where minmod is applied entrywise. Define the nonnegative velocity splits

$$v_j^+ = \max(v_j, 0), \qquad v_j^- = \max(-v_j, 0), \tag{11.197}$$

so that

$$v_j = v_j^+ - v_j^-. \tag{11.198}$$

We introduce the diagonal matrices

$$\mathsf{V}^+ = \text{diag}\left(v_1^+, \ldots, v_{N_v}^+ \right), \qquad \mathsf{V}^- = \text{diag}\left(v_1^-, \ldots, v_{N_v}^- \right). \tag{11.199}$$

At the interface $x_{i+1/2}$, the reconstructed left and right states are

$$f^-_{i+1/2,j} = Y_{j,i} + \frac{h_x}{2} \Sigma(Y)_{j,i} \tag{11.200}$$

and

$$f^+_{i+1/2,j} = Y_{j,i+1} - \frac{h_x}{2} \Sigma(Y)_{j,i+1}, \tag{11.201}$$

respectively. The corresponding upwind numerical flux is

$$\widehat{q}_{i+1/2,j} = v_j^+ f^-_{i+1/2,j} - v_j^- f^+_{i+1/2,j}. \tag{11.202}$$

Hence,

$$\widehat{q}_{i+1/2,j} = v_j^+ \left(Y_{j,i} + \frac{h_x}{2} \Sigma(Y)_{j,i} \right) - v_j^- \left(Y_{j,i+1} - \frac{h_x}{2} \Sigma(Y)_{j,i+1} \right). \tag{11.203}$$

The conservative finite-volume dynamics are

$$\dot{Y}_{j,i}(t) = \frac{\widehat{q}_{i-1/2,j}(t) - \widehat{q}_{i+1/2,j}(t)}{h_x}. \tag{11.204}$$

Expanding the two numerical fluxes in Equation (11.204) gives

$$\begin{aligned}\dot{Y}_{j,i} = \frac{1}{h_x}&\left[v_j^+ Y_{j,i-1} - \left(v_j^+ + v_j^-\right) Y_{j,i} + v_j^- Y_{j,i+1}\right] \\ &+ \frac{1}{2}\left[v_j^+ \Sigma(Y)_{j,i-1} + \left(v_j^- - v_j^+\right)\Sigma(Y)_{j,i} - v_j^- \Sigma(Y)_{j,i+1}\right].\end{aligned} \tag{11.205}$$

We now store the numerical flux through the interface $x_{i+1/2}$ in column i of the matrix

$$\Phi(Y) \in \mathbb{R}^{N_v \times N_x}, \qquad \Phi(Y)_{j,i} = \widehat{q}_{i+1/2,j}. \tag{11.206}$$

The interface-flux matrix is therefore

$$\Phi(Y) = \mathsf{V}^+\left(Y + \frac{h_x}{2}\Sigma(Y)\right) - \mathsf{V}^-\left(Y - \frac{h_x}{2}\Sigma(Y)\right)P_+. \tag{11.207}$$

The conservative finite-volume update can then be written as

$$\dot{Y}(t) = \frac{1}{h_x}\Phi(Y(t))\left(P_- - I_{N_x}\right). \tag{11.208}$$

Indeed, the (j,i) entry of Equation (11.208) is precisely Equation (11.204).

Using $P_+P_- = I_{N_x}$ and expanding Equation (11.208), we obtain

$$\begin{aligned}\dot{Y}(t) = \frac{1}{h_x}&\left[\mathsf{V}^+ Y(t) P_- - \left(\mathsf{V}^+ + \mathsf{V}^-\right) Y(t) + \mathsf{V}^- Y(t) P_+\right] \\ &+ \frac{1}{2}\left[\mathsf{V}^+\Sigma(Y(t))P_- + \left(\mathsf{V}^- - \mathsf{V}^+\right)\Sigma(Y(t)) - \mathsf{V}^-\Sigma(Y(t))P_+\right].\end{aligned} \tag{11.209}$$

This is the matrix form of the componentwise dynamics in Equation (11.205).

To obtain a more compact representation, define the periodic one-sided difference matrices

$$D_x^- = \frac{P_- - I_{N_x}}{h_x}, \qquad D_x^+ = \frac{P_+ - I_{N_x}}{h_x}. \tag{11.210}$$

The finite-volume matrix ODE can then be written as

$$\dot{Y}(t) = \mathsf{V}^+\left(Y(t) + \frac{h_x}{2}\Sigma(Y(t))\right) D_x^- + \mathsf{V}^-\left(Y(t) - \frac{h_x}{2}\Sigma(Y(t))\right) D_x^+. \tag{11.211}$$

Thus, the semidiscrete finite-volume problem has the matrix-ODE form

$$\dot{Y}(t) = \mathcal{R}_{\text{tr}}(Y(t)), \qquad Y(0) = Y_0, \tag{11.212}$$

where

$$\mathcal{R}_{\text{tr}}(Y) = \mathsf{V}^+ \left(Y + \frac{h_x}{2}\Sigma(Y) \right) D_x^- + \mathsf{V}^- \left(Y - \frac{h_x}{2}\Sigma(Y) \right) D_x^+. \tag{11.213}$$

The state matrix Y may admit an accurate low-rank representation. However, the map

$$Y \longmapsto \Sigma(Y)$$

is nonlinear and acts entrywise through the minmod function. Consequently, even when Y is accurately represented by a low-rank matrix, the limited-slope matrix $\Sigma(Y)$ need not have comparably low rank. The numerical experiments below examine how this nonlinear reconstruction affects the accuracy, rank structure, and computational cost of the dynamical low-rank approximation.

11.3.2 Projector-Splitting Time Integration

We advance the semidiscrete transport matrix ODE derived above with the projector-splitting integrator from Definition 11.28. The factor updates use the finite-volume right-hand side directly. Because the fixed-rank manifold, tangent-space projection, and generic factor equations were established earlier in this chapter, we do not repeat them here.

11.3.3 Rank-Growth and Accuracy Experiments

We test the projector-splitting integrator on the collisionless kinetic equation

$$\partial_t f(x, v, t) + v\, \partial_x f(x, v, t) = 0, \qquad (x, v) \in [-L, L] \times [-V_{\text{max}}, V_{\text{max}}], \tag{11.214}$$

with periodic boundary conditions in the spatial variable. For initial data

$$f(x, v, 0) = f_0(x, v), \tag{11.215}$$

the method of characteristics gives the exact solution

$$f(x, v, t) = f_0(x - vt, v), \tag{11.216}$$

where $x - vt$ is interpreted periodically on $[-L, L]$. We consider three initial conditions. For midpoint initialization, the corresponding discrete matrix is defined by

$$(Y_0)_{j,i} = f_0(x_i, v_j), \qquad 1 \le i \le N_x, \qquad 1 \le j \le N_v. \tag{11.217}$$

Thus, the rows of Y_0 correspond to velocity cells and the columns correspond to spatial cells. This midpoint value approximates the double cell average defined earlier. It is second-order accurate for smooth data, but that order need not hold in cells intersecting a discontinuity.

11.3.3.1 Initial Conditions

Smooth rank-one data.

The first initial condition is

$$f_0^{(1)}(x, v) = \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi v}{2V_{\max}}\right). \tag{11.218}$$

This function is separable in x and v. Indeed, if

$$a_1(x) = \sin\left(\frac{\pi x}{L}\right), \qquad b_1(v) = \cos\left(\frac{\pi v}{2V_{\max}}\right), \tag{11.219}$$

then

$$f_0^{(1)}(x, v) = a_1(x) b_1(v). \tag{11.220}$$

Consequently, the midpoint-sampled initial matrix has the outer-product form

$$Y_0^{(1)} = \boldsymbol{b}_1 \boldsymbol{a}_1^{\top} \tag{11.221}$$

and therefore has rank one.

The exact transported solution is

$$f^{(1)}(x, v, t) = \sin\left(\frac{\pi(x - vt)}{L}\right) \cos\left(\frac{\pi v}{2V_{\max}}\right). \tag{11.222}$$

Using the angle-difference identity, this solution can be written as a sum of two separable functions:

$$f^{(1)}(x, v, t) = \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi v t}{L}\right) \cos\left(\frac{\pi v}{2V_{\max}}\right)$$

$$- \cos\left(\frac{\pi x}{L}\right) \sin\left(\frac{\pi v t}{L}\right) \cos\left(\frac{\pi v}{2V_{\max}}\right). \tag{11.223}$$

Hence, although the initial condition has rank one, its exact transported solution has separated rank at most two.

Smooth rank-two data.

The second initial condition is

$$f_0^{(2)}(x, v) = \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi v}{2V_{\max}}\right) + \frac{1}{2} \cos\left(\frac{2\pi x}{L}\right) \sin\left(\frac{\pi v}{2V_{\max}}\right). \tag{11.224}$$

This function is the sum of two separable terms. Its midpoint-sampled matrix therefore has the form

$$Y_0^{(2)} = \boldsymbol{b}_1 \boldsymbol{a}_1^\top + \frac{1}{2} \boldsymbol{b}_2 \boldsymbol{a}_2^\top, \tag{11.225}$$

where

$$\begin{aligned} a_1(x) &= \sin\left(\frac{\pi x}{L}\right), \\ b_1(v) &= \cos\left(\frac{\pi v}{2V_{\max}}\right), \\ a_2(x) &= \cos\left(\frac{2\pi x}{L}\right), \\ b_2(v) &= \sin\left(\frac{\pi v}{2V_{\max}}\right). \end{aligned} \tag{11.226}$$

For a nondegenerate spatial and velocity grid, the initial matrix has rank two.

The corresponding exact solution is

$$f^{(2)}(x, v, t) = \sin\left(\frac{\pi(x - vt)}{L}\right) \cos\left(\frac{\pi v}{2V_{\max}}\right) + \frac{1}{2} \cos\left(\frac{2\pi(x - vt)}{L}\right) \sin\left(\frac{\pi v}{2V_{\max}}\right). \tag{11.227}$$

Expanding the translated sine and cosine shows that each initial rank-one term produces at most two separable terms. Therefore, the exact transported solution has separated rank at most four.

Discontinuous rank-one data.

The third initial condition is

$$f_0^{(3)}(x,v) = \chi_{(-L/4,L/4)}(x)\cos\left(\frac{\pi v}{2V_{\max}}\right), \tag{11.228}$$

where $\chi_{(-L/4,L/4)}$ is the indicator function

$$\chi_{(-L/4,L/4)}(x) = \begin{cases} 1, & |x| < L/4, \\ 0, & |x| \geq L/4. \end{cases} \tag{11.229}$$

The initial condition is separable and therefore has rank one. In particular,

$$f_0^{(3)}(x,v) = a_3(x)b_3(v), \tag{11.230}$$

where

$$a_3(x) = \chi_{(-L/4,L/4)}(x), \qquad b_3(v) = \cos\left(\frac{\pi v}{2V_{\max}}\right). \tag{11.231}$$

The exact transported solution is

$$f^{(3)}(x,v,t) = \chi_{(-L/4,L/4)}(x - vt)\cos\left(\frac{\pi v}{2V_{\max}}\right), \tag{11.232}$$

with the translated spatial argument interpreted periodically. Unlike the first two examples, the velocity-dependent translation of the discontinuities generally does not admit a uniformly low-rank separated representation. Consequently, the discrete solution may undergo substantial rank growth even though the initial matrix has rank one. This example is therefore particularly useful for investigating slope limiting and dynamical low-rank approximation in the presence of discontinuities.

11.3.3.2 Experimental Setup and Diagnostics

For the numerical experiments, we set $L = 1$ and $V_{\max} = 1$, so that $x, v \in [-1, 1]$. The initial time is $t_{\text{init}} = 0$, and the final time is $t_{\text{fin}} = 0.5$. The experiments use prescribed ranks $r = 1, 2, 4, 8, 12, 16$ on a phase-space grid with $N_x = 64$ spatial cells and $N_v = 32$ velocity cells. Consequently, $Y(t) \in \mathbb{R}^{32\times 64}$. The Forward Euler and DLRA calculations use the same slope-limited finite-volume operator and the same timestep, $\tau = 2.60417 \times 10^{-3}$. Thus the full Forward Euler solution is a same-grid full-order comparator, not a grid-converged reference; the analytic solution measures the total discretization plus low-rank error.

On this uniform grid, define

$$\|Z\|_{2,h} = \left(h_x h_v \sum_{j=1}^{N_v} \sum_{i=1}^{N_x} |Z_{j,i}|^2 \right)^{1/2} = \sqrt{h_x h_v}\, \|Z\|_{\mathrm{F}}. \tag{11.233}$$

The reported relative errors are

$$E_{\mathrm{full}}(t) = \frac{\|Y_r(t) - Y_{\mathrm{full}}(t)\|_{2,h}}{\|Y_{\mathrm{full}}(t)\|_{2,h}}, \qquad E_{\mathrm{exact}}(t) = \frac{\|Y_r(t) - Y_{\mathrm{exact}}(t)\|_{2,h}}{\|Y_{\mathrm{exact}}(t)\|_{2,h}}. \tag{11.234}$$

If Z has singular values $\sigma_1 \geq \sigma_2 \geq \cdots$, its numerical rank at tolerance $\tau_{\mathrm{rank}} = 10^{-8}$ is

$$\mathrm{rank}_{\tau_{\mathrm{rank}}}(Z) = \#\left\{k \mid \sigma_k/\sigma_1 > \tau_{\mathrm{rank}}\right\}. \tag{11.235}$$

The same-rank compression error is

$$E_{\mathrm{comp},r}(Z) = \frac{\|Z - Z_r\|_{\mathrm{F}}}{\|Z\|_{\mathrm{F}}}, \tag{11.236}$$

where Z_r is the truncated rank-r singular-value decomposition. These definitions apply to both the state and the limiter matrix. To measure conservation, we define the discrete total mass of a state matrix $Y \in \mathbb{R}^{N_v \times N_x}$ by

$$M_h(Y) = h_x h_v \sum_{j=1}^{N_v} \sum_{i=1}^{N_x} Y_{j,i} = h_x h_v\, \mathbf{1}_{N_v}^{\top} Y \mathbf{1}_{N_x}. \tag{11.237}$$

The signed mass change at the final time is

$$\Delta M = M_h(Y(T)) - M_h(Y(0)). \tag{11.238}$$

Because some of the initial conditions contain both positive and negative values and may have small signed total mass, we normalize the mass change by the discrete L^1-size of the initial condition. The scaled mass drift reported below is therefore

$$\delta_M(Y) = \frac{|M_h(Y(T)) - M_h(Y(0))|}{h_x h_v \sum_{j=1}^{N_v} \sum_{i=1}^{N_x} |Y_{j,i}(0)|}. \tag{11.239}$$

For a nonnegative initial condition, the denominator is equal to the initial discrete mass, so δ_M is the relative change in total mass. The full finite-volume method conserves this quantity

to machine precision under periodic boundary conditions, whereas the projected low-rank evolution does not automatically preserve the conservative flux cancellation.

11.3.3.3 Smooth Rank-One Results

Table 11.1: Final-time accuracy and conservation results for the smooth rank-one collisionless-transport experiment. Here, E_{full} is the relative Frobenius error with respect to the full Forward Euler solution, and E_{exact} is the relative Frobenius error with respect to the exact transported solution.

Method	Prescribed rank	E_{full}	E_{exact}	Scaled mass drift
Full Forward Euler	—	—	2.2449×10^{-3}	3.9518×10^{-17}
PSI	1	5.5027×10^{-1}	5.4983×10^{-1}	3.1082×10^{-16}
PSI	2	2.3427×10^{-3}	2.4660×10^{-3}	9.8417×10^{-9}
PSI	4	2.4302×10^{-3}	2.8596×10^{-3}	6.3326×10^{-9}
PSI	8	2.2544×10^{-3}	3.3558×10^{-3}	2.5439×10^{-8}
PSI	12	2.4625×10^{-3}	4.0071×10^{-3}	7.9791×10^{-9}
PSI	16	2.5070×10^{-3}	4.2705×10^{-3}	1.5375×10^{-8}

Table 11.2: Final numerical ranks and same-rank compression errors of the slope-limiter matrix for the smooth rank-one experiment. Numerical ranks are computed using the relative tolerance 10^{-8}.

Prescribed rank r	$\text{rank}(Y_r(T))$	$\text{rank}(\Sigma(Y_r(T)))$	Final compression error
1	1	1	3.2019×10^{-16}
2	2	32	8.3861×10^{-4}
4	4	32	2.3638×10^{-3}
8	8	32	7.0249×10^{-3}
12	12	32	7.2567×10^{-3}
16	16	32	6.1158×10^{-3}

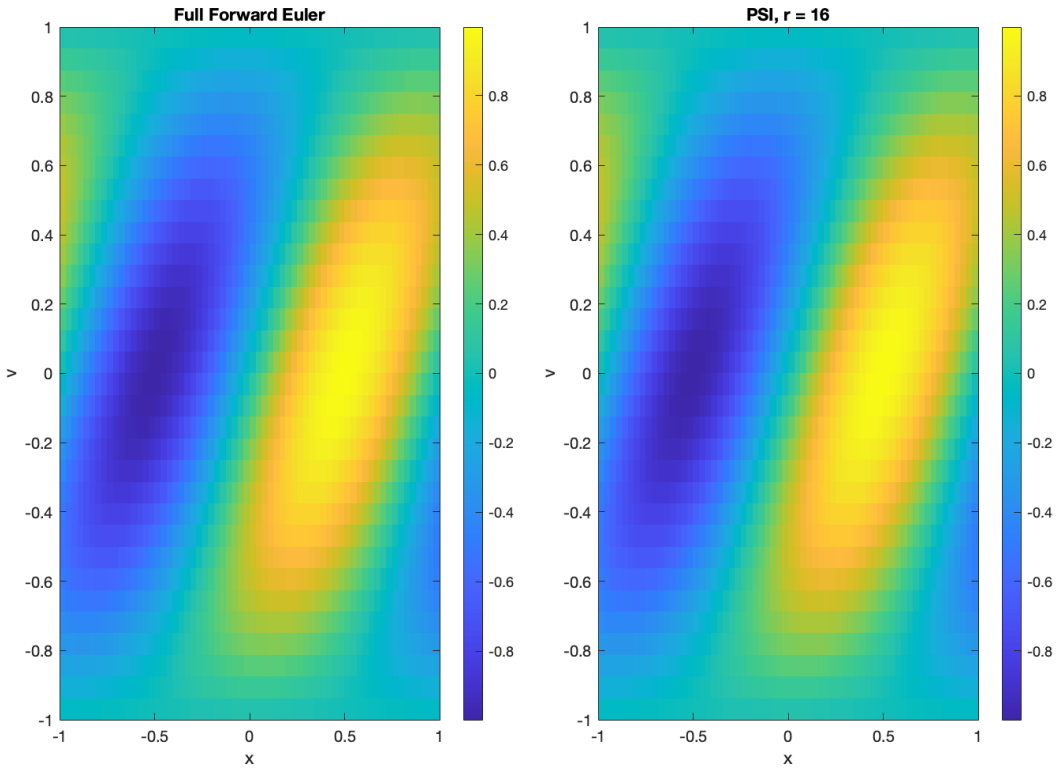


Figure 11.6: Final-time solutions for the smooth rank-one initial condition. The full-matrix Forward Euler solution is shown on the left, and the rank-16 projector-splitting approximation is shown on the right. Both calculations use the same slope-limited finite-volume spatial discretization and timestep.

At each diagnostic time, the same-rank compression error is defined by

$$e_r^{\Sigma}(t) = \frac{\|\Sigma(Y_r(t)) - [\Sigma(Y_r(t))]_r\|_{\mathrm{F}}}{\|\Sigma(Y_r(t))\|_{\mathrm{F}}}, \tag{11.240}$$

where $[\Sigma(Y_r(t))]_r$ denotes the best rank-r approximation obtained by truncating the singular value decomposition.

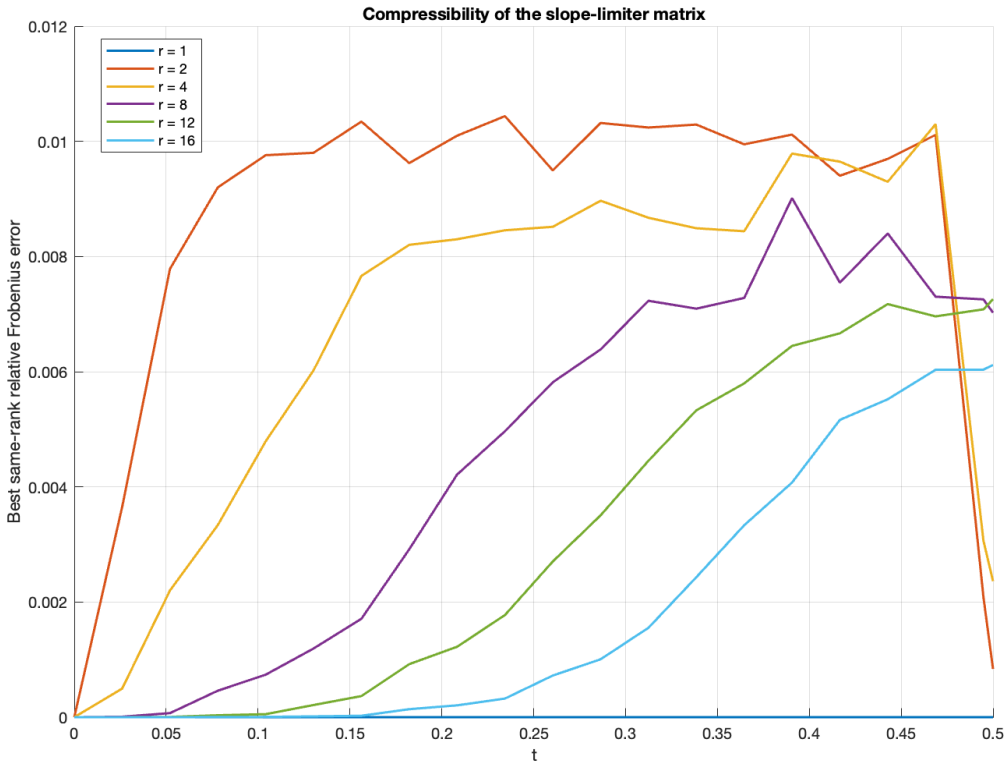


Figure 11.7: Same-rank compression error of the slope-limiter matrix for the smooth rank-one experiment. Each curve shows Equation (11.240) for a different prescribed DLRA rank. Smaller values indicate greater compressibility at the corresponding state rank.

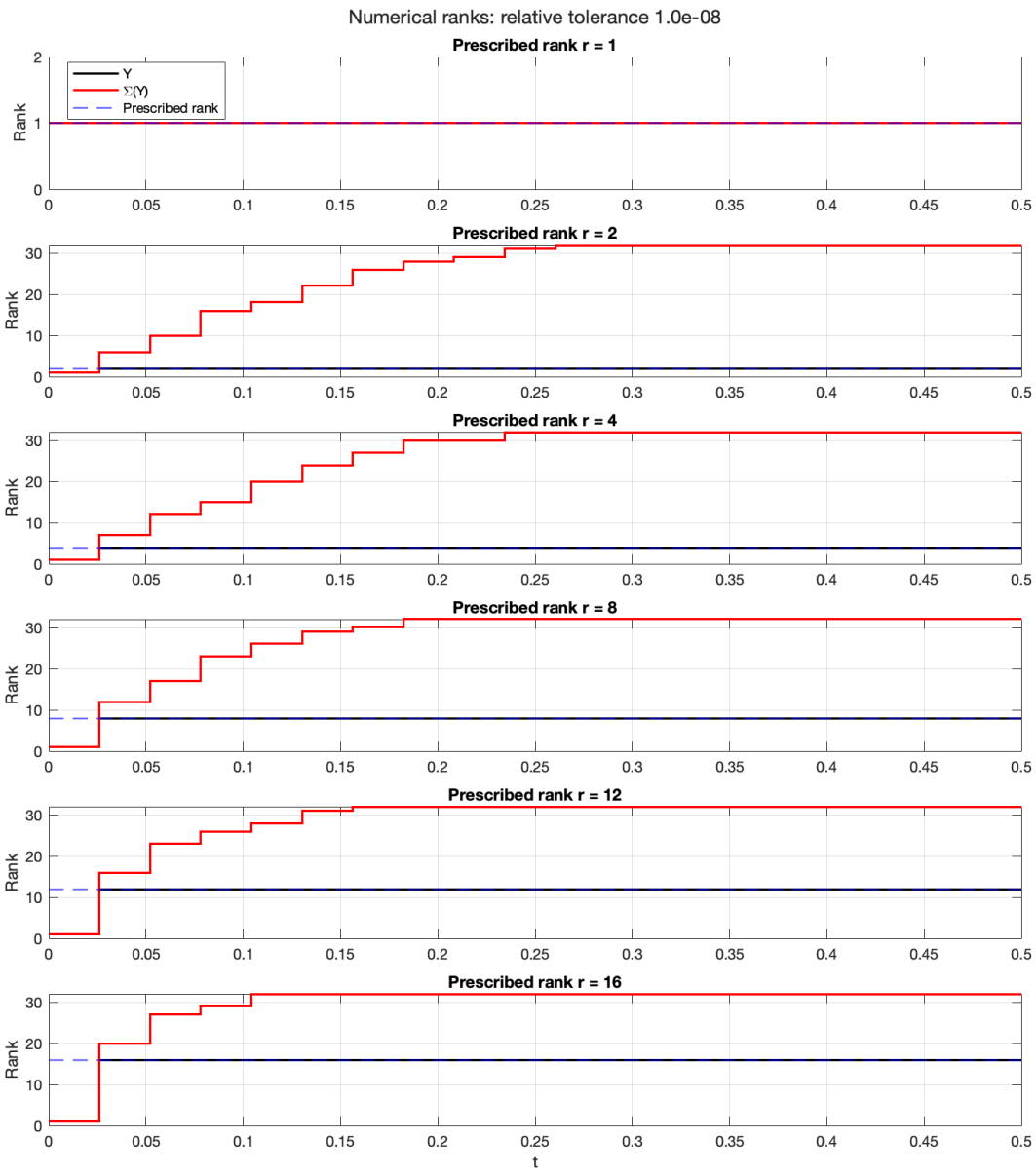


Figure 11.8: Numerical-rank histories for the state matrix $Y_r(t)$ and the corresponding slope-limiter matrix $\Sigma(Y_r(t))$ in the smooth rank-one experiment. Numerical ranks are computed using the relative criterion $\sigma_k/\sigma_1 > 10^{-8}$. Although the projector-splitting approximation maintains the prescribed state rank, the entrywise minmod operation generally produces a slope-limiter matrix of substantially higher numerical rank.

11.3.3.4 Smooth Rank-Two Results

Table 11.3: Final-time accuracy and conservation results for the smooth rank-two collisionless-transport experiment. Here, E_{full} is the relative Frobenius error with respect to the full Forward Euler solution, and E_{exact} is the relative Frobenius error with respect to the exact transported solution.

Method	Prescribed rank	E_{full}	E_{exact}	Scaled mass drift
Full Forward Euler	—	—	1.0631×10^{-2}	1.7418×10^{-17}
PSI	1	6.6675×10^{-1}	6.6472×10^{-1}	1.1032×10^{-16}
PSI	2	6.4172×10^{-1}	6.3879×10^{-1}	6.6048×10^{-16}
PSI	4	1.5162×10^{-2}	1.8665×10^{-2}	6.5760×10^{-7}
PSI	8	1.5170×10^{-2}	2.1143×10^{-2}	3.9726×10^{-6}
PSI	12	1.4957×10^{-2}	2.2204×10^{-2}	3.4377×10^{-6}
PSI	16	1.4991×10^{-2}	2.3183×10^{-2}	1.0849×10^{-5}

Table 11.4: Numerical rank and same-rank compression error of the slope-limiter matrix in the smooth rank-two experiment. Numerical ranks are computed using the relative tolerance 10^{-8}.

r	$\text{rank}(Y_r(T))$	$\text{rank}(\Sigma(Y_r(T)))$	Maximum error	Final error
1	1	1	3.6961×10^{-16}	3.1686×10^{-16}
2	2	32	2.0155×10^{-2}	1.1388×10^{-2}
4	4	32	2.0778×10^{-2}	1.2034×10^{-2}
8	8	32	1.7672×10^{-2}	1.6738×10^{-2}
12	12	32	1.4077×10^{-2}	1.1970×10^{-2}
16	16	32	1.3616×10^{-2}	1.3119×10^{-2}

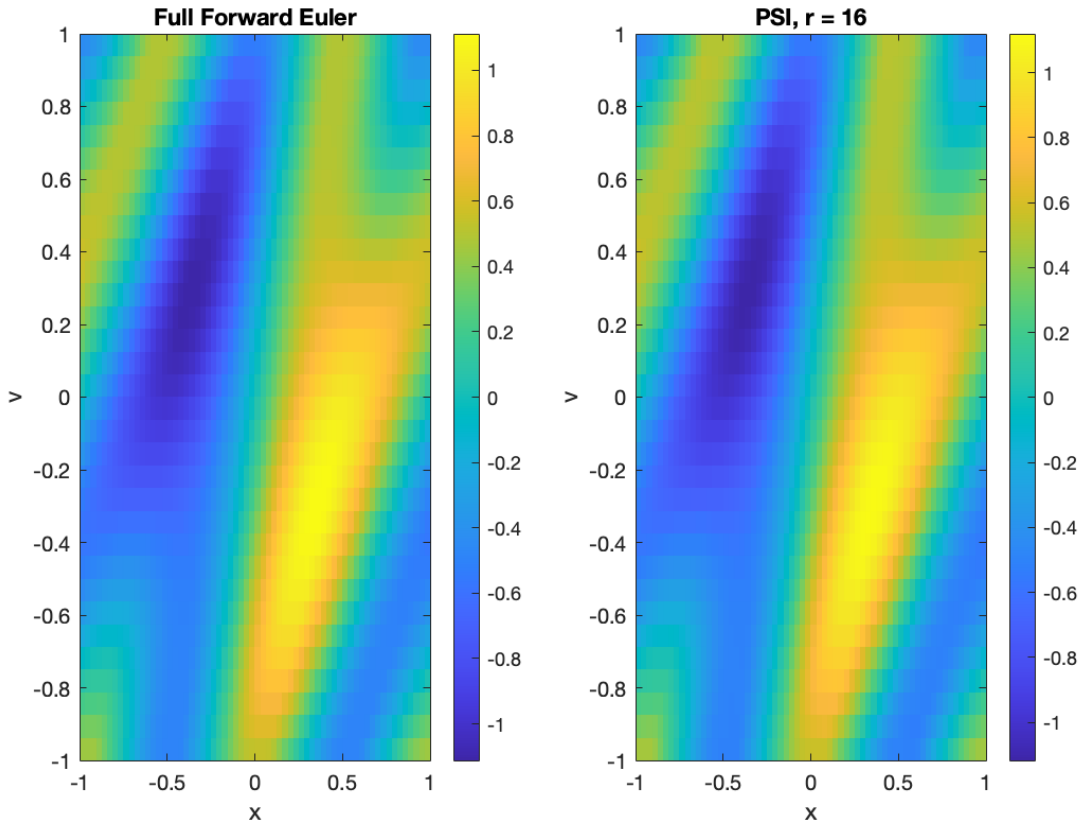


Figure 11.9: Final-time solutions for the smooth rank-two initial condition. The full-matrix Forward Euler solution is shown on the left, and the rank-16 projector-splitting approximation is shown on the right. Both calculations use the same slope-limited finite-volume spatial discretization and timestep.

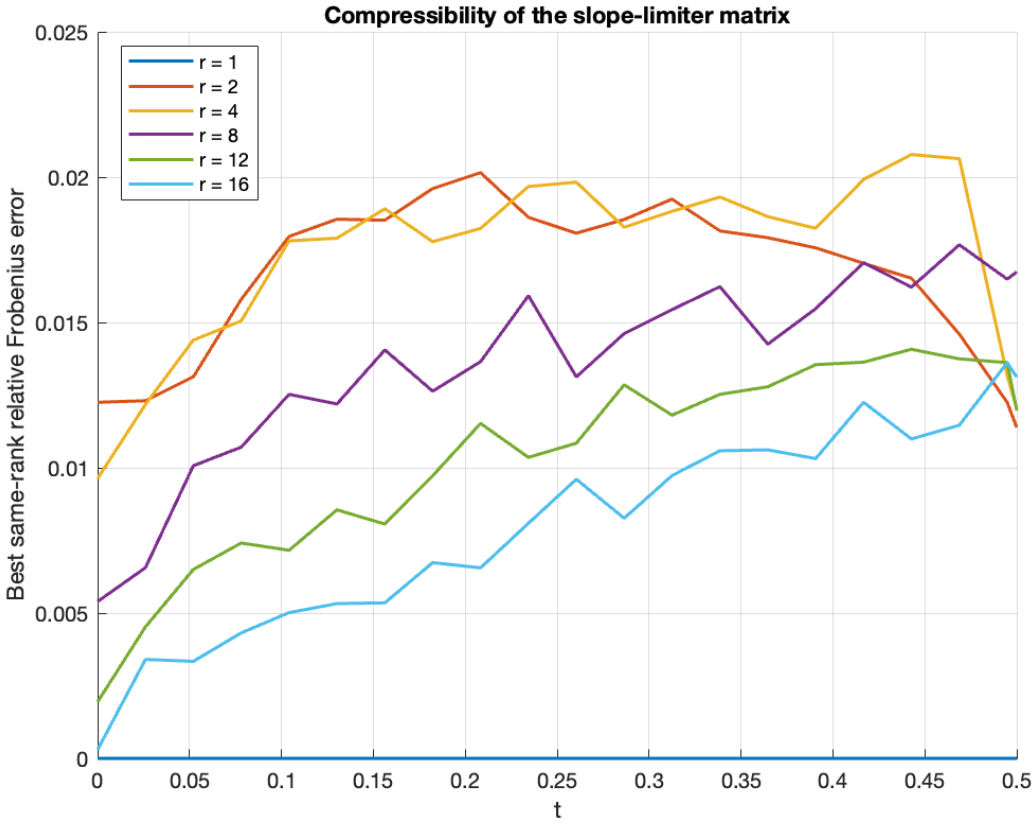


Figure 11.10: Same-rank compression error of the slope-limiter matrix for the smooth rank-two experiment. Each curve shows the best rank-r relative Frobenius approximation error of $\Sigma(Y_r(t))$ for the corresponding prescribed DLRA rank. Smaller values indicate greater compressibility at the state rank.

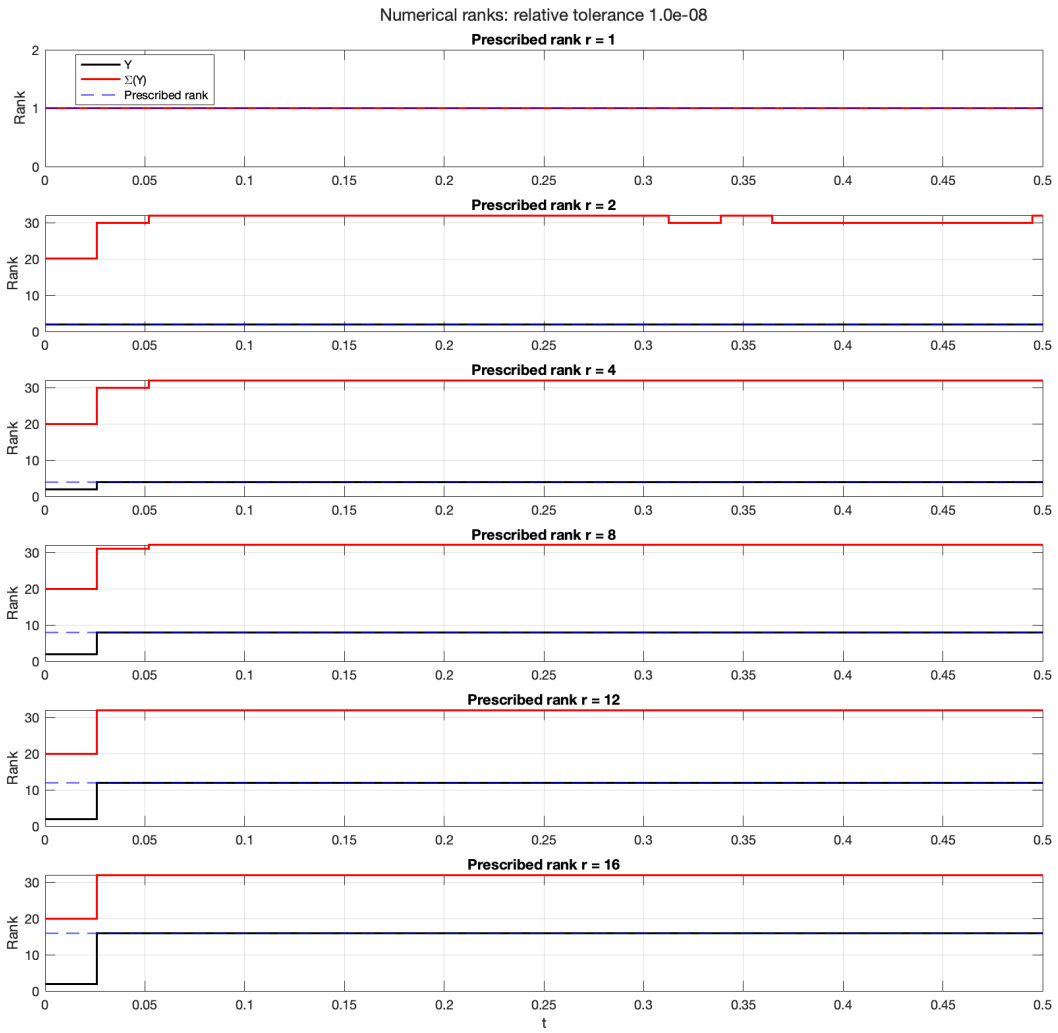


Figure 11.11: Numerical-rank histories for the state matrix $Y_r(t)$ and the slope-limiter matrix $\Sigma(Y_r(t))$ in the smooth rank-two experiment. Numerical ranks are computed using the relative criterion $\sigma_k/\sigma_1 > 10^{-8}$. The state retains its prescribed rank, whereas the slope-limiter matrix rapidly reaches substantially higher numerical rank.

11.3.3.5 Discontinuous Rank-One Results

Table 11.5: Final-time accuracy and conservation results for the discontinuous rank-one collisionless-transport experiment. Here, E_{full} is the relative Frobenius error with respect to the full Forward Euler solution, and E_{exact} is the relative Frobenius error with respect to the exact transported solution. PSI denotes the projector-splitting integrator.

Method	Prescribed rank	E_{full}	E_{exact}	Scaled mass drift
Full Forward Euler	—	—	1.9779×10^{-1}	8.7162×10^{-16}
PSI	1	7.0177×10^{-1}	7.4682×10^{-1}	8.5803×10^{-3}
PSI	2	4.3420×10^{-1}	4.9541×10^{-1}	9.6212×10^{-2}
PSI	4	2.7535×10^{-1}	3.6006×10^{-1}	5.4053×10^{-2}
PSI	8	1.0761×10^{-1}	2.3685×10^{-1}	7.7799×10^{-3}
PSI	12	5.1881×10^{-2}	2.0725×10^{-1}	1.8025×10^{-3}
PSI	16	2.8045×10^{-2}	1.9768×10^{-1}	5.2763×10^{-4}

Table 11.6: Numerical rank and same-rank compression error of the slope-limiter matrix in the discontinuous rank-one experiment. Numerical ranks are computed using the relative tolerance 10^{-8}.

r	$\text{rank}(Y_r(T))$	$\text{rank}(\Sigma(Y_r(T)))$	Maximum error	Final error
1	1	1	3.3149×10^{-16}	2.4560×10^{-16}
2	2	14	7.6741×10^{-2}	1.8045×10^{-2}
4	4	32	1.5475×10^{-1}	5.8315×10^{-2}
8	8	32	1.5538×10^{-1}	1.2345×10^{-1}
12	12	32	1.3679×10^{-1}	1.2270×10^{-1}
16	16	32	1.1168×10^{-1}	1.1157×10^{-1}

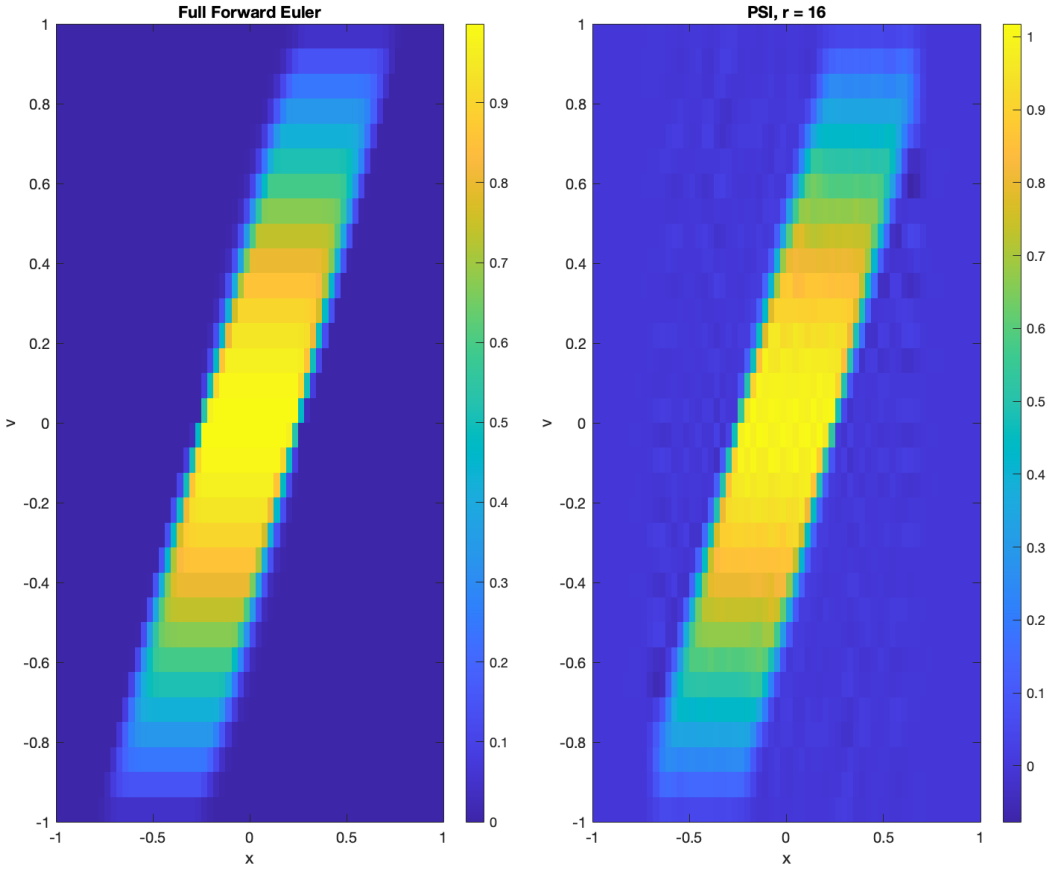


Figure 11.12: Final-time solutions for the discontinuous rank-one initial condition. The full-matrix Forward Euler solution is shown on the left, and the rank-16 projector-splitting approximation is shown on the right. The discontinuity produces considerably greater rank growth and a more demanding low-rank approximation problem than the smooth initial conditions.

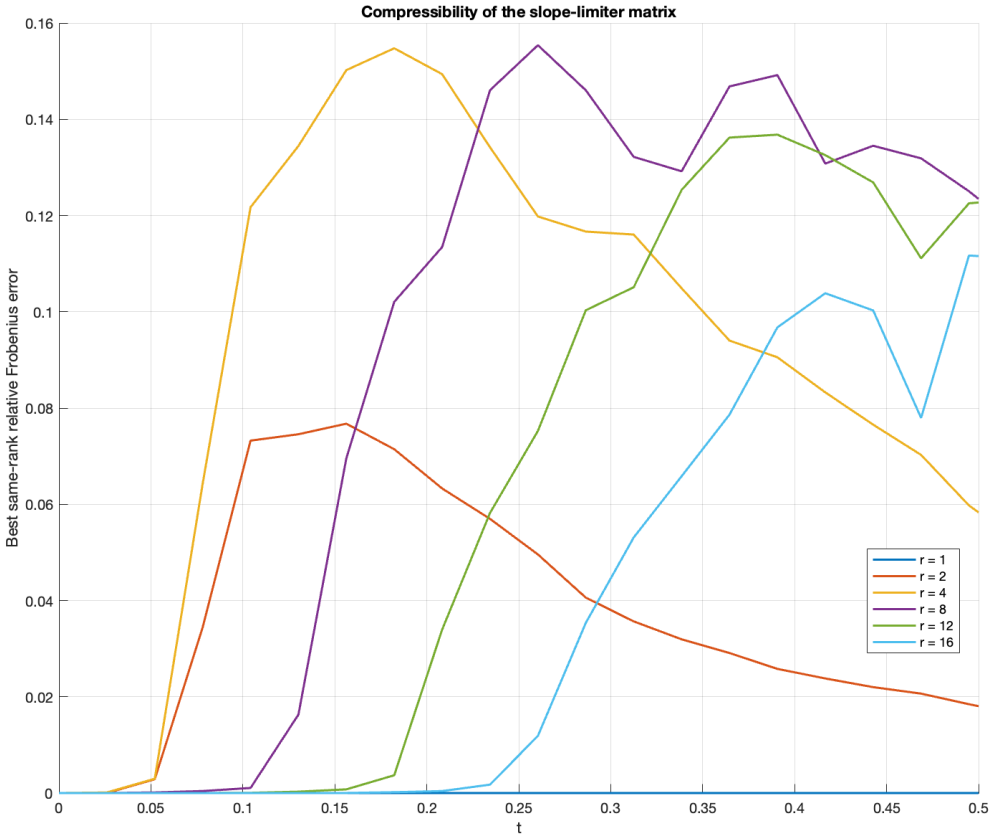


Figure 11.13: Same-rank compression error of the slope-limiter matrix for the discontinuous rank-one experiment. Compared with the smooth experiments, the larger errors show that the slope-limiter matrix is less accurately representable at the prescribed state rank.

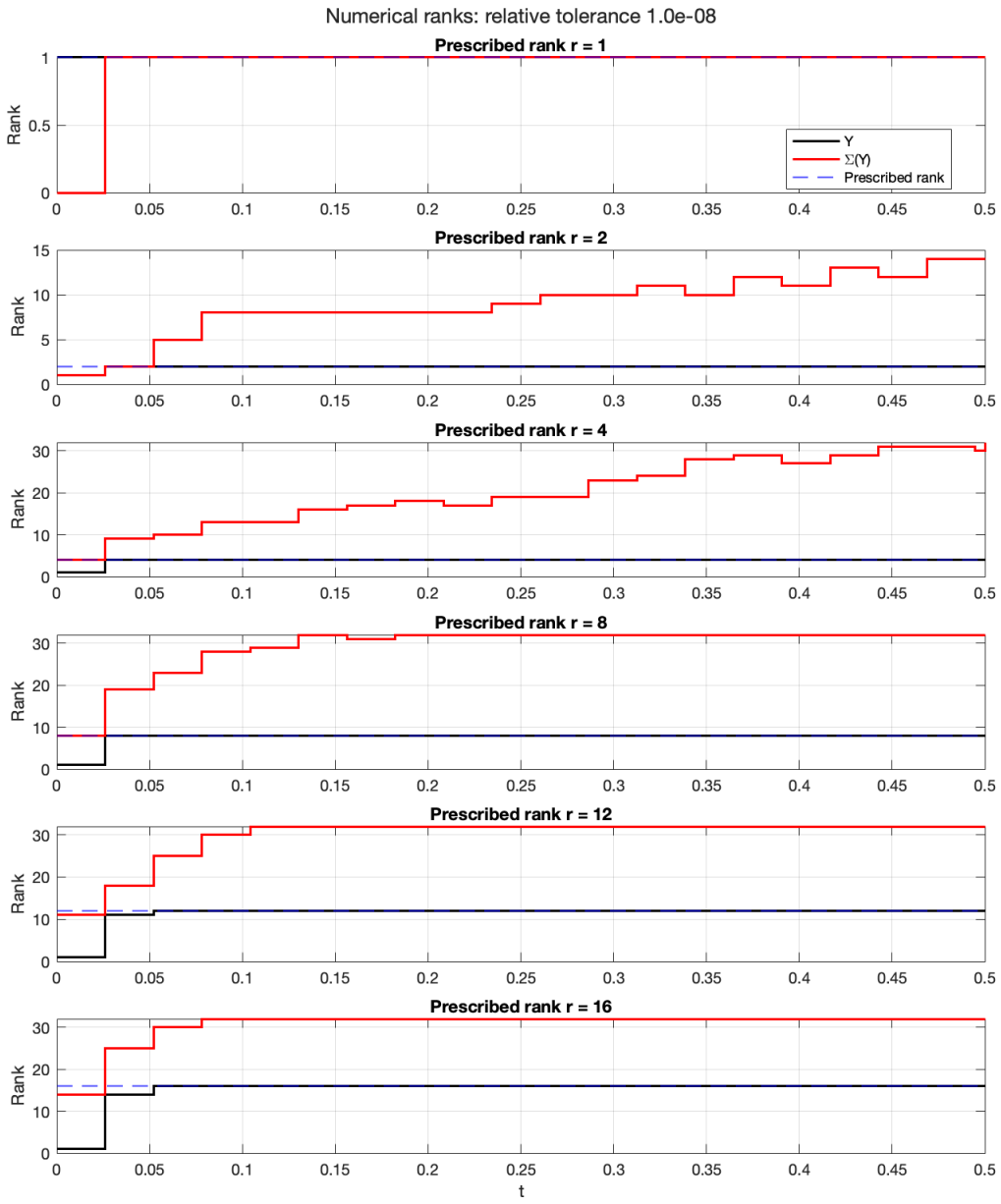


Figure 11.14: Numerical-rank histories for the state matrix $Y_r(t)$ and the slope-limiter matrix $\Sigma(Y_r(t))$ in the discontinuous rank-one experiment. Numerical ranks are computed using the relative criterion $\sigma_k/\sigma_1 > 10^{-8}$. The numerical rank of the limiter matrix grows rapidly, particularly for the higher-rank state approximations.

11.4 Conclusions

The simulations demonstrate that projector-splitting DLRA can reproduce the principal macroscopic features of the kinetic solutions when the prescribed rank is sufficiently large. In the Sod experiments, the higher-rank approximations closely track the full finite-volume density,

velocity, pressure, and temperature profiles for both $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$, while the phase-density error remains localized near the evolving wave structure. At the same time, the sampled Maxwellian and slope-limiter matrices attain numerical ranks substantially larger than the state rank. Thus accurate low-rank evolution of the distribution does not imply that every nonlinear intermediate quantity is itself low rank.

The collisionless-transport experiments make the dependence on solution structure especially clear. For the smooth data, accuracy improves sharply once the prescribed rank reaches the separated rank of the analytic solution: $r = 2$ for the first test and $r = 4$ for the second. The discontinuous solution has no comparable uniformly low-rank representation, so its error decreases more gradually with increasing rank. Moreover, although the limited-slope matrix $\Sigma(Y_r)$ frequently has full numerical rank, it remains accurately compressible in the smooth tests; near discontinuities its singular values decay more slowly, its same-rank compression errors are larger, and low-rank truncation is accompanied by greater mass drift.

Taken together, these results identify nonlinear reconstruction near nonsmooth features—rather than the low-rank state alone—as the principal obstacle to an efficient and conservative DLRA finite-volume method. A promising next step is therefore a hybrid representation in which smooth-region slope information is compressed globally and the less-compressible contribution near shocks is retained as a localized correction. Schematically, we seek an approximation of the form

$$\Sigma(Y) \approx \Sigma_{\mathrm{lr}}(Y) + D_{\Sigma}(Y),$$

where $D_{\Sigma}(Y)$ is supported on an active set containing the cells near nonsmooth features.

Chapter 12

Progress Summary and Next Steps

12.1 Year-Five Progress

The Year Four Report established the basic organization of this cumulative document: introductory material on the multispecies Boltzmann and BGK equations, finite-volume and discrete-velocity methods for the unary BGK model, the multispecies BGK moment system, an implicit iterative moment solver, and a preliminary treatment of ES-BGK models. The Year-Five report retains that foundation and makes four principal changes: it expands the multispecies Boltzmann chapter, adds stand-alone Chapman–Enskog and moment-method chapters, substantially updates the ES-BGK chapter, and adds a new DLRA chapter with numerical simulations.

First, Chapter 2, which was already present in the Year Four Report, has been substantially expanded and reorganized from the revised *BoltzmannOnly* notes. The chapter now gives a more self-contained treatment of elastic binary collisions, weak collision identities, conservation laws, the H-theorem, velocity moments, Maxwellian equilibria, entropy maximization, and the space-homogeneous multispecies problem. The revised presentation distinguishes exact kinetic identities from structural consequences of collision symmetry and from formal closure assumptions. It also fixes the notation used throughout the remainder of the report: n_i and $\rho_i = m_i n_i$ denote species number and mass densities, $\boldsymbol{u}_i$ denotes a species bulk velocity, $\hat{\theta}_i$ denotes a species temperature in energy units, and $\boldsymbol{u}$, ρ, P, and $\boldsymbol{q}$ denote barycentric mixture quantities.

Second, the material connecting kinetic and fluid descriptions has been expanded into two stand-alone chapters. Chapter 3 gives a detailed formal Chapman–Enskog derivation of the compressible Euler and Navier–Stokes–Fourier equations from the unary Boltzmann equation. It identifies the exact balance laws, the solvability conditions for the linearized col-

lision operator, and the constitutive origins of viscous stress and heat conduction. Chapter 4 then places that calculation within the broader setting of finite moment closure, including Maxwellian, Grad, maximum-entropy, quadrature, and discrete-velocity closures. Together, these chapters make explicit which statements are exact and which depend on a reconstruction or asymptotic approximation.

Third, Chapter 10 has been substantially updated. It proves the single-species conservation and entropy-production properties, derives the Prandtl correction with the Chapman–Enskog formalism of Chapter 3, gives a closed positivity-preserving implicit collision update, and separates the automatically conservative parts of a mixture extension from the additional hypotheses needed for positivity and information dissipation.

Fourth, Chapter 11 adds a substantial new direction based on Jack Haight's work on dynamical low-rank approximation. It develops the fixed-rank matrix-manifold framework, tangent-space projection, projector-splitting and basis-update/Galerkin integrators, a low-rank treatment of a $1x1v$ BGK discretization, and numerical simulations that expose the interaction between high-resolution finite-volume reconstruction and low-rank compressibility. This material directly addresses the phase-space cost that motivates reduced representations of kinetic equations.

The theoretical and numerical work on the space-homogeneous multispecies BGK moment equations was already reported in the Year Four Report. The analysis of positivity, global well-posedness, equilibrium, and exponential relaxation appeared in [33], and the nonlinear Gauss–Seidel-type iteration for the backward-Euler moment update appeared in [34]. Chapter 9 retains and reorganizes these results and the representative numerical tests as foundations for future transport solvers; they are not counted as new Year-Five material.

The Year-Five report also preserves and reorganizes the earlier unary BGK, finite-volume, discrete-velocity, and multispecies BGK material. Chapter 8 now gives a self-contained derivation of the conservative, entropic model of Haack, Hauck, and Murillo [32]. It connects the Boltzmann pair operator to the BGK ansatz, derives the common pair velocity and temperature, proves positivity and the conservation laws, establishes the H-theorem and equilibrium characterization, and records the leading fluid-limit Maxwellian. The presentation uses the notation and entropy sign convention fixed in Chapter 2.

This cumulative structure serves a purpose beyond archival completeness. Each cycle of consolidation, correction, and editorial polishing turns the report into a more effective teaching resource for future graduate students who join the project. By assembling the required kinetic theory, model derivations, numerical methods, published results, and open problems in one coherent document, the report provides a common technical foundation and a practical path

from initial study to active research. Its continued refinement should therefore be viewed as a project deliverable in its own right.

12.2 Numerical Prototypes at a Glance

Table 12.1 distinguishes analytical guarantees, numerical evidence, code availability, and unresolved validation needs. It does not infer quantitative results from plots when the underlying output data are unavailable.

Model or method	Analytical status	Numerical evidence	Code status	Principal limitation
Unary BGK finite-volume IMEX	Collision invariants are exact in the continuum formulation; discrete-velocity quadrature introduces a stated conservation caveat.	Relaxation, Sod, rotation, two-stream, and Landau-damping tests in Chapter 6; archived evidence is audited in Table 6.1.	Appendix F contains a prototype associated only with the Sod workflow.	The archived driver configuration does not exactly match the displayed Sod figures; norm-based refinement and invariant-drift studies remain pending.
Multispecies BGK moment solver	Published positivity, well-posedness, relaxation, and convergence results under the hypotheses stated in Chapter 9.	Solver-iteration and time-step counts are reported for the Sod and Ar–Kr–Xe tests.	The simulation code is not included in Appendix F.	Integration and validation within full higher-order transport schemes remain future work.
ES-BGK collision and IMEX update	Unary conservation, entropy, admissibility, and a closed implicit collision update are established; mixture results require additional stated hypotheses.	Representative computed moment profiles are reported in Chapter 10.	No ES-BGK implementation is distributed with this report.	General multispecies positivity, information dissipation, and transport validation remain conditional or incomplete.
DLRA methods	Fixed-rank geometry and integrators are developed from established theory; preservation under truncation is not automatic.	BGK and high-resolution finite-volume experiments report prescribed ranks and comparisons in Chapter 11.	Preliminary MATLAB test codes are discussed but are not included in Appendix F.	Conservation, positivity, rank adaptivity, and nonlinear-operator compression require further study.

Table 12.1: Validation matrix for the principal computational methods. "Analytical status" identifies what is proved or structurally established; "numerical evidence" records only results documented in this report.

12.3 Next Steps

The priorities below are ordered by dependency rather than assigned fixed completion dates. Each item states a concrete deliverable and the evidence that would support completion.

1. **Integrate the implicit solver with transport. Deliverable:** a full IMEX transport implementation for multispecies BGK that uses the published backward-Euler moment iteration, beginning with slab geometry closed in three-dimensional velocity through the Chu reduction. **Completion evidence:** documented conservation and positivity diagnostics, asymptotic tests over a range of Knudsen numbers, nonlinear-solver tolerance studies, and comparison with a fully explicit or otherwise established reference calculation. **Dependencies and risks:** consistent stage coupling, discrete moment matching, and solver tolerances that remain effective in stiff regimes.

2. **Advance multispecies ES-BGK and hybrid kinetic–fluid models. Deliverable:** a multispecies ES-BGK formulation and preliminary numerical solvers that reproduces the intended Prandtl behavior under explicitly stated admissibility hypotheses, together with an initial hybrid kinetic–fluid coupling test inspired by [37, 72]. **Completion evidence:** conservation and entropy checks, positive-definiteness tests for the Gaussian tensors, and a benchmark demonstrating stable switching and consistent interface fluxes. **Dependencies and risks:** mixture-level positivity and information dissipation are not automatic, and switching indicators must be robust near regime interfaces.

3. **Combine DLRA with multispecies collision models. Deliverable:** a rank-adaptive prototype for multispecies BGK or ES-BGK transport that avoids full phase-space reconstruction for the dominant operations whenever possible. **Completion evidence:** rank histories, memory and accuracy comparisons against a same-grid full-order calculation, and measured conservation and positivity defects before and after truncation. **Dependencies and risks:** nonseparable Maxwellians and nonlinear slope limiters may erase the low-rank advantage; invariant-preserving truncation remains an open design constraint. Higher-dimensional discrete-velocity methods and block-structured adaptive mesh refinement for $1x3v$, $2x3v$, and ultimately multidimensional configurations follow this prototype stage.

Appendix A

An Important Integral

We need to compute the following generalized Gaussian integral in several places in the report.

Lemma A.1 (Gaussian Second-Moment Integral)**.**

$$\int_{\mathbb{R}^d} |\boldsymbol{s}|^2 e^{-|\boldsymbol{s}|^2} \, \mathrm{d}\boldsymbol{s} = \frac{d}{2}\pi^{\frac{d}{2}}. \tag{A.1}$$

Proof.

$$\begin{aligned}
\int_{\mathbb{R}^d} |\boldsymbol{s}|^2 e^{-|\boldsymbol{s}|^2} \, \mathrm{d}\boldsymbol{s} &= \int_{\mathbb{R}^d} (s_1^2 + \cdots + s_d^2) e^{-(s_1^2+\cdots+s_d^2)} \, \mathrm{d}\boldsymbol{s} \\
&= \int_{\mathbb{R}^d} \left(\sum_{j=1}^{d} s_j^2 \right) \prod_{k=1}^{d} e^{-s_k^2} \, \mathrm{d}\boldsymbol{s} \\
&= \int_{\mathbb{R}^d} s_1^2 \prod_{k=1}^{d} e^{-s_k^2} \, \mathrm{d}\boldsymbol{s} + \cdots + \int_{\mathbb{R}^d} s_d^2 \prod_{k=1}^{d} e^{-s_k^2} \, \mathrm{d}\boldsymbol{s} \\
&= \left(\prod_{\substack{k=1 \\ k\neq 1}}^{d} \int_{\mathbb{R}} e^{-s_k^2} \, \mathrm{d}s_k \right) \int_{\mathbb{R}} s_1^2 e^{-s_1^2} \, \mathrm{d}s_1 + \cdots + \left(\prod_{\substack{k=1 \\ k\neq d}}^{d} \int_{\mathbb{R}} e^{-s_k^2} \, \mathrm{d}s_k \right) \int_{\mathbb{R}} s_d^2 e^{-s_d^2} \, \mathrm{d}s_d \\
&= \left(\pi^{\frac{1}{2}} \right)^{d-1} \int_{\mathbb{R}} s_1^2 e^{-s_1^2} \, \mathrm{d}s_1 + \cdots + \left(\pi^{\frac{1}{2}} \right)^{d-1} \int_{\mathbb{R}} s_d^2 e^{-s_d^2} \, \mathrm{d}s_d \\
&= \pi^{\frac{d-1}{2}} \sum_{i=1}^{d} \int_{\mathbb{R}} s_i^2 e^{-s_i^2} \, \mathrm{d}s_i,
\end{aligned} \tag{A.2}$$

where we have used many elementary integration techniques, and the fact that $\int_{\mathbb{R}} e^{-x^2} \, \mathrm{d}x = \pi^{\frac{1}{2}}$. It remains to determine the value of the terms of the form $\int_{\mathbb{R}} s^2 e^{-s^2} \, \mathrm{d}s$. To that end,

utilizing integration by parts, with

$$u = s_i \implies \mathrm{d}u = \mathrm{d}s_i \tag{A.3}$$

$$\mathrm{d}v = 2s_i e^{-s_i^2}\,\mathrm{d}s_i \implies v = -e^{-s_i^2}, \tag{A.4}$$

we have

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}} 2s_i^2 e^{-s_i^2}\,\mathrm{d}s_i &= \frac{1}{2}\left[\left[-s_i e^{-s_i^2}\right]_{-\infty}^{\infty} - \int_{\mathbb{R}} -e^{-s_i^2}\,\mathrm{d}s_i\right] \\ &= \frac{1}{2}\int_{\mathbb{R}} e^{-s_i^2}\,\mathrm{d}s_i \\ &= \frac{1}{2}\pi^{\frac{1}{2}}.\end{aligned} \tag{A.5}$$

Thus, Equation (A.2) gives us

$$\int_{\mathbb{R}^d} |\boldsymbol{s}|^2 e^{-|\boldsymbol{s}|^2}\,\mathrm{d}\boldsymbol{s} = \pi^{\frac{d-1}{2}} \sum_{i=1}^{d} \frac{1}{2}\pi^{\frac{1}{2}} = \pi^{\frac{d-1}{2}}\left(\frac{d}{2}\pi^{\frac{1}{2}}\right) = \frac{d}{2}\pi^{\frac{d}{2}}, \tag{A.6}$$

as desired. ■

Appendix B

Picard–Lindelöf Theorem

We will make use of the following version of the Picard–Lindelöf Theorem for autonomous systems:

Theorem B.1. *Suppose that* $\boldsymbol{u}_0 \in \mathbb{R}^d$ *and* $\boldsymbol{f} \in C(\overline{B}(\boldsymbol{u}_0, \beta); \mathbb{R}^d)$*, where* $B(\boldsymbol{u}_0, \beta)$ *(*$\overline{B}(\boldsymbol{u}_0, \beta)$*) denotes the open (closed) Euclidean ball of radius* β *centered at* $\boldsymbol{u}_0$*. Suppose further that* $M > 0$*,* $\|\boldsymbol{f}(\boldsymbol{v})\|_2 \leq M$ *for all* $\boldsymbol{v} \in \overline{B}(\boldsymbol{u}_0, \beta)$*, and* $\boldsymbol{f}$ *is Lipschitz continuous, with constant* $L > 0$*, on* $\overline{B}(\boldsymbol{u}_0, \beta)$*. Set* $\delta = \min\left(\frac{1}{2L}, \frac{\beta}{M}\right)$*. Then the IVP* $\boldsymbol{u}'(t) = \boldsymbol{f}(\boldsymbol{u}(t))$*,* $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$ *has a unique solution on the interval* $[t_0 - \delta, t_0 + \delta]$*.*

Appendix C

Bounding the Eigenvalues of Z and $\widehat{Z}$

The matrices Z and $\widehat{Z}$, defined in (9.20), play a key role in the dynamics of (9.17) and the equivalent system (9.19). In this appendix, we prove bounds on the eigenvalues of these matrices. With the collision frequencies defined in (9.11), the matrices A and B, defined in (9.14a), take the form

$$A_{i,j} = \frac{16}{3}\sqrt{\frac{\pi}{2}}\frac{m_i m_j (d_i + d_j)^2}{(m_i + m_j)^3}\rho_i \rho_j \sqrt{\frac{\hat{\theta}_i}{m_i} + \frac{\hat{\theta}_j}{m_j}}, \tag{C.1}$$

$$B_{i,j} = \frac{8}{3}\sqrt{\frac{\pi}{2}}\frac{(d_i + d_j)^2}{(m_i + m_j)^2}\rho_i \rho_j \sqrt{\frac{\hat{\theta}_i}{m_i} + \frac{\hat{\theta}_j}{m_j}}. \tag{C.2}$$

Since each value used to define the matrix elements is positive and bounded above and below (importantly, the temperature), then their values can be bounded as follows:

$$A_{\min}(t) \coloneqq \min_{i,j} A_{i,j}(t) \leq A_{i,j}(t) \leq \max_{i,j} A_{i,j}(t) =: A_{\max}(t) \tag{C.3}$$

$$B_{\min}(t) \coloneqq \min_{i,j} B_{i,j}(t) \leq B_{i,j}(t) \leq \max_{i,j} B_{i,j}(t) =: B_{\max}(t) \tag{C.4}$$

An earlier version of this appendix used the factor $N - 1$ in the two upper bounds below. The complete-graph identity gives the factor N. This correction does not change the exponential relaxation rates in Chapter 9, which depend only on the unchanged lower bounds $z_{\min}$ and $\widehat{z}_{\min}$.

Theorem C.1 (Eigenvalue Bounds)**.** *Let $A, B, D, F \in \mathbb{R}^{N\times N}$ be the matrices defined in* (9.14)*, and let $A_{\min} = \min_{i,j} A_{i,j}$, $B_{\min} = \min_{i,j} B_{i,j}$, $A_{\max} = \max_{i,j} A_{i,j}$, and $B_{\max} = \max_{i,j} B_{i,j}$. Then*

the eigenvalues of the matrices $Z = P^{-\frac{1}{2}}(D-A)P^{-\frac{1}{2}}$ *and* $\widehat{Z} = Q^{-\frac{1}{2}}(F-B)Q^{-\frac{1}{2}}$ *satisfy*

$$0 = z_0 < z_{\min} \le z_1 \le \cdots \le z_{N-1} \le z_{\max}, \tag{C.5}$$

$$0 = \widehat{z}_0 < \widehat{z}_{\min} \le \widehat{z}_1 \le \cdots \le \widehat{z}_{N-1} \le \widehat{z}_{\max}, \tag{C.6}$$

where

$$z_{\min} = \frac{A_{\min}N}{\max\{\rho_k\}}, \quad z_{\max} = \frac{A_{\max}N}{\min\{\rho_k\}}, \tag{C.7}$$

$$\widehat{z}_{\min} = \frac{B_{\min}N}{\max\{n_k\}}, \quad \widehat{z}_{\max} = \frac{B_{\max}N}{\min\{n_k\}}. \tag{C.8}$$

Proof. The proofs of (C.5) and (C.6) are analogous, so we prove (C.5). Because the interaction graph is complete and its weights are positive, $D-A$ is a weighted graph Laplacian with null space $\operatorname{span}\{\mathbf{1}\}$. Consequently,

$$\mathcal{N}(Z) = \operatorname{span}\{P^{1/2}\mathbf{1}\}.$$

Let $\boldsymbol{z} \perp P^{1/2}\mathbf{1}$ and set $\boldsymbol{y} = P^{-1/2}\boldsymbol{z}$. The relevant constraint is therefore the weighted orthogonality condition

$$\boldsymbol{y}^\top P\mathbf{1} = 0,$$

rather than $\boldsymbol{y}^\top\mathbf{1} = 0$. Let $\overline{y} = N^{-1}\sum_i y_i$ and $\boldsymbol{e} = \boldsymbol{y} - \overline{y}\mathbf{1}$, so that $\boldsymbol{e}^\top\mathbf{1} = 0$. The complete-graph identity gives

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - y_j)^2 = N\|\boldsymbol{e}\|_2^2.$$

Because $\sum_i e_i = 0$, the value $c = 0$ minimizes $c \mapsto \sum_i (e_i + c)^2$. Therefore $\boldsymbol{y}^\top P\boldsymbol{y} = \sum_i \rho_i (e_i + \overline{y})^2 \ge \min\{\rho_k\}\|\boldsymbol{e}\|_2^2$. Weighted orthogonality also gives $\overline{y} = -(\sum_i \rho_i e_i)/(\sum_i \rho_i)$, and consequently $\boldsymbol{y}^\top P\boldsymbol{y} = \sum_i \rho_i e_i^2 - (\sum_i \rho_i e_i)^2/(\sum_i \rho_i) \le \max\{\rho_k\}\|\boldsymbol{e}\|_2^2$. Thus

$$\min\{\rho_k\}\|\boldsymbol{e}\|_2^2 \le \boldsymbol{y}^\top P\boldsymbol{y} \le \max\{\rho_k\}\|\boldsymbol{e}\|_2^2. \tag{C.9}$$

Using the Laplacian quadratic form and (C.9), we obtain

$$\frac{NA_{\min}}{\max\{\rho_k\}} \le \frac{\boldsymbol{z}^\top Z\boldsymbol{z}}{\|\boldsymbol{z}\|_2^2} = \frac{\boldsymbol{y}^\top(D-A)\boldsymbol{y}}{\boldsymbol{y}^\top P\boldsymbol{y}} \le \frac{NA_{\max}}{\min\{\rho_k\}}. \tag{C.10}$$

The Courant–Fischer theorem now yields (C.5). Replacing (A, D, P, ρ_k) by (B, F, Q, n_k) proves (C.6). ∎

Appendix D

Null Space Components of $\widetilde{W}$ and $\widetilde{\xi}$

In this appendix, we show that $\widetilde{W}_{\mathcal{N}}$ (the projection of $\widetilde{W}$ onto $\mathcal{N}(Z)$) and $\widetilde{\xi}_{\mathcal{N}}$ (the projection of $\widetilde{\xi}$ onto $\mathcal{N}(\widehat{Z})$) are both zero. These results are used in the proofs of Theorems 9.13 and 9.14, respectively.

Lemma D.1 (Vanishing Velocity Null-Space Component)**.** $\widetilde{W}_{\mathcal{N}} \equiv \mathbf{0}$.

Proof. Since the null space of Z is spanned by $P^{\frac{1}{2}}\mathbf{1}$, it is sufficient to show that $\widetilde{W}^{\top} P^{\frac{1}{2}}\mathbf{1} = \mathbf{0}$. Write $\boldsymbol{u}^{\infty} = \frac{U^{\top}P\mathbf{1}}{\mathbf{1}^{\top}P\mathbf{1}}$ and $(\boldsymbol{U}^{\infty})^{\top} = \boldsymbol{u}^{\infty}\mathbf{1}^{\top}$, so that

$$\widetilde{W}^{\top} P^{\frac{1}{2}}\mathbf{1} = (W - W^{\infty})^{\top} P^{\frac{1}{2}}\mathbf{1} = U^{\top}P\mathbf{1} - (U^{\infty})^{\top} P\mathbf{1} = U^{\top}P\mathbf{1} - \frac{U^{\top}P\mathbf{1}}{\mathbf{1}^{\top}P\mathbf{1}}\mathbf{1}^{\top}P\mathbf{1} = \mathbf{0}. \quad \text{(D.1)}$$

■

Lemma D.2 (Vanishing Energy Null-Space Component)**.** $\widetilde{\xi}_{\mathcal{N}} \equiv \mathbf{0}$.

Proof. Since the null space of $\widehat{Z}$ is spanned by $Q^{\frac{1}{2}}\mathbf{1}$, it is sufficient to show that $\widetilde{\xi}^{\top} Q^{\frac{1}{2}}\mathbf{1} = 0$. Recall that $\widetilde{\xi} = Q^{-\frac{1}{2}}(\omega - \omega^{\infty})$; hence

$$\widetilde{\xi}^{\top} Q^{\frac{1}{2}}\mathbf{1} = (\omega - \omega^{\infty})^{\top}\mathbf{1} = \sum_{i=1}^{N} \omega_i - \sum_{i=1}^{N} \omega_i^{\infty} = 0, \quad \text{(D.2)}$$

where the fact that $\sum_i \omega_i = \sum_i \omega_i^{\infty}$ is a consequence of (9.50) in the proof of Proposition 9.7.

■

Appendix E

Some Theory of Manifolds

This appendix collects the differential-geometric definitions and results used in Chapter 11.

E.1 Definitions

Definition E.1 (Manifold, [50])**.** *Let $(\mathcal{M}, \tau)$ be a second-countable, Hausdorff topological space. If for each $p \in \mathcal{M}$, there exists an open $U \in \tau$ containing p and a homeomorphism $\varphi : U \to V$ for some open subset $V \subseteq \mathbb{R}^n$, then $\mathcal{M}$ is called an n-**dimensional manifold** or n-**manifold** for short. The tuple (U, φ) is called a **chart** or **coordinate patch**. A collection of charts $(U_\gamma)_{\gamma \in \Gamma}$ such that $\mathcal{M} \subseteq \cup_{\gamma \in \Gamma} U_\gamma$ is called an **atlas**.*

Definition E.2 (Smooth Map, [50])**.** *Let $\mathcal{M}, \mathcal{N}$ be smooth manifolds of dimension m and n, respectively. A function $F : \mathcal{M} \to \mathcal{N}$ is **smooth** if for every $p \in \mathcal{M}$, there exist smooth charts (U, φ) containing p and (V, ψ) containing $F(p)$ such that $F(U) \subseteq V$ and the coordinate representation $\psi \circ F \circ \varphi^{-1} : \varphi(U) \subseteq \mathbb{R}^m \to \psi(V) \subseteq \mathbb{R}^n$ is smooth.*

Definition E.3 (Curve, [50])**.** *Let $\mathcal{M}$ be a manifold. A continuous map $\gamma : I \to \mathcal{M}$, where $I \subseteq \mathbb{R}$ is an interval, is called a **curve**. It is a smooth curve when γ is smooth.*

Definition E.4 (Diffeomorphism, [50])**.** *$\mathcal{M}$, $\mathcal{N}$ smooth manifolds with or without boundary. A **diffeomorphism** from $\mathcal{M}$ to $\mathcal{N}$ is a smooth bijective map $F : \mathcal{M} \to \mathcal{N}$ that has a smooth inverse. We say $\mathcal{M}$ and $\mathcal{N}$ are diffeomorphic if such a function exists.*

Definition E.5 (Regular value, [50])**.** *Let $\mathcal{M}, \mathcal{N}$ be two smooth manifolds and $F : \mathcal{M} \to \mathcal{N}$ a smooth map. A point $y \in \mathcal{N}$ is a **regular value** if, for every $x \in F^{-1}(y)$, the derivative map $DF[x] : T_x\mathcal{M} \to T_y\mathcal{N}$ is surjective.*

Definition E.6 (Projection)**.** *A **projection** on a vector space V is a linear map $P : V \to V$ such that $P^2 = P$.*

Definition E.7 (Orthogonal Complement, Orthogonal Projection)**.** *Let X be a Hilbert space and $Y \subset X$ a closed subspace. Then*

$$X = Y \bigoplus Y^{\perp}, \qquad Y^{\perp} = \{x \in X : \langle x, y \rangle = 0 \text{ for every } y \in Y\}.$$

*The space $Y^{\perp}$ is called the **orthogonal complement of** Y. Any $x \in X$ can be uniquely represented as $x = y + w$, where $y \in Y$ and $w \in Y^{\perp}$. Define $P_Y : X \to X$ by $P_Y x = y$. Then P_Y is called the **orthogonal projection of** X **onto** Y.*

E.2 Theorems

The following result is usually called the **regular level set theorem**; it is also known as the **preimage theorem**. The formulation below is Corollary 5.14 in [50].

Theorem E.8 (Regular level set theorem (preimage theorem))**.** *Let $\mathcal{M}$ and $\mathcal{N}$ be smooth manifolds of dimensions p and q, respectively, and let $F : \mathcal{M} \to \mathcal{N}$ be smooth. If $y \in \mathcal{N}$ is a regular value of F, then $F^{-1}(y)$ is a closed smooth embedded submanifold of $\mathcal{M}$ of codimension q. In particular,*

$$\dim F^{-1}(y) = p - q.$$

Proposition E.9 (Proposition 3.23 in [50])**.** *Suppose $\mathcal{M}$ is a smooth manifold with or without boundary and $p \in \mathcal{M}$. Every $v \in T_p\mathcal{M}$ is the velocity of some smooth curve in $\mathcal{M}$.*

Theorem E.10 (Hilbert Projection Theorem)**.** *Let X be a Hilbert space over $\mathbb{R}$ or $\mathbb{C}$, $Y \subset X$ a closed subspace of X, and $x \in X$. Then, there exists a unique $y \in Y$ such that $\|x - y\| = \min_{z \in Y} \|x - z\|$. Another way to phrase this is $y = \arg\min_{z \in Y} \|x - z\|$.*

Proof. Let $d = \inf_{z \in Y} \|x - z\|$. This infimum exists because the set of distances is nonempty and bounded below by zero. There exists a sequence $(y_n) \subset Y$ such that $d_n = \|x - y_n\| \to d$. We argue that this sequence converges in Y to a point y, y attains this minimum, and that y is the unique minimizer. By the parallelogram law,

$$\begin{aligned} 2\|x - y_m\|^2 + 2\|x - y_n\|^2 &= \|y_n - y_m\|^2 + \|2x - (y_m + y_n)\|^2 \\ &= \|y_n - y_m\|^2 + 4\left\|x - \frac{y_m + y_n}{2}\right\|^2 \end{aligned}$$

$$\geq \|y_n - y_m\|^2 + 4d^2. \tag{E.1}$$

The last inequality follows from the fact that $\frac{1}{2}(y_m+y_n) \in Y$. This expression can be rewritten as

$$0 \leq \|y_m - y_n\|^2 \leq 2\|x - y_m\|^2 + 2\|y_n - x\|^2 - 4d^2. \tag{E.2}$$

Since the right-hand side tends to zero as $m, n \to \infty$, it follows that $\|y_m - y_n\| \to 0$. Thus (y_n) is a Cauchy sequence. Completeness of X gives $y_n \to y$ for some $y \in X$, and closedness of Y gives $y \in Y$. By continuity of the norm,

$$\begin{aligned} \|x - y\|^2 &= \lim_{n\to\infty} \|x - y_n\|^2 \\ &= d^2. \end{aligned} \tag{E.3}$$

Taking the square root of both sides yields the result we want. So, $\|x-y\| = \min_{z\in Y} \|x-z\|$. Now, we show that y is the unique minimizer of $\|x-z\|$ for $z \in Y$. Let $y^* \in Y$ be an element such that $\|x - y^*\| = \min_{z\in Y} \|x - z\|$. Applying the parallelogram law again yields

$$\begin{aligned} \|y - y^*\|^2 + 4\left\|x - \frac{y+y^*}{2}\right\|^2 &= 2\|x - y\|^2 + 2\|x - y^*\|^2 \\ &= 4d^2, \\ \|y - y^*\|^2 &= 4d^2 - 4\left\|x - \frac{y+y^*}{2}\right\|^2. \end{aligned} \tag{E.4}$$

Since $\frac{1}{2}(y + y^*) \in Y$,

$$\begin{aligned} \|y - y^*\|^2 &= 4d^2 - 4\left\|x - \frac{y+y^*}{2}\right\|^2 \\ &\leq 0. \end{aligned} \tag{E.5}$$

So, $y = y^*$ and it follows that y is the unique minimizer of $\|x - z\|$ for $z \in Y$. ■

Corollary E.11. *In the setting of Theorem E.10,* $P_Y x = y = \arg\min_{z\in Y} \|x - z\|$.

Proof. As previously mentioned, $X = Y \bigoplus Y^\perp$. So, $x = y^* + w$ for some $y^* \in Y$ and $w \in Y^\perp$. Note that $\|x - y\| \leq \|x - y^*\|$. Since $x - y^* \in Y^\perp$ and $y - y^* \in Y$, $x - y^* \perp y - y^*$. By the

Pythagorean theorem,

$$\|x - y^*\|^2 + \|y^* - y\|^2 = \|(x - y^*) + (y^* - y)\|^2 = \|x - y\|^2, \tag{E.6}$$

implying that $\|x - y^*\| \leq \|x - y\|$. It follows that $\|x - y\| = \|x - y^*\|$. Since y is the unique minimizer of $\|x - z\|$ for $z \in Y$, $y^* = y$. So, $P_Y x = y$. ■

Corollary E.12. *Let X be a Hilbert space, $Y \subset X$ a closed subspace, $x \in X$. Then $y = \arg\min_{z \in Y} \|x - z\|$ if and only if $\langle x - y, z \rangle = 0$ for all $z \in Y$.*

Proof. The proof is a string of if and only ifs. By Corollary E.11, $y = \arg\min_{z \in Y} \|x - z\| \iff x = y + w$ for $y \in Y, w \in Y^\perp$ unique $\iff x - y \in Y^\perp \iff \langle x - y, z \rangle = 0$ for all $z \in Y$. ■

Theorem E.13 (Quotient Manifold Theorem, 21.10 in [50])**.** *Suppose G is a Lie group acting smoothly, freely, and properly on a smooth manifold M. Then the orbit space M/G is a topological manifold of dimension equal to $\dim M - \dim G$, and has a unique smooth structure with the property that the quotient map $\pi : M \to M/G$ is a smooth submersion.*

Appendix F

$1x1v$ Prototype Code

The MATLAB code listed in this appendix is retained as the prototype associated with the Sod shock-tube workflow in Section 6.5.3. It implements the same $1x1v$ BGK/IMEX method and Euler-comparison workflow, but it is included for inspection, training, and subsequent verification rather than presented as production-quality software. The archived driver currently uses $x \in [-0.5, 0.5]$ and $N_v = 258$, whereas the figure description in Section 6.5.3 records $x \in [-0.75, 0.75]$. The listing is therefore linked to that run family, not claimed as the exact source of the displayed figure. The configuration mismatch must be reconciled during the planned code review before the listing is cited as an exact reproduction package.

The source files are available in the companion GitHub repository

https://github.com/stevenmwise/BGKSampleCodes.

They are also collected in the report's `code/` directory, whose `README.md` records the file roles, current run configuration, known verification gaps, and required validation measurements. The full listings remain in this appendix so that the report is self-contained and preserves the code version reviewed with this release.

F.1 Code Availability, Run Connection, and Verification Checklist

The appendix code is connected only to the Sod shock-tube material in Section 6.5.3; it is not the source for the relaxation, square-pulse, two-stream, Landau-damping, multispecies, ES-BGK, or DLRA figures elsewhere in the report. This explicit scope avoids implying code provenance that the archived materials do not establish.

A future archival release of the repository should include a tagged version, MATLAB release requirements, a license, a permanent archive identifier, and scripts that regenerate every reported table and figure. Before the repository is cited as an exact reproduction package, the prototype will be checked for array orientation, ghost-cell indexing, boundary conditions, parameter consistency, conservation of mass, momentum, and energy, positivity of density and temperature, and mesh/time-step convergence. Numerical values added to this report will be generated by those checked scripts rather than estimated from plotted curves.

F.2 Main Driver: `vlasovPoissonBGKMain.m`

```
%VLASOVPOISSONBGKMAIN Configure and run the archived 1x1v BGK prototype.
%
% Model:
%
% f_t + v * f_x + a * f_v = 1/tau * (M-f)
%
% The current configuration solves the Sod shock-tube problem with a = 0.
% In the limit tau -> 0 (equivalently lambda -> infinity), BGK approaches
% the Euler system. See README.md for the archived parameters, known report
% configuration mismatch, and verification work required before release.
%
% Scheme:
% IMEX RK: Explicit Advection, Implicit Collision.
%
clear;
clc;
tic
%
% Number of ghost cells:
del = 1;
%
% Collision time:
tau = 10^(-4);
%
% Theta values for minmod:
thetaX = 2;
thetaV = 2;
```

```
%
% Spatial domain:
xL = -0.5;
xR =  0.5;
%
% Number of cells:
Nx = 256;
%
% Cell edge points:
x = linspace(xL,xR,Nx+1);
%
% Cell center points:
xC = x(1:end-1) + 0.5 * (x(2:end) - x(1:end-1));
xCenter = [xC(1:del) , xC , xC(end-del+1:end)];
hX = xC(2) - xC(1);
%
% Velocity domain:
vMin = -10;
vMax =  10;
%
% Number of cells:
Nv = 258;
%
% Cell edge points:
v = linspace(vMin,vMax,Nv+1);
%
% Cell center points:
vC = v(1:end-1) + 0.5 * (v(2:end) - v(1:end-1));
vCenter = [vC(1:del) , vC , vC(end-del+1:end)];
hV = vC(2) - vC(1);
%
% Define time levels:
tInit = 0;
tFin  = 0.2;
%
% Vectors for ghost cells
gVx = [1,length(xC)]+del;
gVv = [1,length(vC)]+del;
%
tableNum = 8;
[Ae,be,ce,Ai,bi,ci] = butcherTable(tableNum);
butcher.Ae = Ae; butcher.Ai = Ai;
```

```
butcher.be = be; butcher.bi = bi;
butcher.ce = ce; butcher.ci = ci;
%
% Stiffness matrix for the Poisson solve...zero Dirichlet BC
% (as per the test problems). This is not needed for the Sod problem,
% but I kept it so I didn't have to change all my function
% dependencies.
%
temp = zeros(1,Nx);
temp(1)=2; temp(2)=-1;
A = toeplitz(temp);
A(1,1) = 3; A(end,end) = 3;
%
grid.del = del;
grid.gVx = gVx; grid.Nx = Nx; grid.xC = xC; grid.xCenter = xCenter;
grid.gVv = gVv; grid.Nv = Nv; grid.vC = vC; grid.vCenter = vCenter;
grid.hX = hX;
grid.hV = hV;
grid.xL   = xL  ; grid.xR   = xR  ;
grid.vMin = vMin; grid.vMax = vMax;
grid.tInit = tInit; grid.tFin = tFin;
grid.thetaX = thetaX;
grid.thetaV = thetaV;
%
testNum = 1;
[f] = initialCondition(testNum,grid);
%
[tVec,y,MAX] = vlasovPoissonBGKSolver(f,butcher,grid,tau,A,f);
%
figure(107)
pcolor(xC,vC,y);
title(['Contour plot of computed f, in the Velocity/Space', ...
  ' Domain, for Time t = ', num2str(tFin), ' '])
xlabel('x')
ylabel('v')
shading interp;
colormap(jet);
colorbar;
%
toc
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
%
% Embedded functions below:
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
function [f] = initialCondition(testNum,grid)
%
xC = grid.xC;
vC = grid.vC;
%
% Initialize solution array:
f = zeros(length(vC),length(xC),1);
%
switch testNum
  case 1 % tFin = 0.250
    nL = 1.00000; uL =  0.00000; thetaL = 0001.000;
    nR = 0.12500; uR =  0.00000; thetaR = 0000.800;
  case 2 % tFin = 0.150
    nL = 1.00000; uL = -2.00000; thetaL = 0000.400;
    nR = 1.00000; uR =  2.00000; thetaR = 0000.400;
  case 3 % tFin = 0.012
    nL = 1.00000; uL =  0.00000; thetaL = 1000.000;
    nR = 1.00000; uR =  0.00000; thetaR = 0000.010;
  case 4 % tFin = 0.035
    nL = 1.00000; uL =  0.00000; thetaL = 0000.010;
    nR = 1.00000; uR =  0.00000; thetaR = 0100.000;
  case 5 % tFin = 0.035
    nL = 5.99924; uL = 19.59750; thetaL = 0460.894;
    nR = 5.99242; uR = -6.19633; thetaR = 046.0950;
end
%
% Initial condition (make a separate function?):
%
for j = 1:length(vC)
  for i = 1:length(xC)
    if xC(i) <= 0.0
      f(j,i) = nL / sqrt(2*pi*thetaL) * exp(-(vC(j)-uL)^2 ...
        / (2*thetaL));
    else
      f(j,i) = nR / sqrt(2*pi*thetaR) * exp(-(vC(j)-uR)^2 ...
        / (2*thetaR));
    end
```

```
  end
end
%
% %Plot the initial conditions....
% figure(110)
% % pcolor(x_c,v_c,f);
% surf(x_c,v_c,f); view(-30,50); zlim([0,2.1]);
% % title(['Contour plot of f, in the Velocity/Space Domain,' ...
% ' for Time t = ',num2str(0),' '])
% xlabel('x')
% ylabel('v')
% shading interp;
% colormap(jet);
% colorbar;
%
end
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
function [Ae,be,ce,Ai,bi,ci] = butcherTable(table)
%
% Butcher Tableaux for the IMEX-RK scheme. Tables 2-6 come from
% the Pareschi & Russo paper. Table 7 is Backward Euler. Table 8
% (needs reference, from Cory).
%
switch table
  case 2
    Ae = [0,0 ; 1,0];
    be = [0.5 ; 0.5];
    ce = [0 ; 1];
%
    gam = 1-1/sqrt(2);
    Ai = [gam,0 ; 1-2*gam,gam];
    bi = [0.5 ; 0.5];
    ci = [gam ; 1-gam];
%
  case 3
    Ae = [0,0,0 ; 0,0,0 ; 0,1,0];
    be = [0 ; 0.5 ; 0.5];
    ce = [0 ; 0 ; 1];
%
    Ai = [0.5,0,0 ; -0.5,0.5,0 ; 0,0.5,0.5];
```

```
    bi = [0 ; 0.5 ; 0.5];
    ci = [0.5 ; 0 ; 1];
%
  case 4
    Ae = [0,0,0 ; 0.5,0,0 ; 0.5,0.5,0];
    be = [1/3 ; 1/3 ; 1/3];
    ce = [0 ; 0.5 ; 1];
%
    Ai = [0.25,0,0 ; 0,0.25,0 ; 1/3,1/3,1/3];
    bi = [1/3 ; 1/3 ; 1/3];
    ci = [0.25 ; 0.25 ; 1];
%
  case 5
    Ae = [0,0,0 ; 1,0,0 ; 0.25,0.25,0];
    be = [1/6 ; 1/6 ; 2/3];
    ce = [0 ; 1 ; 0.5];
%
    gam = 1-1/sqrt(2);
    Ai = [gam,0,0 ; 1-2*gam,gam,0 ; 0.5-gam,0,gam];
    bi = [1/6 ; 1/6 ; 2/3];
    ci = [gam ; 1-gam ; 0.5];
%
  case 6
    Ae = [0,0,0,0 ; 0,0,0,0 ; 0,1,0,0 ; 0,0.25,0.25,0];
    be = [0 ; 1/6 ; 1/6 ; 2/3];
    ce = [0 ; 0 ; 1 ; 0.5];
%
    alpha = 0.24169426078821;
    beta  = 0.06042356519705;
    eta   = 0.12915286960590;
    Ai = [alpha,0,0,0; -alpha,alpha,0,0; 0,1-alpha,alpha,0; ...
      beta,eta,0.5-beta-eta-alpha,alpha];
    bi = [0 ; 1/6 ; 1/6 ; 2/3];
    ci = [alpha ; 0 ; 1 ; 0.5];
%
  case 7
%
% Forward Euler:
    Ae = 0;
    be = 1;
    ce = 0;
%
```

```
% Backward Euler
    Ai = 1;
    bi = 1;
    ci = 1;
%
  case 8
    gam = 1 - 1/sqrt(2);
    delt = 1 - 1/(2*gam);
%
    Ae = [0,0,0 ; gam,0,0 ; delt,1-delt,0];
    be = [delt;1-delt;0];
    ce = [0;gam;1];
%
    Ai = [0,0,0 ; 0,gam,0 ; 0,1-gam,gam];
    bi = [0;1-gam;gam];
    ci = [0;gam;1];
end
end
```

Listing F.1: Main driver: `vlasovPoissonBGKMain.m`.

F.3 vlasovPoissonBGKSolver.m

```
function [t,Z,M] = vlasovPoissonBGKSolver(y0,butcher,grid,tau,A,f)
%VLASOVPOISSONBGKSOLVER Advance the prototype kinetic solution in time.
%   [T,Z,M] = VLASOVPOISSONBGKSOLVER(Y0,BUTCHER,GRID,TAU,A,F) applies an
%   IMEX Runge--Kutta method: advection is explicit and BGK relaxation is
%   implicit. Y0 and F store the initial grid function with indexing
%   (velocity,space). BUTCHER contains the explicit and implicit tableaux;
%   GRID contains mesh and boundary data; TAU is the collision time.
%
%   A is the retained Poisson stiffness matrix. It is not active in the Sod
%   configuration because the electric field is set to zero in DIVFLUX.
%   Prototype dimensions and invariants must be verified before release.
%
Ai = butcher.Ai;
bi = butcher.bi;
ci = butcher.ci;
Ae = butcher.Ae;
be = butcher.be;
ce = butcher.ce;
%
gVx = grid.gVx; Nx = grid.Nx; hX = grid.hX;
gVv = grid.gVv; Nv = grid.Nv; hV = grid.hV;
%
% A is the stiffness matrix for the Poisson solve.
%
% Number of ghost cells
del = grid.del;
%
% Spatial domain:
xL = grid.xL; xR = grid.xR;
%
% Velocity domain:
vMin = grid.vMin; vMax = grid.vMax;
%
% Define CFL(s)
CFLMaxX = min(hX) / max(abs(vMax),abs(vMin));
CFLMaxV = min(hV) / max( abs(xL) , abs(xR) );
%
% Define initial/final time levels:
```

```
tCurrent = grid.tInit;
tFinal   = grid.tFin ;
%
% size of grid function array y including ghost cells:
d = size(y0)+2*del;
%
% Initialize array of solutions:
y = zeros([d,2]);
%
% Input initial conditions into solution vector
y(gVv(1):gVv(2),gVx(1):gVx(2),1) = f;
y(:,:,1) = applyBoundaryConditions(y(:,:,1),grid);
%
% Number of stages of RK scheme:
s = length(ce);
%
% Initial vectors for RK scheme:
uo  = zeros([d,s]);
u   = zeros([d,s]);
fI  = zeros([d,s]);
fE  = zeros([d,s]);
M   = zeros([d,s]);
%
redo = 0;
EMax = 1;
t = 1;
%
while tCurrent < tFinal
%
  dt = min(0.24 * min(CFLMaxX,CFLMaxV),tFinal-tCurrent) ...
    / (1 + (redo > 0));
%
% This flag indicates that the electric field gets bigger than E_max.
% If so, the step needs to recalculate.
  redo = 0;
%
  uo(:,:,:) = 0;
  u(:,:,:) = 0;
  fI(:,:,:) = 0;
  fE(:,:,:) = 0;
  M(:,:,:) = 0;
  uo(:,:,1) = y(:,:,1);
```

```
  [~,M(:,:,1)] = BGKCollision(tCurrent,y(:,:,1),grid,tau,tCurrent);
  u(:,:,1) = (tau*uo(:,:,1) + dt*Ai(1,1)*M(:,:,1)) ...
    / (tau + dt*Ai(1,1));
  u(:,:,1) = applyBoundaryConditions(u(:,:,1),grid);
  fI(:,:,1) = BGKCollision(tCurrent+ci(1)*dt,u(:,:,1),grid, ...
    tau,tCurrent);
%
  [fE(:,:,1),re] = divFlux(tCurrent+ce(1)*dt,u(:,:,1),grid,EMax,A);
  redo = redo+re;

  for r = 2:s
    previousStages = 1:r-1;
    uo(:,:,r) = y(:,:,1) ...
      + dt * reshape(Ae(r,previousStages) ...
      * reshape(permute(fE(:,:,previousStages),[3,1,2]),r-1, ...
      (Nv+2*del)*(Nx+2*del)), Nv+2*del, Nx+2*del) ...
      + dt * reshape(Ai(r,previousStages) ...
      * reshape(permute(fI(:,:,previousStages),[3,1,2]),r-1, ...
      (Nv+2*del)*(Nx+2*del)), Nv+2*del, Nx+2*del);
    uo(:,:,r) = applyBoundaryConditions(uo(:,:,r),grid);
%
    [~,M(:,:,r)] = BGKCollision(tCurrent+ci(r)*dt,uo(:,:,r), ...
      grid,tau,tCurrent);
%
    u(:,:,r) = (tau*uo(:,:,r) + dt*Ai(r,r)*M(:,:,r)) ...
      / (tau + dt*Ai(r,r));
    u(:,:,r) = applyBoundaryConditions(u(:,:,r),grid);

    [fE(:,:,r),re] = divFlux(tCurrent+ce(r)*dt,u(:,:,r),grid,EMax,A);
    redo = redo+re;

    fI(:,:,r) = BGKCollision(tCurrent+ci(r)*dt,u(:,:,r), ...
      grid,tau,tCurrent);
  end
%
  y(:,:,2) = y(:,:,1) ...
    + (dt * reshape(be' * reshape(permute(fE,[3,1,2]),s, ...
    (Nv+2*del)*(Nx+2*del)), Nv+2*del,Nx+2*del) ...
    +  dt * reshape(bi' * reshape(permute(fI,[3,1,2]),s, ...
    (Nv+2*del)*(Nx+2*del)), Nv+2*del,Nx+2*del) ...
    ) * (1 - (redo > 0));
  y(:,:,2) = applyBoundaryConditions(y(:,:,2),grid);
```

```
%
  y(:,:,1) = y(:,:,2);
  tCurrent = tCurrent + dt*(1 - (redo>0));
%
% This computes the moments and Maxwellian at the final time step.
  if tCurrent == tFinal
    BGKCollision(tCurrent+ci(r)*dt,y(:,:,1),grid,tau,tCurrent);
  end
end
%
% Return solution at final time
Z = y(gVv(1):gVv(2),gVx(1):gVx(2),1);
%
end

function y = applyBoundaryConditions(y,grid)
%APPLYBOUNDARYCONDITIONS Fill zero-gradient spatial and velocity ghosts.
del = grid.del;
gVx = grid.gVx;
gVv = grid.gVv;

for i = 1:del
  y(:,gVx(1)-i) = y(:,gVx(1));
  y(:,gVx(2)+i) = y(:,gVx(2));
end
for j = 1:del
  y(gVv(1)-j,:) = y(gVv(1),:);
  y(gVv(2)+j,:) = y(gVv(2),:);
end
end
```

Listing F.2: `vlasovPoissonBGKSolver.m`.

F.4 BGKCollision.m

```
function [z,M] = BGKCollision(t,y,grid,tau,tCurrent)
%BGKCOLLISION Evaluate the BGK relaxation term and local Maxwellian.
%   [Z,M] = BGKCOLLISION(T,Y,GRID,TAU,TCURRENT) computes velocity moments
%   from the phase-space grid function Y, constructs the corresponding
%   Maxwellian M, and returns Z = (M-Y)/TAU. GRID supplies cell centers,
%   quadrature spacing, and ghost-cell indices. The time arguments are kept
%   for compatibility with the prototype solver and its diagnostic plots.
%
%   Prototype note: array orientation and ghost-cell indexing must be checked
%   as part of the verification checklist in README.md before release.
%
del = grid.del;
hx = grid.hX; gVx = grid.gVx; x = grid.xC;
hv = grid.hV; gVv = grid.gVv; v = grid.vC;
%
% Grid function array must have ghost cells already.
M = zeros(length(x)+2*del,length(v)+2*del);
nDens = zeros(1,length(x));
mDens = zeros(1,length(x));
EDens = zeros(1,length(x));

uDens = zeros(1,length(x));
theta = zeros(1,length(x));
%
% Moment integrals: n, n*u, and E.
nDens(1:length(x)) = hv * sum(y(gVv(1):gVv(2),1+del:length(x)+del));
mDens(1:length(x)) = hv * sum(v(:).*y(gVv(1):gVv(2), ...
  1+del:length(x)+del));
EDens(1:length(x)) = 0.5 * hv * sum(v(:).^2 .* y(gVv(1):gVv(2), ...
  1+del:length(x)+del));

uDens(:) = mDens(:)./nDens(:);

theta(:) = 2*EDens(:)./nDens(:)-uDens(:).^2;

M(gVx(1):gVx(2),gVv(1):gVv(2)) = nDens(:) ./ sqrt(2*pi*theta(:)) ...
  .* exp(-0.5 * abs(v - uDens(:)).^2 ./ theta(:));
%
```

```
% Ghost cells: zero flow boundaries.
for i = 1:del
  M(:,gVv(1)-i) = M(:,gVv(1));
  M(:,gVv(2)+i) = M(:,gVv(2));
end
for j = 1:del
  M(gVx(1)-j,:) = M(gVx(1),:);
  M(gVx(2)+j,:) = M(gVx(2),:);
end
%
if tCurrent == 0.2
%
% initial shock happens here.
  x0 = 0;
  gam = 3;
  uL = [1,0,1]; uR = [0.125,0,0.1];
%
  pDens = nDens .* theta;
%
  [nSod,uSod,pSod,eSod] = sodSoln(x,tCurrent,uL,uR,gam,x0);
%
  figure(1)
  plot(x,nDens,'b','LineWidth',3)
  hold on
  plot(x,nSod,'LineWidth',3)
  xlim([-0.5,0.5])
  ylim([0,1])
  set(gca,'fontsize',24)
  legend('BGK','Euler')
  hold off
%
  figure(2)
  plot(x,uDens,'b','LineWidth',3)
  hold on
  plot(x,uSod,'LineWidth',3)
  xlim([-0.5,0.5])
  ylim([0,1])
  set(gca,'fontsize',24)
  legend('BGK','Euler')
  hold off
%
  figure(3)
```

```
  plot(x,theta,'b','LineWidth',3)
  hold on
  plot(x,2*eSod,'LineWidth',3)
  xlim([-0.5,0.5])
  ylim([0,2]) % may need to shift this
  set(gca,'fontsize',24)
  legend('BGK','Euler')
  hold off
end
%
z = (M'-y)/tau;
%
% M is a [Nv]X[Nx] array after transpose
M = M';
%
end
```

Listing F.3: `BGKCollision.m`

F.5 divFlux.m

```
function [z,redo] = divFlux(t,y,grid,EMax,A)
%DIVFLUX Compute the finite-volume advection operator.
%   [Z,REDO] = DIVFLUX(T,Y,GRID,EMAX,A) reconstructs limited interface
%   states and applies upwind fluxes in position and velocity. The Sod
%   prototype sets the electric field to zero, so the velocity-advection
%   contribution vanishes. T, EMAX, and A are retained for compatibility
%   with the more general Vlasov--Poisson prototype interface.
%
% y = grid function of f_{ij}^k values for fixed k
%
del = grid.del;
hx = grid.hX; gVx = grid.gVx; x = grid.xCenter; thetaX = grid.thetaX;
hv = grid.hV; gVv = grid.gVv; v = grid.vCenter; thetaV = grid.thetaV;
%
fP = zeros(size(y));
fM = zeros(size(y));
gP = zeros(size(y));
gM = zeros(size(y));
%
% construct slopes
[sX(:,:),sV(:,:)] = slopeReconstruction(y,grid);
%
% % Spatial fluxes
fP(:,gVx(1):gVx(2)) = (y(:,gVx(1):gVx(2)) ...
  + 0.5*hx*sX(:,gVx(1):gVx(2))) .* v(:).*(v>0)' ...
  +(y(:,gVx(1)+1:gVx(2)+1) - 0.5*hx*sX(:,gVx(1)+1:gVx(2)+1)) ...
  .* v(:).*(v<=0)' ;
fM(:,gVx(1):gVx(2)) = (y(:,gVx(1)-1:gVx(2)-1) ...
  + 0.5*hx*sX(:,gVx(1)-1:gVx(2)-1)) .* v(:).*(v>0)' ...
  +(y(:,gVx(1):gVx(2)) - 0.5*hx*sX(:,gVx(1):gVx(2))) ...
  .* v(:).*(v<=0)' ;
%
% The electric field is unnecessary for this test, so we set it to
% zero here.
E = zeros(1,length(x));
redo = 0;

% % Velocity fluxes
```

```
gP(gVv(1):gVv(2),:) = (y(gVv(1):gVv(2),:) ...
  + 0.5*hv*sV(gVv(1):gVv(2),:)) .* E.*(E>0) ...
  + (y(gVv(1)+1:gVv(2)+1,:) - 0.5*hv*sV(gVv(1)+1:gVv(2)+1,:)) ...
  .* E.*(E<=0) ;
gM(gVv(1):gVv(2),:) = (y(gVv(1)-1:gVv(2)-1,:) ...
  + 0.5*hv*sV(gVv(1)-1:gVv(2)-1,:)) .* E.*(E>0) ...
  + (y(gVv(1):gVv(2),:) - 0.5*hv*sV(gVv(1):gVv(2),:)) ...
  .* E.*(E<=0) ;

z = - (fP - fM)/hx - (gP - gM)/hv;

end
```

Listing F.4: `divFlux.m`

F.6 slopeReconstruction.m

```
function [sX,sV] = slopeReconstruction(y,grid)
%SLOPERECONSTRUCTION Form limited slopes in position and velocity.
%   [SX,SV] = SLOPERECONSTRUCTION(Y,GRID) builds centered and one-sided
%   slope candidates and applies MINMOD. GRID supplies spacings, limiter
%   parameters, and interior/ghost-cell index ranges.
%
%   Prototype note: the mixed use of spatial and velocity index ranges is a
%   priority item in the verification checklist in README.md.
%
sX = zeros(size(y));
sV = zeros(size(y));
%
del = grid.del;
hx = grid.hX; gVx = grid.gVx; thetaX = grid.thetaX;
hv = grid.hV; gVv = grid.gVv; thetaV = grid.thetaV;
%
SV(:,:,1) = (y(gVv(1)+1:gVv(2)+1,gVx(1):gVx(2)) ...
  - y(gVv(1)-1:gVv(2)-1,gVx(1):gVx(2))) / 2 ;
SV(:,:,2) = thetaV * (y(gVx(1):gVv(2),gVx(1):gVx(2)) ...
  - y(gVv(1)-1:gVv(2)-1,gVx(1):gVx(2))) ;
SV(:,:,3) = thetaV * (y(gVx(1)+1:gVv(2)+1,gVx(1):gVx(2)) ...
  - y(gVv(1):gVv(2),gVx(1):gVx(2))) ;
%
sV(gVv(1):gVv(2),gVx(1):gVx(2)) = minMod(SV)/hv;
%
SX(:,:,1) = (y(gVv(1):gVv(2),gVx(1)+1:gVx(2)+1) ...
  - y(gVv(1):gVv(2),gVx(1)-1:gVx(2)-1)) / 2 ;
SX(:,:,2) = thetaX * (y(gVx(1):gVv(2),gVx(1):gVx(2)) ...
  - y(gVv(1):gVv(2),gVx(1)-1:gVx(2)-1)) ;
SX(:,:,3) = thetaX * (y(gVx(1):gVv(2),gVx(1)+1:gVx(2)+1) ...
  - y(gVv(1):gVv(2),gVx(1):gVx(2))) ;
%
sX(gVv(1):gVv(2),gVx(1):gVx(2)) = minMod(SX)/hx;
%
for i = 1:del
  sX(:,gVx(1)-i) = sX(:,gVx(1));
  sX(:,gVx(2)+i) = sX(:,gVx(2));
end
```

```
for j = 1:del
  sV(gVv(1)-j,:) = sV(gVv(2),:);
  sV(gVv(2)+j,:) = sV(gVv(1),:);
end
%
end
```

Listing F.5: `slopeReconstruction.m`

F.7 minMod.m

```
function M = minMod(x)
%MINMOD Apply the three-argument minmod limiter along the third dimension.
%   M = MINMOD(X) expects the three candidate slopes in X(:,:,1:3).
%
M(:,:) = 0.25*abs(sign(x(:,:,1))+sign(x(:,:,2))) ...
  .*(sign(x(:,:,1))+sign(x(:,:,3))).* min(abs(x),[],3);
%
end
```

Listing F.6: minMod.m

F.8 sodSoln.m

```
function [n,u,p,e] = sodSoln(x,tCurrent,phiL,phiR,gam,x0)
%SODSOLN Evaluate an exact ideal-gas Euler Riemann solution.
%   [N,U,P,E] = SODSOLN(X,T,PHIL,PHIR,GAM,X0) returns density, velocity,
%   pressure, and specific internal energy at the points X and time T.
%   PHIL and PHIR contain (density, velocity, pressure), and X0 is the
%   initial discontinuity location.

nL = phiL(1); uL = phiL(2); pL = phiL(3);
nR = phiR(1); uR = phiR(2); pR = phiR(3);

if tCurrent < 0
  error('sodSoln:InvalidTime','The solution time must be nonnegative.');
end

n = zeros(size(x));
u = zeros(size(x));
p = zeros(size(x));

if tCurrent == 0
  left = x <= x0;
  n(left) = nL; u(left) = uL; p(left) = pL;
  n(~left) = nR; u(~left) = uR; p(~left) = pR;
  e = p./((gam-1)*n);
  return
end

pStar = pStarSolve(phiL,phiR,gam,1e-10);
aL = sqrt(gam*pL/nL);
aR = sqrt(gam*pR/nR);

[fL,~] = pressureFunction(pStar,nL,pL,aL,gam);
[fR,~] = pressureFunction(pStar,nR,pR,aR,gam);
uStar = 0.5*(uL+uR+fR-fL);
similarityCoordinate = (x-x0)/tCurrent;

for index = 1:numel(x)
  waveSpeed = similarityCoordinate(index);
  if waveSpeed <= uStar
```

```
    [n(index),u(index),p(index)] = sampleLeftState( ...
      waveSpeed,nL,uL,pL,aL,pStar,uStar,gam);
  else
    [n(index),u(index),p(index)] = sampleRightState( ...
      waveSpeed,nR,uR,pR,aR,pStar,uStar,gam);
  end
end

e = p./((gam-1)*n);
end

function [n,u,p] = sampleLeftState(speed,nK,uK,pK,aK,pStar,uStar,gam)
%SAMPLELEFTSTATE Sample the left shock or rarefaction.
if pStar > pK
  shockSpeed = uK-aK*sqrt((gam+1)*pStar/(2*gam*pK) ...
    +(gam-1)/(2*gam));
  if speed <= shockSpeed
    n = nK; u = uK; p = pK;
  else
    pressureRatio = pStar/pK;
    n = nK*(pressureRatio+(gam-1)/(gam+1)) ...
      / ((gam-1)*pressureRatio/(gam+1)+1);
    u = uStar; p = pStar;
  end
else
  aStar = aK*(pStar/pK)^((gam-1)/(2*gam));
  headSpeed = uK-aK;
  tailSpeed = uStar-aStar;
  if speed <= headSpeed
    n = nK; u = uK; p = pK;
  elseif speed >= tailSpeed
    n = nK*(pStar/pK)^(1/gam);
    u = uStar; p = pStar;
  else
    u = 2*(aK+(gam-1)*uK/2+speed)/(gam+1);
    soundSpeed = 2*(aK+(gam-1)*(uK-speed)/2)/(gam+1);
    n = nK*(soundSpeed/aK)^(2/(gam-1));
    p = pK*(soundSpeed/aK)^(2*gam/(gam-1));
  end
end
end
```

```
function [n,u,p] = sampleRightState(speed,nK,uK,pK,aK,pStar,uStar,gam)
%SAMPLERIGHTSTATE Sample the right shock or rarefaction.
if pStar > pK
  shockSpeed = uK+aK*sqrt((gam+1)*pStar/(2*gam*pK) ...
    +(gam-1)/(2*gam));
  if speed >= shockSpeed
    n = nK; u = uK; p = pK;
  else
    pressureRatio = pStar/pK;
    n = nK*(pressureRatio+(gam-1)/(gam+1)) ...
      / ((gam-1)*pressureRatio/(gam+1)+1);
    u = uStar; p = pStar;
  end
else
  aStar = aK*(pStar/pK)^((gam-1)/(2*gam));
  headSpeed = uK+aK;
  tailSpeed = uStar+aStar;
  if speed >= headSpeed
    n = nK; u = uK; p = pK;
  elseif speed <= tailSpeed
    n = nK*(pStar/pK)^(1/gam);
    u = uStar; p = pStar;
  else
    u = 2*(-aK+(gam-1)*uK/2+speed)/(gam+1);
    soundSpeed = 2*(aK-(gam-1)*(uK-speed)/2)/(gam+1);
    n = nK*(soundSpeed/aK)^(2/(gam-1));
    p = pK*(soundSpeed/aK)^(2*gam/(gam-1));
  end
end
end

function [f,df] = pressureFunction(p,nK,pK,aK,gam)
%PRESSUREFUNCTION Evaluate one side of the exact pressure equation.
if p > pK
  coefficientA = 2/((gam+1)*nK);
  coefficientB = (gam-1)*pK/(gam+1);
  rootTerm = sqrt(coefficientA/(p+coefficientB));
  f = (p-pK)*rootTerm;
  df = rootTerm*(1-(p-pK)/(2*(p+coefficientB)));
else
  exponent = (gam-1)/(2*gam);
  pressureRatio = p/pK;
```

```
  f = 2*aK/(gam-1)*(pressureRatio^exponent-1);
  df = pressureRatio^(-(gam+1)/(2*gam))/(nK*aK);
end
end
```

Listing F.7: `sodSoln.m`

F.9 pStarSolve.m

```
function pStar = pStarSolve(phiL,phiR,gam,TOL)
%PSTARSOLVE Compute the pressure in the star region of a Riemann problem.
%   PSTAR = PSTARSOLVE(PHIL,PHIR,GAM,TOL) solves the exact Euler pressure
%   equation by Newton iteration. PHIL and PHIR contain density, velocity,
%   and pressure; GAM is the ratio of specific heats.

rhoL = phiL(1); uL = phiL(2); pL = phiL(3);
rhoR = phiR(1); uR = phiR(2); pR = phiR(3);

if rhoL <= 0 || rhoR <= 0 || pL <= 0 || pR <= 0 || gam <= 1
  error('pStarSolve:InvalidState', ...
    'Densities and pressures must be positive and gamma must exceed one.');
end

aL = sqrt(gam*pL/rhoL);
aR = sqrt(gam*pR/rhoR);

% Positive primitive-variable Riemann-solver pressure estimate.
pOld = max(TOL,0.5*(pL+pR) ...
  - 0.125*(uR-uL)*(rhoL+rhoR)*(aL+aR));

maxIterations = 100;
for iteration = 1:maxIterations
  [fL,dfL] = pressureFunction(pOld,rhoL,pL,aL,gam);
  [fR,dfR] = pressureFunction(pOld,rhoR,pR,aR,gam);
  pNew = max(TOL,pOld-(fL+fR+uR-uL)/(dfL+dfR));
  relativeChange = 2*abs(pNew-pOld)/(pNew+pOld);
  pOld = pNew;
  if relativeChange <= TOL
    pStar = pNew;
    return
  end
end

error('pStarSolve:NoConvergence', ...
  'Star-region pressure iteration did not converge.');
end
```

```
function [f,df] = pressureFunction(p,rho,pK,aK,gam)
%PRESSUREFUNCTION Evaluate one side of the exact pressure equation.
if p > pK
  coefficientA = 2/((gam+1)*rho);
  coefficientB = (gam-1)*pK/(gam+1);
  rootTerm = sqrt(coefficientA/(p+coefficientB));
  f = (p-pK)*rootTerm;
  df = rootTerm*(1-(p-pK)/(2*(p+coefficientB)));
else
  exponent = (gam-1)/(2*gam);
  pressureRatio = p/pK;
  f = 2*aK/(gam-1)*(pressureRatio^exponent-1);
  df = pressureRatio^(-(gam+1)/(2*gam))/(rho*aK);
end
end
```

Listing F.8: `pStarSolve.m`

Appendix G

Bulk-Quantity Notation Map

This appendix consolidates the macroscopic quantities introduced in Chapter 2 and fixes the conventions used throughout the report. Unless stated otherwise, $i \in \{1, \ldots, N\}$ indexes a species, summation over species is written explicitly, and d is the velocity-space dimension. Chapter 2 has $d = 3$; later reduced models may use another value of d.

The central convention is frame based. A hat denotes a central moment measured relative to the species velocity $\boldsymbol{u}_i$. Its unhatted counterpart is measured relative to the barycentric mixture velocity $\boldsymbol{u}$. A quantity without a species index is a mixture total, except for $\hat{\theta}$ and θ, which are the indicated number-density-weighted mixture temperatures. Total kinetic-energy density is always denoted by ω; internal-energy density is denoted by e. Momentum density is written explicitly as $\rho_i \boldsymbol{u}_i$ or $\rho \boldsymbol{u}$, without introducing another symbol.

The report uses $K_n = \lambda_{\text{mfp}}/L$ in the introductory scale discussion and ε in the nondimensional kinetic equations; both denote the Knudsen ratio. The symbol λ without the subscript mfp denotes a collision frequency, and $\tau = \lambda^{-1}$ denotes its relaxation time in the BGK chapters. Chapter 11 also uses τ locally for a numerical time step, always introduced by $t_{n+1} = t_n + \tau$; Δt is used for time steps elsewhere. The bold symbol $\boldsymbol{\tau}$ denotes the viscous-stress tensor and is distinct from either scalar use.

G.1 Kinetic Inputs, Densities, and Velocity Frames

Symbol	Definition	Meaning and convention
$f_i(\boldsymbol{x}, \boldsymbol{v}, t)$	Kinetic distribution of species i	Its velocity moments generate all species bulk fields.
m_i	Particle mass of species i	A parameter, not a density.
n_i	$\int f_i \, \mathrm{d}\boldsymbol{v}$	Species number density.
ρ_i	$m_i n_i$	Species mass density.
n	$\sum_i n_i$	Mixture number density.
ρ	$\sum_i \rho_i$	Mixture mass density.
$\boldsymbol{u}_i$	$n_i^{-1} \int f_i \boldsymbol{v} \, \mathrm{d}\boldsymbol{v}$	Species bulk velocity.
$\rho_i \boldsymbol{u}_i$	$m_i \int f_i \boldsymbol{v} \, \mathrm{d}\boldsymbol{v}$	Species momentum density; this explicit product is the standard notation.
$\boldsymbol{u}$	$\rho^{-1} \sum_i \rho_i \boldsymbol{u}_i$	Barycentric or mixture velocity.
$\rho \boldsymbol{u}$	$\sum_i \rho_i \boldsymbol{u}_i$	Total mixture momentum density.
$\boldsymbol{V}_i$	$\boldsymbol{u}_i - \boldsymbol{u}$	Diffusion velocity; $\sum_i \rho_i \boldsymbol{V}_i = \boldsymbol{0}$.
$\boldsymbol{c}_i$	$\boldsymbol{v} - \boldsymbol{u}_i$	Peculiar velocity in the species frame.
$\boldsymbol{c}$	$\boldsymbol{v} - \boldsymbol{u}$	Peculiar velocity in the barycentric frame.

G.2 Energy Densities and Temperatures

Energy densities.

Symbol	Definition	Meaning and convention
ω_i	$\frac{m_i}{2}\int f_i\lvert\boldsymbol{v}\rvert^2\,\mathrm{d}\boldsymbol{v}$	Species total kinetic-energy density in the laboratory frame.
ω	$\sum_i \omega_i$	Mixture total kinetic-energy density.
$\hat{e}_i$	$\frac{m_i}{2}\int f_i\lvert\boldsymbol{c}_i\rvert^2\,\mathrm{d}\boldsymbol{v}$	Species-frame internal-energy density.
$\hat{e}$	$\sum_i \hat{e}_i$	Sum of the species-frame internal energies; excludes diffusion energy.
e_i	$\frac{m_i}{2}\int f_i\lvert\boldsymbol{c}\rvert^2\,\mathrm{d}\boldsymbol{v}$	Contribution of species i to barycentric internal energy.
e	$\sum_i e_i$	Total barycentric internal-energy density; includes diffusion energy.
e_{sp}	e/ρ, or e/n in explicitly unit-mass models	Specific internal energy; the normalization must be stated locally.

Temperatures and temperature maps.

Symbol	Definition	Meaning and convention
$\hat{\theta}_i$	$\frac{2\hat{e}_i}{dn_i}$	Species-frame temperature in energy units.
θ_i	$\frac{2e_i}{dn_i}$	Barycentric temperature contribution assigned to species i.
$\hat{\theta}$	$\frac{2\hat{e}}{dn} = \frac{1}{n}\sum_i n_i\hat{\theta}_i$	Species-frame mixture temperature.
θ	$\frac{2e}{dn} = \frac{1}{n}\sum_i n_i\theta_i$	Barycentric mixture temperature.
ϑ	θ/m	Thermal variance (Maxwellian velocity variance) in the unary ES–BGK chapter.
$\vartheta_i(\boldsymbol{u}_i,\omega_i)$	$\frac{2\omega_i}{dn_i} - \frac{m_i}{d}\lvert\boldsymbol{u}_i\rvert^2$	Chapter 9 recovery map whose value is $\hat{\theta}_i$; it is distinct from unary thermal variance.

In particular,

$$e_i = \hat{e}_i + \frac{\rho_i}{2}|\boldsymbol{V}_i|^2, \qquad e = \hat{e} + \frac{1}{2}\sum_i \rho_i|\boldsymbol{V}_i|^2, \tag{G.1}$$

$$\theta_i = \hat{\theta}_i + \frac{m_i}{d}|\boldsymbol{V}_i|^2, \qquad \theta = \hat{\theta} + \frac{1}{dn}\sum_i \rho_i|\boldsymbol{V}_i|^2. \tag{G.2}$$

Thus a hat does not mean "mixture average": it identifies the species velocity frame.

G.3 Stress and Flux Quantities

Symbol	Definition	Meaning and convention
M_i	$m_i \int f_i \boldsymbol{v} \otimes \boldsymbol{v} \, \mathrm{d}\boldsymbol{v}$	Species momentum-flow tensor.
M	$\sum_i \mathsf{M}_i$	Total momentum-flow tensor.
$\widehat{\mathsf{P}}_i$	$m_i \int f_i \boldsymbol{c}_i \otimes \boldsymbol{c}_i \, \mathrm{d}\boldsymbol{v}$	Species-frame pressure tensor.
P_i	$m_i \int f_i \boldsymbol{c} \otimes \boldsymbol{c} \, \mathrm{d}\boldsymbol{v}$	Species contribution to the barycentric pressure tensor.
$\widehat{\mathsf{P}}$	$\sum_i \widehat{\mathsf{P}}_i$	Sum of species-frame pressure tensors.
P	$\sum_i \mathsf{P}_i$	Total barycentric pressure tensor.
p	$d^{-1}\mathrm{tr}(\mathsf{P})$ in isotropic contexts	Scalar thermodynamic pressure; lower-case p distinguishes it from P.
Π	$\mathsf{P} - p\mathsf{I}$	Deviatoric pressure tensor; $\mathrm{tr}(\Pi) = 0$.
$\widehat{\Pi}_i$	$\widehat{\mathsf{P}}_i - n_i\hat{\theta}_i\mathsf{I}$	Species-frame deviatoric pressure tensor used in the multispecies ES–BGK model; $\mathrm{tr}(\widehat{\Pi}_i) = 0$.
$\boldsymbol{\tau}$	$2\mu\mathsf{D}^{\circ}(\boldsymbol{u})$	Newtonian viscous-stress tensor; at first Chapman–Enskog order, $\Pi = -\varepsilon\boldsymbol{\tau} + \mathcal{O}(\varepsilon^2)$.
$\boldsymbol{r}_i$	$\frac{m_i}{2} \int f_i \boldsymbol{v} \lvert\boldsymbol{v}\rvert^2 \, \mathrm{d}\boldsymbol{v}$	Species energy flux.
$\boldsymbol{r}$	$\sum_i \boldsymbol{r}_i$	Total energy flux.
$\boldsymbol{\chi}_i$	$\frac{m_i}{2} \int f_i \boldsymbol{c}_i \lvert\boldsymbol{c}_i\rvert^2 \, \mathrm{d}\boldsymbol{v}$	Species-frame heat flux.
$\boldsymbol{\chi}$	$\sum_i \boldsymbol{\chi}_i$	Sum of species-frame heat fluxes.
$\boldsymbol{q}_i$	$\frac{m_i}{2} \int f_i \boldsymbol{c} \lvert\boldsymbol{c}\rvert^2 \, \mathrm{d}\boldsymbol{v}$	Species contribution to barycentric heat flux.
$\boldsymbol{q}$	$\sum_i \boldsymbol{q}_i$	Total barycentric heat flux.

The main decompositions are

$$\omega_i = \frac{\rho_i}{2}|\boldsymbol{u}_i|^2 + \hat{e}_i, \qquad \mathsf{M}_i = \rho_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i + \widehat{\mathsf{P}}_i, \tag{G.3}$$

$$\boldsymbol{r}_i = \omega_i \boldsymbol{u}_i + \widehat{\mathsf{P}}_i \boldsymbol{u}_i + \boldsymbol{\chi}_i, \qquad \mathsf{P}_i = \widehat{\mathsf{P}}_i + \rho_i \boldsymbol{V}_i \otimes \boldsymbol{V}_i, \tag{G.4}$$

$$\mathsf{M} = \rho \boldsymbol{u} \otimes \boldsymbol{u} + \mathsf{P}, \qquad \omega = \frac{\rho}{2}|\boldsymbol{u}|^2 + e, \tag{G.5}$$

$$\mathsf{P} = p\mathsf{I} + \Pi, \qquad \mathrm{tr}(\Pi) = 0, \tag{G.6}$$

$$\boldsymbol{r} = \omega \boldsymbol{u} + \mathsf{P}\boldsymbol{u} + \boldsymbol{q}, \qquad \frac{1}{2}\mathrm{tr}(\widehat{\mathsf{P}}_i) = \hat{e}_i, \qquad \frac{1}{2}\mathrm{tr}(\mathsf{P}_i) = e_i. \tag{G.7}$$

G.4 Information and Entropy Quantities

These Chapter 2 bulk fields obey a separate sign convention: information decreases and physical entropy increases.

Symbol	Definition or relation	Meaning
h	$\sum_i \int f_i \ln f_i \, \mathrm{d}\boldsymbol{v}$	Information density.
H	$\int_\Omega h \, \mathrm{d}\boldsymbol{x}$	Total information.
s	$-h$	Entropy density.
S	$-H = \int_\Omega s \, \mathrm{d}\boldsymbol{x}$	Total entropy.
$\boldsymbol{J}_h$	$\sum_i \int \boldsymbol{v} f_i \ln f_i \, \mathrm{d}\boldsymbol{v}$	Information flux.
$\boldsymbol{J}_s$	$-\boldsymbol{J}_h$	Entropy flux.
σ_h	$\sum_i \int Q_i \ln f_i \, \mathrm{d}\boldsymbol{v} \le 0$	Collisional information rate; here $Q_i = Q_i[f_i, \boldsymbol{f}]$.
σ_s	$-\sigma_h \ge 0$	Entropy production rate.

G.5 Conserved Variables and Special Cases

The species conserved-variable triple is

$$(\rho_i,\ \rho_i \boldsymbol{u}_i,\ \omega_i), \tag{G.8}$$

while collisions conserve the mixture totals

$$(\rho,\ \rho \boldsymbol{u},\ \omega). \tag{G.9}$$

Species momentum and energy may be exchanged, so the individual collision sources for $\rho_i \boldsymbol{u}_i$ and ω_i need not vanish even though their sums do.

For a unary gas, the species and barycentric frames coincide:

$$\boldsymbol{u}_1 = \boldsymbol{u}, \quad \hat{e}_1 = e_1 = e, \quad \hat{\theta}_1 = \theta_1 = \theta, \quad \widehat{\mathsf{P}}_1 = \mathsf{P}_1 = \mathsf{P}, \quad \boldsymbol{\chi}_1 = \boldsymbol{q}_1 = \boldsymbol{q}. \tag{G.10}$$

In an isotropic Maxwellian, $\mathsf{P} = p\mathsf{I}$ and $p = n\theta$. Local italic capital letters such as the matrices P and Q in Chapter 9 are algebraic matrices, not pressure tensors; physical pressure tensors always use the sans-serif symbol P.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] Pierre Andries, Patrick Le Tallec, Jean-Philippe Perlat, and Benoît Perthame. The Gaussian-BGK model of Boltzmann equation with small Prandtl number. *European Journal of Mechanics - B/Fluids*, 19(6):813–830, 2000.

[3] Philippe Arnault and Sébastien Guisset. Chapman–Enskog derivation of multicomponent Navier–Stokes equations. *Physics of Plasmas*, 29(9):090901, 2022.

[4] Markus Bachmayr, Henrik Eisenmann, Emil Kieri, and André Uschmajew. Existence of dynamical low-rank approximations to parabolic problems. *Mathematics of Computation*, 90(330):1799–1830, 2021.

[5] Claude Bardos, François Golse, and C. David Levermore. Fluid dynamic limits of kinetic equations. I. formal derivations. *Journal of Statistical Physics*, 63(1–2):323–344, 1991.

[6] 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. *Physical Review*, 94(3):511–525, 1954.

[7] Graeme A. Bird. *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Clarendon Press, Oxford, 1994.

[8] Alexander V Bobylev, Marzia Bisi, Maria Groppi, Giampiero Spiga, and Irina F Potapenko. A general consistent BGK model for gas mixtures. *Kinetic and Related Models*, 11(6):1377–1393, 2018.

[9] Alberto Bressan. Notes on the boltzmann equation: Lecture notes for a summer course at sissa. Technical report, Pennsylvania State University, 2005.

[10] Carlo Cercignani. *The Boltzmann Equation and Its Applications*, volume 67 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1988.

[11] Carlo Cercignani. *Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations*, volume 21 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge, 2000.

[12] Carlo Cercignani, Reinhard Illner, and Mario Pulvirenti. *The Mathematical Theory of Dilute Gases*, volume 106 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994.

[13] Gianluca Ceruti, Jonas Kusch, and Christian Lubich. A rank-adaptive robust integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 62(4):1149–1174, 2022.

[14] Gianluca Ceruti and Christian Lubich. An unconventional robust integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 62(1):23–44, 2022.

[15] 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 Mathematical Library. Cambridge University Press, 1990.

[16] C. K. Chu. Kinetic-theoretic description of the formation of a shock wave. *The Physics of Fluids*, 8(1):12–22, 1965.

[17] F. Coron and B. Perthame. Numerical passage from kinetic to fluid equations. *SIAM Journal on Numerical Analysis*, 28(1):26–42, 1991.

[18] Anaïs Crestetto, Christian Klingenberg, and Marlies Pirner. Kinetic/fluid micro-macro numerical scheme for a two component gas mixture. *Multiscale Modeling & Simulation*, 18(2):970–998, 2020.

[19] Alec Dektor and Lukas Einkemmer. Interpolatory dynamical low-rank approximation for the 3+3d Boltzmann–BGK equation. *Journal of Computational Physics*, 547:114515, 2026.

[20] Giacomo Dimarco and Lorenzo Pareschi. Numerical methods for kinetic equations. *Acta Numerica*, 23:369–520, 2014.

[21] Shuyu Dong, Bin Gao, Wen Huang, and Kyle A. Gallivan. On the analysis of optimization with fixed-rank matrices: A quotient geometric view, 2022.

[22] Lukas Einkemmer. A low-rank algorithm for weakly compressible flow. *SIAM Journal on Scientific Computing*, 41(5):A2795–A2814, 2019.

[23] Lukas Einkemmer, Jingwei Hu, and Lexing Ying. An efficient dynamical low-rank algorithm for the Boltzmann–BGK equation close to the compressible viscous flow regime. *SIAM Journal on Scientific Computing*, 43(5):B1057–B1080, 2021.

[24] Lukas Einkemmer and Christian Lubich. A low-rank projector-splitting integrator for the Vlasov–Poisson equation. *SIAM Journal on Scientific Computing*, 40(5):B1330–B1360, 2018.

[25] Francis Filbet and Shi Jin. An Asymptotic Preserving Scheme for the ES-BGK Model of the Boltzmann Equation. *Journal of Scientific Computing*, 46(2):204–224, 2011.

[26] Rodney O. Fox, Frédérique Laurent, and Aymeric Vié. Conditional hyperbolic quadrature method of moments for kinetic equations. *Journal of Computational Physics*, 365:269–293, 2018.

[27] C. Kristopher Garrett and Cory D. Hauck. A fast solver for implicit integration of the Vlasov-Poisson system in the Eulerian framework. *SIAM Journal on Scientific Computing*, 40(2):B483–B506, 2018.

[28] François Golse. The Boltzmann equation and its hydrodynamic limits. In C. M. Dafermos and Eduard Feireisl, editors, *Handbook of Differential Equations: Evolutionary Equations*, volume 2, pages 159–301. Elsevier, 2005.

[29] Harold Grad. On the kinetic theory of rarefied gases. *Communications on Pure and Applied Mathematics*, 2(4):331–407, 1949.

[30] Harold Grad. Asymptotic theory of the Boltzmann equation. *Physics of Fluids*, 6(2):147–181, 1963.

[31] Jeffrey Haack, Cory D Hauck, Christian F Klingenberg, Marlies Pirner, and Sandra Warnecke. Numerical schemes for a multi-species BGK model with velocity-dependent collision frequency. *Journal of Computational Physics*, 473:111729, Jan 2023.

[32] Jeffrey R. Haack, Cory D. Hauck, and Michael S. Murillo. A conservative, entropic multispecies BGK model. *Journal of Statistical Physics*, 168(4):826–856, 2017.

[33] Evan Habbershaw, Ryan S. Glasby, Jeffrey R. Haack, Cory D. Hauck, and Steven M. Wise. Asymptotic relaxation of moment equations for a multi-species, homogeneous BGK model. *SIAM Journal on Applied Mathematics*, 85(1):294–313, 2025.

[34] Evan Habbershaw, Cory D. Hauck, and Steven M. Wise. Implicit update of the moment equations for a multi-species, homogeneous BGK model. *SIAM Journal on Numerical Analysis*, 63(2):881–907, 2025.

[35] Evan Habbershaw and Steven M. Wise. Year four report. Progress report, University of Tennessee, Knoxville, October 2025. Unpublished progress report.

[36] Stewart Harris. *Introduction to the Theory of the Boltzmann Equation*. Dover Publications, 2004.

[37] Cory D. Hauck and Ryan G. McClarren. A collision-based hybrid method for time-dependent, linear, kinetic transport equations. *Multiscale Modeling & Simulation*, 11(4):1197–1227, 2013.

[38] Cory D. Hauck and Stefan Schnake. A predictor-corrector strategy for adaptivity in dynamical low-rank approximations. *SIAM Journal on Matrix Analysis and Applications*, 44(3):971–1005, 2023.

[39] Lowell H. Holway, Jr. New statistical models for kinetic theory: Methods of construction. *The Physics of Fluids*, 9(9):1658–1673, 09 1966.

[40] Jingwei Hu and Yubo Wang. An adaptive dynamical low rank method for the nonlinear Boltzmann equation. *Journal of Scientific Computing*, 92(2):75, 2022.

[41] J.H. Jeans. LXX. The persistence of molecular velocities in the kinetic theory of gases. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(48):700–703, 1904.

[42] J.H. Jeans. *The Dynamical Theory of Gases*. Cambridge University Press, 1904(1921).

[43] Doheon Kim, Myeong-Su Lee, and Seok-Bae Yun. Entropy production estimate for the ES-BGK model with the correct Prandtl number. *Journal of Mathematical Analysis and Applications*, 514(2):126323, 2022.

[44] Christian Klingenberg and Marlies Pirner. Existence, uniqueness and positivity of solutions for BGK models for mixtures. *Journal of Differential Equations*, 264(2):702–727, 09 2018.

[45] Christian Klingenberg, Marlies Pirner, and Gabriella Puppo. A consistent kinetic model for a two-component mixture with an application to plasma. *Kinetic and Related Models*, 10(2):445–465, 2017.

[46] Christian Klingenberg, Marlies Pirner, and Gabriella Puppo. Kinetic ES–BGK models for a multi-component gas mixture. In *Theory, Numerics and Applications of Hyperbolic Problems II*, volume 237 of *Springer Proceedings in Mathematics & Statistics*, pages 195–208. Springer, 2018.

[47] Othmar Koch and Christian Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007.

[48] Gilberto Medeiros Kremer. *An Introduction to the Boltzmann Equation and Transport Processes in Gases*. Interaction of Mechanics and Mathematics. Springer, Berlin, 2010.

[49] Jonas Kusch, Steffen Schotthöfer, and Alexandra Walter. An augmented backward-corrected projector splitting integrator for dynamical low-rank training. *SIAM Journal on Mathematics of Data Science*, 8(3):820–849, 2026.

[50] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2 edition, 2012.

[51] C. David Levermore. Moment closure hierarchies for kinetic theories. *Journal of Statistical Physics*, 83(5–6):1021–1065, 1996.

[52] Shuangqian Liu, Tong Yang, and Huijiang Zhao. Compressible Navier–Stokes approximation to the Boltzmann equation. *Journal of Differential Equations*, 256(11):3770–3816, 2014.

[53] Christian Lubich and Ivan V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 54(1):171–188, 2014.

[54] L. Mieussens. Discrete velocity model and implicit scheme for the BGK equation of rarefied gas dynamics. *Mathematical Models and Methods in Applied Sciences*, 10(8):1121–1149, 2000.

[55] L. Mieussens. Schemes for Boltzmann-BGK equation in plane and axisymmetric geometries. *Journal of Computational Physics*, 162:429–466, 2000.

[56] Stéphane Mischler. Uniqueness for the BGK-equation in $\mathbb{R}^N$ and rate of convergence for a semi-discrete scheme. *Differential and Integral Equations*, 9(5):1119 – 1138, 1996.

[57] Clément Mouhot and Robert M. Strain. Spectral gap and coercivity estimates for linearized Boltzmann collision operators without angular cutoff. *Journal de Mathématiques Pures et Appliquées*, 87(5):515–535, 2007.

[58] Lorenzo Pareschi and Giovanni Russo. Implicit–explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *Journal of Scientific Computing*, 25:129–155, 2005.

[59] B. Perthame and M. Pulvirenti. Weighted L^∞ bounds and uniqueness for the Boltzmann BGK model. *Archive for Rational Mechanics and Analysis*, 125(3):289–295, 1993.

[60] Benoît Perthame. Global existence to the BGK model of Boltzmann equation. *Journal of Differential Equations*, 82(1):191–205, 1989.

[61] Sandra Pieraccini and Gabriella Puppo. Implicit–explicit schemes for BGK kinetic equations. *Journal of Scientific Computing*, 32(1):1–28, 2007.

[62] Marlies Pirner. *Kinetic Modelling of Gas Mixtures*. PhD thesis, Julius-Maximilians-Universität Würzburg, 2018.

[63] Gabriella Puppo. Kinetic models of BGK type and their numerical integration, 2019. arXiv:1902.08311.

[64] Laure Saint-Raymond. From the BGK model to the Navier–Stokes equations. *Annales Scientifiques de l'École Normale Supérieure*, 36(2):271–317, 2003.

[65] Steffen Schotthöfer, Timon Klein, and Jonas Kusch. A geometric framework for momentum-based optimizers for low-rank training. In *Advances in Neural Information Processing Systems*, volume 38, pages 100015–100045, 2025.

[66] Steffen Schotthöfer and M. Paul Laiu. Federated dynamical low-rank training with global loss convergence guarantees, 2024.

[67] Steffen Schotthöfer, M. Paul Laiu, Martin Frank, and Cory D. Hauck. Structure-preserving neural networks for the regularized entropy-based closure of a linear, kinetic, radiative transport equation. *Journal of Computational Physics*, 533:113967, 2025.

[68] Steffen Schotthöfer, Tianbai Xiao, Martin Frank, and Cory D. Hauck. Structure preserving neural networks: A case study in the entropy closure of the Boltzmann equation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19406–19433. PMLR, 2022.

[69] Steffen Schotthöfer, H. Lexie Yang, and Stefan Schnake. Dynamical low-rank compression of neural networks with robustness under adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 38, pages 155754–155783, 2025.

[70] Steffen Schotthöfer, Emanuele Zangrando, Gianluca Ceruti, Francesco Tudisco, and Jonas Kusch. GeoLoRA: Geometric integration for parameter efficient fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[71] Steffen Schotthöfer, Emanuele Zangrando, Jonas Kusch, Gianluca Ceruti, and Francesco Tudisco. Low-rank lottery tickets: Finding efficient low-rank neural networks via matrix differential equations. In *Advances in Neural Information Processing Systems*, volume 35, pages 20051–20063, 2022.

[72] Minwoo Shin, Cory D. Hauck, and Ryan G. McClarren. A collision-based hybrid method for the BGK equation. *Journal of Computational Physics*, 501:112784, 2024.

[73] Gary A Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of Computational Physics*, 27(1):1–31, 1978.

[74] H. Struchtrup. *Macroscopic Transport Equations for Rarefied Gas Flows: Approximation Methods in Kinetic Theory*. Interaction of Mechanics and Mathematics. Springer, Berlin, 2005.

[75] Henning Struchtrup and Manuel Torrilhon. Regularization of Grad's 13 moment equations: Derivation and linear analysis. *Physics of Fluids*, 15(9):2668–2680, 2003.

[76] Pierre-Yves C. R. Taunay and Michael E. Mueller. Quadrature-based moment methods for kinetic plasma simulations. *Journal of Computational Physics*, 473:111700, 2023.

[77] Eleuterio Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*. Springer, Germany, 2nd edition, 1999.

[78] Cédric Villani. A review of mathematical topics in collisional kinetic theory. In Susan Friedlander and Denis Serre, editors, *Handbook of Mathematical Fluid Dynamics*, volume 1, pages 71–305. Elsevier, 2002.

[79] W.G. Vincenti and C.H. Kruger. *Introduction to Physical Gas Dynamics*. Krieger, Malabar, Florida, 1986.

[80] Homer F Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

[81] K. Xu. A gas-kinetic BGK scheme for the Navier-Stokes equations and its connection with artificial dissipation and Godunov methods. *Journal of Computational Physics*, 171:289–335, 2001.

[82] Emanuele Zangrando, Steffen Schotthöfer, Gianluca Ceruti, Jonas Kusch, and Francesco Tudisco. Geometry-aware training of factorized layers in tensor Tucker format. In *Advances in Neural Information Processing Systems*, volume 37, pages 129743–129773, 2024.

[83] Y. Zheng and H. Struchtrup. Burnett equations for the ellipsoidal statistical BGK model. *Continuum Mechanics and Thermodynamics*, 16(1):97–108, 2004.