ABSTRACT

Title of dissertation: ENTROPY-BASED MOMENT CLOSURES IN SEMICONDUCTOR MODELS

Cory David Hauck, Doctor of Philosophy, 2006

Dissertation directed by: Professor C. David Levermore
Applied Mathematics and Scientific Computation

We investigate aspects of *entropy-based moment closures* which are used to simplify kinetic models of particle systems. Closures of this type use variational principles to formally generate balance laws for velocity moments of a kinetic density. These balance laws form a symmetric hyperbolic system of partial differential equations that satisfies an analog of Boltzmann's famous *H-Theorem*. However, in spite of this elegant structure, practical implementation of entropy-based closures requires that several analytical and computational issues be settled.

Our presentation is devoted to the development of electron transport models in semiconductor devices. In this context, balance laws for velocity moments are generally referred to as *hydrodynamic models*. Such models provide a reasonable alternative to kinetic and Monte Carlo approaches, which are usually expensive, and the

well-known *drift-diffusion* model, which is much simpler but a has a limited range of validity.

We first analyze the minimization problem that defines the entropy closure. It is known that there are physically relevant cases for which this problem is ill-posed. Using a dual formulation, we find so-called *complementary slackness conditions* which give a geometric interpretation of ill-posed cases in terms of the Lagrange multipliers of the minimization problem. Under reasonable assumptions, we show that these cases are rare in a very precise sense.

We also develop pertubations of well-posed entropy-based closures, thereby making them useful for modeling systems with heat flux and anisotropic stress. Heat flux has long been known to be an important component of electron transport in semiconductors. However, we also observe that anisotropy in the stress tensor also plays an important role in regions of high electric field. This conclusion is made based on our simulations of two different devices.

Finally, we devise a new split scheme for hydrodynamic models. The splitting is based on the balance of forces in the hydrodynamic model that recovers the drift-diffusion equation in the asymptotic limit of small mean-free-path. This scheme removes numerically stiffness and excessive dissipation typically associated with standard shock-capturing schemes in the drift-diffusion limit. In addition, it significantly reduces numerical current oscillations near material junctions.

ENTROPY-BASED MOMENT CLOSURES IN SEMICONDUCTOR MODELS

by

Cory David Hauck

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:

Professor C. David Levermore, Chair
Professor Stuart S. Antman
Professor Jian-Guo Liu
Professor Martin Peckerar
Professor André L. Tits

DEDICATION

To the memories of my grandfather, R. D. Driskill
and my friend, C. I. Kwon;
To my mother, Karen, and my father, Bob;
To my son, Miles; and most of all,
To my wife, Heather.

ACKNOWLEDGEMENTS

I thank Dave Levermore for serving as my advisor and for introducing me to the field of kinetic theory. I am grateful for his support and guidance over the last five years. Throughout the many ups and downs of my graduate career, Dave has been a guidepost for me—not only with his words, but also by his example. I will carry the lessons he has taught with me long after I have left Maryland.

I thank Stuart Antman for his many excellent lectures, both in the classroom and in various seminars. I thank him for his support and advice, for the individual instruction he gave me in his office, and for many friendly discussions about mathematics, physics, and teaching.

I thank Andre Tits for an excellent course in optimal control theory. Although technically an engineering course, it was one of the best math classes I've ever taken. I would also like to thank Andre for his consultation on various topics in optimization and for his insights and suggestions which were immensely helpful in writing Chapter 4 of this dissertation.

I thank Jian-Guo Liu for serving on my committee and for his continuing interest in my work.

I thank Professor Martin Peckerar for agreeing to serve on my committee and for providing an engineering perspective to modeling semiconductors.

I thank Eitan Tadmor for many helpful conversations about the numerics of hyperbolic PDE.

I thank Professor Orazio Muscato for giving me the Monte Carlo data that was used for comparisons in Chapter 5.

I thank Jack Calcut, Henry King, and Elmar Winkelnkemper for their assistance in understanding concepts in topology and algebra with which I was unfamiliar.

Contents

List of Abbreviations **xiv**

1 Introduction **1**

1.1 A Kinetic Model of Electron Transport 3

1.2 Drift-Diffusion 6

1.3 Moment Systems 10

1.4 Closures Based on Entropy Methods 12

1.5 Entropy Closures for Well-Posed Minimization Problems 13

1.6 Issues in Numerical Simulation 17

1.7 Preview of Results 19

2 From Semi-Classical Transport to the Drift-Diffusion Model **24**

2.1 Basic Physics of Semiconductors 24

2.1.1 Crystal Structure 25

2.1.2 Energy Bands 26

2.1.3 Doping 27

2.2 Mathematical Formulation 28

2.2.1 Collision Operators 29

2.2.2 Generation/Recombination Operators 31

2.2.3 Notions of Equilibrium 32
2.3 Simplifications 37
2.3.1 Unipolar Model 38
2.3.2 Low Density Approximation 39
2.3.3 Parabolic Band Approximation 41
2.4 Drift-Diffusion Equations 43
2.4.1 Scaling 45
2.4.2 Chapman-Enskog Expansion 49

3 Hydrodynamic Models **52**

3.1 Mathematical Background 53
3.1.1 Formal Kinetic Properties 54
3.1.2 Moment Systems 58
3.1.3 Evaluation of the Collision Operator 60
3.2 The Bløtekjær, Baccarani, Wordemann (BBW) Model 62
3.2.1 Closure 64
3.2.2 Reduction to Second Order 67
3.2.3 The Baccarani-Wordemann Expressions 70
3.2.4 Discussion of the BBW Model 72
3.3 The Anile and Pennisi (AP) Model 75
3.3.1 Extended Thermodynamics 76
3.3.2 Closure 79
3.3.3 Reduction to Second Order 81

3.3.4 Discussion of the AP Closure 83
3.4 Entropy-Based Closures 84
3.4.1 Relationship with Extended Thermodynamics 85
3.4.2 Relative Entropy Formulation 88
3.4.3 Well-Posedness of Entropy-Based Closures 93
3.4.4 Generalized BGK Collision Operators 95
3.4.5 Examples 105
3.5 Perturbations of Entropy-Based Moment Closures 113
3.5.1 General Setting 114
3.5.2 Balance 117
3.5.3 Entropy Dissipation 118
3.5.4 Examples 119

4 Entropy Minimization and Realizability 124

4.1 Preliminaries 127
4.1.1 Admissible Spaces 127
4.1.2 Construction of Admissible Spaces 128
4.1.3 The Entropy Functional 134
4.1.4 Cones 135
4.1.5 Semi-algebraic Sets 141
4.2 Entropy Minimization 143
4.2.1 Formulation 143
4.2.2 The Dual Function 145

4.2.3 Duality Theorems 152

4.3 The Relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ 155

4.3.1 Justification of the Formal Legendre Duality 155

4.3.2 Examples 158

4.3.3 Degenerate Densities 159

4.3.4 Geometry of $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ 163

4.3.5 Examples 170

4.4 Appendix: Duality Theorems 174

5 Simulation of an n^+-n-n^+ Diode **181**

5.1 Reduction to One Dimension 182

5.1.1 The Case $\mathbf{m} = (1, v, \frac{1}{2}|v|^2)^T$ 183

5.1.2 The Case $\mathbf{m} = (1, v, v \vee v)^T$ 184

5.2 The Models 186

5.2.1 Bløtekjær-Type Models. 186

5.2.2 The Benchmark Device 194

5.2.3 Boundary Conditions 195

5.3 The Numerical Scheme 196

5.3.1 Finite Volume Formulation 197

5.3.2 Flux Evaluation 198

5.3.3 Remaining Discretization 205

5.3.4 Remarks 207

5.4 Numerical Results 209

5.4.1 Bløtekjær-Type Models 210

5.4.2 Anile-Pennisi Models 212

5.4.3 Entropy-Based Models 213

6 Computational Issues: Stiffness and Balance 249

6.1 The Benchmark Problem 253

6.2 Drift-Diffusion 255

6.2.1 Non-Dimensionalization 255

6.2.2 The Drift-Diffusion Scaling 258

6.2.3 The Drift-Diffusion Limit 260

6.2.4 Physical Validity 260

6.2.5 Preview of Numerical Issues 262

6.3 Numerical Background 264

6.3.1 A Model Problem 264

6.3.2 Systems of Balance Laws 268

6.3.3 Previous Computations of the Hydrodynamic Model 270

6.4 A New Splitting Approach to the Hydrodynamic Model 274

6.4.1 Two Step Splitting 277

6.4.2 Three Step Splitting 278

6.5 Details of the Scheme 280

6.5.1 Spatial Discretization 282

6.5.2 Time Discretization 288

6.6 Numerical Results. 289

6.6.1 The Transition Regime 290
6.6.2 The Drift-Diffusive Regime 290
6.6.3 Convergence Analysis 299
6.7 Conclusions and Discussion 302

7 Simulation of a Unipolar MESFET Device **304**
7.1 Modeling Two Dimensional Transport 308
7.1.1 Equations in Two Dimensions 308
7.1.2 The Benchmark Device 313
7.2 Numerical Scheme 316
7.2.1 Discretization of Convective Terms 316
7.2.2 Discretization of Diffusive Terms 318
7.2.3 Multigrid Poisson Solver 320
7.2.4 Discretization of Field and Collision Terms 328
7.3 Numerical Results 328

Bibliography **353**

List of Tables

6.1 Convergence rate of scheme $S3S$-I-2. 301

6.2 Convergence rate of scheme $S1$-2. 302

6.3 Convergence rate of scheme $S2$-2. 302

List of Figures

5.1 The n-n^+-n diode. 195

5.2 Electron concentration for Bløtekjær-type models. 215

5.3 Electron concentration for Bløtekjær-type models, magnified view. . . 216

5.4 Electron current for Bløtekjær-type models. 217

5.5 Electron velocity for Bløtekjær-type models. 218

5.6 Electron velocity for Bløtekjær-type models, magnified view. 219

5.7 Thermal energy for Bløtekjær-type models. 220

5.8 Electron energy for Bløtekjær-type models. 221

5.9 Electric field for Bløtekjær-type models. 222

5.10 Heat flux for Bløtekjær-type models. 223

5.11 Heat flux for Bløtekjær-type models, magnified view. 224

5.12 Electron energy flux for Bløtekjær-type models 225

5.13 Electron energy flux for Bløtekjær-type models, magnidified view. . 226

5.14 Electron concentration for AP models. 227

5.15 Electron concentration for AP models, magnified view. 228

5.16 Electron current for AP models. 229

5.17 Electron velocity for AP models. 230

5.18 Electron velocity for AP models, magnified view 231

5.19 Electron thermal energy for AP models. 232

5.20 Electron energy for AP models. 233

5.21 Electric field for AP models. 234

5.22 Heat flux for AP models. 235

5.23 Electron energy flux for AP models. 236

5.24 Electron energy flux for AP models, magnidified view. 237

5.25 Electron concentration for entropy-based models. 238

5.26 Electron concentration for entropy-based models, magnified view. . . 239

5.27 Electron current for entropy-based models. 240

5.28 Electron velocity for entropy-based models. 241

5.29 Electron velocity for entropy-based models, magnified view. 242

5.30 Electron thermal energy for entropy-based models 243

5.31 Electric field for entropy-based models. 244

5.32 Electron heat flux for entropy-based models. 245

5.33 Electron energy flux for entropy-based models. 246

5.34 Anistropic stress $\frac{m_e \sigma}{n}$ for AP and PEB models. 247

5.35 Components of thermal energy for perturbed Gaussian closures. . . . 248

6.1 The one dimensional n-n^+-n diode of length L. 254

6.2 Steady state current oscillations for Scheme $S1$. 275

6.3 Steady state current oscillations for Scheme $S2$. 276

6.4 Steady state results for $S1$ and $S3E$-1. 291

6.5 Steady state current oscillations for scheme $S3$. 292

6.6 Steady state results for $S3E$-1 vs. drift-diffusion results with 1600 meshpoints. 295

6.7 Steady state results for $S3E$-1 vs. drift-diffusion results with 200 meshpoints. 296

6.8 Steady state current oscillations in the drift diffusion regime. 297

6.9 Current oscillations for various values of ε. 298

7.1 Schematic representation of the MESFET device. 315

7.2 Steady-state electron concentration. 331

7.3 Steady-state momentum, x-component. 332

7.4 Steady-state momentum, y-component. 333

7.5 Steady-state velocity, x-component. 334

7.6 Steady-state velocity, y-component. 335

7.7 Steady-state potential. 336

7.8 Steady-state electric field, x-component. 337

7.9 Steady-state electric field, y-component. 338

7.10 Steady-state energy profile. 339

7.11 Steady-state temperature. 340

7.12 Anisotropy in the Maxwellian closure 341

7.13 Anisotropy in the Gaussian closure. 342

List of Abbreviations

$\mathbb{R}^d$	Space of real d-dimensional vectors
$\mathbb{S}^{d-1}$	Unit sphere in $\mathbb{R}^d$
k_B	Boltzmann constant
m_e^*	effective electron mass
q_e	magnitude of electron charge
$\epsilon = \epsilon(x)$	electric permittivity
θ_ℓ, T_ℓ	semiconductor lattice temperature, $T_\ell = \frac{m_e^* \theta_\ell}{k_B}$
$\mu = \mu(x)$	electron mobility
$a = a(x)$	electron diffusivity
x, ∇_x	position vector, spatial gradient
v, ∇_v	velocity vector, velocity gradient
t, ∂_t	time, time derivative
$F = F(x, v, t)$	kinetic density
M_ℓ	lattice Maxwellian distribution
$\mathcal{C}$	collision operator
$\Phi = \Phi(x, t)$	electric potential
$E = E(x, t)$	electric field, $E = -\nabla_x \Phi$
$D = D(x)$	doping profile
$\mathbf{m} = \mathbf{m}(v)$	vector of polynomials in v

$\langle\cdot\rangle$	integration over velocity space
$\boldsymbol{\rho} = \boldsymbol{\rho}(x,t)$	spatial density, $\langle \mathbf{m}F \rangle$
$n = n(x,t)$	electron concentration
$u = u(x,t)$	electron mean velocity
$\theta = \theta(x,t)$	electron temperature
$\Theta = \Theta(x,t)$	electron temperature matrix
$\Sigma = \Sigma(x,t)$	electron anisotropic stress tensor
$q = q(x,t)$	electron heat flux vector
$Q = Q(x,t)$	electron heat flux tensor
$\mathcal{M}_{n,u,\theta}$	Maxwellian distribution
$\mathcal{G}_{n,u,\Theta}$	Gaussian distribution
$U \vee V$	symmetric tensor product of tensors U and V
$U^{\vee s}$	symmetric tensor power, $U^{\vee s} = U^{\vee(s-1)} \vee U$

Chapter 1

Introduction

This work address several mathematical and computational issues on the topic of moment systems in kinetic theory, particularly as they pertain to modeling electron transport in semiconductor devices. In a kinetic description, the distribution of a large number of particles is interpreted mathematically by a kinetic density F, which is a non-negative function in phase space. Moment systems are models that simplify the kinetic description of these particles by tracking the evolution of only a handful of physically relevant statistical averages of F, called moments. Moment systems require a closure, meaning that assumptions about the functional form of F must be made in order to make up for the loss of information that occurs in the averaging process.

Moment systems have several important functions. They can be used as stand-alone models, presumably with the flexibility to improve accuracy by adding more moments. They can be used in highly efficient hybrid schemes for modeling multi-scale phenomenon. Such schemes combine a variety of different models in such a way as to maximize efficiency for a given level of accuracy. Thus, expensive models are used only in regimes where they are absolutely needed. Finally, moments systems

can be used as preconditioners for more complicated models that may suffer from numerical stiffness.

One of the great challenges of creating moment systems is to find an appropriate closure—one that retains the fundamental physical and mathematical structure of the original kinetic description. As the name suggests, *entropy-based* moment closures impose that structure through the optimization of a convex functional which is thermodynamic potential directly related to the kinetic entropy. In [51, 53], it is shown that entropy-based closures formally generate moment systems which are symmetric hyperbolic systems of partial differential equations (usually referred to as balance laws). Furthermore, it is shown that these moment systems satisfy an analog of Boltzmann's H-Theorem, i.e., that solutions dissipate a Lyapunov function derived from the thermodynamic potential and that the dissipation vanishes only for closures that assume the distribution of particles is in thermodynamic equilibrium.

Below, Section 1.1 gives the kinetic formulation of electron transport in semiconductors that will be the focus of this dissertation. Section 1.2 is about the drift-diffusion model that accurately approximates the kinetic description in certain physical regimes. Section 1.3 is an introduction to moment systems in general, and Section 1.4 describes entropy-based closures. Sample closures are given in Section 1.5. In Section 1.6, numerical challenges related to stiffness and asymptotic limits are discussed. Finally, Section 1.7 gives a preview of results and lays out a map for the remainder of the dissertation.

1.1 A Kinetic Model of Electron Transport

Semiconductors are crystalline materials composed of atoms that are bound together in a periodic lattice. Because the number of atoms is very large, their common energy levels decouple into many closely spaced levels which can be treated as a continuous band. Rather than being identified with a particular shell of a particular atom, electrons in a semiconductor are characterized by the energy band in which they are found. Charge is transported in semiconductors by the flow of *carrier electrons*, which are unbound electrons in the conduction band of a semiconductor, and *holes*, which are vacancies in the valence band.

In the model which is the focus of this dissertation, we consider only the flow of carrier electrons, which are treated as classical particles with effective mass m_e^* [59] and charge $-q_e$. These electrons exist in a semiconductor material which is represented mathematically by a bounded domain $\Omega \subset \mathbb{R}^3$, and their distribution in position-momentum phase space is given by a kinetic density $F = F(x, v, t)$ that is defined for positions $x \in \Omega \subset \mathbb{R}^3$, velocities $v \in \mathbb{R}^3$, and times $t \geq 0$. The interpretation of F is that for any $\Lambda \subset \Omega \times \mathbb{R}^3$, the integral

$$\int \int_{\Lambda} F(x, v, t)\, dvdx$$

gives the number of particles at time t with positions x and velocities v such that $(x, v) \in \Lambda$. The evolution of F is governed by the so-called Boltzmann transport

equation:

$$\partial_t F + v \cdot \nabla_x F + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \nabla_v F = \mathcal{C}(F)\,. \tag{1.1}$$

The left-hand side of (1.1) describes the action of carrier electrons under their own inertia and by the force derived from the electric potential Φ that satisfies the Poisson equation

$$-\nabla_x \cdot (\epsilon \nabla_x \Phi) = q_e(D - \langle F \rangle)\,. \tag{1.2}$$

Here $\epsilon = \epsilon(x)$ is the electric permittivity of the semiconductor material and $D = D(x)$ is a fixed concentration of charge called the *doping profile*. The bracket notation used in (1.2) is a shorthand for integration over all velocity space—that is, for any function $g = g(v)$,

$$\langle g \rangle \equiv \int_{\mathbb{R}^3} g(v)\, dv\,.$$

The collision operator $\mathcal{C}$ on the right-hand side of (1.1) is an integral operator that describes collisions (energy-momentum exchanges) between carrier electrons and the vibrations of the semiconductor lattice known as phonons. It can be generalized to include other type of interactions such as electron-electron and three-particle Auger collisions [74], but electron-phonon collisions are usually the dominant mechanism. For Maxwell-Boltzmann statistics, $\mathcal{C}$ is linear and takes the form

$$\mathcal{C}(f) = \int_{\mathbb{R}^3} \sigma(x, v, v') \left(M_\ell f' - M_\ell' f \right)\, dv'\,, \tag{1.3}$$

where

$$M_\ell(v) \equiv \frac{1}{(2\pi\theta_\ell)^{3/2}} \exp\left(-\frac{|v|^2}{2\theta_\ell}\right) \tag{1.4}$$

is the *lattice Maxwellian* and primes in (1.3) denote evaluation at v' rather than v. The transition kernel σ describes the rate at which incoming particles with velocity v emerge from a phonon collision with velocity v'. It satisfies the *detailed balance relation*, $\sigma(x, v, v') = \sigma(x, v', v)$.

Together (1.1) and (1.2) form the *Boltzmann-Poisson* system. We leave unspecified the boundary conditions that give the flux of electrons into Ω along with whatever external potential is applied to drive the system. Rigorous results concerning the existence and uniqueness of solutions of this system can be found in [66, 68].

The collision operator satisfies several important properties which impose structure on solutions to (1.1). One such property is that $\langle \mathcal{C}(f) \rangle = 0$ for all $f \in \mathrm{Dom}(\mathcal{C})$. Thus, upon integrating (1.1) over all of velocity space, one finds a conservation law for the electron concentration $n = n(x,t) \equiv \langle F(x, \cdot, t) \rangle$:

$$\partial_t n + \nabla_x \cdot \langle vF \rangle = 0\,. \tag{1.5}$$

This law reflects the fact that electrons are conserved by phonon collisions.

Another property of the collision operator which relates $\mathcal{C}$ to the relative entropy density

$$\kappa(f) \equiv f \log\left(\frac{f}{M_\ell}\right) - f\ , \tag{1.6}$$

is that

$$
\begin{aligned}
(i)\ \langle \partial_f \kappa(f) \mathcal{C}(f) \rangle &\leq 0, \quad \forall\ f \in \mathrm{Dom}(\mathcal{C})\,, \\
(ii)\ \langle \partial_f \kappa(f) \mathcal{C}(f) \rangle &= 0 \iff \mathcal{C}(f) = 0\,.
\end{aligned}
\tag{1.7}
$$

The condition $\mathcal{C}(f) = 0$ defines the equilibria of $\mathcal{C}$. For $\mathcal{C}$ given by (1.3), the equilibria are positive multiples of M_ℓ. The relationship between $\mathcal{C}$ and κ implies that solutions of the Boltzmann-Poisson system satisfy the local dissipation relation [53]

$$
\begin{aligned}
&\partial_t \left(\mathcal{K}(F) + \frac{\epsilon}{2 m_e^* \theta_\ell} |\nabla_x \Phi|^2 \right) \\
&\quad + \nabla_x \cdot \left(\mathcal{I}(F) - \frac{\epsilon}{\theta_\ell} \Phi \partial_t \nabla_x \Phi - \frac{q_e}{\theta_\ell} \Phi \langle v F \rangle \right) = \langle \partial_f \kappa(F) \mathcal{C}(F) \rangle \leq 0\,,
\end{aligned}
\tag{1.8}
$$

where the *relative entropy* and the *relative entropy flux* are given by

$$
\mathcal{K}(f) \equiv \langle \kappa(f) \rangle \qquad \text{and} \qquad \mathcal{I}(f) \equiv \langle v\, \kappa(f) \rangle\ , \tag{1.9}
$$

respectively. In addition, the dissipation in (1.8) vanishes if and only if $\mathcal{C}(F) = 0$.

1.2 Drift-Diffusion

Approximate solutions for the Boltzmann-Poisson system have be computed using Monte Carlo methods [48, 87] as well as direct discretizations [18–20] of (1.1)-(1.2). Both methods can be prohibitively expensive. For this reason, electron transport in semiconductors has traditionally been modeled by the much simpler drift-diffusion-

Poisson system [35, 62] which describes the evolution of n and Φ according to

$$\partial_t n - \nabla_x \cdot (a\nabla_x n - \mu n \nabla_x \Phi) = 0\,, \tag{1.10a}$$

$$-\nabla_x \cdot (\epsilon \nabla_x \Phi) = q_e(D - n)\,. \tag{1.10b}$$

Here the electron diffusivity $a = a(x)$ and the electron mobility $\mu = \mu(x)$ are positive transport coefficients that come from the collision operator and satisfy the so-called Einstein relations [59]:

$$\frac{a}{\mu} = \frac{m_e^*}{q_e}\theta_\ell\,.$$

Often (1.10) is referred to simply as the drift-diffusion system and (1.10a) as the drift-diffusion equation. The flux in the drift-diffusion equation is actually the electric current, modulo the constant factor $-q_e$. Boundary conditions for (1.10) must still be specified.

If the potential and thermal energy of carrier electrons is on the order of the lattice energy $m_e^*\theta_\ell$, then the drift-diffusion system provides an accurate approximation for the Boltzmann-Poisson system in the limit of small mean-free-path. In this limit, F is driven toward local thermal equilibrium. Indeed, it is proved rigorously in [68] that in the interior of Ω,

$$F(x, v, t) = n(x, t)M_\ell(v) + O(\varepsilon)\ , \tag{1.11}$$

where ε is the the ratio of mean-free-path to device length and the evolution of n is governed by (1.10a).

The flux in (1.10a) comes from an asymptotic analysis of (1.1), where the first-order correction to (1.11) is given by

$$-M_\ell g \cdot \left[\theta_\ell \nabla_x n - \frac{q_e}{m_e^*} n \nabla_x \Phi\right] \tag{1.12}$$

and for the collision operator in (1.3), g satisfies

$$\mathcal{C}(M_\ell g) = M_\ell v\,, \qquad \langle M_\ell g \rangle = 0\,. \tag{1.13}$$

It turns out that (1.13) has a unique solution and that

$$\langle vF \rangle \overset{O(\varepsilon^2)}{\simeq} \langle v M_\ell g \rangle = \tau \left[\theta_\ell \nabla_x n - \frac{q_e}{m_e^*} n \nabla_x \Phi\right]\,, \tag{1.14}$$

where

$$\tau \equiv \frac{m_e^*}{q_e}\mu = -\frac{1}{3}\,\mathrm{trace}\,\langle v \otimes M_\ell g \rangle\,.$$

The flux in (1.10a) is recovered upon substituting (1.14) into the conservation law (1.5).

As its name suggests, the drift-diffusion model attributes the evolution of n to the balance between the diffusive term $a\nabla_x n$ and the drift term $\mu n \nabla_x \Phi$ that together make up the flux in (1.10a). A physical explanation is as follows. In the absence of external forces, the random thermal motion of carrier electrons creates a pressure, and any gradient in the pressure causes carrier electrons to diffuse. The charge that is displaced by diffusing electrons induces a potential gradient (i.e. an electric field)

that acts on carrier electrons to exactly counterbalance the diffusion. Current flows only when an external voltage (such as a battery) is applied, in which case carrier electrons move through the semiconductor with mean velocity

$$u = a\frac{\nabla_x n}{n} - \mu\nabla_x\Phi \ .$$

The drift-diffusion model is usually sufficient for simulating the behavior devices on the micron scale. Such devices are large enough that carrier electrons can be treated like a continuum fluid near a local thermal equilibrium with the semiconductor lattice (in the sense of (1.11)). Meanwhile, the net effect of the fast scale dynamics that occur in between collisions can be accurately represented by the perturbation g given in (1.12).

For smaller, more modern devices, the near-equilibrium assumption is no longer valid and the dynamics of carrier electrons in between lattice interactions must be considered in more detail. Furthermore, the external voltage applied to devices does not usually scale with the device size. The result in small devices is the formation of regions where the electric field $E = -\nabla_x\Phi$ is quite large. When potential energy from the electric field in these regions is converted into thermal energy, so-called hot electrons are created. These electrons are characterized by a temperature θ that differs significantly from the lattice temperature, in which case the drift-diffusion system is no longer an accurate model of their behavior. To see how drift-diffusion fails, consider as an example the pressure of the electron distribution, which is given by $p = n\theta$. Electrons undergo force due the pressure gradient which depends on

spatial variations in both n and θ. However, the drift-diffusion model assumes that $\theta = \theta_\ell$ so that any variation in θ is ignored. Only the variation in n, which generates the diffusion term $a\nabla_x n$ in (1.10a), is taken into account. Thus hot electron effects can not be accommodated by the drift-diffusion model.

1.3 Moment Systems

The practical issue for smaller devices is how to improve upon the drift-diffusion model without reverting back to a computationally expensive kinetic description. Moment methods provide an alternative approach to semiconductor modeling in so-called transition regimes, where the electron distribution is no longer in equilibrium, yet still maintains structure at the macroscopic level. Rather than resolve (1.1) in full detail, moment systems track the evolution of a finite set of velocity moments $\langle \mathbf{m}F \rangle$ where $\mathbf{m}$ is a vector of polynomials in v. This approach significantly reduces the complexity of (1.1) by replacing the velocity dependence of F by a finite number of parameters. Moreover, this approach is natural because it is the moments which are experimentally measurable quantities.

For functions F that satisfy (1.1), the system of moment equations with respect to $\mathbf{m}$ is

$$\partial_t \langle \mathbf{m}F \rangle + \nabla_x \cdot \langle v\mathbf{m}F \rangle - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \langle \nabla_v \mathbf{m}F \rangle = \langle \mathbf{m}\mathcal{C}(F) \rangle \ , \tag{1.15}$$

where all integrals are assumed to be well-defined. This system is not closed, meaning that there are more dependent variables than equations. However, if there exists a function $\mathcal{F}$ such that $F = \mathcal{F}[\langle \mathbf{m}F \rangle]$, then the flux terms $\langle v\mathbf{m}F \rangle$, field terms $\langle \nabla_v \mathbf{m}F \rangle$,

and collision terms $\langle \mathbf{m}\mathcal{C}(F)\rangle$ can be related to the spatial densities $\boldsymbol{\rho} = \boldsymbol{\rho}(x,t) \equiv \langle \mathbf{m}F(x,\cdot,t)\rangle$ to provide a closed system the form

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \mathbf{f}(\boldsymbol{\rho}) - \frac{q_e}{m_e^*}\nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) = \mathbf{r}(\boldsymbol{\rho})\,, \tag{1.16}$$

where

$$\mathbf{f}(\boldsymbol{\rho}) = \langle v\mathbf{m}\mathcal{F}[\boldsymbol{\rho}]\rangle\ , \qquad \mathbf{l}(\boldsymbol{\rho}) = \langle \nabla_v \mathbf{m}\mathcal{F}[\boldsymbol{\rho}]\rangle\ , \qquad \mathbf{r}(\boldsymbol{\rho}) = \langle \mathbf{m}\mathcal{C}(\mathcal{F}[\boldsymbol{\rho}])\rangle\ .$$

(The square bracket notation here denotes possible non-local dependence on $\boldsymbol{\rho}$ such as spatial derivatives). However, F is an element of an infinite dimensional vector space and typically cannot be expressed by any finite number of components. Therefore, any closure of (1.2) will require that F be *approximated* by a function of $\boldsymbol{\rho}$ and its spatial derivatives, in which case (1.4) only approximates the evolution of $\boldsymbol{\rho}$. The goal then is to devise an approximation that maintains the key physical and mathematical features of (1.1).

We note that the drift-diffusion equation (1.10a) can be placed into the framework of moment systems with the choice $\mathbf{m} = 1$ and the approximation

$$\mathcal{F} = nM_\ell - M_\ell g \cdot \left[\nabla_x n - \frac{q_e}{\theta_\ell m_e^*} n \nabla_x \Phi\right]$$

where $n = \langle F\rangle$ and g satisfies (1.13). However, to model systems in the transition regime, we will employ variational principles with the relative entropy.

1.4 Closures Based on Entropy Methods

One way to devise a closure for (1.15) in transition regimes is with entropy methods. For many particle-based systems, equilibria can be characterized as minimizers for a physically meaningful convex functional. This characterization of equilibrium is a classical result from statistical mechanics, and it motivates a choice for approximating F, even for systems not in equilibrium.

In the case of electron transport, the appropriate convex functional is the relative entropy $\mathcal{K}$, which is defined in (1.6) and (1.9). In the classical setting, the *entropic projection* is defined as the minimizer for the problem

$$\min_{f \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{K}(f) : \langle \mathbf{m} f \rangle = \boldsymbol{\rho} \right\} \ , \tag{1.17}$$

where $\boldsymbol{\rho} = \langle \mathbf{m} F \rangle$ and

$$\mathbb{F}_{\mathbf{m}} \equiv \left\{ g \in L_1\left(\mathbb{R}^D\right) : g \gneqq 0 \text{ and } \langle |m_s g| \rangle < \infty, \ (s = 0, \ldots, l-1) \right\} .$$

The minimizer for (1.17)—if it exists—is a projection of F (in velocity space) onto a finite-dimensional subspace parametrized by $\boldsymbol{\rho}$, thus providing a candidate for $\mathcal{F}$ that closes (1.2). Such a closure is termed an *entropy-based closure*, and the resulting moment system inherits important structural features from the Boltzmann-Poisson system. In particular, (1.16) becomes a symmetric hyperbolic system with solutions that satisfy the dissipation relation (1.8) evaluated at $F = \mathcal{F}$. In addition, the dissipation vanishes if and only if $\mathcal{F}$ is an equilibrium density—that is $\mathcal{C}(\mathcal{F}) = 0$.

The main advantage of entropy-based closures is their formal structure. Another advantage is their ability to recover the drift-diffusion system for F near equilibrium. The main challenge of entropy-based closures is that for most choices of $\mathbf{m}$, there exists a set $\mathcal{D}$ of physically realizable values of $\boldsymbol{\rho}$ for which a minimizer (1.17) does not exist. Such densities are termed *degenerate*. Understanding the geometry of $\mathcal{D}$ is an important step in modifying entropy-based closures in cases where (1.17) is ill-posed. In practice, it may be that an entropy closure will naturally avoid the set $\mathcal{D}$ as $\boldsymbol{\rho}$ evolves in time according to (1.16) or that the moment system can be adapted to avoid $\mathcal{D}$ in a way that is physically reasonable.

1.5 Entropy Closures for Well-Posed Minimization Problems

There are choices of $\mathbf{m}$ for which (1.17) is well-posed, meaning that there exists a minimizer for all physically realizable values of $\boldsymbol{\rho}$. Indeed, in cases where the polynomial components of $\mathbf{m}$ are of degree two or less, entropy-based closures generate several well-known models. A simple example is the drifted-diffusion model, which is generated by an entropy-closure for the choice $\mathbf{m} = \{1, v\}$. The model is

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{1.18a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\theta_\ell I) - \frac{q_e}{m_e^*} n \nabla_x \Phi = \langle v\, \mathcal{C}(\mathcal{M}_{n,u,\theta_\ell}) \rangle\ , \tag{1.18b}$$

where the '$\vee$' notation denotes the symmetric tensor product [1]. The variables n and u are defined by the relations

$$n = \langle F \rangle \qquad \text{and} \qquad nu = \langle vF \rangle \ , \tag{1.19}$$

and the entropic projection is

$$\mathcal{F}[\boldsymbol{\rho}](v) = \mathcal{M}_{n,u,\theta_\ell}(v) \equiv \frac{n}{(2\pi\theta_\ell)^{3/2}} \exp\left(-\frac{|v-u|^2}{2\theta_\ell} \right) .$$

This model replaces the constitutive relation for the current in the drift-diffusion system with equation (1.18b). The evolution of the current is now driven by the isotropic stress tensor $n\theta_\ell I$, the electric field, and the (yet to be evaluated) collision term on the right-hand side. Often this collision term is evaluated using a relaxation approximation of the operator $\mathcal{C}$ that gives

$$\langle v\, \mathcal{C}(\mathcal{M}_{n,u,\theta_\ell}) \rangle \simeq -\frac{q_e}{\mu m_e^*} nu \, .$$

With this expression, the drift-diffusion current can be recovered from (1.18b) by neglecting the time derivative and the flux term $nu \vee u$ on the left-hand side—an argument can be formalized using an asymptotic analysis of (1.18). However, because $\theta = \theta_\ell$, drifted-diffusion is still insufficient for modeling hot electrons.

The next example is a second-order model (i.e. one based on polynomial compo-

[1] Given an s-fold tensor U and an r-fold tensor V, the *symmetric tensor product* $W = U \vee V$ is an $(s+r)$-fold tensor $W_{a_1,\ldots,a_{s+r}} = |\Pi|^{-1} \sum_{\pi\in\Pi} U_{\pi(a_1),\ldots,\pi(a_s)} V_{\pi(a_{s+1}),\ldots,\pi(a_{s+r})}$, where Π is the set of all permutations of $a_1, \ldots, a_{s+r}$ and $|\Pi|$ is the cardinality of Π.

nents of maximum degree two). It is generated by the *Maxwellian* closure, which is based on the choice $\mathbf{m} = \left\{1, v, \frac{1}{2}|v|^2\right\}$. This closure takes its name from the fact that the entropic projection is a Maxwellian distribution,

$$\mathcal{F}[\boldsymbol{\rho}](v) = \mathcal{M}_{n,u,\theta}(v) \equiv \frac{n}{(2\pi\theta)^{3/2}} \exp\left(-\frac{|v-u|^2}{2\theta}\right) ,$$

that is traditionally expressed in terms of the variables (n, u, θ). These variables correspond to the moments $\langle \mathbf{m} F \rangle$ via the relation

$$n\theta = \frac{1}{3} \left\langle |v-u|^2 F \right\rangle \tag{1.20}$$

and the relations already given in (1.19).

The moment system generated by the Maxwellian closure is

$$\partial_t n + \nabla_x \cdot (nu) = 0 , \tag{1.21a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\theta I) - \frac{q_e}{m_e^*} n \nabla_x \Phi = \langle v\, \mathcal{C}(\mathcal{M}_{n,u,\theta}) \rangle , \tag{1.21b}$$

$$\partial_t \left(\frac{1}{2} n |u|^2 + \frac{3}{2} n\theta\right) + \nabla_x \cdot \left(\frac{1}{2} n |u|^2 u + \frac{5}{2} n\theta u\right) - \frac{q_e}{m_e^*} nu \cdot \nabla_x \Phi = \left\langle \frac{1}{2} |v|^2 \mathcal{C}(\mathcal{M}_{n,u,\theta}) \right\rangle , \tag{1.21c}$$

where the evaluation of the collision terms on the right-hand side of (1.21b) and (1.21c) depends on the explicit form of $\mathcal{C}$. Like the drifted-diffusion model, (1.21) recovers the drift-diffusion equation near thermal equilibrium; *and* it has the added benefit of tracking the electron temperature ($\theta \neq \theta_\ell$) via the addition of the energy equation

(1.21c). Even so, the stress tensor $n\theta I$ for the Maxwellian closure is isotropic. In other words, the anisotropic stress tensor

$$\Sigma \equiv \left\langle \left((v-u) \vee (v-u) - \frac{1}{3}|v|^2 \right) F \right\rangle$$

vanishes identically when $F = \mathcal{M}_{n,u,\theta}$. This means that, in the reference frame of the mean velocity u, the stress that electrons undergo is assumed to be independent of direction—a property that will be violated in the presence of a strong electric field. Another shortcoming of the closure is that the heat flux q, which is known to be a critical component of modeling electron transport, is zero for the Maxwellian closure—that is,

$$q \equiv \frac{1}{2} \left\langle |v-u|^2 (v-u) F \right\rangle$$

vanishes identically when $F = \mathcal{M}_{n,u,\theta}$.

The third and final well-posed example is the model generated with the *Gaussian* closure, which uses $\mathbf{m} = \{1, v, v \vee v, \}$. The entropic projection in this case is a Gaussian distribution,

$$\mathcal{F}[\boldsymbol{\rho}](v) = \mathcal{G}_{n,u,\Theta} \equiv \frac{n}{\sqrt{\det(2\pi\Theta)}} \exp\left(-\frac{1}{2} (v-u)^T \Theta^{-1} (v-u) \right) ,$$

where the additional variable Θ is defined by the relation

$$n\Theta = \langle (v-u) \vee (v-u) F \rangle \ .$$

Therefore $n\Theta = \Sigma + n\theta I$ and $\operatorname{trace}\Theta = 3\theta$. The moment system is

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{1.22a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\Theta) - \frac{q_e}{m_e} n \nabla_x \Phi = \langle v\, \mathcal{C}(\mathcal{G}_{n,u,\Theta}) \rangle\ , \tag{1.22b}$$

$$\partial_t (nu \vee u + n\Theta) + \nabla_x \cdot (nu \vee u \vee u + 3n\Theta \vee u)$$
$$-2\frac{q_e}{m_e} nu \vee \nabla_x \Phi = \langle v \vee v\, \mathcal{C}(\mathcal{G}_{n,u,\Theta}) \rangle\ , \tag{1.22c}$$

where, like (1.21), the collision terms on the right-hand side have yet to explicitly evaluated.

The Gaussian model enjoys all the benefits of the Maxwellian model. In addition, Σ is generally non-zero. However, the heat flux q still vanishes identically when $F = \mathcal{G}_{n,u,\Theta}$.

Devising systems that properly model the anisotropic stress and heat flux is very challenging. In order to recover non-zero values for Σ (in the Maxwellian case) and q (in either case), many closures have been posed for systems beyond second order. Strictly speaking, however, entropy-based closures are not well-posed in these cases.

1.6 Issues in Numerical Simulation

As mentioned previously, one of the positive aspects of entropy-based closures is that they formally recover the drift-diffusion equation (1.10a) in the drift-diffusion regime, i.e., when F is near local thermal equilibrium. However, when computing numerical solutions for entropy-based systems, one must take great care to ensure that the

choice of numerical scheme preserves this asymptotic behavior. This fact is true for all moment systems, regardless of the closure that is employed.

Standard discretization techniques for hyperbolic balance laws are often plagued by numerical stiffness and excessive dissipation in the drift-diffusion regime. Stiffness is due to the fact that the diffusion time scale is much longer than the convective time scale in the drift-diffusion regime. However, the typical discretization of a hyperbolic system requires time steps on the order of the convective scale in order to ensure numerical stability. Such restrictions are physically unnatural, and they require many time steps in order to observe the slower dynamics associated with the drift-diffusion balance. Excessive dissipation is also the result of stability restrictions that are commonly found in shock-capturing schemes. These schemes introduce numerical dissipation on the order of the hyperbolic wave speeds in order to prevent numerical oscillations. In the drift-diffusion regime, these speeds are quite large, and the resulting numerical dissipation can overshadow the real physical diffusion in (1.10a). In such cases, the discretization of the moment system will not be an accurate approximation of (1.10a) in the drift-diffusion regime.

In addition to poor asymptotic behavior, standard discretizations often have difficulty capturing the delicate balance of forces found at the continuum level which give rise to key physical properties of a given system. Therefore, one must devise intelligent techniques to mimic these balances. This problem arises in semiconductor models due to the presence of the source term $n\nabla_x\Phi$ in the momentum equation. (See, for example, equations (1.21b) and (1.22b)). It is not unusual for simulations to display to find heavy, non-physical oscillations near material boundaries due to a

lack of proper numerical balance.

1.7 Preview of Results

In this dissertation, I have addressed some of the unresolved issues discussed in the proceeding sections. My main results are the following

1. **A geometrical description of $\mathcal{D}$.** I have analyzed a modified minimization problem that was first introduced in [73] by relaxing the constraints in (1.17), thereby ensuring the existence of a minimizer. In applying a dual formulation to this problem, I have found *complementary slackness conditions* which I use to describe $\mathcal{D}$. Under reasonable assumptions, $\mathcal{D}$ is the finite union of fiber bundles, each of codimension one or greater in the space of physically realizable densities. The fibers of these bundles are cones given by the complementary slackness conditions. This characterization of $\mathcal{D}$ recovers results found in [42, 43], where the largest degree polynomial in $\mathbf{m}$ is radially symmetric. However, it also applies to more general choices of $\mathbf{m}$, which can be useful in devising systems for capturing the anisotropic behavior of F. Numerical results (discussed below) provide evidence that anisotropy plays an important role in modeling electron transport in semiconductors.

2. **A new hierarchy of closures**. I have derived a new hierarchy of closures based on a combination of entropy-based variational principles and perturbative analysis. To simplify evaluation of the collision terms $\mathbf{r}(\boldsymbol{\rho})$ in (1.4), I have constructed a generalized Bhatnagar-Gross-Krook (GBGK) to approximate $\mathcal{C}$.

This operator is an algebraic function of the state variables that allows each state variable to relax to its equilibrium value at a different rate. It therefore strikes a balance between the macroscopic simplicity and microscopic detail in a way that is consistent with the philosophy of the moment approach.

At the base of this new hierarchy is the drift-diffusion equation, but more importantly, the hierarchy includes pertubations of the Maxwellian closure (PM) that introduces anisotropic stress and heat flux terms in (1.21) and perturbations of the Gaussian closure (PG) which introduce heat flux terms in (1.22). Heat flux has long been known to be an important component in models describing the flow of hot electrons in semiconductors, and more recently, direct kinetic simulations have shown that an electron distribution will be highly anisotropic in regions where the electric field is strong [19, 20]. The benefit of the PM and PG closures is that they may be able to capture such behavior without the burden of expensive kinetic or Monte Carlo simulations. With standard entropy-based closures, these corrections could only be introduced through the additional of higher degree polynomials in $\mathbf{m}$, in which case the minimizer in (1.17) does not always exist.

I have used PM and PG models to simulate the behavior of a unipolar n^+-n-n^+ diode with slab symmetry and a unipolar n^+-n-n^+ MESFET device with translational symmetry, and I have compared the numerical results with other hydrodynamic models found in the literature. In doing so, I have observed that anisotropy plays an important role in the velocity and temperature profiles of

both devices and that differences between the PM and PG models are nontrivial. This is not entirely surprising, since Σ is a diffusive term in the PM closure that comes directly from the perturbation analysis, whereas in the PG closure, Σ is a convective term whose evolution is determined by the addition of state variables as in (1.22c). Numerical results show that introducing non-zero values for Σ has a significant impact on the simulated behavior of the MESFET near regions of high-electric field. They also show that the perturbed Gaussian closure provides a more consistent approximation of Σ than the perturbed Maxwellian closure does.

3. **A new numerical scheme for simulating hydrodynamic models.** I have adapted a split scheme which was originally introduced in [40] to simulate a stiff 2×2 hyperbolic system much like the drifted-diffusion system. This new scheme can be applied to any hydrodynamic model with the form given in (1.4). The splitting is based on the balance of forces in (1.4) that dominate the behavior of solutions in the drift-diffusion regime.

   I have tested the new scheme with PM model to simulate an n^+-n-n^+ diode with slab symmetry. In addition to being accurate in the transition regime, the scheme lacks the stiffness and excessive dissipation of standard shock capturing schemes in the drift-diffusion regime. Furthermore, the splitting significantly reduces the size of the numerical current oscillations at the diode junctions when compared to other methods.

We now lay out the organization of this dissertation. Chapter 2 is a review of selected topics, with most mathematical results being due to Frédéric Poupaud. We begin with a review of the basic physics of semiconductors and a semi-classical description of electron-hole transport. The semi-classical description is then simplified through a series of assumptions to arrive at the classical Boltzmann-Poisson system (1.1)-(1.2). Equation (1.1) will serve as our master equation from which all simplified models will be derived. We then present the formal asymptotic analysis of (1.1) and formally derive the drift-diffusion equation.

In Chapter 3, we present a more detailed formulation of moment systems for semiconductor transport along the lines of Section 1.1. We discuss several popular models from the literature and then derive the hierarchy of perturbed entropy-based (PEB) models.

Chapter 4 is a detailed analysis of the minimization problem upon which entropy closures are based. In it, we introduce the dual function associated with (1.17), state and prove duality theorems, and show how the complementary slackness condition can be used to characterize $\mathcal{D}$.

Chapter 5 is a return to the old and new hydrodynamic models from Chapter 3. In it, we compute numerical solutions for three families of models for the axially symmetric n^+-n-n^+ diode. These families are: (i) variations on Bløtekjær's model [13, 14], (ii) variations on the Anile-Pennisi model [4], and (iii) pertubations of (1.21) and (1.22). We compare various models using an n^+-n-n^+ diode as a benchmark device.

Chapter 6 is a computational study of the new splitting method. After describing the scheme, we perform simulations of n^+-n-n^+ diodes of different lengths, some of which are accurately described by the drift-diffusion model and others which require more detailed models. We validate the effectiveness of the scheme in both situations.

Finally, in Chapter 7, we compute numerical solutions of perturbed Maxwellian and Gaussian models used to simulate the behavior of a MESFET device with translational symmetry. We compare our results to a typical Bløtekjær-type model, which has the form of (1.21) with the ad-hoc addition of a diffusive heat flux.

Chapter 2

From Semi-Classical Transport to the Drift-Diffusion Model

This chapter is a review of concepts that will serve as a backdrop for work in later chapters. Section 1 gives a brief introduction to the physics of semiconductors. In Section 2, a mathematical formulation of semi-classical charge transport is presented, and in Section 3, simplifying assumptions are introduced to establish the classical Boltzmann equation of electron transport, from which all subsequent models will be derived. This includes drift-diffusion, which is the traditional model of choice for simulating the behavior of semiconductor devices and is derived in Section 4. A large portion of Sections 2, 3, and 4 is based on the original work of Poupaud [66,68].

2.1 Basic Physics of Semiconductors

The ability to transport charge through a solid material depends primarily on the energy required to free electrons from their bound states. For conductors, the required energy is very small and therefore electrons flow freely when subjected to an electrical potential. For insulators, the required energy is prohibitively large, thereby preventing substantial electron flow. The term *semiconductor* describes a class of materials for which electron transport is possible, but the energy needed to excite

bound electrons and make them available for transport is substantially larger than that of conductive materials. As a consequence, the conductive properties of semiconductors can be easily manipulated to create devices with a various, highly nonlinear current-voltage characteristics. It is for this reason that semiconductors are used in the fabrication of nearly all modern electronic devices.

This section contains a very brief introduction to the physics of semiconductors. Material is presented here only at the level required for the development of subsequent chapters. For a thorough treatment of the physical aspects of semiconductors, the reader is referred to [44]. For an engineering perspective that includes a discussion of devices see [62, 84]. For a summary of the mathematical theory, see [59].

2.1.1 Crystal Structure

Semiconductors are crystalline solids. Their crystal structure consists of a lattice

$$L = \{ia_1 + ja_2 + la_3 : i, j, l \in \mathbb{Z}\} ,$$

where a_1, a_2, a_3 are linearly independent vectors in $\mathbb{R}^3$, and a basis that is attached to each point in L. The basis that may be a single atom or a collection of atoms. The crystal is held together by bonds between these atoms.

Associated with L is the reciprocal lattice

$$\hat{L} = \left\{ia^1 + ja^2 + la^3 : i, j, l \in \mathbb{Z}\right\} ,$$

where a^1, a^2, a^3 are vectors in $\mathbb{R}^3$ such that $a_i \cdot a^j = 2\pi\delta_{ij}$. The structure of the reciprocal lattice is important because it determines the *dispersion relation*—the relationship between the energy e and the momentum $\hbar k$ of an electron moving through the crystal. (Here k is the electron wave number and $\hbar$ is the reduced Planck's constant.) For example, the energy is periodic with respect to the reciprocal lattice. Therefore, the momentum component of electron transport can be restricted to the Brillouin zone, defined as the set of all points in momentum space that lie closer to the origin that to any other point of the reciprocal lattice:

$$\mathbb{B} \equiv \left\{ k \in \mathbb{R}^3 : |k| < |k - k_0| \ \forall k_0 \in \hat{L} \backslash \{0\} \right\} .$$

2.1.2 Energy Bands

Recall from basic chemistry that the electrons in an individual, isolated atom are found in discrete energy levels called shells. The outermost shell is called the valence shell, and electrons in unfilled valence shells often interact with other atoms or ions to form various molecular structures. In crystalline solids, the common energy level of N different atoms will actually split in N different levels separated by very narrow gaps. This is because the Pauli exclusion principle limits the number of electrons that can be found in a given level. As N becomes large, these levels form—practically speaking—a continuum of possible energies called an energy band. The shape of an energy band in energy-momentum space is determined by the graph of the dispersion relation, and for each different band, the dispersion relation changes.

What makes a semiconductor unique is the energy gap between the valence band

(the highest energy band of bound electrons) and the conduction band (the lowest energy band of free electrons). For conductive materials, these bands are very close together or even overlap. Thus a large number of electrons will flow freely under the influence of external forces. For insulators, the band gap is too large for an electron in the valence band to be excited into the conduction band. Therefore the application of an external forces will not result in the flow of current. For semiconductors, the gap is somewhere in between so that thermal energy can excite *some* electrons into the conduction band where they are free to move. When this happens, a positively charged ion is left behind in the valence band which creates a "hole" in the lattice. This hole is effectively filled in by other electrons in the valence band which in turn leaves holes elsewhere in the lattice. It turns out that these holes can be described mathematically as freely moving, positively charged particles.

2.1.3 Doping

At room temperature, thermal energy induces the creation of approximately $10^{10}\,\mathrm{cm}^{-3}$ electron-hole pairs which is much too small to provide a usable operating current. Therefore the creation of free electrons and holes is augmented through a process known as doping, in which atoms or molecules, called dopants, are injected into a semiconductor material. Dopants can have either too many valence electrons (n-type) or too few valence electrons (p-type) to fit naturally into the bonding pattern of the crystal lattice structure. Those with too few valence electrons will cause ionization of atoms in the lattice to fill in the gaps, thus creating holes in the valence band. Those with too many valence electrons will ionize, thereby sending extra

electrons into the conduction band.

The effect of the doping process is two-fold. First, it significantly increases the concentration of holes and free electrons: typical concentrations range from $10^{16}\,\mathrm{cm}^{-3}$ to $10^{18}\,\mathrm{cm}^{-3}$. Second, because electrons and holes are no longer created in pairs, ionized dopants create a distribution of fixed charge in the spatial domain. This distribution of charge is known as the doping profile. It gives rise to an internal electric field that plays an important role in the dynamics of electron transport.

2.2 Mathematical Formulation

We consider a system of particles with electrons in a single valley of an conduction band and holes in a single peak of a valence band. The semiclassical description of the kinetic densities $F_1(x,k,t)$ for electrons and $F_2(x,k,t)$ for holes—defined for positions $x \in \Omega \subset \mathbb{R}^3$, wave number $k \in \mathbb{B}$, and time $t \geq 0$—is given by the Boltzmann transport equations

$$\partial_t F_1 + v_1(k) \cdot \nabla_x F_1 + q_e \nabla_x \Phi \cdot \nabla_k F_1 = Q_1(F_1) + R_1(F_1, F_2)\,, \tag{2.1a}$$

$$\partial_t F_2 + v_2(k) \cdot \nabla_x F_2 - q_e \nabla_x \Phi \cdot \nabla_k F_2 = Q_2(F_2) + R_2(F_1, F_2)\,. \tag{2.1b}$$

The dispersion relations for the conduction and valence bands are given by $e_1(k)$ and $e_2(k)$, respectively, and $v_i(k) = \nabla_k e_i(k)$ is the group velocity [34] for electrons in each band. The constant q_e is the magnitude of an electron charge and Q_1, Q_2, R_1, and R_2 are integral operators that model collisions and generation/recombination processes.

The quantity Φ is the electrical potential that satisfies Poisson's equation

$$-\nabla_x \cdot (\epsilon \nabla_x \Phi) = q_e \left(D - \langle F_1 \rangle + \langle F_2 \rangle \right) , \tag{2.2}$$

where $\epsilon = \epsilon(x)$ is the electric permittivity of the semiconductor material, $D = D(x)$ is the doping profile, and the notation $\langle \cdot \rangle$ indicates integration over all $k \in \mathbb{B}$.

Together (2.1) and (2.2) will be referred to as the Boltzmann-Poisson system. It still requires boundary conditions, which for Φ are usually specified by separating Ω into two parts. Artificial and insulating boundaries typically take Neumann conditions, while Ohmic and Schottky contacts take Dirichlet conditions [74]. Conditions for F_i are specified according to the characteristics of (2.1). For any $x \in \partial\Omega$, let $\nu(x)$ be the outward normal vector to $\partial\Omega$ at x. Then conditions for F_i must be given at all $(x, k) \in \Omega \times \mathbb{B}$ such that $v(k) \cdot \nu(x) < 0$—where the characteristics of (2.1) enter the domain. This can be done by specifying boundary data for F_i at these points or by providing a rule that relates the incoming and outgoing data according to some physical process at the boundary of the spatial domain.

It should be noted that a quantum version of (2.1) exists that is derived directly from the Schrödinger equation. See, for example, Sections 1.4-1.5 of [59].

2.2.1 Collision Operators

The collision operators Q_i considered here model particle-phonon scattering, which is the exchange of momentum and energy between particles (carrier electrons or holes) and quantum vibrations in the crystal lattice of the semiconductor known as phonons.

These phonons are assumed to be in a state of thermal equilibrium that is characterized by the lattice temperature T_ℓ.

The collision operators are expressed mathematically as integral operators of the form

$$Q_i(f) = \int_{\mathbb{B}} [s_i(x, k', k) f'(1 - f) - s_i(x, k, k') f(1 - f')] \, dk' \,, \tag{2.3}$$

where the local scattering rate s is a periodic function of k and k'. The prime notation attached to f in (2.3) and elsewhere implies dependence on k' rather than k. For a fixed position $x \in \Omega$, $s(x, k', k)$ gives the rate at which particles with initial wave vector k emerge from a phonon collision with wave vector k'. Although such interactions typically do not conserve energy or momentum, they do preserve particle number. This property is confirmed by the fact that $\langle Q_i(f) \rangle = 0$, which follows from the symmetry in the right-hand side of (2.3) with respect to the k and k' variables.

An important concept in scattering is the *principle of detailed balance* which asserts that the transition probabilities between any two states must be equal for a system in equilibrium—that is, for any equilibrium density F_{eq}, the local scattering rate satisfies

$$s(x, k', k) F'_{\text{eq}} (1 - F_{\text{eq}}) = s(x, k, k') F_{\text{eq}} (1 - F'_{\text{eq}}) \,. \tag{2.4}$$

In other words, equilibrium *cannot* be maintained by cyclical processes.

Although not considered here, there are more general collision operators that model additional physical processes. This includes particle-particle scattering and Auger scattering, the latter of which occurs when a carrier electron is absorbed by the lattice and its energy is transferred to bound electron that escapes the lattice. We

refer the interest reader to [59, 69] and references therein for detailed mathematical expressions for operators describing each type of collision and to [44] for a discussion of the physics involved.

2.2.2 Generation/Recombination Operators

The generation/recombination operators model the creation and annihilation of electron-hole pairs. They have the form

$$R_1(f_1, f_2) = \int_{\mathbb{B}} [g(x, k', k)(1 - f_1')(1 - f_2) - r(x, k, k')f_1 f_2'] \, dk' \,,$$
$$R_2(f_1, f_2) = \int_{\mathbb{B}} [g(x, k, k')(1 - f_1')(1 - f_2) - r(x, k', k)f_1' f_2] \, dk' \,,$$

where g and r and periodic functions of k and k' that, in analogy with (2.4), satisfy

$$g(x, k', k)(1 - F_{1,\mathrm{eq}}')(1 - F_{2,\mathrm{eq}}) = r(x, k, k')F_{1,\mathrm{eq}}F_{2,\mathrm{eq}}') \,. \tag{2.5}$$

By symmetry, $\langle R_2(f_1, f_2)\rangle = \langle R_1(f_2, f_1)\rangle$, and since $\langle Q_i(f)\rangle = 0$, the difference between the $L^1(dk)$ norms of the electron and hole kinetic densities is preserved by the flow described by (2.1):

$$\partial_t \left(\langle F_2\rangle - \langle F_1\rangle\right) + \nabla_x \cdot \left(\langle v_2 F_2\rangle - \langle v_1 F_1\rangle\right) = 0 \,.$$

Physically, this is just a statement of charge conservation.

2.2.3 Notions of Equilibrium

The Boltzmann-Poisson system is said to be in local equilibrium at a point $x \in \Omega$ if

$$Q_1(F_1) + R_1(F_1) = Q_2(F_2) + R_2(F_1, F_2) = 0\,,$$

If local equilibrium holds for every x, then the densities F_1 and F_2 are constant along particle trajectories that are characteristics of (2.1). Physically this means that collision and generation/recombination processes do not contribute to the evolution of the system.

We now restrict our attention to the case where Q_i models particle-phonon collisions. The *relative entropy* plays an important role here in characterizing equilibria of the Boltzmann-Poisson system. It also describes the trend of solutions toward such equilibria. For $i = 1$ (electrons) and $i = 2$ (holes), the relative entropy is $\mathcal{K}_i(f) = \langle \kappa_i(f, \cdot) \rangle$, where

$$\kappa_i(z, k) \equiv z \log z - z + (1 - z) \log(1 - z) + (-1)^{i+1} \frac{e_i(k) - e_i^0}{k_B T_\ell}\, z\,. \tag{2.6}$$

Here k_B is Boltzmann's constant and T_ℓ is the temperature of the lattice, which is assumed to be in equilibrium. The constant e_i^0 gives the value of the respective band edge. For electrons, it is the conduction band minimum ($e_1^0 \leq e_1(k)$), and for holes it is the valence band maximum ($e_2^0 \geq e_2(k)$). In the language of thermodynamics, $\mathcal{K}_i$ is the Massieu function corresponding to the Helmholtz free energy [17].

The relationship between the relative entropies and equilibria of (2.1) is based on

Legendre duality. In general, the Legendre transform of a convex scalar function $\xi(z)$, $z \in \mathbb{R}^n$, is defined by the implicit relation

$$\xi(z) + \xi^*(y) = y^T z\,, \qquad y = \partial_z \xi \in \mathbb{R}^n\,. \tag{2.7}$$

Differentiating (2.7) shows that $z = \partial_y \xi^*$ and, consequently, $(\xi^*)^* = \xi$. The Legendre transform of κ can be computed explicitly:

$$\kappa_i^*(y,k) = \log\left[1 + \exp\left((-1)^{i+1}\frac{e_i(k) - e_i^0}{k_B T_\ell} - y\right)\right]\,, \qquad y = \partial_z \kappa_i(z,k)\,, \tag{2.8}$$

and z can be computed in terms of y:

$$z = \partial_y \kappa_i^*(y,k) = \left[1 + \exp\left((-1)^{i+1}\frac{e_i(k) - e_i^0}{k_B T_\ell} - y\right)\right]^{-1}\,. \tag{2.9}$$

In [57, 58] it is shown that, when restricted to particle-phonon collisions, the null-spaces Q_i are given by functions of the form

$$F_{i,eq} = \frac{1}{1 + (-1)^{i+1} g(e) \exp\left(\frac{e_1(k)}{k_B T_\ell}\right)}\,, \tag{2.10}$$

where $g(e + \hbar\nu) = g(e)$ and ν is the frequency of the phonon involved in a collision. However, in practice, there are collisions of many non-commensurate frequencies (i.e. ν_1/ν_2 is not rational). Therefore, g is a constant and the expression in (2.10) reduces

to the well-known Fermi-Dirac distribution:

$$F_{i,eq} = \left[1 + \exp\left((-1)^{i+1}\frac{e(k)-\omega}{k_B T_\ell}\right)\right]^{-1} , \tag{2.11}$$

where ω is the chemical potential. By comparing (2.9) and (2.11), we see that

$$F_{i,eq} = \partial_y \kappa_i^* \left(\phi_i, k\right) , \qquad \phi_i \equiv (-1)^i \frac{e_i^0 - \omega}{k_B T_\ell}.$$

With the form of $F_{i,eq}$ given in (2.9), relations (2.4) and (2.5) become

$$s(x,k,k') = s(x,k',k)\exp\left((-1)^i \frac{e(k)-e(k')}{k_B T_\ell}\right) , \tag{2.12a}$$

$$r(x,k,k') = g(x,k',k)\exp\left(\frac{e_1(k)-e_2(k')}{k_B T_\ell}\right) . \tag{2.12b}$$

In [66], it is shown that (2.12a) and (2.12b) are necessary and sufficient conditions to prove the following Theorems:

Theorem 1 (Poupaud) *Suppose that $s > 0$ is a bounded function that satisfies (2.12a). Then*

$$\langle \partial_z \kappa_i(f,\cdot) Q_i(f) \rangle \leq 0$$

for any measurable function f. Furthermore, the following are equivalent

1. $Q_i(f) = 0$;
2. $\langle \partial_z \kappa_i(f,\cdot) Q_i(f) \rangle = 0$;

3. There exists a constant ϕ_i such that

$$f(k) = \partial_y \kappa_i^* \left(\phi_i, k\right) \, .$$

Theorem 2 (Poupaud) *Suppose that r and g are bounded functions related by (2.12b). Then*

$$\langle \partial_z \kappa_1(f_1, \cdot) \ \left(Q_1(f_1) + R_1(f_1, f_2)\right)\rangle + \langle \partial_z \kappa_2(f_2, \cdot) \ \left(Q_2(f_2) + R_2(f_1, f_2)\right)\rangle \leq 0$$

for any measurable functions f_1 and f_2. Furthermore, the following are equivalent:

1. $R_1(f_1, f_2) = R_2(f_1, f_2) = 0$;
2. $Q_1(f_1) + R_1(f_1, f_2) = Q_2(f_2) + R_2(f_1, f_2) = 0$;
3. $\langle \partial_z \kappa_1(f_1, \cdot) \ [Q_1(f_1) + R_1(f_1, f_2)]\rangle + \langle \partial_z \kappa_2(f_2, \cdot) \ [Q_2(f_2) + R_2(f_1, f_2)]\rangle = 0$;
4. *If ω is the chemical potential, then*

$$f_i(k) = \partial_y \kappa_i^* \left(\phi_i, k\right) \, ,$$

where

$$\phi_i = (-1)^i \frac{(e_i^0 - \omega)}{k_B T_\ell} \, .$$

Theorem (2) implies the following corollary, which relates equilibria to the dissipation of an entropy-based Lyapunov functional.

Corollary 3 *If (2.12) holds, then the Boltzmann-Poisson system locally dissipates the quantity*

$$\langle \kappa(F_1, \cdot) \rangle + \langle \kappa(F_2, \cdot) \rangle + \frac{\epsilon}{2k_B T_\ell} |\nabla_x \Phi|^2 .$$

The dissipation rate is zero if and only if

$$F_1 = \frac{1}{1 + \exp\left(\frac{e_1(k) - \omega}{k_B T_\ell}\right)} \qquad \textit{and} \qquad F_2 = \frac{1}{1 + \exp\left(\frac{\omega - e_2(k)}{k_B T_\ell}\right)}$$

for some constant ω.

Proof. The proof is a calculation. We sketch the details. Multiplying (2.1) by κ_i and integrating over the Brillouin zone gives,

$$\begin{aligned} &\partial_t \langle \kappa_i(F_i, \cdot) \rangle + \nabla_x \cdot \langle \kappa_i(F_i, \cdot) \rangle \\ &+ q_e \nabla_x \Phi \cdot \langle \partial_z \kappa_i(F_i, \cdot) \nabla_k F_i \rangle = \langle \partial_z \kappa_i(F_i, \cdot) \left(Q_i(F_i) + R_i(F_1, F_2) \right) \rangle , \end{aligned} \tag{2.13}$$

where, by the periodicity of $\mathbb{B}$,

$$\begin{aligned} \nabla_x \Phi \cdot \langle \partial_z \kappa(F_i, \cdot) \nabla_k F_i \rangle &= -\nabla_x \Phi \cdot \langle \partial_k \kappa(F_i, \cdot) \rangle \\ &= \frac{(-1)^{i+1} \hbar}{k_B T_\ell} \nabla_x \Phi \cdot \langle v F_i \rangle \\ &= \frac{(-1)^{i+1} \hbar}{k_B T_\ell} \left(\nabla_x \cdot \left(\Phi \langle v F_i \rangle \right) - \Phi \nabla_x \cdot \langle v F_i \rangle \right) \\ &= \frac{(-1)^{i+1} \hbar}{k_B T_\ell} \Big(\nabla_x \cdot \left(\Phi \langle v F_i \rangle \right) \\ &\qquad - \Phi \left(\partial_t \langle F_i \rangle - \langle R_i(F_1, F_2) \rangle \right) \Big) . \end{aligned} \tag{2.14}$$

Combining (2.13) and (2.14) with the fact that $\langle R_2(f_1, f_2)\rangle = \langle R_1(f_2, f_1)\rangle$ gives

$$\begin{aligned}
&\partial_t \left(\langle \kappa(F_1, \cdot)\rangle + \langle \kappa(F_2, \cdot)\rangle + \frac{\epsilon}{2k_B T_\ell} |\nabla_x \Phi|^2 \right) \\
&+ \nabla_x \cdot \left(\langle \kappa(F_1, \cdot)\rangle + \langle \kappa(F_2, \cdot)\rangle + \frac{q_e}{k_B T_\ell} \Phi \left(\langle vF_1\rangle - \langle vF_2\rangle \right) + \frac{\epsilon}{k_B T_\ell} \Phi \partial_t (\nabla_x \Phi) \right) \\
&= \langle \partial_z \kappa_1(F_1, \cdot)\ (Q_1(F_1) + R_1(F_1, F_2))\rangle + \langle \partial_z \kappa_2(F_2, \cdot)\ (Q_2(F_2) + R_2(F_1, F_2))\rangle\ .
\end{aligned}$$

The result now follows directly from Theorem (2). ■

In light of (2.12a), the collision operator can be written in the form

$$Q_i(f) = \int_{\mathbb{B}} [\tilde{s}_i(x, k', k) M_i f'(1 - f) - M_i' f(1 - f')]\ dk'\,, \tag{2.15}$$

where

$$M_\ell^i \equiv \frac{1}{N_\ell^i} \exp\left((-1)^i \frac{e_i(k)}{k_B T_\ell} \right)\,, \qquad N_\ell^i \equiv \left\langle \exp\left((-1)^i \frac{e_i(k)}{k_B T_\ell} \right) \right\rangle$$

and $\tilde{s}_i > 0$ is symmetric in the k and k' variables with

$$\tilde{s}_i(x, k', k) \equiv \frac{s_i(x, k', k)}{M_\ell^i} = \frac{s_i(x, k, k')}{(M_\ell^i)'} \equiv \tilde{s}_i(x, k, k')\,,$$

2.3 Simplifications

In this section, we present a series of successive approximations that simplify the Boltzmann-Poisson system. Our goal is to reduce (2.1) to a unipolar (electron only) model with a parabolic dispersion relation. The order in which approximations are presented is not intended to imply any type of hierarchical structure, but rather to

reach the simplified model as quickly as possible.

2.3.1 Unipolar Model

In some cases, $F_1 \gg F_2$ (or vice-versa), and in such cases, it is common to set $F_2 = 0$ and $R_1 = R_2 = 0$ and to drop the remaining subscripts in (2.1). The Boltzmann-Poisson system for $F = F_1$ becomes

$$\partial_t F + v(k) \cdot \nabla_x F + q_e \nabla_x \Phi \cdot \nabla_k F = Q(F) \tag{2.16a}$$

$$-\nabla \cdot (\epsilon \nabla_x \Phi) = q_e \left(D - \langle F \rangle \right) , \tag{2.16b}$$

where $Q = Q_1$ is given by (2.15). Since $\langle Q \rangle = 0$, the $L^1(dk)$ norm of F is preserved by the flow of (2.1):

$$\partial_t \langle F \rangle + \nabla_x \cdot \langle vF \rangle = 0 . \tag{2.17}$$

The relative entropy density $\kappa = \kappa_1$ and its Legendre transform $\kappa^* = \kappa_1^*$ are given by (2.6) and (2.8), respectively. The following is a corollary of Theorem 1.

Corollary 4 *If (2.12a) holds for $i = 1$, then the unipolar Boltzmann-Poisson system (2.16) locally dissipates the quantity*

$$\langle \kappa(F, \cdot) \rangle + \frac{\epsilon}{2 k_B T_\ell} |\nabla_x \Phi|^2 .$$

The dissipation rate is zero if and only if

$$F = \frac{1}{1 + \exp\left(\frac{e(k)-\omega}{k_B T_\ell}\right)}$$

for some constant ω*.*

Proof. The proof is a calculation along the lines of the proof of Corollary 3. It shows that solutions of (2.16) satisfy

$$\partial_t \left(\langle \kappa(F, \cdot) \rangle + \frac{\epsilon}{2k_B T_\ell} |\nabla_x \Phi|^2 \right) + \nabla_x \cdot \left(\langle \kappa(F_1, \cdot) \rangle + \frac{q_e}{k_B T_\ell} \Phi \langle vF \rangle + \frac{\epsilon}{k_B T_\ell} \Phi \partial_t (\nabla_x \Phi) \right) = \langle \partial_z \kappa(F, \cdot)\, Q(F) \rangle \ .$$

The result now follows immediately from Theorem 1. ■

2.3.2 Low Density Approximation

For $0 < F \ll 1$ the collision operator can be linearized, giving

$$\mathcal{C}(f) \equiv \int_{\mathbb{B}} \tilde{s}(x, k', k)(M_\ell f' - M'_\ell f)\, dk' \,,$$

where now

$$M_\ell(k) = \frac{1}{N_\ell} \exp\left(\frac{-e(k)}{k_B T_\ell}\right) \,, \qquad N_\ell = \left\langle \exp\left(\frac{-e(k)}{k_B T_\ell}\right) \right\rangle \,. \tag{2.18}$$

Let $\mathbb{H}_{1/M_\ell}$ be the Hilbert space with inner product

$$(f,g)_{1/M_\ell} = \left\langle \frac{fg}{M_\ell} \right\rangle .$$

Using symmetry in the k and k' variables, one may readily show that $\mathcal{C}$ is self-adjoint with respect this inner product:

$$\begin{aligned}(g,\mathcal{C}(f))_{1/M_\ell} &= \int_{\mathbb{B}}\int_{\mathbb{B}} \left[\tilde{s}(x,k',k)\left(gf' - \frac{M_\ell'}{M_\ell}gf\right)\right] dk'dk \\ &= \int_{\mathbb{B}}\int_{\mathbb{B}} \left[\tilde{s}(x,k',k)\left(g'f - \frac{M_\ell'}{M_\ell}gf\right)\right] dk'dk \\ &= \langle f\mathcal{C}(Mg)\rangle \ ,\end{aligned}$$

and also negative definite:

$$\begin{aligned}\langle f\mathcal{C}(Mf)\rangle &= -\frac{1}{2}\int_{\mathbb{B}}\int_{\mathbb{B}} \left[\tilde{s}(x,k',k)\left(\frac{M_\ell'}{M_\ell}f^2 - 2ff' + \frac{M_\ell}{M_\ell'}(f')^2\right)\right] dk'dk \\ &= -\frac{1}{2}\int_{\mathbb{B}}\int_{\mathbb{B}} \left[\tilde{s}(x,k',k)\left(\frac{M_\ell'}{M_\ell}f - \frac{M_\ell}{M_\ell'}f'\right)^2\right] dk'dk \le 0\,.\end{aligned} \tag{2.19}$$

The entropy density and its Legendre transform in the low density approximation are

$$\begin{aligned}\kappa(z,k) &= z\log z - z + \frac{e(k)-e^0}{k_BT_\ell}\,z\,, \\ \kappa^*(y,k) &= \exp\left(-\frac{e(k)-e^0}{k_BT_\ell} - y\right) .\end{aligned}$$

with equilibria are given by

$$F_{eq} = \partial_y \kappa^*(\phi, k)\,, \qquad \phi = -\frac{(e^0 - \omega)}{k_B T_\ell}.$$

2.3.3 Parabolic Band Approximation

If the energy of carrier electrons lies near the conduction band minimum at k_0, then $\nabla_k e(k_0) = 0$ and the dispersion relation can be expanded to second order as

$$e(k) - e(k^0) = \frac{1}{2}\left(\nabla_k^2 e\right)(k) : (k - k^0)^{\vee 2} + O(|k - k_0|^3)\,.$$

Upon a rotational change of coordinates, we may assume that the Hessian of e is diagonal with positive entries. Thus, if we ignore terms that are $O(|k - k_0|^3)$, then

$$e(k) - e(k^0) = \frac{1}{2}\sum_{i=1}^{3} \frac{\partial^2 e}{\partial k_i^2}(k_i - k_i^0)^2\,. \tag{2.20}$$

Equation (2.20) is called the parabolic approximation

For spherical bands, the diagonal entries of the Hessian will all be equal, but in general, this is not the case. Still, it is common to introduce a spherical approximation, thereby expressing curvature of the band with a single scalar quantity. Typically, the value of this quantity is chosen in such a way that preserves the *density of states*, which is defined as the number of different momentum states consistent with a given energy. The Hessian of e enters the formula for the density of states via

the determinant

$$\det\left(\nabla_k^2 e\right) = \prod_{i=1}^{3}\left(\frac{\partial^2 e}{\partial k_i^2}\right) .$$

Therefore, (2.20) is replaced by the expression

$$e(k) - e(k^0) = \frac{\hbar}{2m_e^*}|k - k^0|^2 , \qquad \frac{1}{m_e^*} = \left(\prod_{i=1}^{3}\left(\frac{\partial^2 e}{\partial k_i^2}\right)\right)^{1/3} .$$

The dispersion relation now has the form of a classical particle with mass m_e^*. This value is called the effective electron mass and is usually expressed as a fraction of the true electron mass m_e.

In the spherical, parabolic band approximation, the group velocity is

$$v(k) = \nabla_k e(k) = \frac{\hbar}{m_e^*}(k - k^0) . \tag{2.21}$$

and the function M in (2.18) reduces to

$$M = \frac{1}{(2\pi\theta_\ell)^{3/2}} \exp\left(-\frac{|v|^2}{2\theta_\ell}\right) , \tag{2.22}$$

where $\theta_\ell = k_B T_\ell / m_e^*$. (Because k_B/m_e^* is constant, it is common to refer to θ_ℓ as the lattice temperature when, in fact, is has units of velocity squared.) In light of (2.21), the kinetic equation (2.16a) is typically rewritten with velocity replacing wave

number as an independent variable:

$$\partial_t F + v \cdot \nabla_x F + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \nabla_v F = \mathcal{C}(F) \tag{2.23a}$$

$$-\nabla \cdot (\epsilon \nabla_x \Phi) = q_e \left(D - \langle F \rangle \right) , \tag{2.23b}$$

where

$$\mathcal{C}(f) = \int_{\mathbb{R}^3} \left[\sigma(x, v', v) \left(M f' - M' f \right) \right] dv \, , \tag{2.24}$$

and the new scattering rate σ is an approximation of $\tilde{s}$. Because the parabolic approximation is local in momentum space, consistency requires that the Brillouin zone be extended to all of $\mathbb{R}^3$. Moreover, any dependence of the dispersion relation on the reciprocal lattice—beyond the numerical value of m_e^*—is removed. This means, in particular, that the scattering rate σ is rotationally invariant—that is, for any orthogonal matrix $O \in \mathbb{R}^{3\times 3}$,

$$\sigma(x, O^T v', O^T v) = \sigma(x, v', v) \, . \tag{2.25}$$

2.4 Drift-Diffusion Equations

In this section, we formally derive the drift-diffusions equations, beginning with the simplified Boltzmann-Poisson system given in (2.23). A rigorous statement and proof concerning the relation between the Boltzmann-Poisson system and the drift-diffusion-Poisson system can be found in [68]. The drift-diffusion equation [35, 59]

is

$$\partial_t n - \nabla_x \cdot (a \nabla_x n - \mu n \nabla_x \Phi) = 0 \,, \tag{2.26}$$

where $n \equiv \langle F \rangle$ is the electron concentration, $\mu = \mu(x)$ is the electron mobility, and $a = a(x)$ is the electron diffusivity. It must be supplemented by Poisson's equation (2.23b) for the potential Φ. Boundary conditions for (2.26) are usually specified by separating the boundary of Ω into two parts: at Ohmic and Schottky contacts, Dirichlet conditions for n are given; and at insulating and artificial boundaries, the flux in (2.26) is set to zero.

The drift-diffusion equation can be derived from the Boltzmann equation, beginning with the conservation law (2.17):

$$\partial_t n + \nabla_x \cdot \langle v F \rangle = 0 \,. \tag{2.27}$$

This law requires a closure that expresses the flux $\langle vF \rangle$ in terms of n. The drift-diffusion equations are based on a closure that assumes F is near equilibrium. Before deriving it, let us first review the properties [59] of the collision operator $\mathcal{C}$ given in (2.24).

1. $\mathcal{C}$ is self-adjoint with respect to the inner product

$$\langle f, g \rangle_{1/M} = \int_{\mathbb{R}^3} f(v)\, g(v)\, \frac{dv}{M(v)} \,,$$

where M is given by (2.22). $\mathcal{C}$ is a Fredholm operator with a null-space composed entirely of multiples of M. The equation $\mathcal{C}(f) = g$ has a solution if and

only if $\langle g \rangle = 0$. This solution is unique given the restriction that $\langle fM \rangle = 0$. We denote this solution by $f = \mathcal{C}^{-1}(g)$.

2. For any orthogonal matrix $O \in \mathbb{R}^{3\times 3}$, define the operator $\mathcal{O}$ by

$$(\mathcal{O}g)(v) \equiv g(Ov).$$

Then (2.25) implies that $\mathcal{C}$ and $\mathcal{O}$ commute, i.e.,

$$\mathcal{O}\mathcal{C}(g) = \mathcal{C}(\mathcal{O}g).$$

As a consequence,

$$\langle g_1 \mathcal{C}(g_2) \rangle = \langle \mathcal{O}(g_1 \mathcal{C}(g_2)) \rangle = \langle \mathcal{O}g_1 \mathcal{C}(\mathcal{O}g_2) \rangle \ , \quad \forall\ g_1, g_2 \in \mathrm{Dom}(\mathcal{C}) \ , \tag{2.28a}$$

for all $g_1, g_2 \in \mathrm{Dom}(\mathcal{C})$ and

$$\left\langle g_1 \mathcal{C}^{-1}(g_2) \right\rangle = \left\langle \mathcal{O}\left(g_1 \mathcal{C}^{-1}(g_2)\right) \right\rangle = \left\langle \mathcal{O}g_1 \mathcal{C}^{-1}(\mathcal{O}g_2) \right\rangle \ , \quad \forall\ g_1, g_2 \in \mathrm{Dom}\left(\mathcal{C}^{-1}\right) \tag{2.28b}$$

for all $g_1, g_2 \in \mathrm{Dom}(\mathcal{C}^{-1})$.

2.4.1 Scaling

The behavior of solutions to the Boltzmann-Poisson system depends heavily on the ratio of the mean-free-path between collisions to the length of a device and also on

the size of the electrons' thermal velocity relative to the drift velocity induced by the electric field $E = -\nabla_x \Phi$. To understand how these parameters play a role, one must first non-dimensionalize (2.23a). First, we rescale the independent variables

$$x = x_0 \hat{x}\,, \qquad v = v_0 \hat{v}\,, \qquad t = t_0 \hat{t}\,. \tag{2.29}$$

Here a "naught" subscript denotes the magnitude of the associated dimensional variable and a carat denotes the new dimensionless variable. The value x_0 is the physical device scale, v_0 is the free velocity (set to capture the slowest dynamics of problem), and $t_0 = x_0/v_0$ is the time it takes a particle at velocity v_0 to traverse the distance x_0. Next, we rescale the dependent variables:

$$F\left(x, v, t\right) = F_0 \hat{F}\left(\hat{x}, \hat{v}, \hat{t}\right)\,, \qquad \mathcal{C}(F) = \mathcal{C}_0 \hat{C}(\hat{F})\,, \tag{2.30}$$
$$\Phi\left(x, v, t\right) = [\Phi_0] \hat{\Phi}\left(\hat{x}, \hat{v}, \hat{t}\right)\,.$$

Here $[\Phi_0]$ is the voltage drop across the device. With the non-dimensional variables given in (2.29) and (2.30), the transport equation is (with hats removed)

$$\frac{1}{t_0}\partial_t F + \frac{v_0}{x_0} v \cdot \nabla_x F - \frac{1}{v_0 x_0}\frac{q_e}{m}[\Phi_0]\nabla_x \Phi \cdot \nabla_v F = \frac{\mathcal{C}_0}{F_0}\mathcal{C}(F)\,. \tag{2.31}$$

In addition to the free velocity, there are two other important velocity scales: the thermal velocity $\theta_0^{1/2}$ and the drift velocity v_E due to the electric field. The thermal

velocity is given by

$$\theta_0^{1/2} = \frac{\text{mean-free-path}}{\tau_0},$$

where the relaxation time scale $\tau_0 \equiv \mathcal{C}_0/F_0$ characterizes the average time between electron-phonon collisions and the mean-free-path is the average distance between collisions. A reference value for the drift velocity is found by considering a particle initially at rest at time zero that is accelerated by a constant electric field $E_0 = -\frac{[\Phi_0]}{x_0}$. Just before a collision at time τ_0, the particle has drift velocity

$$v_E \equiv \frac{qE_0}{m_e^*}\tau_0 = \frac{q_e}{m_e^*}\frac{[\Phi_0]}{x_0}.$$

In order to capture the dynamics at both of these velocity scales, we set $v_0 = \min(\theta_0^{1/2}, v_E)$.

We now identify several non-dimensional parameters. First is the scaled Knudsen number ε, which gives the ratio of mean-free-path to device length:

$$\varepsilon \equiv \frac{\text{mean-free-path}}{x_0} = \frac{v_0\tau_0}{x_0},$$

and also relates the reference times t_0 and τ_0:

$$\frac{t_0}{\tau_0} = \varepsilon\frac{\theta_0^{1/2}}{u}.$$

There are also two velocity ratios, η and δ, which measure the ratio of the free to

thermal velocity and the drift to thermal velocity, respectively:

$$\eta = \frac{v_0}{\theta_0^{1/2}}\,, \qquad \delta = \frac{v_E}{\theta_0^{1/2}} = \frac{q_e \tau_0}{m_e^*} \frac{[\Phi_0]}{v_0 x_0}\,.$$

In terms of these three ratios, (2.31) becomes

$$\eta \partial_t F + \eta v \cdot \nabla_x F - \frac{\delta}{\varepsilon \eta} \nabla_x \Phi \cdot \nabla_v F = \frac{1}{\varepsilon} \mathcal{C}(F)\,. \tag{2.32}$$

Several asymptotic limits can be realized based on the size of ε and δ.

1. **The drift-diffusion balance**. Assume that collision processes dominate the dynamics of electron transport, so the drift velocity is small compared to the thermal velocity. Then $\varepsilon = \delta = \eta$ and (2.32) becomes

$$\varepsilon \partial_t F + v \cdot \nabla_x F + \nabla_x \Phi \cdot \nabla_v F = \frac{1}{\varepsilon} \mathcal{C}(F)\,. \tag{2.33}$$

2. **The drift-collision balance**. Assume that the electric field force is large enough to balances the collision forces, so the drift velocity and thermal velocity are of comparable size. Therefore $\varepsilon \ll \delta = 1$, and (2.32) becomes

$$\partial_t F + v \cdot \nabla_x F + \frac{1}{\varepsilon} \nabla_x \Phi \cdot \nabla_v F = \frac{1}{\varepsilon} \mathcal{C}(F)\,.$$

Other possible limits exist. For example, one can consider a high field ballistic scaling [1, 2, 18, 77] in which a nontrivial proportion of carrier electrons pass through

the semiconductor material without being slowed by collisions. There are also variations on the drift-collision balance [67, 70]. However, we will not consider these cases here. In fact, in what follows, we will concentrate entirely on the small-field approximation that leads to the drift-diffusion balance.

2.4.2 Chapman-Enskog Expansion

We now use a Chapman-Enskog expansion to approximate F in the asymptotic limit $\varepsilon \to 0$ of the drift-diffusion scaling, thereby re-deriving the closure used in (2.27). We formally expand F to first order ε:

$$F = n\hat{M}_\ell + \varepsilon \hat{M}_\ell \mathcal{F}^{(1)}[n]\,, \tag{2.34}$$

where $n = \langle F \rangle$, $\left\langle \mathcal{F}^{(1)}[n] \right\rangle = 0$, and $\hat{M}_\ell$ is the rescaled version of (2.22) given by

$$\hat{M}_\ell = \frac{1}{(2\pi)^{3/2}} \exp\left(-\frac{|v|^2}{2}\right)\,.$$

(Note that square brackets around dependent variables indicate non-local dependence.) Plugging (2.34) into (2.33) and comparing powers of ε gives

$$\begin{aligned} \mathcal{C}(\hat{M}_\ell \mathcal{F}^{(1)}) &= -v \cdot \nabla_x (n\hat{M}_\ell) + \nabla_x \Phi \cdot \nabla_v n \hat{M}_\ell \\ &= -\hat{M}_\ell v \cdot [\nabla_x n + n \nabla_x \Phi]\,. \end{aligned} \tag{2.35}$$

Because $\langle vM \rangle = 0$, (2.35) is solvable for $\mathcal{F}^{(1)}$:

$$\mathcal{F}^{(1)} = -\frac{1}{\hat{M}_\ell} \mathcal{C}^{-1}(\hat{M}_\ell v) \cdot [\nabla_x n + n \nabla_x \Phi] \,,$$

whereby

$$\langle vF \rangle = -\left\langle v \otimes \mathcal{C}^{-1}(\hat{M}_\ell v) \right\rangle \cdot [\nabla_x n + n \nabla_x \Phi] \,. \tag{2.36}$$

We note that upon rescaling that $\mathcal{C}$ retains the properties given in (2.28) with M_ℓ replaced by $\hat{M}_\ell$, Using these properties, one can show that $\left\langle v \otimes \mathcal{C}^{-1}(\hat{M}_\ell v) \right\rangle$ is a multiple of the identity. This multiple defines the mobility μ:

$$\mu \equiv \left\langle v_1 \otimes \mathcal{C}^{-1}(\hat{M}_\ell v_1) \right\rangle = \left\langle v_2 \otimes \mathcal{C}^{-1}(\hat{M}_\ell v_2) \right\rangle = \left\langle v_3 \otimes \mathcal{C}^{-1}(\hat{M}_\ell v_3) \right\rangle \,,$$

Setting (2.36) into (2.27) gives the drift-diffusion equation in non-dimensional form:

$$\partial_t n - \nabla_x \cdot (\mu \, (\nabla_x n - n \nabla_x \Phi)) = 0 \,.$$

In dimensional variables,

$$\partial_t n - \nabla_x \cdot \left(\mu \left(\frac{m_e^* \theta_\ell}{q_e} \nabla_x n - n \nabla_x \Phi \right) \right) = 0 \,, \tag{2.37}$$

and by comparing (2.37) with (2.26), we recover the so-called Einstein relations [84]:

$$a = \frac{m_e^* \theta_\ell}{q_e} \mu = \frac{k_B T_\ell}{q_e} \mu \,.$$

In [66], it is proven rigorously that

$$F = nM_\ell + O(\varepsilon)\,,$$

where F solves (2.23a) and n solves (2.37). This result depends on appropriate specification of boundary conditions and the introduction of a boundary layer.

Chapter 3

Hydrodynamic Models

In this chapter, we analyze several moment models of electron transport, which in the context of semiconductors are generally referred to as *hydrodynamic models* or sometimes *extended hydrodynamic models*. These models provide a reasonable alternative to highly complex kinetic equations, and unlike drift-diffusion, their derivation is not based on any near-equilibrium assumptions. Our main result is the derivation of a new hierarchy of models based on perturbations of standard entropy-based closures. The motivation for these new closures is to incorporate a description of the heat flux (which plays an important role in electron transport) into well-posed entropy-based models in a way that is justifiable at the kinetic level.

Extensive studies have been done on a variety hydrodynamic models by electrical engineers, physicists and applied mathematicians. Although many models exist in the literature (see [33] for a survey), most of them are variations or extensions of the works by Bløtekjær [13, 14] and Stratton [81]. The continuing challenge of creating credible models is to find an accurate description of electron transport in high-field, hot-electron regimes. Our approach is to derive models beginning from a kinetic description.

The chapter is laid out as follows. In Section 3.1, we give mathematical formulation of electron transport at the kinetic level, which will form a foundation for the remainder of the chapter. This includes a general recipe for constructing moments models from a kinetic description. In Section 3.2, we review the widely used model attributed to Bløtekjær [13,14] and Baccarani and Wordemann [8]. In Section 3.3, we review the model of Anile and Pennisi [4] in the context of extended thermodynamics. In Section 4, we lay out the formal framework for closures based on entropy minimization [51–53]. Finally in Section 3.5, we derive the new perturbed entropy-based closures and present several examples.

3.1 Mathematical Background

Let $F = F(x, v, t)$ be the kinetic density of free electrons in a single conduction band of a semiconductor material. In a classical description, F evolves according to the Boltzmann transport equation

$$\partial_t F + v \cdot \nabla_x F + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \nabla_v F = \mathcal{C}(F) \,. \tag{3.1}$$

Here $v \in \mathbb{R}^3$ is the velocity coordinate, $x \in \Omega \subset \mathbb{R}^3$ is the spatial coordinate, and $t \geq 0$ is time. The constant q_e is the magnitude of the electron charge, and m_e^* is the effective electron mass that characterizes the conduction band in the classical parabolic limit [59].

The left-hand side of (3.1) describes the evolution of particles under their own inertia and by the force derived from the electric potential Φ that satisfies Poisson's

equation

$$-\nabla_x \cdot (\epsilon \nabla_x \Phi) = q_e \left(D - \langle F \rangle \right) . \tag{3.2}$$

Together (3.1) and (3.2) are the *Boltzmann-Poisson* system. The new quantities in (3.2) are the doping profile $D = D(x)$ and the electric permittivity $\epsilon = \epsilon(x)$. The angle brackets denote integration with respect to the velocity variable—that is for any function $g = g(v)$

$$\langle g \rangle \equiv \int_{\mathbb{R}^3} g(v) dv \, .$$

The collision operator $\mathcal{C}$ on the right-hand side of (3.1) is the integral operator introduced in (2.24) of Chapter 2 that describes collisions between particles and phonons in the semiconductor lattice which are assumed to be in thermal equilibrium with lattice temperature θ_ℓ.

3.1.1 Formal Kinetic Properties

We recall several important properties of the collision operator. First, for any function $\xi = \xi(v)$, the following are equivalent:

$$(i) \, \langle \xi \mathcal{C}(f) \rangle = 0 \text{ for all } f \in \mathrm{Dom}(\mathcal{C}) \, ; \tag{3.3a}$$

$$(ii) \, \xi \text{ is constant.} \tag{3.3b}$$

In particular, $\xi = 1$ gives the conservation law for electron concentration

$$\partial_t \langle F \rangle + \nabla_x \cdot \langle vF \rangle = 0 \, . \tag{3.4}$$

Second, for all $f \in \mathrm{Dom}(\mathcal{C})$,

$$\left\langle \log\left(\frac{f}{M_\ell}\right) \mathcal{C}(f) \right\rangle \leq 0\,, \tag{3.5}$$

where

$$M_\ell(v) \equiv \frac{1}{(2\pi\theta_\ell)^{3/2}} \exp\left(-\frac{|v|^2}{2\theta_\ell}\right) \tag{3.6}$$

is the *lattice Maxwellian.* Finally, there is an H-Theorem. For all $f \in \mathrm{Dom}(\mathcal{C})$, the following statements are equivalent characterizations of the equilibria of $\mathcal{C}$:

$$(i)\ \mathcal{C}(f) = 0\,. \tag{3.7a}$$

$$(ii)\ f = \langle f \rangle M_\ell. \tag{3.7b}$$

$$(iii)\ \left\langle \log\left(\frac{f}{M_\ell}\right) \mathcal{C}(f) \right\rangle = 0\,, \tag{3.7c}$$

The *kinetic entropy* $\mathcal{H}$ and the *relative kinetic entropy* $\mathcal{K}$ are defined as

$$\mathcal{H}(f) = \langle f \log(f) - f \rangle\ , \tag{3.8a}$$

$$\mathcal{K}(f) = \left\langle f \log\left(\frac{f}{M_\ell}\right) - f \right\rangle\ . \tag{3.8b}$$

The kinetic entropy arises in the study of dilute gases where a *different* collision

operator, $\mathcal{C}_{gas}$, satisfies an H-Theorem similar to (3.7) with

$$(ii')\ f = \mathcal{M}_{n,u,\theta}\,, \tag{3.9a}$$

$$(iii')\ \langle \log(f)\mathcal{C}_{gas}(f)\rangle = 0\,, \tag{3.9b}$$

where $\mathcal{M}_{n,u,\theta}$ is a *Maxwellian*:

$$\mathcal{M}_{n,u,\theta}(v) \equiv \frac{n}{(2\pi\theta)^{3/2}} \exp\left(\frac{|v-u|^2}{2\theta}\right)$$

for some parameters $(n, u, \theta) \in \mathbb{R}^+ \times \mathbb{R}^3 \times \mathbb{R}^+$ depending on x and t. The evolution equation for $\mathcal{H}$ is

$$\partial_t \mathcal{H}(F) + \nabla_x \cdot \mathcal{J}(F) = \mathcal{D}(F)\,, \tag{3.10}$$

where the kinetic entropy flux $\mathcal{J}$ and the entropy dissipation $\mathcal{D}$ are given by

$$\mathcal{J}(F) = \langle v(F\log(F) - F)\rangle \qquad \text{and} \qquad \mathcal{D}(F) = \langle \log(F)\,\mathcal{C}(F)\rangle\ . \tag{3.11a}$$

In general, $\mathcal{H}$ will be locally dissipated by solutions of (3.1) if

$$\langle \log(M_\ell)\,\mathcal{C}(f)\rangle = \frac{1}{2\theta}\left\langle |v|^2\mathcal{C}(f)\right\rangle \le 0\,.$$

The relative kinetic entropy derives its name from the fact that it is essentially the difference between the kinetic entropy of a function f and the kinetic entropy of

M_ℓ:

$$\mathcal{K}(f) + M_\ell = \mathcal{H}(f) - \mathcal{H}(M_\ell)\,. \tag{3.12}$$

In fact, the left-hand side of (3.12) is often defined as the kinetic entropy. This makes little difference in any of the subsequent results and we use (3.8b) as a matter of convenience.

We stress that $\mathcal{K}$ is the appropriate thermodynamic potential for describing the Boltzmann-Poisson system. To see this fact, consider that the left-hand side of (3.7c) can be interpreted as a statement that collision processes do not change $\mathcal{K}$ at equilibrium. Indeed, for a general functional $\mathcal{T}$, define the formal differential

$$\delta\mathcal{T}(f;g) \equiv \lim_{\delta\to 0} \frac{\partial}{\partial\delta}\mathcal{T}(f+\delta g)\,,$$

then by (3.5)

$$\delta(\mathcal{K};\mathcal{C}(f)) \doteq \left\langle \log\left(\frac{f}{M_\ell}\right)\mathcal{C}(f)\right\rangle \leq 0$$

with equality, by (3.7c), if and only if $f = M_\ell$. (Above we use the notation $\doteq$ to show that equality is formal. The differential does not always exist). In conjunction with (3.3), equation (3.7c) implies (3.7b) automatically. On the other hand, the statement

$$\delta(\mathcal{H};\mathcal{C}(f)) \doteq \langle \log(f)\,\mathcal{C}(f)\rangle = 0$$

implies, also in conjunction with (3.3), that f is a positive constant. This is clearly not the appropriate characterization of the equilibria of $\mathcal{C}$, because f in that case would not even be integrable. Thus from a mathematical point of view, it is $\mathcal{K}$,

not $\mathcal{H}$, that is the correct object to describe the distribution F that satisfies (3.1). Indeed, $\mathcal{K}$ satisfies the local dissipation law

$$\partial_t \left(\mathcal{K}(F) + \frac{\epsilon \left| \nabla_x \Phi \right|^2}{2 m_e^* \theta_\ell} \right) \tag{3.13}$$
$$+ \nabla_x \cdot \left(\mathcal{J}(F) - \frac{\epsilon \Phi \partial_t \nabla_x \Phi - q_e \Phi \left\langle v F \right\rangle}{m_e^* \theta_\ell} \right) = \left\langle \log\left(\frac{F}{M_\ell} \right) \mathcal{C}(F) \right\rangle < 0 \,,$$

which includes the contribution of the electric potential Φ. (The reader is referred to the appendix of [17] for an explanation of how electric and magnetic fields are incorporated in a thermodynamic description of physical system.)

From a physical point of view, $\mathcal{K}$ should be written in the form

$$\mathcal{K} = \mathcal{H} - \left\langle \frac{|v|^2}{2\theta_\ell} f \right\rangle \,,$$

which makes it more recognizable as the Massieu function corresponding to the non-equilibrium version of the Helmholtz free energy [17]. Because electrons are in contact with the thermal bath that is the semiconductor lattice, we see that $\mathcal{K}$ is the correct object to describe the physical system.

3.1.2 Moment Systems

Rather than attempt to resolve (3.1) in full detail, one may instead track the vector $\langle \mathbf{m} F \rangle$, where $\mathbf{m} = (m_0, \ldots, m_{l-1})^T$ is a vector whose l components are linearly independent polynomials in v. (Often $\mathbf{m}$ will be referred to as a *polynomial vector*.) This significantly reduces the complexity of the problem by replacing the velocity

dependence of F by a finite number of macroscopic variables that depend on x and t. For functions F in the set

$$\mathbb{F}_{\mathbf{m}} = \left\{ g \in L_1\left(\mathbb{R}^D\right) : g \gneqq 0 \text{ and } \langle |m_s g| \rangle < \infty,\ (s = 0, \ldots, l-1) \right\} \tag{3.14}$$

that satisfy (3.1) and decay sufficiently fast for large $|v|$, the evolution of *spatial densities* $\boldsymbol{\rho} = \boldsymbol{\rho}(x,t) \equiv \langle \mathbf{m}F \rangle$ is given by

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \langle v\mathbf{m}F \rangle - \nabla_x \Phi \cdot \langle \nabla_v \mathbf{m}F \rangle = \langle \mathbf{m}\mathcal{C}(F) \rangle \,. \tag{3.15}$$

For any function $\xi = \xi(v)$, the integral $\langle \xi F \rangle$ will be referred to as the *moment of F with respect to ξ*. Thus components of $\langle \mathbf{m}F \rangle$ are moments with respect to the polynomials m_s. The *order* of $\langle m_s F \rangle$ is the degree of m_s. The set of moment equations that make up (3.15) is called *the moment system for F with respect to m*, and the *order* of this system is the degree of the highest degree polynomial component of $\mathbf{m}$. We note that in the context of semiconductors, these moment systems are generally referred to as hydrodynamic models.

In general, (3.15) is not closed, meaning that there are more dependent variables than equations. However, if we can find a function $\mathcal{F} : \mathbb{R}^l \to \mathbb{F}_{\mathbf{m}}$ such that $F = \mathcal{F}[\boldsymbol{\rho}]$, (3.15) becomes a closed system of the form

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \mathbf{f}(\boldsymbol{\rho}) - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) = \mathbf{r}(\boldsymbol{\rho}) \,, \tag{3.16}$$

where

$$\mathbf{f}(\boldsymbol{\rho}) = \langle v\mathbf{m}\mathcal{F}[\boldsymbol{\rho}]\rangle \ , \qquad \mathbf{l}(\boldsymbol{\rho}) = \langle \nabla_v \mathbf{m}\mathcal{F}[\boldsymbol{\rho}]\rangle \ , \qquad \mathbf{r}(\boldsymbol{\rho}) = \langle \mathbf{m}\mathcal{C}(\mathcal{F}[\boldsymbol{\rho}])\rangle \ .$$

(The bracket notation here denotes possibly non-local dependence on $\boldsymbol{\rho}$). The components of $\mathbf{f}$ will be referred to as *flux terms* or simply *fluxes*; components of $\mathbf{l}$ will be referred to as *field terms*; and components of $\mathbf{r}$ will be referred to as *collision terms*. In practical situations it is $\boldsymbol{\rho}$, and not F, that is a measurable quantity.

Because $\mathbb{F}_{\mathbf{m}}$ is an infinite dimensional vector space, a generic function in $\mathbb{F}_{\mathbf{m}}$ cannot be expressed by any finite number of components. Therefore, any closure for (3.15) will require that F be *approximated* by a function $\mathcal{F}[\boldsymbol{\rho}]$, in which case (3.16) only approximates the evolution of $\boldsymbol{\rho}$. The goal then is to identity candidates for $\mathcal{F}$ for which (3.16) maintains the key physical and mathematical features of (3.15) as well as the original Boltzmann-Poisson system.

3.1.3 Evaluation of the Collision Operator

Once a suitable candidate for $\mathcal{F}$ is found, evaluation of $\mathbf{f}$ and $\mathbf{l}$ is, in theory, straightforward. However, evaluation of the full collision operator $\mathcal{C}$ is a nontrivial computation. For detailed kinetic and Monte Carlo simulations, the amount of work required for such a computation is justifiable. However, for hydrodynamic models that track only a handful of velocity moments, a more sensible approach is to use simple approximations that are easier to calculate yet still maintain the key features of full the collision operator. In particular, an approximation should dissipate the relative ki-

netic entropy $\mathcal{K}$, satisfy an H-Theorem as described by (3.7), and capture the correct relaxation properties in the equilibrium limit. The standard BGK model [12] is a popular choice. It takes the form

$$\tilde{\mathcal{C}}(f) = -\frac{1}{\tau}\left(f - \frac{\langle \tau^{-1} f\rangle}{\langle \tau^{-1} M_\ell\rangle} M_\ell\right),$$

where $\tau = \tau(x, v)$ is the microscopic relation time and M_ℓ is given in (3.7c). In the parabolic limit, τ is usually modeled by a power law with the factored form [33]

$$\tau(x, v) = \bar{\tau}(x)\left(\frac{|v|^2}{2\theta_\ell}\right)^\gamma, \tag{3.17}$$

where the exponent γ is a fitting parameter. When multiple scattering processes are involved γ is given by some average representative value.

Since the point of moment models is to average out the velocity dependence of f, the microscopic relaxation time is often replaced by a macroscopic relaxation time τ_1 that depends on the macroscopic variables and is velocity independent, in which case

$$\tilde{\mathcal{C}}(f) = -\frac{1}{\tau_1}\left(f - \langle f\rangle M_\ell\right).$$

However, by removing the local velocity dependence from τ, moments of the kinetic distribution are forced to relax to their equilibrium value at the same rate, which is very inaccurate.

In [13], a generalized BGK operator is introduced in which a kinetic density is assumed to relax to equilibrium through a sequence of intermediate states. The states

are not given explicitly; rather, they are specified only by their relevant moments—that is, only moments of interest in the hydrodynamic model are given. This relaxation operator maintains the conservation properties of the full collision operator and allows different moments to relax to equilibrium at different rates. In this way, some degree of microscopic information from τ is retained.

A similar multi-stage operator is presented in [51] in the context of neutral fluids that uses a sequence of fully specified intermediate states. The benefit is a relaxation operator that satisfies the same conservation *and* entropy dissipation properties as the full collisional operator.

A drawback to the approach in [51] for the neutral fluid case is that it fails to capture the correct transport coefficients in the incompressible Navier-Stokes limit. At issue there is the ability to obtain the correct Prandtl number—essentially the ratio of viscosity to thermal conductivity. However, in the context of electron-lattice collisions, the analog of the Navier-Stokes limit is the drift-diffusion limit. In this limit, there is only one transport coefficient: the mobility. Thus the approach in [51] will be satisfactory for the electron transport model.

3.2 The Bløtekjær, Baccarani, Wordemann (BBW) Model

The use of moment equations to describe electron transport was pioneered by Bløtekjær in [13, 14]. We briefly review the derivation in [13], beginning with a third-order system based on the vector

$$\mathbf{m} = (1, v, v \vee v, v \vee v \vee v)^T \ . \tag{3.18}$$

The densities associated with these moments are the

$$\begin{aligned} &\text{concentration:} && \langle F \rangle \ , \\ &\text{momentum:} && \langle vF \rangle \ , \\ &\text{velocity flux tensor:} && \langle v \vee v\, F \rangle \ , \end{aligned}$$

and the unnamed third-order tensor $\langle v \vee v \vee v\, F \rangle$. Typically moment equations are expressed in terms of the concentration n, the bulk velocity u, the temperature tensor Θ, the heat flux tensor Q, and the unnamed fourth-order tensor R. They are related to F by

$$n = \langle F \rangle \ , \qquad u = \frac{1}{n} \langle vF \rangle \ , \qquad \Theta = \frac{1}{n} \langle (v-u) \vee (v-u)\, F \rangle \ , \tag{3.19}$$

$$Q = \left\langle (v-u)^{\vee 3}\, F \right\rangle \ , \qquad R = \left\langle (v-u)^{\vee 4}\, F \right\rangle .^{1}$$

Often Θ is split into its trace and traceless parts

$$\Theta = n\theta I + \Sigma$$

where $\theta = \text{trace}(\Theta)$ is the (scalar) temperature and

$$\Sigma = \left\langle \left(v \vee v - \frac{1}{3}|v|^2 \right) F \right\rangle$$

is the anisotropic stress. With the variables defined in (3.19), the moment equations for F with respect to $\mathbf{m}$ are given by (3.16) with

$$\boldsymbol{\rho} = \begin{pmatrix} n \\ nu \\ nu \vee u + n\Theta \\ nu^{\vee 3} + 3n\Theta \vee u + Q) \end{pmatrix},$$

$$\mathbf{f}(\boldsymbol{\rho}) = \begin{pmatrix} nu \\ nu \vee u + n\Theta \\ nu^{\vee 3} + 3n\Theta \vee u + Q \\ nu^{\vee 4} + 4nQ \vee u + 6n\Theta \vee u \vee u + R \end{pmatrix},$$

$$\nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ n\nabla_x \Phi \\ 2nu \vee \nabla_x \Phi \\ (3nu \vee u + n\Theta) \vee \nabla_x \Phi \end{pmatrix}.$$

Note that R and the collision terms have yet to be specified.

3.2.1 Closure

The closure process consists in approximating the flux terms and the collision terms. The field terms are already given in terms of the densities.

3.2.1.1 Flux Terms Finding a closure for $\mathbf{f}$ amounts to specifying R. In [13], this is done by replacing F in the expression for R by a *Gaussian* distribution,

$$\mathcal{G}_{n,u,\Theta}(v) \equiv \frac{n}{\sqrt{\det(2\pi\Theta)}} \exp\left(-\frac{1}{2}(v-u)^T \Theta^{-1}(v-u)\right),$$

which gives

$$R = 3n\Theta \vee \Theta\,. \tag{3.20}$$

It should be noted that $\mathcal{G}$ is constructed in order to recover the correct values for n, u, and Θ. However, $\mathcal{G}$ cannot be used evaluate the heat flux Q. Since, $\mathcal{G}(v+u) = \mathcal{G}(-(v+u))$, a simple symmetry argument shows that the heat flux tensor associated with $\mathcal{G}$ is identically zero.

3.2.1.2 Collision Terms The next step is to find an expression for the collision terms $\mathbf{r}(\boldsymbol{\rho})$. Bløtekjær's approach here is to approximate the collision operator by a series of relaxation terms,

$$\tilde{\mathcal{C}}(F) = -\sum_{s=0}^{2} \eta_s (F - f_s)\,, \tag{3.21}$$

where the relaxation rates $\eta_s > 0$ depend on the moments $\langle \mathbf{m}F \rangle$. The quantity f_s is a distribution with concentration n_s, bulk velocity u_s, and temperature θ_s that is spherically symmetric about u_s, meaning

$$f_s\left(O^T(v-u_s)\right) = f_s(v-u_s)$$

for any orthogonal matrix O. For example, f_s could be a *Maxwellian* distribution of the form

$$\mathcal{M}_{n_s,u_s,\theta_s}(v) \equiv \frac{n_s}{(2\pi\theta_s)^{3/2}} \exp\left(-\frac{|v-u_s|^2}{2\theta_s}\right),$$

but it need not be.

Approximating the collision operator in this fashion allows each of the spatial densities for F to relax to their corresponding equilibrium values independently . In [13], the choice

$$(n_0, u_0, \theta_0) = (n, 0, \theta_\ell)$$

models the relaxation of energy to the thermal energy of the lattice; the choice

$$(n_1, u_1, \theta_1) = \left(n, 0, \frac{1}{3}|u|^2 + n\theta_\ell\right)$$

models relaxation of momentum to zero while energy is conserved; and finally, the choice

$$(n_2, u_2, \theta_2) = \left(n, u, n|u|^2 + 3n\theta\right)$$

models relaxation to an isotropic density while both energy and momentum are conserved. These choices produce a multi-stage relaxation approximation to $\mathcal{C}$ that,

when substituted into (3.21), yields the following collision terms:

$$\left\langle v\,\tilde{\mathcal{C}}(F)\right\rangle = -\frac{1}{\tau_p}nu \tag{3.22a}$$

$$\left\langle v\vee v\,\tilde{\mathcal{C}}(F)\right\rangle = -\frac{1}{\tau_p}\left(nu\vee u-\frac{n|u|^2}{3}I\right)-\frac{1}{\tau_w}\left(\frac{n|u|^2}{3}I+n(\theta-\theta_\ell)I\right) \tag{3.22b}$$
$$-\frac{1}{\tau_\sigma}\left(n\Theta-n\theta I\right)$$

$$\left\langle v\vee v\vee v\,\tilde{\mathcal{C}}(F)\right\rangle = -\frac{1}{\tau_\sigma}\left(nu^{\vee 3}+3n\Theta\vee u+Q\right)\,, \tag{3.22c}$$

where

$$\frac{1}{\tau_w}=\eta_0\,,\qquad \frac{1}{\tau_p}=\eta_1+\eta_0\,,\qquad \frac{1}{\tau_\sigma}=\eta_2+\eta_1+\eta_0\,.$$

Here τ_w is the energy relaxation time, τ_p is the momentum relaxation time, and τ_σ is the relaxation time Σ.

3.2.2 Reduction to Second Order

With the collision terms in (3.22), the evolution for the heat flux tensor is

$$\partial_t Q+u\cdot\nabla_x Q+\nabla_x\cdot(nQ\vee u+3n\Theta\vee u\vee u)-3n\Theta\vee(\nabla\cdot\Theta)$$
$$=-\frac{1}{\tau_\sigma}Q-\frac{1}{\tau_p}\left(n|u|^2I\vee u\right)+\frac{1}{\tau_w}\left(n|u|^2+3n(\theta-\theta_\ell)\right)I\vee u\,. \tag{3.23}$$

Several simplifying assumptions can be made to express Q in terms of lower-order moments. If a stationary balance is assumed, then the time derivative in (3.23) disappears; and if the kinetic energy of the system is assumed to be small relative to the thermal energy, then terms involving the bulk velocity u can be neglected. By

employing these assumptions and the approximation (3.20) for R, one finds a simple expression for Q:

$$Q = -3\tau_\sigma n \left(\Theta \cdot \nabla\right) \vee \Theta \,. \tag{3.24}$$

In his original work [13], Bløtekjær further reduces his model by limiting the dynamics of the problem to one dimension. In such cases, Θ reduces to a diagonal matrix

$$\Theta = \mathrm{diag}\left(\theta_L, \theta_T, \theta_T\right) \,.$$

where θ_L is the temperature component along the dynamic axis and θ_T is the temperature component in each of the two transverse directions. Thus,

$$\theta = \frac{1}{3}\left(\theta_L + 2\theta_T\right) \,.$$

After further simplifications, θ_T is expressed in terms of θ_L, which leads to a closed system in terms of the variables n, u, and θ_L. However, a more natural approach is to take one-half times the trace of the stress equation and then use θ as a fundamental variable rather than θ_L. The resulting system is a closed set of equations that

describes the evolution of concentration, momentum, and energy associated with F:

$$\partial_t n + \nabla_x \cdot (nu) = 0 \tag{3.25a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu^2 + n\theta) - \frac{q_e}{m_e^*} n \nabla_x \Phi = -\frac{1}{\tau_p} nu \tag{3.25b}$$

$$\begin{aligned} \partial_t \left(\frac{n|u|^2}{2} + \frac{3n\theta}{2} \right) & \\ + \nabla_x \cdot \left(\frac{n|u|^2 u}{2} + \frac{5n\theta u}{2} + q \right) & \\ - \frac{q_e}{m_e^*} nu \cdot \nabla_x \Phi = -\frac{1}{\tau_w} \left(\frac{n|u|^2}{2} + \frac{3n}{2} (\theta - \theta_\ell) \right), & \end{aligned} \tag{3.25c}$$

where the variable $q \equiv \frac{1}{2} \operatorname{trace}(Q)$ is the heat flux vector. For Q given by (3.24),

$$q = -\frac{5}{2} \tau_\sigma n \theta \nabla_x \theta \,.$$

Because (3.25) suppresses any non-isotropic features of $\mathcal{G}_{n,u,\Theta}$, one might as well start with a Maxwellian distribution,

$$\mathcal{M}_{n,u,\theta}(v) = \frac{n}{(2\pi\theta)^{3/2}} \exp\left(\frac{|v-u|^2}{2\theta} \right) ,$$

and an approximate collision operator

$$\tilde{\mathcal{C}}(F) = -\eta_1 (F - f_1) - -\eta_0 (F - f_0) \tag{3.26}$$

that is the same as (3.21), but without the non-isotropic relaxation. (As with the Gaussian, the heat flux cannot be evaluated with $\mathcal{M}_{n,u,\theta}$ directly). The result is

the same expression for R as in (3.20) and the same collision terms as in (3.22). Indeed, the advantage of using a Gaussian distribution is lost when the content of Θ is retained in only one component. The Maxwellian distribution is the starting point for the model investigated in [14], where the heat flux is simply given by

$$q = -\kappa \nabla_x \theta \,, \tag{3.27}$$

but the heat diffusivity κ is left unspecified. With the approximate collision operator given in (3.21) and the choice of intermediate states that follow,

$$\kappa = \frac{5}{2} \tau_\sigma n \theta \,.$$

However, in [14], κ is never actually specified—only the form of q in (3.27) and the requirement that $\kappa > 0$.

3.2.3 The Baccarani-Wordemann Expressions

The most popular version of the Bløtekjær model (3.25) is that of Baccarani and Wordemann [8] who find analytical formulas for the heat conduction κ and the relaxation times τ_p and τ_w. The heat conduction is expressed with the Wiedmann-Franz law [44],

$$\kappa = \left(\frac{5}{2} + \gamma \right) \tau_p n \theta, \qquad -5/2 \leq \gamma \leq 0 \,, \tag{3.28}$$

where the parameter γ is the exponent found in the expression for the microscopic relaxation time (3.17). However, as a practical matter, the choice of γ has become

a fit parameter that is chosen to fit Monte Carlo or experimental data. Two values commonly found in the literature are $\gamma = -1.0$ [25, 28] and $\gamma = -2.1$ [25, 32].

The momentum relaxation time, τ_p, is assumed to vary inversely with temperature:

$$\tau_p = \frac{m_e^*}{q}\frac{\theta_\ell}{\theta}\mu_0\,. \tag{3.29}$$

Here μ_0 is the low field mobility that depends on the doping profile and, to a lesser extent, the temperature. To model the energy relaxation time, the mobility $\mu = \frac{q}{m_e^*}\tau_p$ is assumed to vary according to the Caughley-Thomas formula [74],

$$\mu = \mu_0 \left[1 + \left(\frac{\mu_0 \left|\nabla_x \Phi\right|}{v_s}\right)^2\right]^{-1/2}, \tag{3.30}$$

where v_s is the saturation velocity [44, 84] of the electrons. Then both the momentum and energy relaxation times are required to be consistent with the stationary, space-homogeneous form of (3.25)—that is, they are assumed to satisfy

$$\frac{q_e}{m_e^*} n \cdot \nabla_x \Phi = \frac{1}{\tau_p} nu\,, \tag{3.31a}$$

$$\frac{q_e}{m_e^*} nu \cdot \nabla_x \Phi = \frac{1}{\tau_w}\left(\frac{n|u|^2}{2} + \frac{3n\left(\theta - \theta_\ell\right)}{2}\right), \tag{3.31b}$$

Combining (3.29)-(3.31) yields an expression for the relaxation time τ_w:

$$\tau_w = \frac{1}{2}\frac{\mu_0 m_e^*}{q}\frac{\theta_\ell}{\theta}\left(1 + \frac{3\theta}{v_{sat}^2\left(\theta + \theta_\ell\right)}\right). \tag{3.32}$$

The system (3.25), with the heat flux given in (3.27), the heat diffusion in (3.28),

and the relaxation times in (3.29) and (3.32), will henceforth be referred to as the Bløtekjær-Baccarani-Wordeman (BBW) model.

3.2.4 Discussion of the BBW Model

The system (3.25) possesses a great deal of structure because—except for the heat flux—it is completely derivable from the kinetic equations using the approximate collision operator in (3.26) and setting $\mathcal{F}[\boldsymbol{\rho}] = \mathcal{M}_{n,u,\theta}$. In this case $\tilde{\mathcal{C}}$ satisfies (3.5) since

$$\left\langle \log(\mathcal{M}_{n,u,\theta})\tilde{\mathcal{C}}(\mathcal{M}_{n,u,\theta}) \right\rangle = -\frac{1}{2}\frac{n}{\theta\theta_\ell}\left((\theta+\theta_\ell)\,|u|^2 + 3\,(\theta-\theta_\ell)^2\right) < 0\,.$$

Therefore, the transport equation (3.1) implies formally that

$$\partial_t \mathcal{H}(\mathcal{M}_{\rho,u,\theta}) + \nabla_x \cdot \mathcal{J}(\mathcal{M}_{\rho,u,\theta}) < 0\,.$$

and if q locally dissipates the quantity

$$\mathcal{H}(\mathcal{M}_{\rho,u,\theta}) = n\log\left(\frac{n}{(2\pi\theta)^{3/2}}\right) - \frac{5}{2}n\,,$$

then $\mathcal{H}(\mathcal{M}_{\rho,u,\theta})$ will be dissipated by solutions of (3.25). To show that q is dissipative, introduce the energy variable

$$e = \frac{1}{2}n|u|^2 + \frac{3}{2}n\theta\,.$$

A short calculation shows that

$$\frac{\partial}{\partial e}\left(\mathcal{H}(\mathcal{M}_{\rho,u,\theta})\right) = -\frac{1}{\theta}$$

and then

$$\begin{aligned} -\frac{1}{2\theta}\nabla_x \cdot q &= \nabla_x \cdot \left(\frac{1}{2\theta}\kappa\nabla_x\theta\right) - \nabla_x\left(\frac{1}{2\theta}\right)\cdot\kappa\nabla_x\theta \\ &= \nabla_x \cdot \left(\frac{1}{2\theta}\kappa\nabla_x\theta\right) + \frac{\kappa}{2\theta^2}|\nabla_x\theta|^2 . \end{aligned}$$

Since $\kappa > 0$, q dissipates the entropy $\mathcal{H}(\mathcal{M}_{\rho,u,\theta})$. Moreover, as shown in [53], (3.5) implies that (3.25) also dissipates the quantity

$$\mathcal{K}(\mathcal{M}_{\rho,u,\theta}) + \frac{\epsilon}{2m_e^*\theta_\ell}\left|\nabla_x\Phi\right|^2 ,$$

which takes into account interactions between electrons and the lattice as well as presence of the electrical potential Φ. As explained in Section 3.1.1, $\mathcal{K}$ is a more appropriate object than $\mathcal{H}$ for studying the Boltzmann-Poisson system.

From a computational viewpoint, it is important to note that the convective part of (3.25) is just the Euler equations for a compressible fluid; in particular, (3.25) is hyperbolic. It has wave speeds u and $u \pm a$, where $a = \sqrt{\frac{5}{3}\theta}$ is the sound speed. It is possible that solutions to (3.25) possess shocks, in which case special shock capturing methods must be used in numerical simulations. Computation of the BBW model can be found in several places. For time-dependent solutions, see [25] for an ENO

method, [72] for a central scheme approach, and [41] for a relaxation scheme approach. Stationary schemes can be found in [28, 29].

A general criticism of Bløtekjær's model is that the flux closure lacks consistency. As stated in [14], the inclusion of the term $q = -\kappa \nabla_x \theta$, is to incorporate the most "important effect of a non-Maxwellian distribution function". However, using a Gaussian or Maxwellian distribution to then evaluate R in (3.20) is inconsistent with the non-zero heat flux assumption. Thus, the Bløtekjær's approach for expressing fluxes does not fall into the framework presented in the introduction that leads from (3.15) to (3.22). More consistent approaches for deriving q will be presented later in this chapter.

The Baccarani and Wordemann expression for heat diffusion and relaxation times are also subject to criticism. The expression of the heat diffusivity is based on a phenomenological argument rather than a kinetic based derivation, and the use of r as a fitting parameter is very suspect. Its value is varied to compensate for inaccuracies in the model, and it is usually the case that there is a trade-off. Choices of r that produce "good" results for one macroscopic variable (say, the bulk velocity) invariably produce "bad" results for another (say, the temperature). The expressions for the relaxation times are also phenomenological, and the assumptions in (3.31) used to derive the relaxation times lead to significant errors, especially where convective gradients are known to be large.

3.3 The Anile and Pennisi (AP) Model

In [4], a moment model is formulated based on the vector

$$\mathbf{m} = \left(1, v, v \vee v, v|v|^2\right)^T .$$

The moment system takes the form (3.16) where

$$\boldsymbol{\rho} = \begin{pmatrix} n \\ nu \\ nu \vee u + n\Theta \\ nu^2 + 2n\Theta \cdot u + 3n\theta u + 2q) \end{pmatrix}, \tag{3.33a}$$

$$\mathbf{f}(\boldsymbol{\rho}) = \begin{pmatrix} nu \\ nu \vee u + n\Theta \\ nu^{\vee 3} + 3n\Theta \vee u + Q \\ \left(n|u|^2 + 3n\theta\right) u \vee u + |u|^2\Theta + 4(\Theta \cdot u) \vee u \\ +2Q \cdot u + 4q \vee u + r \end{pmatrix}, \tag{3.33b}$$

$$\nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ n\nabla_x \Phi \\ 2nu \vee \nabla_x \Phi \\ 2n\left(u \cdot \nabla_x \Phi\right) u + n\nabla_x \Phi |u|^2 \end{pmatrix}. \tag{3.33c}$$

All of the variables used here have been previously defined except for $r \equiv \text{trace}\,(R)$. The collision terms have yet to be specified.

3.3.1 Extended Thermodynamics

The closure in [4] is based on the principles of extended thermodynamics [61], which impose an entropy structure at the continuum level. For an arbitrary vector $\mathbf{m}$, this formulation is based on two assumptions: first, that the moments $\langle v\mathbf{m}f\rangle$ and $\langle \mathbf{m}\mathcal{C}(f)\rangle$ can be expressed in terms of the densities $\boldsymbol{\rho} = \langle \mathbf{m}f\rangle$ to provide a closure of the form (3.16) and, second, that there exists a strictly convex entropy $h = h(\boldsymbol{\rho})$, an entropy flux $j = j(\boldsymbol{\rho})$, and a dissipation term $d(\boldsymbol{\rho})$ such that

$$\partial_t h(\boldsymbol{\rho}) + \nabla_x \cdot j(\boldsymbol{\rho}) = d(\boldsymbol{\rho})\,, \tag{3.34}$$

where $d(\boldsymbol{\rho}) \leq 0$. (Actually, the sign convention for physical entropy is minus the mathematical entropy, so that the inequality in reversed in most physics texts and $-d$ is called a production term.)

The existence of a strictly convex entropy provides a great deal of structure to the system (3.16). It turns out that h^*, the Legendre transform of h, plays an important role. If h is sufficiently smooth, then h^* is defined through the implicit relation

$$h(\boldsymbol{\rho}) + h^*(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{\rho} \tag{3.35}$$

where

$$\boldsymbol{\alpha} = \left[\frac{\partial h}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho})\right]^T \quad . \tag{3.36}$$

One may readily verify that

$$\boldsymbol{\rho} = \left[\frac{\partial h^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right]^T \quad , \tag{3.37}$$

whereby $(h^*)^* = h$. Because of (3.37), h^* is frequently referred to as the *density potential*.

Multiplying (3.16) by $\boldsymbol{\alpha}$ gives

$$\partial_t h(\boldsymbol{\rho}) + \boldsymbol{\alpha}^T \left(\nabla_x \cdot \mathbf{f}(\boldsymbol{\rho}) - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) - \mathbf{r}(\boldsymbol{\rho}) \right) = 0$$

whenever $\boldsymbol{\rho}$ is continuously differentiable. If we identify

$$d(\boldsymbol{\rho}) = \boldsymbol{\alpha}^T \left(\nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) + \mathbf{r}(\boldsymbol{\rho}) \right) ,$$

it follows then that

$$\frac{\partial h}{\partial \boldsymbol{\rho}} \frac{\partial \mathbf{f}}{\partial \boldsymbol{\rho}} = \frac{\partial j}{\partial \boldsymbol{\rho}} \tag{3.38}$$

and, by taking derivatives with respect to $\boldsymbol{\rho}$, that

$$\frac{\partial^2 h}{\partial \boldsymbol{\rho}^2} \frac{\partial \mathbf{f}}{\partial \boldsymbol{\rho}} + \frac{\partial h}{\partial \boldsymbol{\rho}} \frac{\partial^2 \mathbf{f}}{\partial \boldsymbol{\rho}^2} = \frac{\partial^2 j}{\partial \boldsymbol{\rho}^2} \, .$$

For clarity, we repeat this last relation using indical notation:

$$\frac{\partial^2 h}{\partial \rho_a \partial \rho_s}\frac{\partial f_{ab}}{\partial \rho_r} + \frac{\partial h}{\partial \rho_a}\frac{\partial^2 f_{ab}}{\partial \rho_r \partial \rho_s} = \frac{\partial^2 j_b}{\partial \rho_r \partial \rho_s}$$

Because the Hessians of $\mathbf{f}$ and j are symmetric in the r and s indices, it follows that

$$\frac{\partial^2 h}{\partial \boldsymbol{\rho}^2}\frac{\partial \mathbf{f}}{\partial \boldsymbol{\rho}} = \left[\frac{\partial^2 h}{\partial \boldsymbol{\rho}^2}\frac{\partial \mathbf{f}}{\partial \boldsymbol{\rho}}\right]^T . \tag{3.39}$$

Since h is convex, its Hessian is symmetric and positive definite. By (3.39), it therefore symmetrizes $\partial \mathbf{f}/\partial \boldsymbol{\rho}$, which makes (3.16) a symmetric hyperbolic system [27]. If the *flux potential* j^* is defined by

$$j^*(\boldsymbol{\alpha}) \equiv \boldsymbol{\alpha}^T \mathbf{f}(\boldsymbol{\rho}) - j(\boldsymbol{\rho}) , \tag{3.40}$$

then (3.38) implies

$$\mathbf{f}(\boldsymbol{\rho}) = \left[\frac{\partial j^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right]^T , \tag{3.41}$$

This relation justifies the name for j^* and allows one to write (3.16) in the potential form

$$\partial_t \left(\left[\frac{\partial h^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right]^T\right) + \nabla_x \cdot \left[\frac{\partial j^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right]^T - \nabla_x \Phi \cdot \mathbf{l}\left(\left[\frac{\partial h^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right]^T\right) = \mathbf{r}\left(\left[\frac{\partial h^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\right]^T\right) .$$

3.3.2 Closure

The closure process consists of approximating the flux terms and the collision terms. The field terms are already given in terms of the densities.

3.3.2.1 Flux Terms In theory, the closure of $\mathbf{f}$ is computed as follows. Given the densities $\boldsymbol{\rho}$ and the entropy $h(\boldsymbol{\rho})$, one first computes the variable $\boldsymbol{\alpha}$ using (3.36). Then $\boldsymbol{\alpha}$ is used to evaluate the flux $\mathbf{f}(\boldsymbol{\rho}) = j^*_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$. However in practice, an analytical expression for the entropy is often lacking and finding $\boldsymbol{\alpha}$ is non-trivial. The standard approach in extended thermodynamics is to expand $\boldsymbol{\alpha}$ around some fixed value for which $h(\boldsymbol{\rho})$ has a known analytical expression. Often this is at thermal equilibrium or at what is referred to as partial thermal equilibrium [4]. Expansions are then truncated to provide analytical expressions with which to approximate $\boldsymbol{\alpha}$ and $\mathbf{f}$.

For the system (3.33a)-(3.33c) a closure for $\mathbf{f}$ amounts to specifying Q and r. Using the extended thermodynamics approach, these terms are given in [4] by

$$Q = \frac{2}{5} q \vee I, \qquad r = n\theta \left(7\Theta - \frac{16}{3}\theta I \right) .$$

3.3.2.2 Collision Terms The procedure for closing collision terms is more ambiguous. Extended thermodynamics places restrictions on these terms, but does not always specify them completely. Therefore, based on physical considerations, Anile

and Pennisi use collision terms borrowed from Bløtekjær's model (3.22)

$$\langle v\,\mathcal{C}(F)\rangle = -\frac{1}{\tau_p}nu \tag{3.42a}$$

$$\langle v\vee v\,\mathcal{C}(F)\rangle = -\frac{1}{\tau_p}\left(nu\vee u-\frac{1}{3}|u|^2I\right)-\frac{1}{\tau_w}\left(\frac{1}{3}n|u|^2I+n(\theta-\theta_\ell)I\right) \tag{3.42b}$$

$$-\frac{1}{\tau_\sigma}\left(n\left(\Theta-\theta I\right)\right)$$

$$\left\langle v|v|^2\mathcal{C}(F)\right\rangle = -\frac{1}{\tau_q}\left(2q+5n\theta u\right)\,. \tag{3.42c}$$

Note that (3.42c) is almost the trace of the collision term in (3.22c) with $\Theta=\theta I$ and $\tau_q=\tau_\sigma$. The difference is a term $n|u|^2$ that, because u is considered small in some sense, will be neglected anyway in the reduction that follows. The parameter τ_q is the relaxation time for the energy flux, which is the moment with respect to $\frac{1}{2}|v|^2v$. The remaining relaxation times are the same as for the Bløtekjær derivation. It turns out [4] that the entropy dissipation relation places restrictions on their relative values. In practice, they are usually some functional form in terms of the average electron energy or the energy flux-to-energy ratio. The parameters for these forms are fit according to Monte-Carlo data. It is generally accepted that such fits are more accurate than the Bacarrani-Wordeman expressions. However, they are device dependent, meaning that change in physical specifications or applied voltage requires a new fit. Using these Monte Carlo calculations, it has been found that $0<\tau_w<\tau_p<\tau_\sigma<\tau_q$ for all relevant ranges of the energy [63].

3.3.3 Reduction to Second Order

As in the Bløtekjær case, the model of Anile and Pennisi is reduced to a smaller system, this time by an asymptotic analysis known as Maxwellian iteration [61] that proceeds as follows. First, the stress tensor Θ is separated into its trace and traceless parts:

$$\Theta = \theta I + \frac{1}{n}\Sigma\,.$$

The equation for Σ is

$$\partial_t\Sigma + \nabla_x\cdot(\Sigma u) + \frac{2}{5}\left(\nabla_x \vee q\right) - \frac{2}{15}\nabla_x\cdot q + 2\left(n\Theta\cdot\nabla_x\right)\vee u - \frac{2}{3}n\Theta : \nabla_x u I = -\frac{1}{\tau_\sigma}\Sigma\,. \quad (3.43)$$

Beginning with $s = 0$, the Maxwellian iteration is performed by placing the s iterate of Σ on the left hand side of (3.43), and then solving for the $s+1$ iterate on the right-hand side. Terms that are nonlinear in the $s+1$ iterate are neglected (which is why neglecting $n|u|^2$ in (3.42c) is not an issue). The zeroth iterate, $\Sigma^{(0)} = 0$, corresponds to the value of Σ at thermal equilibrium. The result at the first iteration is a balance between the right hand side of (3.43) and the last two terms of the left hand side evaluated at $\Theta = \theta I$. If $\Sigma^{(1)}$ is used to approximate Σ, then

$$\begin{aligned}\Sigma &= -2\tau_\sigma n\theta\left(\nabla_x \vee u - \frac{1}{3}\left(\nabla\cdot u\right) I\right) \quad (3.44)\\ &= -\tau_\sigma n\theta\left(\nabla_x u + \left(\nabla_x u\right)^T - \frac{2}{3}\left(\nabla\cdot u\right) I\right)\end{aligned}$$

The next step is to find a simple expression for q, whose evolution is given by

$$\begin{aligned}
&\partial_t q + u\cdot\nabla_x q + \frac{7}{5}q\left(\nabla_x\cdot u\right) + \frac{5}{2}n\theta\nabla_x\theta \\
&\quad - \frac{7}{2}n\theta\Sigma\cdot\nabla_x n + \frac{2}{5}\left(\nabla_x u\right)\cdot q + \theta\nabla_x\cdot\Sigma \\
&\quad + \frac{5}{2}\Sigma\cdot\nabla_x\left(n\theta\right) + \frac{7}{5}q\cdot\nabla_x u - \Sigma\cdot\left(\nabla_x\cdot\Sigma\right) \\
&\qquad = -\frac{1}{\tau_q}\left(q + \frac{5}{2}nu\theta\right) + \frac{1}{\tau_p}\left(\frac{5}{2}n\theta u + \Sigma\cdot u\right) .
\end{aligned}$$

Maxwell iteration for q gives the first-order balance

$$\frac{5}{2}n\theta\nabla_x\theta = -\frac{1}{\tau_q}\left(q + \frac{5}{2}nu\theta\right) + \frac{1}{\tau_p}\frac{5}{2}n\theta v$$

which implies that

$$q = -\frac{5}{2}\tau_q n\theta\nabla_x\theta + \frac{5}{2}n\theta u\left(\frac{\tau_q}{\tau_p} - 1\right) . \tag{3.45}$$

The balances for Σ and q reduce (3.16),(3.33) to a second-order, closed system of equations for the concentration, momentum, and energy:

$$\partial_t n + \nabla_x\cdot\left(nu\right) = 0 \tag{3.46a}$$

$$\partial_t(nu) + \nabla_x\cdot\left(nu^2 + n\theta + \Sigma\right) + \frac{q_e}{m_e^*}n\nabla_x\Phi = -\frac{1}{\tau_p}nu \tag{3.46b}$$

$$\begin{aligned}
&\partial_t\left(\frac{n|u|^2}{2} + \frac{3n\theta}{2}\right) \\
&\qquad + \nabla_x\cdot\left(\frac{n|u|^2u}{2} + \frac{5n\theta u}{2} + q\right) \\
&\qquad\qquad + \frac{q_e}{m_e^*}nu\cdot\nabla_x\Phi = -\frac{1}{\tau_w}\left(\frac{n|u|^2}{2} + \frac{3n}{2}\left(\theta - \theta_\ell\right)\right) ,
\end{aligned} \tag{3.46c}$$

where Σ and q are given by (3.44) and (3.45), respectively. This system (3.46) will henceforth be referred to as the Anile-Pennisi (AP) model.

3.3.4 Discussion of the AP Closure

One of the primary objectives of the AP closure was to have a more rigorous derivation of the heat flux that includes a convective component as seen in (3.45) and argued for in [49, 80] that the heat flux should have a convective component . However, it remains to be seen how much the convective component really affects the accuracy of lower-order moments. Difficulties with any of these models usually occur at material junctions where spatial gradients are large. Thus is it reasonable to expect that diffusive contributions to the heat flux will dominate in these areas. In the next chapter, we will check numerically if this is indeed the case.

The extended thermodynamic approach assumes the existence of a strictly convex entropy as a fundamental principle from which hyperbolicity results. However, it is not clear if either property survives the expansion process used to approximate the Lagrange multipliers or the process of Maxwell iteration used to deduce (3.34) from the original third-order system. Moreover, the extended thermodynamic approach is a bit awkward when considering the potential Φ. In fact, nowhere in (3.46) is it actually clear where Φ plays a role. As we shall see later, the formal presentation must be altered slightly to allow for more general situations.

Numerical results for the AP model can be found in [63, 72]. However, neither of references uses the exact form of the closure given in (3.46). In [72], the authors use a

splitting method with central schemes [54,65] to compute a hydrodynamic model that uses the expression for q from the AP closure, but sets $\Sigma = 0$. In [63], an iterative Newton-type scheme is used to find steady-state solutions for a model that uses q and Σ in the energy equation, but ignores Σ in the momentum equation. In both cases, the relaxation times are computed with Monte Carlo simulations as functions of the average energy. Although it is generally accepted that this approach is more accurate than the Baccarani-Wordemann expressions, it is still subject to criticism since it discounts the effect of other macroscopic variables or variations in the electric field.

3.4 Entropy-Based Closures

Given a polynomial vector $\mathbf{m}$ and densities $\boldsymbol{\rho} = \langle \mathbf{m} F \rangle$, there are an infinite number of functions $f \in \mathbb{F}_{\mathbf{m}}$ such that $\langle \mathbf{m} f \rangle = \boldsymbol{\rho}$. The *minimum entropy principle* (or maximum entropy principle if you are a physicist) provides a criterion for selecting the appropriate function $\mathcal{F}[\boldsymbol{\rho}]$ to approximate F. It states that the most likely distribution that is consistent with the constraints $\langle \mathbf{m} f \rangle = \boldsymbol{\rho}$ is the distribution that minimizes the kinetic entropy of the system. This discovery of this principle in the context of equilibrium thermodynamics is usually attributed to E.T. Jaynes, although in his first paper on the subject [37], Jaynes states, "The mathematical facts concerning the maximization of entropy ... were pointed out long ago by Gibbs." He continues by crediting also C.E. Shannon, whose work in information theory [75, 79] showed that "the expression for entropy has a deeper meaning, quite independent

of thermodynamics." Thus "the fact that a probability distribution maximizes the entropy subject to certain constraints becomes the essential fact which justifies use of that distribution for inference."

3.4.1 Relationship with Extended Thermodynamics

In [61], it is proved that kinetic closures based on entropy minimization are formally equivalent to the systems derived from extended thermodynamics. Given a density F and a vector of polynomials $\mathbf{m}$, let $\boldsymbol{\rho} = \langle \mathbf{m} F \rangle$ and let $\mathcal{F}[\boldsymbol{\rho}]$ be the minimizer that solves

$$h(\boldsymbol{\rho}) \equiv \min \{ \mathcal{H}(f) : \langle \mathbf{m} f \rangle = \boldsymbol{\rho} \} \ . \tag{3.47}$$

If the minumun in (3.47) exists, and if $\mathcal{H}$ is differentiable at the solution, then standard Lagrange multiplier theory implies that

$$\mathcal{F}[\boldsymbol{\rho}] = \exp(\boldsymbol{\alpha}^T \mathbf{m}) \, . \tag{3.48}$$

It is clear from (3.47) that $\boldsymbol{\alpha} \in \mathbb{R}^n$ is related to $\boldsymbol{\rho}$ through the constraints:

$$\left\langle \mathbf{m} \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle = \boldsymbol{\rho} \, . \tag{3.49}$$

We now that this relation is invertible for $\boldsymbol{\alpha}$ as a function of $\boldsymbol{\rho}$.

Following [51], one can identify explicitly the density and flux potentials

$$h^*(\boldsymbol{\alpha}) \equiv \left\langle \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle \qquad \text{and} \qquad j^*(\boldsymbol{\alpha}) \equiv \left\langle v \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle \ .$$

Formally differentiating h^* with respect to $\boldsymbol{\alpha}$ recovers the constraint relations in (3.48). Hence

$$h(\boldsymbol{\rho}) + h^*(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \boldsymbol{\rho} \tag{3.50}$$

which implies that h^* is, in fact, the Legendre dual of h. Moreoever, because

$$\frac{\partial^2 h^*}{\partial \boldsymbol{\alpha}^2}(\boldsymbol{\alpha}) = \left\langle \mathbf{m}\mathbf{m}^T \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle$$

is positive definite, the relation (3.49) may be inverted for $\boldsymbol{\alpha}$ as a function of $\boldsymbol{\rho}$. As a result, the closure

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \langle v\mathbf{m}\mathcal{F}[\boldsymbol{\rho}] \rangle - \nabla_x \Phi \cdot \langle \nabla_v \mathbf{m}\mathcal{F}[\boldsymbol{\rho}] \rangle = \langle \mathbf{m}\mathcal{C}(\mathcal{F}[\boldsymbol{\rho}]) \rangle \tag{3.51}$$

possesses an auxiliary entropy equation of the form (3.34), where

$$h(\boldsymbol{\rho}) = \mathcal{H}(\mathcal{F}[\boldsymbol{\rho}])\,, \qquad j(\boldsymbol{\rho}) = \mathcal{J}(\mathcal{F}[\boldsymbol{\rho}])\,, \qquad d(\boldsymbol{\rho}) = \mathcal{D}(\mathcal{F}[\boldsymbol{\rho}])\,.$$

Furthermore, differentiating (3.50) with respect to $\boldsymbol{\rho}$ shows that

$$\boldsymbol{\alpha} = \left[\frac{\partial h}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}) \right]^T . \tag{3.52}$$

and that

$$\frac{\partial^2 h}{\partial \boldsymbol{\rho}^2} = \left[\frac{\partial^2 h^*}{\partial \boldsymbol{\alpha}^2} \right]^{-1}$$

is positive definite. Thus h is convex.

In order to fit into the framework of extended thermodynamics, it must one must also show that h is dissipated by solution of (3.51). This can be done by rewriting (3.51) in the form

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot j^*_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \left\langle \nabla_v \mathbf{m} \, \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle = \left\langle \mathbf{m} \mathcal{C}(\exp(\boldsymbol{\alpha}^T \mathbf{m})) \right\rangle$$

where $\boldsymbol{\alpha}$ is given by (3.52). Multiplying this equation by $\boldsymbol{\alpha}^T$ gives

$$\begin{aligned} &\partial_t h(\boldsymbol{\rho}) + \boldsymbol{\alpha}^T \nabla_x \cdot j^*(\boldsymbol{\alpha}) \\ &\qquad - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \left\langle \nabla_v (\boldsymbol{\alpha}^T \mathbf{m}) \, \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle = \left\langle \boldsymbol{\alpha}^T \mathbf{m} \, \mathcal{C}(\exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle \end{aligned}$$

Using the fact that

$$\begin{aligned} \alpha^T \nabla_x \cdot j^*(\boldsymbol{\alpha}) &= \nabla_x \cdot (\boldsymbol{\alpha}^T j^*(\boldsymbol{\alpha})) - (\nabla_x^T \boldsymbol{\alpha}) \cdot j^*(\boldsymbol{\alpha}) \\ &= \nabla_x \cdot (\boldsymbol{\alpha}^T j^*(\boldsymbol{\alpha}) - j^*(\boldsymbol{\alpha})) \end{aligned}$$

and

$$\left\langle \nabla_v (\boldsymbol{\alpha}^T \mathbf{m}) \, \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle = \left\langle \nabla_v \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle = 0 \,,$$

it follows that h satisfies (3.34), where

$$j(\boldsymbol{\rho}) = \boldsymbol{\alpha}^T j^*(\boldsymbol{\alpha}) - j^*(\boldsymbol{\alpha}) \,, \qquad d(\boldsymbol{\rho}) = \left\langle \boldsymbol{\alpha}^T \mathbf{m} \, \mathcal{C}(\exp(\boldsymbol{\alpha}^T \mathbf{m})) \right\rangle . \tag{3.53}$$

However, since $\mathcal{H}$ is not always dissipated by solutions of (3.1), there is no guarantee

that $d(\boldsymbol{\rho}) < 0$ in (3.53).

In addition to an explicit, kinetic-based formulation, the minimum entropy method also provides an algorithm for computing $\boldsymbol{\alpha}$ that is not readily available in the extended thermodynamics theory. Instead of resorting to approximations near local equilibrium, one can instead solve the dual problem for (3.47)

$$h^*(\hat{\boldsymbol{\alpha}}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\{ \boldsymbol{\alpha}^T \boldsymbol{\rho} - h(\boldsymbol{\rho}) \right\} . \tag{3.54}$$

However, an algorithm based on (3.54) is still lacking. Issues that preclude a standard implementation of (3.54) will be discussed in the following chapter.

3.4.2 Relative Entropy Formulation

Even though the entropy $\mathcal{H}$ gives rise to an extended thermodynamic closure, its validity at the kinetic level is questionable because it does not give rise to an H-Theorem. Indeed, as discussed in Section 3.1, the condition

$$\delta(\mathcal{H}; \mathcal{C}(f)) \doteq \langle \log(f) \mathcal{C}(f) \rangle = 0$$

does not correctly characterize the manifold of equilibria for $\mathcal{C}$. (Recall that $\doteq$ denotes a formal calculation.) Moreover, there is no guarantee that $h(\boldsymbol{\rho})$ will be dissipated. The discrepancy here is due to the fact that $\mathcal{H}$ neglects the presence of the electrostatic potential Φ and the interaction of the electrons with the lattice.

Rather than $\mathcal{H}$, the relevant object to consider [53] is the entropy relative to the

lattice, given by

$$\mathcal{K}(f) \equiv \left\langle f \log\left(\frac{f}{M_\ell}\right) - f \right\rangle .$$

and the dissipation law (3.7) gives rise to an extended thermodynamic description. The entropy minimization problem for $\mathcal{K}$ is

$$k(\boldsymbol{\rho}) \equiv \min\left\{\mathcal{K}(f) : \langle \mathbf{m} f \rangle = \boldsymbol{\rho}\right\} , \tag{3.55}$$

and its formal solution is given by

$$\mathcal{F}[\boldsymbol{\rho}] = M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) , \tag{3.56}$$

where M_ℓ is given by (3.6) and $\boldsymbol{\beta}$ is determined by the constraint equations

$$\left\langle \mathbf{m} M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle = \boldsymbol{\rho} . \tag{3.57}$$

The minimization problems (3.47) and (3.55) are equivalent if and only if $|v|^2 \in \text{span}\{\mathbf{m}\}$, in which case their minimizer is the same and

$$\boldsymbol{\beta}^T \mathbf{m} + \frac{|v|^2}{2\theta_\ell} = \boldsymbol{\alpha}^T \mathbf{m} .$$

All of the formal structure in the previous section remains modulo this simple change

of variables between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Plugging (3.56) into (3.55) gives

$$k(\boldsymbol{\rho}) = \mathcal{K}(M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m})) = h(\boldsymbol{\rho}) - \left\langle \frac{|v|^2}{2\theta_\ell} F \right\rangle ,$$

and it easy to see from (3.48) and (3.56) that

$$k^*(\boldsymbol{\beta}) \equiv \left\langle M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle = h^*(\boldsymbol{a}) .$$

Thus, k and k^* are strictly convex Legendre duals that satisfy

$$k(\boldsymbol{\rho}) + k^*(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{\rho} ,$$

where

$$\boldsymbol{\beta} = \left[\frac{\partial k}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}) \right]^T , \qquad \boldsymbol{\rho} = \left[\frac{\partial k^*}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \right]^T .$$

When $\mathbf{m} = \{1\}$ or $\mathbf{m} = \{1, v\}$, then (3.47) and (3.55) will differ since $\mathcal{H}$ is no longer bounded below on the set $\mathbb{F}_{\mathbf{m}}$. However (3.55) still possess a finite solution and, with it, all of the formal structure described here.

Following [53], one can identify a locally dissipation law for the moment closure based on (3.55) that is an analog for the kinetic law stated in (3.13).

Proposition 5 (Levermore) *If $\mathcal{F}[\boldsymbol{\rho}]$ is defined as the minimizer of (3.55), then solutions of (3.51) formally a local dissipation law for the quantity*

$$K(\boldsymbol{\rho}, \nabla_x \Phi) \equiv k(\boldsymbol{\rho}) + \frac{\epsilon}{2 m_e^* \theta_\ell} |\nabla_x \Phi|^2 \tag{3.58}$$

The dissipation vanishes if and only if $\mathcal{F}[\boldsymbol{\rho}] = n$.

Proof. The closure (3.51) takes the form

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot i^*_\beta(\boldsymbol{\beta}) - \frac{q_e}{m^*_e} \nabla_x \Phi \cdot \left\langle \nabla_v \mathbf{m}\, M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle = \left\langle \mathbf{m}\, \mathcal{C}\left(M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right) \right\rangle , \quad (3.59)$$

where

$$i^*(\boldsymbol{\beta}) = \left\langle v M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle .$$

Multiplying (3.59) by $\boldsymbol{\beta}^T$ gives

$$\partial_t k(\boldsymbol{\rho}) + \boldsymbol{\beta}^T \nabla_x \cdot i^*_\beta(\boldsymbol{\beta}) \quad (3.60)$$

$$- \frac{q_e}{m^*_e} \nabla_x \Phi \cdot \left\langle \nabla_v (\boldsymbol{\beta}^T \mathbf{m})\, M_\ell \exp\left(\boldsymbol{\beta}^T \mathbf{m}\right) \right\rangle = \left\langle \boldsymbol{\beta}^T \mathbf{m}\, \mathcal{C}\left(M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right) \right\rangle . \quad (3.61)$$

We first put the flux term into divergent form:

$$\begin{aligned} \boldsymbol{\beta}^T \nabla_x \cdot i^*(\beta) &= \nabla_x \cdot (\boldsymbol{\beta}^T i^*(\boldsymbol{\beta})) - (\nabla_x^T \boldsymbol{\beta}) \cdot i^*(\boldsymbol{\beta}) \\ &= \nabla_x \cdot (\boldsymbol{\beta}^T i^*(\boldsymbol{\beta}) - i^*(\boldsymbol{\beta})) \\ &\equiv i(\boldsymbol{\rho}) . \end{aligned}$$

Now to handle the field term, we compute

$$\begin{aligned}\left\langle \nabla_v(\boldsymbol{\beta}^T\mathbf{m})\, M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle &= \left\langle M_\ell \nabla_v \left(\exp(\boldsymbol{\beta}^T\mathbf{m})\right)\right\rangle \\ &= \left\langle \nabla_v \left(M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right)\right\rangle - \left\langle (\nabla_v M_\ell)\exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle \\ &= -\frac{1}{\theta_\ell}\left\langle v M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle\end{aligned}$$

so that (3.60) becomes

$$\begin{aligned}&\partial_t k(\boldsymbol{\rho}) + \nabla_x \cdot i(\boldsymbol{\rho}) \\ &\quad - \frac{q_e}{m_e^*\theta_\ell}\nabla_x\Phi \cdot \left\langle v M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle = \left\langle \boldsymbol{\beta}^T\mathbf{m}\, \mathcal{C}(M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m}))\right\rangle .\end{aligned} \tag{3.62}$$

To put the field term into divergence form, we differentiate by parts:

$$\begin{aligned}\nabla_x\Phi \cdot \left\langle v M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle =& \nabla_x \cdot \Phi \left\langle v M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle \\ &- \Phi\nabla_x \cdot \left\langle v M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle ,\end{aligned} \tag{3.63}$$

then use Poisson's equation with fact that $\left\langle M_\ell \exp\left(\boldsymbol{\beta}^T\mathbf{m}\right)\right\rangle = \langle F\rangle$ to manipulate the non-divergent term on the right-hand side of (3.63):

$$\begin{aligned}\Phi\nabla_x \cdot \left\langle v M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle &= -\Phi\partial_t \left\langle M_\ell \exp(\boldsymbol{\beta}^T\mathbf{m})\right\rangle \\ &= -\frac{1}{q_e}\Phi\partial_t \left(\nabla_x \cdot (\epsilon\nabla_x\Phi)\right) \\ &= \frac{1}{q_e}\partial_t\left(\frac{\epsilon}{2}|\nabla_x\Phi|^2\right) - \nabla_x \cdot (\epsilon\Phi\partial_t\nabla_x\Phi) .\end{aligned} \tag{3.64}$$

Together (3.62), (3.63), and (3.64) give

$$\begin{aligned}
&\partial_t \left(k(\boldsymbol{\rho}) + \frac{\epsilon}{2m_e^*\theta_\ell} |\nabla_x \Phi|^2 \right) \\
&\quad + \nabla_x \cdot \left(i(\boldsymbol{\rho}) - \frac{\epsilon}{\theta_\ell} \Phi \partial_t \nabla_x \Phi) - \frac{q_e}{m_e^* \theta_\ell} \Phi \left\langle v M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle \right) \\
&\quad - \frac{q_e}{m_e^* \theta_\ell} \nabla_x \Phi \cdot \left\langle v M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle \\
&\qquad = \left\langle \boldsymbol{\beta}^T \mathbf{m} \, \mathcal{C}(M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m})(\right\rangle \, ,
\end{aligned}$$

and according to (3.5),

$$\begin{aligned}
\left\langle \boldsymbol{\beta}^T \mathbf{m} \, \mathcal{C}(M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m}) \right\rangle &= \left\langle \log(\exp(\boldsymbol{\beta}^T \mathbf{m})) \, \mathcal{C}(M_\ell \exp(\boldsymbol{\beta}^T \mathbf{m})) \right\rangle \\
&\leq 0 \, .
\end{aligned}$$

with equality if and only if $\exp(\boldsymbol{\beta}^T \mathbf{m}) = n$. This concludes the proof. ■

3.4.3 Well-Posedness of Entropy-Based Closures

So far, the discussion of entropy-based closure has been completely formal, and the issue of when solutions to (3.47) and (3.55) exist has been largely suppressed. Indeed, it is known that there are functions $f \in \mathbb{F}_{\mathbf{m}}$ whose moments cannot be realized by *any* exponential function—that is, for *no* value of $\boldsymbol{\alpha} \in \mathbb{R}^n$ does

$$\left\langle \mathbf{m} \exp(\boldsymbol{\alpha}^T \mathbf{m}) \right\rangle = \langle \mathbf{m} f \rangle \, .$$

In such degenerate cases, neither (3.47) or (3.55) will yield a solution [42, 43, 73]. Roughly speaking, these cases occur because neither $\mathcal{H}$ or $\mathcal{K}$ inflicts a strong enough penalty on distributions with too much mass in the tails. When this happens, a minimizing sequence $\{g_s\}_{s=1}^{\infty}$ for $\mathcal{H}$ or $\mathcal{K}$ will still converge in L^1 to a function g. However, g will no longer satisfy the constraint equation, i.e., $\langle \mathbf{m} g \rangle \neq \langle \mathbf{m} f \rangle$.

The issue of non-realizability leads one to question whether entropy minimization is equivalent to extended thermodynamics. This objection was first raised in [42]. The problem is that the extended thermodynamic approach *assumes* the existence of the entropy h which is explicitly provided by the minimizer of either $\mathcal{H}$ or $\mathcal{K}$. When a minimizer does not exist, extended thermodynamics still finds coefficients $\boldsymbol{\alpha}$ that are obtained through an approximate expansion. In such cases, the two approaches are significantly different.

In [3] (and references therein), higher-order closures (greater than two) are based on the entropy minimization problem (3.47) for $\mathcal{H}$. Since these models include $|v|^2$ (or the energy equivalent in the non-parabolic case), the discrepancies between $\mathcal{H}$ and $\mathcal{K}$ do not matter here. However, the closures derived in these studies continue to follow the extended thermodynamics practice of approximate expansions that result in analytical closures, even when the entropy minimization problem does not have a solution. In such degenerate cases, it is not clear whether these systems are hyperbolic or even dissipate an entropy.

In the next chapter, we will investigate the minimization problem in detail using the machinery of optimization theory [56]. The goal there is to characterize the set of non-realizable functions as completely as possible. In doing so we will recover and

extend many of the results found in [42, 43] and in [73].

3.4.4 Generalized BGK Collision Operators

In this subsection, we will develop generalized BGK (GBGK) operators for a full collision operator $\mathcal{C}$. These operators were first introduced in [51] in the context of neutral fluids. Specifying a collision operator is important not only for evaluating macroscopic collision terms, but also for the pertubation procedure that will be introduced in the next section. We begin with some notation.

We define an *admissible* space $\mathbb{M}$ as any linear space of polynomials in v such that:

(i) $\mathbb{M} \supset \mathbb{M}_0 \equiv \text{span}\{1\}$;

(ii) $\mathbb{M}$ is invariant with respect to rotation;

(iii) The cone $\mathbb{M}_c = \{p \in \mathbb{M} : \langle \exp(p) \rangle < \infty\}$ has nonempty interior.

Suppose that we are building a moment system with a vector $\mathbf{m}$ whose l components are polynomials that form a basis for $\mathbb{M}$, and let

$$\mathbb{M}_0 \subsetneq \mathbb{M}_1 \subsetneq \ldots \subsetneq \mathbb{M}_s \subsetneq \mathbb{M}$$

be a finite sequence of admissible spaces. For each r, $0 \leq r \leq s$, let $\mathbf{m}_{(r)}$ be a vector

whose $l_{(r)}$ components are polynomials that form a basis for $\mathbb{M}_r$ and set

$$\boldsymbol{\rho} = \langle \mathbf{m} F \rangle \ , \qquad \boldsymbol{\rho}_{(r)} = \left\langle \mathbf{m}_{(r)} F \right\rangle , \ 0 \leq r \leq s.$$

(We note that the use of parentheses in the subscripts are to differentiate $\mathbf{m}_{(r)}$ from the quantity $\mathbf{m}_r$ introduced and frequently used in the the next chapter). For a given function $f \in \mathbb{F}_{\mathbf{m}}$, the *entropic projection* of f with respect to $\mathbf{m}$ is defined as the solution of (3.55) and is denoted $\mathcal{E}(f; \mathbf{m})$. Similarly, $\mathcal{E}(f; \mathbf{m}_{(r)})$ is the entropic projection of f with respect to $\mathbf{m}_{(r)}$ and is defined as the solution (3.55)—when it exists—except with the constraints $\left\langle \mathbf{m}_{(r)} f \right\rangle = \boldsymbol{\rho}_{(r)}$. In what follows, we will use the shorthand notation

$$\mathcal{E}_f \equiv \mathcal{E}(f; \mathbf{m}) \qquad \text{and} \qquad \mathcal{E}_f^r \equiv \mathcal{E}(f; \mathbf{m}_{(r)}).$$

With this notation, $\mathcal{E}_f^0$ is the equilibrium associated with f, i.e., $\mathcal{E}_f^0 = \langle f \rangle M_\ell$.

The sequence $\{\mathcal{E}_f^r\}_{r=0}^s$ of entropic projection is used to construct an approximation of the full collision operator $\mathcal{C}$. This approximation is a multi-stage relaxation operator:

$$\tilde{\mathcal{C}}(f) = -\nu_s \left(f - \mathcal{E}_f^s \right) - \sum_{r=0}^{s-1} \nu_r \left(\mathcal{E}_f^{r+1} - \mathcal{E}_f^r \right), \qquad \mathrm{Dom}(\tilde{\mathcal{C}}) = \mathrm{Dom}(\mathcal{C}) \cap \mathbb{F}_{\mathbf{m}} , \quad (3.65)$$

where $\{\nu_r\}_{r=0}^s$ is an increasing sequence of positive numbers, each of which depends on $\boldsymbol{\rho}_{(r)}$. Each ν_r is the rate at which $\mathcal{E}_f^{r+1}$ relaxes to $\mathcal{E}_f^r$ while ν_s is the rate at which f

relaxes to $\mathcal{E}_f^s$. To understand how the relaxation operator behaves at the macroscopic level, take any polynomial $p \in \mathbb{M}_{r+1}$ and compute

$$\left\langle p\mathcal{E}_f^{r+1} \right\rangle = \left\langle p_1\mathcal{E}_f^{r+1} \right\rangle + \left\langle p_2\mathcal{E}_f^{r+1} \right\rangle = \left\langle p_1\mathcal{E}_f^{r} \right\rangle + \left\langle p_2\mathcal{E}_f^{r+1} \right\rangle ,$$

where $p_1 \in \mathbb{M}_r$ and $p_2 \in \mathbb{M}_{r+1} \backslash \mathbb{M}_r$. As a result,

$$\left\langle p \left(\mathcal{E}_f^{r+1} - \mathcal{E}_f^{r}\right) \right\rangle = \left\langle p_2 \left(\mathcal{E}_f^{r+1} - \mathcal{E}_f^{r}\right) \right\rangle , \tag{3.66}$$

which means that the collision operator relaxes moments with respect to functions in the space $\mathbb{M}_{r+1} \backslash \mathbb{M}_r$ to zero at a rate ν_r. Another important implication of (3.66) is that

$$\mathcal{E}(\mathcal{E}(f; \mathbf{m}_{(r_1)}); \mathbf{m}_{(r_2)}) = \mathcal{E}(f; \mathbf{m}_{(r_2)})$$

whenever $r_2 \leq r_1$.

Another way understand $\tilde{\mathcal{C}}$ is to rewrite it in the following way:

$$\tilde{\mathcal{C}}(f) = -\sum_{r=0}^{s} \eta_r \left(f - \mathcal{E}_f^r\right) , \tag{3.67}$$

where $\eta_r = \nu_r - \nu_{r-1} > 0$. In this form, $\tilde{\mathcal{C}}$ is reminiscent the operator used by Bløtekjær (3.21) . It is a generalization of (3.21) in the sense that it allows for an arbitrary number of relaxation processes. It is more specific in the sense that the intermediate distributions are specified exactly. The benefit of this added detail is that $\tilde{\mathcal{C}}$ satisfies an H-Theorem. Also, the choice of the intermediate states is certainly

not arbitrary since they represent the most likely distribution that is consistent with the constraints $\left\langle \mathbf{m}_{(r)} f \right\rangle = \boldsymbol{\rho}_{(r)}$.

Proposition 6 (Levermore) *The collision operator $\tilde{\mathcal{C}}$ satisfies*

1. ***Conservation:*** $\left\langle \tilde{\mathcal{C}}(f) \right\rangle = 0$ *for all* $f \in \mathrm{Dom}(\tilde{\mathcal{C}})$.

2. ***Dissipation:*** $\left\langle f \log\left(\frac{f}{M_\ell}\right) \tilde{\mathcal{C}}(f) \right\rangle \leq 0$ *for all* $f \in \mathrm{Dom}(\tilde{\mathcal{C}})$.

3. ***Characterization of Equilibrium:*** *For all* $f \in \mathrm{Dom}(\tilde{\mathcal{C}})$*, the following statements are equivalent:*

$$(i)\ \tilde{\mathcal{C}}(f) = 0\,; \tag{3.68a}$$

$$(ii)\ \left\langle \log\left(\frac{f}{M_\ell}\right) \tilde{\mathcal{C}}\,(f) \right\rangle = 0\,; \tag{3.68b}$$

$$(iii)\ f = \langle f \rangle M_\ell\,. \tag{3.68c}$$

4. ***Affine Behavior:*** *For any function* $f \in \mathbb{F}_{\mathbf{m}}$

$$\tilde{\mathcal{C}}(f) = \tilde{\mathcal{C}}\left(\mathcal{E}(f;\mathbf{m})\right) - \nu_s \left(f - \mathcal{E}(f;\mathbf{m})\right)\,. \tag{3.68d}$$

Proof.

1. The proof of property (1) is trivial since $\left\langle \mathcal{E}_f^r \right\rangle = \langle f \rangle$ for $0 \leq r \leq s$.

2. We use the general fact that, for any y and $z \in \mathbb{R}$,

$$(z - y) \log\left(\frac{z}{y}\right) \geq 0\,, \tag{3.69}$$

with equality if and only if $y = z$; but first, let $f \in \mathrm{Dom}(\tilde{\mathcal{C}})$. Then

$$\left\langle \log\left(\frac{f}{M_\ell}\right) \tilde{\mathcal{C}}(f) \right\rangle = -\sum_{r=0}^{s-1} \eta_r \left\langle \left(f - \mathcal{E}_f^r\right) \log\left(\frac{f}{M_\ell}\right) \right\rangle .$$

Using the fact that

$$\left\langle \left(f - \mathcal{E}_f^r\right) \log\left(\frac{\mathcal{E}_f^r}{M_\ell}\right) \right\rangle = \boldsymbol{\beta}_r^T \left\langle \left(f - \mathcal{E}_f^r\right) \mathbf{m}_{(r)} \right\rangle = 0 \,,$$

we conclude that

$$\left\langle \left(f - \mathcal{E}_f^r\right) \log\left(\frac{f}{M_\ell}\right) \right\rangle = \left\langle \left(f - \mathcal{E}_f^r\right) \log\left(\frac{f}{\mathcal{E}_f^r}\right) \right\rangle . \tag{3.70}$$

which is non-negative due to (3.69) and, when applied (3.67), gives

$$\left\langle \log\left(\frac{f}{M_\ell}\right) \tilde{\mathcal{C}}(f) \right\rangle \leq 0 \,.$$

3. First, the fact that $(i) \Rightarrow (ii)$ is trivial. Next, if $f = \langle f \rangle M_\ell$, then $\mathcal{E}_f^r = \langle f \rangle M_\ell$ for each r. Hence $(iii) \Rightarrow (i)$. Finally, from (3.69) and (3.70), it follows that, for each r,

$$\left\langle \left(f - \mathcal{E}_f^r\right) \log\left(\frac{f}{\mathcal{E}_f^r}\right) \right\rangle = 0 \;\text{ if and only if }\; f = \mathcal{E}_f^r$$

which implies

$$\left\langle \log\left(\frac{f}{M_\ell}\right)\tilde{\mathcal{C}}(f)\right\rangle = 0 \;\text{ if and only if }\; f=\mathcal{E}_f^r=\mathcal{E}_f^{r-1}=\ldots=\mathcal{E}_f^0\,.$$

Hence $(ii) \Leftrightarrow (iii)$.

4. Since the mapping $\mathcal{E}$ is a projection,

$$\mathcal{E}\left(\mathcal{E}(f;\mathbf{m}_{(r)}),\mathbf{m}_{(r)}\right)=\mathcal{E}(f;\mathbf{m}_{(r)}),\qquad 0\le r\le s\,.$$

The result now follows immediately when plugging f and $\mathcal{E}(f;\mathbf{m})$ into formula (3.65) for $\tilde{\mathcal{C}}$.

■

3.4.4.1 Linearization of $\tilde{\mathcal{C}}$ Define the linear operator $\mathcal{L}$ acting on a function g by

$$\mathcal{L}g=-\frac{1}{M_\ell}D\tilde{\mathcal{C}}(M_\ell)M_\ell g=-\frac{1}{M_\ell}\lim_{\delta\to 0}\frac{\partial}{\partial\delta}\left(\tilde{\mathcal{C}}(M_\ell\,(1+\delta g))\right)\,. \tag{3.71}$$

Using the calculation

$$D_f\mathcal{E}(f;\mathbf{m})g\equiv\lim_{\delta\to 0}\frac{\partial}{\partial\delta}\mathcal{E}(f+\delta g;\mathbf{m})=\mathcal{E}(f;\mathbf{m})\mathbf{m}^T\left\langle\mathbf{m}\mathbf{m}^T\mathcal{E}(f;\mathbf{m})\right\rangle^{-1}\langle\mathbf{m}g\rangle \tag{3.72}$$

with (3.65) and the identity $\mathcal{E}(M_\ell;\mathbf{m})=M_\ell$, $\mathcal{L}$ can be calculated explicitly:

$$\mathcal{L}g=\sum_{r=0}^{s}\nu_r\left(\mathcal{P}_{r+1}-\mathcal{P}_r\right)\,, \tag{3.73}$$

where $\mathcal{P}_{s+1} \equiv \mathcal{I}$ and

$$\mathcal{P}_r g = \mathbf{m}_{(r)}^T \left\langle \mathbf{m}_{(r)} \mathbf{m}_{(r)}^T M_\ell \right\rangle^{-1} \left\langle \mathbf{m}_{(r)} M_\ell g \right\rangle, \qquad 0 \leq r \leq s, \tag{3.74}$$

is the orthogonal projection of g onto $\mathbb{M}_r$ in the Hilbert space $\mathbb{H}_{M}$ with inner product

$$(f, g)_{M} = \int_{\mathbb{R}^3} f(v)\, g(v)\, M_\ell(v) dv\,.$$

Proposition 7 *The operator $\mathcal{L}$ is bounded, positive, and self-adjoint from $\mathbb{H}_{M_\ell}$ to $\mathbb{H}_{M_\ell}$. The null space of $\mathcal{L}$ is $\mathbb{M}_0$, and it has a well defined pseudo-inverse from $\mathcal{R}(\mathcal{L})$ to $\mathcal{R}(\mathcal{L})$, given by*

$$\mathcal{L}^{-1} = \sum_{r=0}^{s} \frac{1}{\nu_r} \left(\mathcal{P}_{r+1} - \mathcal{P}_r\right). \tag{3.75}$$

Proof. Each projection is bounded and self-adjoint in $\mathbb{H}_{M_\ell}$:

$$\begin{aligned} \|\mathcal{P}_r f\|_{M_\ell} &= \|f\|_{M_\ell} - \|(\mathcal{I} - \mathcal{P}_r) f\|_{M_\ell} \leq \|f\|_{M_\ell}\,, \\ \left\langle f M_\ell \mathcal{P}_r g \right\rangle &= \left\langle f M_\ell \mathbf{m}_{(r)}^T \right\rangle \left\langle \mathbf{m}_{(r)} \mathbf{m}_{(r)}^T M_\ell \right\rangle^{-1} \left\langle \mathbf{m}_{(r)} M_\ell g \right\rangle = \left\langle g M_\ell \mathcal{P}_r f \right\rangle\,. \end{aligned}$$

Hence, $\mathcal{L}$ is also bounded and self-adjoint. Moreover, since $\mathbb{M}_r \subset \mathbb{M}_{r+1}$,

$$\left\langle g M_\ell \mathcal{P}_r g \right\rangle \leq \left\langle g M_\ell \mathcal{P}_{r+1} g \right\rangle\,. \tag{3.76}$$

Plugging (3.76) into (3.73) shows that $\mathcal{L}g \geq 0$ and that $\mathcal{L}g = 0$ if and only if $(\mathcal{P}_{r+1} - \mathcal{P}_r)\, g = 0$ for all r. In particular, $\mathcal{L}g = 0$ if and only if $g = \mathcal{P}^0 g$; hence

$\mathcal{N}(\mathcal{L}) = \mathbb{M}_0$. By general Hilbert space theory, $\mathcal{L}$ has a unique pseudo inverse $\mathcal{L}^{-1}$ from $\mathcal{R}(\mathcal{L})$ to $\mathcal{R}(\mathcal{L}^*) = \mathbb{M}_0^\perp$, and since $\mathcal{L}$ is self-adjoint, $\mathcal{R}(\mathcal{L}^*) = \mathcal{R}(\mathcal{L})$. Finally, if g is decomposed into orthogonal components,

$$g = \sum_{r=0}^{s} g_r \qquad , g_r \equiv \sum_{r=0}^{s} \left(\mathcal{P}_{r+1} - \mathcal{P}_r\right) g \,,$$

then

$$\mathcal{L}g = \sum_{r=0}^{s} \nu_r g_r$$

and

$$\frac{1}{\nu_r} \left(\mathcal{P}_{r+1} - \mathcal{P}_r\right) \mathcal{L}g = \left(\mathcal{P}_{r+1} - \mathcal{P}_r\right) g_r. \tag{3.77}$$

Summing (3.77) over r proves the formula for $\mathcal{L}^{-1}$given in (3.75). ■

3.4.4.2 Relaxation Rates The Boltzmann equation with the GBGK operator (3.67) is

$$\partial_t F + v \cdot \nabla_x F + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \nabla_v F = -\sum_{r=0}^{s} \eta_r \left(F - \mathcal{E}_F^r\right) \,.$$

The choice of the intermediate states $\mathcal{E}_F^r$ is based on the relative sizes of the relaxation rates. This is because the property $\eta_r = \nu_r - \nu_{r-1} > 0$ is crucial to proving Proposition 6.2. In addition, $\tilde{\mathcal{C}}$ should recover the correct mobility in the drift-diffusion limit. Recall from Chapter 2 that the mobility is given by

$$\mu = -\frac{1}{3} \frac{q}{m_e^* \theta_\ell} \operatorname{trace} \left(\left\langle v \mathcal{C}^{-1} \left(v \mathcal{M}\right) \right\rangle \right) \tag{3.78}$$

(Here, the operator $\mathcal{C}^{-1}$ is the pseudo-inverse of the full collision operator, which is linear.) Thus when replacing $\mathcal{C}$ by $\tilde{\mathcal{C}}$, one must choose values of ν_r such that

$$\left\langle vM_\ell \mathcal{L}^{-1} v \right\rangle = -\left\langle v\mathcal{C}^{-1}(vM_\ell) \right\rangle .$$

Proposition 8 *Let $\mathbb{M}_a$ be the smallest smallest space containing the polynomial v and suppose that $\left\langle v\mathbf{m}_{(r)}M \right\rangle = 0$ for all $r < a$. Then*

$$\left\langle vM_\ell \mathcal{L}^{-1}\, v \right\rangle = \frac{\theta_\ell}{\nu_a} I\,.$$

Proof. Write out $\langle v\mathcal{L}^{-1}\,(vM_\ell)\rangle$ in three pieces:

$$\begin{aligned}\left\langle vM_\ell \mathcal{L}^{-1}\, v \right\rangle = &\sum_{r=0}^{a-2} \frac{1}{\nu_r} \left\langle vM_\ell \left(\mathcal{P}_{r+1}(v) - \mathcal{P}_r(v)\right)\right\rangle \\ &+ \frac{1}{\nu_a} \left\langle vM_\ell \left(\mathcal{P}_a(v) - \mathcal{P}_{a-1}(v)\right)\right\rangle \\ &+ \sum_{r=a}^{s} \frac{1}{\nu_r} \left\langle vM_\ell \left(\mathcal{P}_{r+1}(v) - \mathcal{P}_r(v)\right)\right\rangle .\end{aligned}$$

All of the terms in the first line of this sum are zero by hypothesis as is the second term of the second line. The terms of the third line are also zero since $\mathcal{P}_r\,(v) = v$ for all $r \geq a$. The only remaining term is

$$\left\langle \frac{1}{\nu_a} vM_\ell \mathcal{P}_a\,(v) \right\rangle = \left\langle \frac{1}{\nu_a}\,(v \vee v)\, M_\ell \right\rangle = \frac{\theta_\ell}{\nu_a} I\,.$$

■

Given this proposition, it is clear that the correct mobility is recovered by an appropriate choice of the relaxation rate ν_a. In practice, $\nu_a = \tau_p^{-1}$, where τ_p is the momentum relaxation time. To recover the correct mobility, we set

$$\tau_p = -\frac{1}{3\theta_\ell} \operatorname{trace} \left\langle v\mathcal{C}^{-1}\left(vM_\ell\right)\right\rangle ,$$

which is consistent with the well-known relation [44, 62]

$$\mu = \frac{q_e}{m_e^*}\tau_p \,. \tag{3.79}$$

The numerical values for μ and the remaining relaxation rates can be determined in several ways. They can be computed using the full collision operator or by Monte Carlo simulations. In the latter case, rates are determined for a device with specific parameters. If physical characteristics such as doping profile or device dimension or external potential are changed, then new rates must be calculated. In some cases—such as the Baccarani-Wordeman model discussed later in this chapter—rates are specified based on phenomenological arguments that combine a mixture of theory, experiment, and approximation to justify analytical formulas for the rates as functions of the macroscopic variables. It is generally understood the such approximations are less accurate than the Monte-Carlo approach. However, they do not require re-calibration for each new device.

3.4.5 Examples

We present several entropy-based closures that are known to be well-posed—that is, for which (3.55) has a solution. These closures all have the form

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \mathbf{f}(\boldsymbol{\rho}) - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \mathbf{l}(\boldsymbol{\rho}) = \mathbf{r}(\boldsymbol{\rho}) \,.$$

3.4.5.1 The Equilibrium Closure. In the trivial case, $\mathbf{m} = \{1\}$. The moment system is

$$\partial_t n + \nabla_x \cdot \langle vF \rangle = 0 \,.$$

where $n = \langle F \rangle$. The entropic projection of F with respect to $\mathbf{m}$ is

$$\mathcal{E}(F; \mathbf{m}) = n M_\ell \,.$$

Approximating F by $\mathcal{F}[\boldsymbol{\rho}] = \mathcal{E}(F; \mathbf{m})$ gives

$$\partial_t n = 0 \,.$$

As is, this closure is quite trivial, and the collision operator does not play a role. However, in the next section, we will show how perturbations of $\mathcal{F}[\boldsymbol{\rho}]$ lead to the drift-diffusions equations.

3.4.5.2 The Drifted Diffusion Closure Let $s = 0$ and set

$$\mathbf{m}_{(0)} = 1, \qquad \mathbf{m} = \begin{pmatrix} 1 \\ \\ v \end{pmatrix} .$$

The moment system is

$$\partial_t n + \nabla_x \cdot (nu) = 0$$

$$\partial_t (nu) + \nabla_x \cdot (v \vee v\, F) - \frac{q_e}{m_e^*} n \nabla_x \Phi = \langle v \mathcal{C}\,(F) \rangle \ ,$$

where

$$n = \langle F \rangle \,, \qquad u = \frac{1}{n} \langle vF \rangle \ . \tag{3.80}$$

The entropic projection for evaluating the flux terms is a *Maxwellian*

$$\mathcal{E}(F; \mathbf{m}) = \mathcal{M}_{n,u,\theta_\ell} \,,$$

which gives

$$(v \vee v\, F) = (nu \vee u + n\theta_\ell) \,.$$

The entropic projection used to evaluate collision terms is

$$\mathcal{E}_F^0 = \mathcal{E}(F; \mathbf{m}_{(0)}) = n M_\ell \,.$$

The GBGK operator reduces to the standard BGK operator,

$$\tilde{\mathcal{C}}\left(F\right) = -\nu_0 \left(F - nM_\ell\right) \,,$$

where ν_0 is the rate at which momentum relaxes to zero. It is to traditional [13, 44] to write

$$\nu_0 = \frac{1}{\tau_p}$$

where τ_p is related to μ by (3.79). (The subscript p here stands for momentum.) Therefore the collision term is

$$\left\langle v\, \tilde{\mathcal{C}}\left(F\right) \right\rangle = -\frac{1}{\tau_p} nu \,.$$

3.4.5.3 The Maxwellian Closure Let $s = 2$ with

$$\mathbf{m}_{(0)} = 1, \quad \mathbf{m}_{(1)} = \begin{pmatrix} 1 \\ \\ \frac{1}{2}|v|^2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} 1 \\ \\ v \\ \\ \frac{1}{2}|v|^2 \end{pmatrix} .$$

The moment system is

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{3.81a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\theta I + \Sigma) - \frac{q_e}{m_e^*} n \nabla_x \Phi = \left\langle v\, \tilde{\mathcal{C}}\,(F) \right\rangle\,, \tag{3.81b}$$

$$\partial_t \left(\frac{n|u|^2}{2} + \frac{3n\theta}{2} \right) + \nabla_x \cdot \left(\frac{n|u|^2 u}{2} + \frac{5n\theta u}{2} + \Sigma u + q \right) - \frac{q_e}{m_e^*} nu \cdot \nabla_x \Phi = \left\langle \frac{|v|^2}{2} \tilde{\mathcal{C}}\,(F) \right\rangle\,, \tag{3.81c}$$

where n and u are given in (3.79) and

$$\theta = \frac{1}{n} \left\langle |v-u|^2 F \right\rangle, \qquad \Sigma = \left\langle \left((v-u) \vee (v-u) - \frac{1}{3}|v-u|^2 \right) F \right\rangle$$

$$q = \left\langle |v-u|^2 (v-u)\, F \right\rangle\,.$$

The entropic projection for evaluating the flux terms is a Maxwellian,

$$\mathcal{E}(F; \mathbf{m}) = \mathcal{M}_{n,u,\theta}\,,$$

which gives

$$\Sigma = 0\,, \qquad q = 0.$$

The entropic projections used to evaluate the collision terms are

$$\mathcal{E}_F^0 = \mathcal{E}(F; \mathbf{m}_{(0)}) = nM_\ell\,,$$

$$\mathcal{E}_F^1 = \mathcal{E}(F; \mathbf{m}_{(1)}) = \mathcal{M}_{n,0,\omega}\,,$$

where

$$\omega = \frac{2}{3n}\left\langle |v|^2 F \right\rangle = \frac{1}{3}|u|^2 + \theta\,.$$

The collision operator,

$$\tilde{\mathcal{C}}\left(F\right) = -\nu_1\left(F - \mathcal{E}_F^1\right) - \nu_0\left(\mathcal{E}_F^1 - \mathcal{E}_F^0\right)\,, \tag{3.82}$$

models a two-part relaxation process. The relaxation from F to $\mathcal{E}_F^1$ corresponds to the relaxation to a state with zero average momentum at a rate ν_1. The relaxation from $\mathcal{E}_F^1$ to $\mathcal{E}_F^0$ corresponds to the relaxation of energy to the thermal energy of the lattice at a rate $\nu_0 < \nu_1$. Traditionally, these rates are defined in terms of their corresponding relaxation times [13]:

$$\nu_0 = \frac{1}{\tau_w} \qquad \text{and} \qquad \nu_1 = \frac{1}{\tau_p}\,.$$

(As before, the subscript p is for momentum while the subscript w stands for energy.) The production terms take the form

$$\begin{aligned}\left\langle v\,\tilde{\mathcal{C}}\left(F\right)\right\rangle &= -\frac{1}{\tau_p}nu\,,\\ \left\langle \frac{1}{2}|v|^2\tilde{\mathcal{C}}\left(F\right)\right\rangle &= -\frac{1}{\tau_w}\left(\frac{1}{2}n|u|^2 + \frac{3}{2}n\left(\theta - \theta_\ell\right)\right)\,.\end{aligned}$$

Remark 9 *At first glance, the choice of the intermediate state $\mathcal{E}_1$ in this example*

may seem a bit strange. In fact, it might seem more natural to replace $\mathbf{m}_{(1)}$ *by*

$$\tilde{\mathbf{m}}_{(1)} = \begin{pmatrix} 1 \\ \\ v \end{pmatrix}$$

and $\mathcal{E}_F^1$ *by* $\tilde{\mathcal{E}}_F^1 = \mathcal{M}_{n,u,\theta_\ell}$. *The resulting collision operator models the combination of a fast(er) process of heat relaxation to the thermal lattice temperature and a slow(er) process of momentum relaxation. Such an operator is inconsistent with the fact that momentum relaxes faster than temperature.*

In gas dynamics, their is no preferred inertial frame, and the choice of $\mathbf{m}_{(1)}$ *would not be appropriate since* $\operatorname{span}\{1, |v|^2\}$ *is not invariant under Galilean shifts. However, for electron-lattice collisions in a semiconductor, a preferred frame is provided by the lattice. Hence invariance under Galilean shift is no longer expected to hold.*

Related to this discussion is the quantity ω, *which serves the role of temperature in the Maxwellian that is the intermediate state* $\mathcal{E}_F^1$. *Since the kinetic density* $\mathcal{E}_F^1$ *has mean velocity zero,* ω *represents the entire energy of the system. It is in this way that momentum and energy relaxation are handled as two distinct processes.*

3.4.5.4 The Gaussian Closure Let $s = 3$ with

$$\mathbf{m}_{(0)} = 1\,, \qquad \mathbf{m}_{(1)} = \begin{pmatrix} 1 \\ \frac{1}{2}|v|^2 \end{pmatrix}\,,$$

$$\mathbf{m}_{(2)} = \begin{pmatrix} 1 \\ v \\ \frac{1}{2}|v|^2 \end{pmatrix}\,, \qquad \mathbf{m} = \begin{pmatrix} 1 \\ v \\ v \vee v \end{pmatrix}\,.$$

The moment system is

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{3.84a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\Theta) - \frac{q_e}{m_e^*} n \nabla_x \Phi = \left\langle v\, \tilde{\mathcal{C}}\,(F) \right\rangle\,, \tag{3.84b}$$

$$\partial_t \left(nu \vee u + n\Theta\right) + \nabla_x \cdot \left(nu^{\vee 3} + 3n\Theta \vee u + Q\right) \tag{3.84c}$$

$$-2\frac{q_e}{m_e^*} nu \vee \nabla_x \Phi = \left\langle \frac{|v|^2}{2} \tilde{\mathcal{C}}\,(F) \right\rangle\,,$$

where n and u are defined in (3.79) and

$$\Theta = \frac{1}{n} \left\langle (v-u) \vee (v-u)\, F \right\rangle\,, \qquad Q = \left\langle (v-u)^{\vee 3} F \right\rangle\,.$$

The entropic projection for evaluating the flux terms is a Gaussian

$$\mathcal{E}(F; \mathbf{m})(v) = \mathcal{G}_{n,u,\Theta}(v)\,,$$

which gives $Q = 0$. The entropic projections used to evaluate $\tilde{\mathcal{C}}$ are

$$\mathcal{E}_F^0 = \mathcal{E}(F; \mathbf{m}_{(0)}) = nM_\ell \,,$$

$$\mathcal{E}_F^1 = \mathcal{E}(F; \mathbf{m}_{(1)}) = \mathcal{M}\,(n, 0, \omega) \;,$$

$$\mathcal{E}_F^2 = \mathcal{E}(F; \mathbf{m}_{(2)}) = \mathcal{M}\,(n, u, \theta) \;.$$

The collision operator,

$$\tilde{\mathcal{C}}(F) = -\nu_2\,(F - \mathcal{E}_2) - \nu_1\,(\mathcal{E}_2 - \mathcal{E}_1) - \nu_0\,(\mathcal{E}_1 - \mathcal{E}_0) \;, \tag{3.85}$$

models a three-part relaxation process: first is relaxation to the isotropic distribution $\mathcal{E}_2$ at the rate v_2; next is momentum relaxation to the distribution $\mathcal{E}_1$ with zero mean velocity at the rate v_1; finally there is relaxation to the local equilibrium $\mathcal{E}_0$ at the temperature θ_ℓ at the rate v_0. Again, it is traditional [13] to write

$$\nu_0 = \frac{1}{\tau_w} \,, \qquad \nu_1 = \frac{1}{\tau_p} \,, \qquad \nu_2 = \frac{1}{\tau_\sigma} \,,$$

where the additional subscript σ connotes Σ, the anisotropic part of the stress tensor. The production terms for the Gaussian system are

$$\begin{aligned}
\left\langle v\,\tilde{\mathcal{C}}\,(F) \right\rangle &= -\frac{1}{\tau_p} nu \,, \\
\left\langle v \vee v\,\tilde{\mathcal{C}}\,(F) \right\rangle &= -\frac{1}{\tau_\sigma}\,(n\Theta - n\theta I) - \frac{1}{\tau_p}\left(nu \vee u - \frac{1}{3} n|u|^2\right) \\
&\qquad - \frac{1}{\tau_w}\left(\frac{1}{3} n|u|^2 I + n\,(\theta - \theta_\ell)\, I\right) \;.
\end{aligned}$$

Remark 10 *As with the Maxwellian closure, there are other possible choices for the vectors* $\mathbf{m}_r$ *that determine the intermediate states. However, just as before, our choice is guiding by the ordering of relaxation rates that ensures that* $\tilde{\mathcal{C}}$ *dissipates entropy. In this case, experiments confirm that* $\tau_\sigma < \tau_p < \tau_w$ *which justifies our choice of* $\mathbf{m}_r$.

3.5 Perturbations of Entropy-Based Moment Closures

A particular drawback of the Maxwellian and Gaussian closures, as compared to the BBW and AP closures, is that they fail to capture heat flow, which experiments have found to be an important aspect in the dynamics of electron transport. Many attempts have been made to extend moment systems to high order to capture these effects more accurately. As mentioned previously, a series of papers (see [3] and references therein), have developed higher-order closures (meaning order greater than two) in the framework of extended thermodynamics that is formally justified by the principle of minimum entropy applied to the kinetic entropy $\mathcal{H}$. However, from a mathematical point of view, entropy closures for systems of order greater than two are not well-posed. Thus we propose a new approach that combines well-posed entropy closures with a perturbative analysis. The basic idea is to assume that the kinetic density is a small perturbation from its entropic projection, and then use this perturbation to derive more accurate expressions for the stress and heat flux. We find that such pertubations lead to convective and diffusive corrections that agree with other closures in some respects and differ in others. A numerical investigation into the effects of these corrective terms is the topic of Chapter 5. We call the new

hierarchy of closures *perturbed entropy-based* (PEB) closures.

3.5.1 General Setting

We begin with Boltzmann transport equation,

$$\partial_t F + v \cdot \nabla_x F + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \nabla_v F = \tilde{\mathcal{C}}(F)\,, \tag{3.87}$$

where $\tilde{\mathcal{C}}$ is given by (3.65). As usual, (3.87) can be integrated against a vector $\mathbf{m}$ of polynomials in v to give

$$\partial_t \langle \mathbf{m} F \rangle + \nabla_x \cdot \langle v \mathbf{m} F \rangle + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \langle \mathbf{m} \nabla_v F \rangle = \left\langle \mathbf{m} \tilde{\mathcal{C}}(F) \right\rangle . \tag{3.88}$$

Here we assume that the components of $\mathbf{m}$ are a basis for an admissible polynomial space $\mathbb{M}$. Let $\mathcal{E}_F = \mathcal{E}(F; \mathbf{m})$ and write F as $\mathcal{E}_F$ plus a perturbation:

$$F = \mathcal{E}_F \big(1 + \tilde{F}\big)\,, \tag{3.89}$$

where

$$\left\langle \mathbf{m} \mathcal{E}_F \tilde{F} \right\rangle = 0\,. \tag{3.90}$$

By plugging (3.89) into (3.88), we find an evolution for the spatial density $\boldsymbol{\rho} = \langle \mathbf{m} F \rangle$:

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \left\langle v \mathbf{m} \mathcal{E}_F \big(1 + \tilde{F}\big) \right\rangle - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \left\langle \nabla_v \mathbf{m}\, \mathcal{E}_F \big(1 + \tilde{F}\big) \right\rangle = \left\langle \mathbf{m}\, \tilde{\mathcal{C}}(\mathcal{E}_F \big(1 + \tilde{F}\big)) \right\rangle .$$

To simplify this expression, note first that if $\mathbb{M}$ is invariant under translations of

v, then the components of $\nabla_v \mathbf{m}$ will be linear combinations of the components of $\mathbf{m}$. Thus (3.90) implies

$$\left\langle \nabla_v \mathbf{m}\, \mathcal{E}_F \tilde{F} \right\rangle = 0\,.$$

Moreover, by (3.68d),

$$\tilde{\mathcal{C}}\big(\mathcal{E}_F\big(1+\tilde{F}\big)\big) = \tilde{\mathcal{C}}(\mathcal{E}_F) - \nu_s \mathcal{E}_F \tilde{F}\,,$$

where ν_s is the largest relaxation rate in (3.65). Therefore

$$\left\langle \mathbf{m}\, \tilde{\mathcal{C}}\big(\mathcal{E}_F\big(1+\tilde{F}\big)\big) \right\rangle = \left\langle \mathbf{m}\, \tilde{\mathcal{C}}(\mathcal{E}_F) \right\rangle .$$

The evolution equation for $\boldsymbol{\rho}$ simplifies to

$$\partial_t \boldsymbol{\rho} + \nabla_x \cdot \langle v\, \mathbf{m}\, \mathcal{E}_F \rangle + \nabla_x \cdot \left\langle v\, \mathbf{m}\, \mathcal{E}_F \tilde{F} \right\rangle - \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \langle \nabla_v \mathbf{m}\, \mathcal{E}_F \rangle = \left\langle \mathbf{m}\, \tilde{\mathcal{C}}(\mathcal{E}_F) \right\rangle . \quad (3.91)$$

So far, everything has been exact except for the approximation $\tilde{\mathcal{C}}$ of the full collision operator $\mathcal{C}$. We now approximate $\tilde{F}$ in order to close this system. (Note that we recover the entropic closure by simply setting $\tilde{F} = 0$). Using (3.87) and (3.88), we find that $\tilde{F}$ satisfies

$$\begin{aligned} \frac{\partial_t\big(\mathcal{E}_F \tilde{F}\big)}{\mathcal{E}_F} + \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v \cdot \nabla_x\big(\mathcal{E}_F\big(1+\tilde{F}\big)\big)}{\mathcal{E}_F}\right) \\ + \frac{q_e}{m_e^*} \nabla_x \Phi \cdot \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\nabla_v\big(\mathcal{E}_F\big(1+\tilde{F}\big)\big)}{\mathcal{E}_F}\right) = \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\tilde{\mathcal{C}}(\mathcal{E}_F)}{\mathcal{E}_F}\right) - \nu_s \tilde{F}\,, \end{aligned} \quad (3.92)$$

where $\tilde{\mathcal{P}}_{\varepsilon_F} = \mathcal{I} - \mathcal{P}_{\varepsilon_F}$ and

$$\mathcal{P}_{\varepsilon_F} g = \frac{1}{\mathcal{E}_F} D_f \mathcal{E}(F; \mathbf{m}) \mathcal{E}_F g = \mathbf{m}^T \left\langle \mathbf{m}\mathbf{m}^T \, \mathcal{E}_F \right\rangle^{-1} \langle \mathbf{m} \, \mathcal{E}_F g \rangle$$

is computed using (3.72). The mapping $\mathcal{P}_{\varepsilon_F}$ is the orthogonal projection of F onto $\mathbb{M}$ is the Hilbert space $\mathbb{H}_{\mathcal{E}_F}$ with inner product

$$(f, g)_{\varepsilon_F} \equiv \int_{\mathbb{R}^3} f(v) \, g(v) \, \mathcal{E}_F(v) \, dv \, .$$

Thus $\tilde{\mathcal{P}}_{\varepsilon_F} g$ isolates the components of the function g that are orthogonal to $\mathbf{m}$ in $\mathbb{H}_{\mathcal{E}_F}$.

The contribution from the electric field to the closure depends on whether or not $v \in \mathbb{M}$. To see this, we compute

$$\left(\frac{\nabla_v \mathcal{E}_F}{\mathcal{E}_F} \right)(v) = \nabla_v (\log \mathcal{E}_F) = \boldsymbol{\beta}^T \nabla_v \mathbf{m}(v) - \frac{v}{\theta_\ell} \, .$$

Thus, if $\mathbb{M}$ is invariant under translations of v *and* if $v \in \mathbb{M}$, then the components of $\nabla_v \mathbf{m}$ will be linear combinations of components in $\mathbf{m}$. This implies that in (3.92),

$$\tilde{\mathcal{P}}_{\varepsilon_F} \left(\frac{\nabla_v \mathcal{E}_F}{\mathcal{E}_F} \right) = 0 \, . \tag{3.93}$$

Note that (3.93) does not hold in the case $\mathbf{m} = 1$.

3.5.2 Balance

We assume that (3.93) holds and that the potential energy associated with the electric fields in a given device is on the same order as the thermal energy, in which case the scaled version of (3.92) is

$$\varepsilon^3 \frac{\partial_t\big(\mathcal{E}_F \tilde{F}\big)}{\mathcal{E}_F} + \varepsilon \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\big(\mathcal{E}_F\big(1+\delta\tilde{F}\big)\big)}{\mathcal{E}_F}\right) + \varepsilon^2 \nabla_x\Phi\cdot\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\nabla_v\big(\mathcal{E}_F\tilde{F}\big)}{\mathcal{E}_F}\right) = \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\tilde{\mathcal{C}}(\mathcal{E}_F)}{\mathcal{E}_F}\right) - \varepsilon\nu_s\tilde{F}\,, \tag{3.94}$$

where ε is the scaled Knudsen number. Although we are no longer in the equilibrium limit, the value of ε is still relatively small. For example, in today's most modern semiconductor devices, ε is roughly 0.01 to 0.1. By retaining terms in (3.92) through order ε, the behavior of $\tilde{F}$ is approximated by the balance

$$\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right) = \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\tilde{\mathcal{C}}(\mathcal{E}_F)}{\mathcal{E}_F}\right) - \nu_s\tilde{F}\,,$$

so that

$$\tilde{F} = -\tau_s\left[\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right) - \tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\tilde{\mathcal{C}}(\mathcal{E}_F)}{\mathcal{E}_F}\right)\right]. \tag{3.95}$$

It should be noted that the scaling (3.94)—and hence the balance in (3.95)—is subject to criticism. The issue here is not so much the fact that $\varepsilon \ll 1$ no longer holds. Rather, it is the fact that, in smaller devices, the electric field can become large, especially around material junctions where the doping concentration may vary by several orders of magnitude. Including electric field effects gives a new balance

for approximating $\tilde{F}$:

$$\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right)+\frac{q_e}{m_e^*}\nabla_x\Phi\cdot\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\nabla_v\left(\mathcal{E}_F\tilde{F}\right)}{\mathcal{E}_F}\right)=\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{\tilde{\mathcal{C}}(\mathcal{E}_F)}{\mathcal{E}_F}\right)-\nu_s\tilde{F}\,.$$

The closure derived from this balance is the subject of future work. For the moment, we continue to work with the balance given in (3.95).

3.5.3 Entropy Dissipation

In general, the expression for $\tilde{F}$ in (3.95) provides two corrective terms to the closure in (3.91): a gradient term that produces diffusive corrections and a collision term that produces convective corrections. Recall that (3.91) dissipates the quantity $K(\boldsymbol{\rho},\nabla_x\Phi)$ given in (3.58) whenever $\tilde{F}=0$. The addition of the diffusive terms only helps the situation. Following the calculation of Section (3.4.2), we need only show the following

Proposition 11 *The diffusive term in (3.95) locally dissipates the entropy k (and hence K).*

Proof. The chain rule gives

$$\begin{aligned}-\frac{\partial k}{\partial\boldsymbol{\rho}}\nabla_x\cdot\left\langle\tau_s v\,\mathcal{E}_F\mathbf{m}\,\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right)\right\rangle \qquad (3.96)\\ =\nabla_x\frac{\partial k}{\partial\boldsymbol{\rho}}\cdot\left\langle\tau_s v\,\mathbf{m}\,\mathcal{E}_F\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right)\right\rangle\\ -\nabla_x\cdot\left(\frac{\partial k}{\partial\boldsymbol{\rho}}\left\langle\tau_s v\,\mathbf{m}\,\mathcal{E}_F\tilde{\mathcal{P}}_{\varepsilon_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right)\right\rangle\right)\,,\end{aligned}$$

and the second term on the right-hand side of (3.96) is in divergence form. Thus, it remains to show then that the first term has the appropriate sign. Note that

$$\frac{v \cdot \nabla_x \mathcal{E}_F}{\mathcal{E}_F} = v \cdot \nabla_x (\log(\mathcal{E}_F)) = v \cdot \nabla_x (\boldsymbol{\beta}^T \mathbf{m}) \,, \tag{3.97}$$

where $\boldsymbol{\beta} = \left(\frac{\partial k}{\partial \boldsymbol{\rho}}\right)^T$. Using the fact that $\tilde{\mathcal{P}}_{\varepsilon_F}$is self-adjoint in $\mathbb{H}_{\mathcal{E}_F}$ gives

$$\begin{aligned}
\nabla_x \frac{\partial k}{\partial \boldsymbol{\rho}} \cdot \left\langle v \, \mathbf{m} \, \mathcal{E}_F \tilde{\mathcal{P}}_{\varepsilon_F} \left(\frac{v \cdot \nabla_x \mathcal{E}_F}{\mathcal{E}_F} \right) \right\rangle &= \left\langle v \cdot \nabla_x (\boldsymbol{\beta}^T \mathbf{m}) \, \mathcal{E}_F \tilde{\mathcal{P}}_{\varepsilon_F} v \cdot \nabla_x (\boldsymbol{\beta}^T \mathbf{m}) \right\rangle \\
&= \left\langle \mathcal{E}_F \left(\tilde{\mathcal{P}}_{\varepsilon_F} (v \cdot \nabla_x (\boldsymbol{\beta}^T \mathbf{m})) \right)^2 \right\rangle \\
&\geq 0 \,,
\end{aligned}$$

which concludes the proof. ■

The key to Proposition 11 is the unique form of (3.97) which is not shared by the convective term in (3.95). It is therefore unclear whether inclusion of this term in (3.91) will preserve entropy dissipation or hyperbolicity. The effects of both diffusive and convection corrections in numerical simulations will be examined in the next chapter.

3.5.4 Examples

3.5.4.1 The Equilibrium Closure For the case $\mathbf{m} = 1$, the entropic projection is just the equilibrium distribution: $\mathcal{E}(F; \mathbf{m}) = M_\ell \langle F \rangle$, and the closure for the density $n = \langle F \rangle$ is trivial:

$$\partial_t n = 0 \,.$$

However, it is instructive to see that the perturbation procedure outlined above reduces to the drift-diffusion model in this simple case. We note first that, since $\tilde{\mathcal{C}}\left(\langle F\rangle M_\ell\right) = 0$, there are no convective corrections. However, since (3.93) does not hold, then there will be an electric field contribution.

The perturbation $\tilde{F}$ satisfies

$$\begin{aligned}\tilde{F} &= -\tau_p\left[\tilde{\mathcal{P}}_{\mathcal{E}_F}\left(\frac{v\cdot\nabla_x\mathcal{E}_F}{\mathcal{E}_F}\right) + \frac{q_e}{m_e^*}\nabla_x\Phi\cdot\tilde{\mathcal{P}}_{\mathcal{E}_F}\left(\frac{\nabla_v\mathcal{E}_F}{\mathcal{E}_F}\right)\right]\\ &= -\tau_p\left[\left(\frac{v\cdot\nabla_x n}{n}\right) - \frac{q}{m_e^*\theta_\ell}\nabla_x\Phi\cdot v\right].\end{aligned}$$

and therefore the perturbed flux is

$$\begin{aligned}\left\langle v\mathcal{E}_F\tilde{F}\right\rangle &= -\tau_p\left[\theta_\ell\nabla_x n - \frac{q_e}{m_e^*}n\nabla_x\Phi\right]\\ &= -\frac{m_e^*\theta_\ell}{q}\mu\nabla_x n - \mu n\nabla_x\Phi\,,\end{aligned}$$

which is the just drift-diffusion flux with the mobility formula (3.79).

3.5.4.2 The Maxwellian Closure Recall that the entropic projection for F is a Maxwellian,

$$\mathcal{E}(F;\mathbf{m}) = \mathcal{M}_{n,u,\theta} = \frac{n}{(2\pi\theta)^{3/2}}\exp\left(-\frac{|v-u|^2}{2\theta}\right), \tag{3.98}$$

and the resulting moment equations are given by (3.81). For simplicity of notation, we will henceforth drop subscripts and set $\mathcal{M}\equiv\mathcal{M}_{n,u,\theta}$. If $F=\mathcal{M}$, then Σ and q are identically zero. In order to find non-vanishing expression for Σ and q, we assume

that F has the form

$$F = \mathcal{M}(1 + \tilde{F}),$$

where $\left\langle \mathbf{m}\mathcal{M}\tilde{F} \right\rangle = 0$. According to (3.95), the perturbation $\tilde{F}$ is given by

$$\tilde{F} = -\tau_p \left[\tilde{\mathcal{P}}_{\mathcal{M}} \left(\frac{v \cdot \nabla_x \mathcal{M}}{\mathcal{M}} \right) - \tilde{\mathcal{P}}_{\mathcal{M}} \left(\frac{\tilde{\mathcal{C}}(\mathcal{M})}{\mathcal{M}} \right) \right]$$

where $\tilde{\mathcal{P}}_{\mathcal{M}} = \mathcal{I} - \mathcal{P}_{\mathcal{M}}$ and

$$\begin{aligned} \mathcal{P}_{\mathcal{M}} g &= \mathbf{m}^T \left\langle \mathbf{m}\mathbf{m}^T \mathcal{M} \right\rangle^{-1} \left\langle \mathbf{m}\, \mathcal{M} g \right\rangle \\ &= \frac{1}{n} \left(\langle \mathcal{M} g \rangle + \frac{c}{\theta} \langle c \mathcal{M} g \rangle + \frac{2}{3} \left(\frac{|c|^2}{2\theta} - \frac{3}{2} \right) \left\langle \left(\frac{|c|^2}{2\theta} - \frac{3}{2} \right) \mathcal{M} g \right\rangle \right). \end{aligned} \tag{3.99}$$

It is convenient to write $\Sigma = \Sigma_1 + \Sigma_2$ and $q = q_1 + q_2$, where

$$\Sigma_1 = -\tau_p \left\langle \left(c \vee c - \frac{1}{3}|c|^2 I \right) \mathcal{M} \tilde{\mathcal{P}}_{\mathcal{M}} \left(\frac{v \cdot \nabla_x \mathcal{M}}{\mathcal{M}} \right) \right\rangle, \tag{3.100a}$$

$$\Sigma_2 = \tau_p \left\langle \left(c \vee c - \frac{1}{3}|c|^2 I \right) \mathcal{M} \tilde{\mathcal{P}}_{\mathcal{M}} \left(\frac{\tilde{\mathcal{C}}(\mathcal{M})}{\mathcal{M}} \right) \right\rangle, \tag{3.100b}$$

and

$$q_1 = \frac{\tau_p}{2} \left\langle |c|^2 c \mathcal{M} \tilde{\mathcal{P}}_{\mathcal{M}} \left(\frac{\tilde{\mathcal{C}}(\mathcal{M})}{\mathcal{M}} \right) \right\rangle, \tag{3.101a}$$

$$q_2 = -\frac{\tau_p}{2} \left\langle |c|^2 c \mathcal{M} \tilde{\mathcal{P}}_{\mathcal{M}} \left(\frac{v \cdot \nabla_x \mathcal{M}}{\mathcal{M}} \right) \right\rangle. \tag{3.101b}$$

The terms Σ_1 and q_1 generate diffusive corrections while Σ_2 and q_2 generate convective corrections. By plugging (3.82),(3.98), and (3.99) into (3.100) and (3.101), we find

that

$$\Sigma_1 = -n\theta\tau_p \left(\nabla_x u + (\nabla_x u)^T - \frac{2}{3} (\nabla_x \cdot u) I \right) ,$$
$$\Sigma_2 = nu \vee u - \frac{1}{3} n|u|^2 I ,$$

and

$$q_1 = -\frac{5}{2} n\theta\tau_p \nabla_x \theta ,$$
$$q_2 = -\frac{4}{3} n|u|^2 u + \frac{\tau_p}{\tau_w} \left(\frac{5}{6} n|u|^2 u + \frac{5}{2} n (\theta - \theta_\ell) u \right) .$$

3.5.4.3 The Gaussian Closure The entropic projection of F in this case is now a Gaussian:

$$\mathcal{E}(F; \mathbf{m}) = \mathcal{G}_{n,u,\Theta} = \frac{n}{\sqrt{\det(2\pi\Theta)}} \exp\left(-\frac{1}{2} (v-u)^T \Theta^{-1} (v-u) \right) \tag{3.102}$$

and the resulting moment equations are given by (3.84). For simplicity of notation, we will henceforth drop the subscripts and set $\mathcal{G}_{n,u,\theta} \equiv \mathcal{G}$. If $F = \mathcal{E}(F; \mathbf{m})$, then Q is identically zero. In order to derive non-vanishing values for Q, we assume that F has the form

$$F = \mathcal{G}\big(1 + \tilde{F}\big) ,$$

where $\left\langle \mathbf{m}\mathcal{G}\tilde{F} \right\rangle = 0$. The perturbation $\tilde{F}$ is given by

$$\tilde{F} = -\tau_s \left[\tilde{\mathcal{P}}_{\mathcal{G}} \left(\frac{v \cdot \nabla_x \mathcal{G}}{\mathcal{G}} \right) - \tilde{\mathcal{P}}_{\mathcal{G}} \left(\frac{\tilde{\mathcal{C}}(\mathcal{G})}{\mathcal{G}} \right) \right] ,$$

where $\tilde{\mathcal{P}}_{\mathcal{g}} = \mathcal{I} - \mathcal{P}_{\mathcal{g}}$,

$$\mathcal{P}_{\mathcal{g}} g = \frac{1}{n}\left(\langle \mathcal{G} g\rangle + \psi \cdot \langle c\mathcal{G} g\rangle + \frac{1}{2}\left(\psi \vee \psi - \Theta^{-1}\right) : \langle (c \vee c - \Theta)\, \mathcal{G} g\rangle\right), \tag{3.103}$$

and $\psi = \Theta^{-1}(v-u)$.

It is convenient to write $Q = Q_1 + Q_2$, where

$$Q_1 = -\tau_\sigma \left\langle (v-u)\vee(v-u)\vee(v-u)\, \mathcal{G}\tilde{\mathcal{P}}_{\mathcal{g}} \left(\frac{v\cdot\nabla_x \mathcal{G}}{\mathcal{G}}\right)\right\rangle, \tag{3.104a}$$

$$Q_2 = \tau_\sigma \left\langle (v-u)\vee(v-u)\vee(v-u)\, \mathcal{G}\tilde{\mathcal{P}}_{\mathcal{g}} \left(\frac{\tilde{\mathcal{C}}(\mathcal{G})}{\mathcal{G}}\right)\right\rangle. \tag{3.104b}$$

Then the term Q_1 is a diffusive correction and the term Q_2 is a convective correction. By plugging (3.84), (3.85), and (3.103) into (3.104), we find that

$$\begin{aligned} Q_1 =& -3\tau_\sigma n\,(\Theta\cdot\nabla_x)\vee\Theta\,, \\ Q_2 =& -\frac{\tau_\sigma}{\tau_p}\left(nu\vee u\vee u + n|u|^2 I\vee u + 3n\,(\theta I-\Theta)\vee u\right) \\ &+\frac{\tau_\sigma}{\tau_w}\left(n|u|^2 + 3n\,(\theta-\theta_\ell)\right) I\vee u\,. \end{aligned}$$

Chapter 4

Entropy Minimization and Realizability

In this Chapter, we will examine the minimization problem upon which entropy-based closures for semiconductor models are based. Recall that the relative kinetic entropy is given by

$$\mathcal{K}(f) \equiv \left\langle f \log\left(\frac{f}{M_\ell}\right) - f \right\rangle$$

and that

$$\mathbb{F}_{\mathbf{m}} \equiv \left\{ g \in L_1\left(\mathbb{R}^D\right) : g \gneqq 0 \text{ and } \langle |m_s g| \rangle < \infty,\ (s = 0, \ldots, l-1) \right\} \ .$$

The entropy minimization problem is then

$$\min_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{K}(g) : \langle \mathbf{m} g \rangle = \boldsymbol{\rho} \right\} \ . \tag{4.1}$$

Our main result is a characterization of the set of $\mathcal{D}$ of *degenerate* densities. These are densities $\boldsymbol{\rho}$ for which $\boldsymbol{\rho} = \langle \mathbf{m} f \rangle$ for some $f \in \mathbb{F}_{\mathbf{m}}$, but the minimizer in (4.1) does not exist. Thus if F is a solution of the Boltzmann equation and if $f = F(x, v, t)$ for some fixed (x, t), then the entropy closure will not be well-defined. There two possible ways to address this issue.

1. Ensure that values $\boldsymbol{\rho} \in \mathcal{D}$ will never be attained by the moment system generated by the entropy closure. One can either (i) show that the set of densities for which (4.1) *does* have a solution is invariant under the dynamics of the moment system or (ii) impose this condition in a way that is physically reasonable and mathematically justifiable.

2. Develop a modified approach that (i) is well-posed for *all* physically realizable values of $\boldsymbol{\rho}$, (ii) agrees with the minimum entropy approach for well-posed cases of (4.1), and (iii) produces closures that generate symmetric hyperbolic systems that dissipate a physically meaningful entropy.

For both cases, it is important—at the very least—to show that $\mathcal{D}$ is small in some sense; and under reasonable conditions, we show that $\mathcal{D}$ is the finite union of lower dimensional fiber bundles. The fibers in each bundle are cones which we describe using the *complementary slackness condition* that comes from the dual formulation of (4.1)

For simplicity of exposition, we actually consider the minimization problem for the functional $\mathcal{H}$ rather than $\mathcal{K}$:

$$h(\boldsymbol{\rho}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m} g \rangle = \boldsymbol{\rho}\} \ , \tag{4.2}$$

where

$$\mathcal{H}(f) \equiv \langle f \log f - f \rangle \ .$$

As mentioned in Chapter 3, (4.1) and (4.2) generate the same closure whenever $\mathbf{m}$

contains polynomial elements of degree two or greater. Previous studies of (4.2) can be found in [42, 43, 73], where the definition of h is altered in an attempt to handle ill-posed cases. In [43] h is redefined by relaxing the minimum in (4.2) to an infimum:

$$h(\boldsymbol{\rho}) = \inf_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m} g \rangle = \boldsymbol{\rho}\} \ . \tag{4.3}$$

Meanwhile in [73], an alternative definition of h is given by

$$h(\boldsymbol{\rho}) = \inf_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}\} \ , \tag{4.4}$$

where the notation $\langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}$ means—roughly speaking—that inequalities between certain components are allowed. Later in the chapter, we will attach a precise meaning to this notation.

It has been shown in [73] that (4.4) has a unique minimizer with a specific form that can easily expressed with Lagrange multipliers, and it turns out that the Lagrange multipliers are intimately related to the question of whether or not (4.2) also has a solution. We analyze this relationship in detail by applying a dual formulation to (4.4) based on the theory of convex optimization. We prove the important complementary slackness condition which is used to characterize the set $\mathcal{D}$. In the process of our investigation, we recover and extend previous results from [42, 43] and [73]. We also show that the definitions of h in (4.4) and (4.3) are equivalent, i.e., that the respective infima are equal. This implies that the minimizer of (4.4) is also the unique minimizer of (4.2) whenever (4.2) has a minimum.

The organization of the chapter is as follows. In Section 2, we introduce preliminary notation and background. In Section 3, we present the minimization problem, formulate its dual, and prove the complementary slackness condition. In Section 4, we analyze the relationship between $\boldsymbol{\rho}$ and the Lagrange multipliers from the dual problem, and two examples are given. Finally, in the Appendix, we include proofs of the duality theorems, including the complementary slackness condition.

4.1 Preliminaries

4.1.1 Admissible Spaces

For a given moment system, the choice of $\mathbf{m}$ must satisfy criteria based on physical considerations such as conservation and invariance under coordinates changes. We require that components of $\mathbf{m}$ form a basis for a linear space $\mathbb{M}$ of multivariate polynomials over the field of real numbers that satisfies the following conditions:

I. $\mathbb{M} \supset \text{span}\{1, |v|^2\}$;

II. $\mathbb{M}$ is invariant under rotation;

III. The cone $\mathbb{M}_c = \{p \in \mathbb{M} : \langle \exp(p) \rangle < \infty\}$ has non-empty interior .

Spaces that satisfy Conditions I-III will be called *admissible*.

In Condition I, the constant functions are included in $\mathbb{M}$ so that any moment closure will include the conservation law for the electron concentration n. Most spaces also include multiples of the polynomial v, which gives a balance law for the

momentum, but we do not explicitly require it here. Multiples of $|v|^2$ give a balance law for the energy, although this is not a requirement for solving (4.4). Rather, we require $|v|^2 \in \mathbb{M}$ to ensure the minimization problems for $\mathcal{H}$ and for $\mathcal{K}$ are equivalent (see Chapter 3). Thus we have intentionally excluded the cases $\mathbb{M} = \text{span}\{1\}$ and $\mathbb{M} = \text{span}\{1, v\}$, even though these spaces are known to produce well-posed closures.

In Condition II, invariance under rotation means that $\mathbb{M}$ is unchanged when $v \mapsto O^T v$ for any orthogonal matrix O. This is a prerequisite of classical dynamics. For many cases, invariance under translation is also necessary, which means that $\mathbb{M}$ is also unchanged when $v \mapsto v - u$ for any $v \in \mathbb{R}^d$ (typically $d = 3$). In such cases, consistency implies that Condition I should include the polynomial v. However, this is *not* the case for semiconductors since the lattice provides a fixed frame of reference. Condition III requires, at a minimum, that $\mathbb{M}$ contain polynomials of even maximal degree to ensure the decay necessary for integrability. The reason for imposing this condition will become clear when we examine the dual problem to (4.4).

4.1.2 Construction of Admissible Spaces

We now discuss the practical issue of constructing an admissible space. Given the integers $N \geq 2$ and $d \geq 1$, let $\mathbb{P}_N$ be the set of polynomials from $\mathbb{R}^d$ to $\mathbb{R}$ of degree less than or equal to N. An admissible space $\mathbb{M} \subset \mathbb{P}_N$ is constructed by choosing homogeneous polynomials of degree $j \leq N$, beginning with N. Let $\mathbb{Q}_j$ be the space of homogeneous polynomials from $\mathbb{R}^d$ to $\mathbb{R}$of degree j. Each polynomial $p_j \in \mathbb{P}_j$ can

be represented by a symmetric j-fold tensor $B^j = B^j(p_j)$ via the tensor dot product

$$p_j(v) = B^j \cdot v^{\vee j},$$

in which case the components of B^j are just the coefficients of p. The tensor dot product is applied by simply summing over all available indices. Meanwhile, the superscript notation is used to denote symmetric tensor power.

The space $\mathbb{Q}_j$ can be decomposed into the direct sum

$$\mathbb{Q}_j = \begin{cases} \mathbb{H}_j \oplus |v|^2 \mathbb{H}_{j-2} \oplus |v|^4 \mathbb{H}_{j-4} \oplus \ldots \oplus |v|^j, & j \text{ even} \\ \mathbb{H}_j \oplus |v|^2 \mathbb{H}_{j-2} \oplus |v|^4 \mathbb{H}_{j-4} \oplus \ldots \oplus |v|^{j-1} \mathbb{H}_1, & j \text{ odd} \end{cases} \tag{4.5}$$

where $\mathbb{H}_i$ is the space of harmonic polynomials of degree i [26]. This series terminates at $|v|^j$ for j. If Y^i is the spherical harmonic tensor of degree i, defined for $\omega \in \mathbb{S}^{d-1}$, then any polynomial $q_i \in \mathbb{H}_i$ can be expressed with a tensor dot product

$$q_i(v) = |v|^i \tilde{B}^i \cdot Y^i(\omega), \qquad \omega = \frac{v}{|v|}, \tag{4.6}$$

where $\tilde{B}^i$ is a symmetric, traceless i-fold tensor. Together (4.5) and (4.6) show that $p_j \in \mathbb{Q}_j$ can be written as the product of the homogeneous term $|v|^j$ times linear combinations of spherical harmonic functions evaluated on the unit sphere. If j is even, then only even spherical harmonics will be included; if j is odd, then only odd

spherical harmonics will be included. If $\omega = v/|v|$, then

$$p_j(v) = |v|^j \sum_{i=0}^{j/2} \tilde{B}^{2i} \cdot Y^{2i}(\omega) , \qquad j \text{ even},$$
$$p_j(v) = |v|^j \sum_{i=0}^{(j-1)/2} \tilde{B}^{2i+1} \cdot Y^{2i+1}(\omega) , \qquad j \text{ odd}.$$

It is known [26] that the decomposition in (4.5) is the minimal decomposition of $\mathbb{P}_j$ into orthogonal, rotationally invariant subspaces (meaning that no proper subspace of $\mathbb{H}_j$ is rotationally invariant). Therefore, in order to satisfy Condition II, an admissible space $\mathbb{M}$ must be a direct sum of some combination of these subspaces taken from each $\mathbb{P}_j$, $j \leq N$. To satisfy Condition III, $\mathbb{M}$ must include polynomials from $\mathbb{P}_N$ to dominate the behavior of odd polynomials of lower degree for large $|v|$. Furthermore, amongst the degree N polynomials, $\mathbb{M}$ must include $|v|^N$. This is because spherical harmonics other than $Y^0 \equiv 1$ take on both positive and negative values on the unit sphere. Excluding $|v|^N$ would therefore lead to polynomials p such that

$$\lim_{r \to \infty} p(r\omega) = \infty$$

for ω contained in a subset of the sphere $\mathbb{S}^{d-1}$ with positive measure. In such cases $\exp(p)$ will not integrable and Condition III will be violated.

For illustration, we construct admissible spaces in the simple case that $N = 2$. The spherical harmonics up to order two are

$$Y^0(\omega) = 1 , \qquad Y^1(\omega) = \omega , \qquad Y^2(\omega) = \omega \vee \omega - \frac{1}{d} |\omega|^2 .$$

Any polynomial $p \in \mathbb{M}$ is the sum of its homogeneous components:

$$p\,(v) = p_2\,(v) + p_1\,(v) + p_0\,(v)\ .$$

We begin with the polynomial $p_2 \in \mathbb{P}_2$:

$$p_2\,(v) = B^2 \cdot v^{\vee 2} = \sum_{i,j=1}^{d} B^2_{ij} v_i v_j \tag{4.7}$$

for some symmetric 2-tensor B^2. In terms of spherical harmonics,

$$p_2\,(v) = |v|^2 \left[\tilde{B}^0 \cdot Y^0 + \tilde{B}^2 \cdot Y^2\right] = \tilde{B}^0 |v|^2 + \sum_{i,j=1}^{d} \tilde{B}^2_{ij} \left(v_i v_j - \frac{1}{d}|v|^2 \delta_{ij}\right)\ , \tag{4.8}$$

and comparing (4.7) with (4.8) gives explicit relations between B^2 and the tensors $\tilde{B}^0$ and $\tilde{B}^2$:

$$\tilde{B}^0 = \frac{1}{d}\,\mathrm{trace}(B^2) \qquad \text{and} \qquad \tilde{B}^2_{ij} = B^2_{ij}, \qquad i \neq j\,.$$

The homogeneous polynomials of lower degree can be trivially expressed in the same way:

$$p_1(v) = B^1 \cdot v\,, \qquad p_0 = B^0 \cdot 1\,.$$

The linear polynomial p_1 is just the usual dot product between the 1-tensor (vector) B^1 and v, and $p_0 = B^0$ is just a constant. For these trivial cases, $\tilde{B}^1 = B^1$ and

$\tilde{B}^0 = B^0$. Combining p_2, p_1, and p_0, we have

$$p(v) = \sum_{i.j}^{d} B^2_{ij} v_i v_j + \sum_{i=1}^{d} B^1_i v_i + B^0_i$$

or, in terms of the harmonic polynomials,

$$p\,(v) = \tilde{B}^0 |v|^2 + \sum_{i,j=1}^{d} \tilde{B}^2_{ij} \left(v_i v_j - \frac{1}{d} |v|^2 \delta_{ij} \right) + \sum_{i=1}^{d} \tilde{B}_i v_i + \tilde{B}^0_i \,.$$

There are four admissible spaces for which $N = 2$:

$$\begin{aligned} \mathbb{M}_1 &= \text{span}\{1, |v|^2\}\,, \\ \mathbb{M}_2 &= \text{span}\{1, v \vee v\}\,, \\ \mathbb{M}_3 &= \text{span}\{1, v, |v|^2\}\,, \\ \mathbb{M}_4 &= \text{span}\{1, v, v \vee v\}\,. \end{aligned} \tag{4.9}$$

The only degree two polynomials in the first and third spaces are constant multiples of the radial component $|v|^2$, in which case $\tilde{B}^2$ is identically zero. Degree two polynomials in the second and fourth spaces, on the other hand, include components of the spherical harmonic Y^2 and are therefore useful for modeling anisotropies. The spaces $\mathbb{M}_3$ and $\mathbb{M}_4$ are used to construct the Maxwellian and Gaussian closures, respectively, discussed in the previous chapter. Notice that they are the only two spaces that are translation invariant. This property is important for neutral fluids where there is no preferred frame of reference. However, with semiconductors, the crystal lattice

provides a preferred frame, and translation invariance is not needed.

For larger values of N, it is simpler to represent polynomials more abstractly. Let $\mathbf{m}$ be an array of $l \equiv \dim(\mathbb{M})$ polynomials that form a basis for $\mathbb{M}$, and introduce the decomposition

$$\mathbf{m} = (\mathbf{m}_0, \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N)^T \,, \tag{4.10}$$

where the l_j components of $\mathbf{m}_j$ are the j^{th} degree polynomials components of $\mathbf{m}$. Consistency requires that $\sum_{j=0}^{N} l_j = l$. The sub-arrays $\mathbf{m}_j$ may be thought of as vector or as tensors. For example, if $\mathbb{M} = \text{span}\{1, v, v \vee v\}$, then

$$\mathbf{m}_0 = 1\,, \qquad \mathbf{m}_1 = v\,, \qquad \mathbf{m}_2 = v \vee v\,.$$

Any polynomial $p \in \mathbb{M}$ is the sum of its homogeneous components:

$$p(v) = \boldsymbol{\alpha}^T \mathbf{m}(v) = \sum_{j=1}^{N} \boldsymbol{\alpha}_j^T \mathbf{m}_j(v) \,,$$

where $\boldsymbol{\alpha}$ is an array of l constant coefficients that decomposes as

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N)^T$$

and $\boldsymbol{\alpha}_j^T \mathbf{m}_j$ is the appropriate inner product. For each j, the array $\boldsymbol{\alpha}_j$ has l_j components. If one considers $\mathbf{m}_j$ and $\boldsymbol{\alpha}_j$ as symmetric j-fold tensors, then $\boldsymbol{\alpha}_j^T \mathbf{m}_j$ is just the usual tensor dot product. Because of the vector representation, we will frequently refer to $\mathbf{m}$ and $\boldsymbol{\alpha}$ and their subarrays as vectors. In particular, $\mathbf{m}$ and its

subarrays are frequently referred to as polynomial vectors because their components are polynomials.

4.1.3 The Entropy Functional

The strictly convex *entropy functional* $\mathcal{H} : \mathbb{F}_{\mathbf{m}} \mapsto \mathbb{R} \cup \{\infty\}$ is given by

$$\mathcal{H}(g) \equiv \langle g \log g - g \rangle .$$

By employing the convention $0 \log 0 = 0$—which is consistent with the fact that $\lim_{z \to 0} z \log z = 0$—one can make sense of the integrand when $g(v) = 0$. It is possible that $\mathcal{H}(g) = +\infty$; however, to ensure that $\mathcal{H}$ is well-defined, it must be shown that the negative contribution to the integral is finite. By convexity of the mapping $z \mapsto z \log z - z$, the following inequality holds for all $z, y > 0$:

$$z \log z - z \geq y \log y - y + (\log y)(z - y) \tag{4.11a}$$

$$= z \log y - y . \tag{4.11b}$$

Identifying $z = g(v)$ and $y = e^{-|v|^2}$ gives—after integration over the set $P = \left\{ v \in \mathbb{R}^d : g(v) \log g(v) - g(v) < 0 \right\}$—

$$\begin{aligned} \mathcal{H}^{-}(g) &\equiv -\int_P (g(v) \log g(v) - g(v))\, dv \\ &\leq \int_P \left(|v|^2 g(v) + e^{-|v|^2} \right) dv \\ &\leq \int_{\mathbb{R}^d} \left(|v|^2 g(v) + e^{-|v|^2} \right) dv. \end{aligned} \tag{4.12}$$

Since $|v|^2 \in \mathbb{M}$, $\int_{\mathbb{R}^d} |v|^2 g(v)\, dv$ is finite for all $g \in \mathbb{F}_{\mathbf{m}}$ and so too is $\mathcal{H}^-$.

4.1.4 Cones

The minimization problem is essentially a study of cones. A subset C of a vector space $\mathbb{X}$ is a *cone* if, for all real numbers $\lambda > 0$, $x \in C$ if and only if $\lambda x \in C$. It is said to be *pointed* if $x \in C$ and $-x \in C$ implies that $x = 0$.

One cone that we have already seen is $\mathbb{F}_{\mathbf{m}}$, which is both convex and pointed. Another important cone is the normal cone. Given a convex set $\Omega \subset \mathbb{X}$ and a point $x \in \partial\Omega$, the *normal cone of* Ω *at* x is

$$\mathcal{NC}(\Omega, x) \equiv \{x^* \in \mathbb{X}^* : x^*(y - x) \leq 0\,, \quad \forall\, y \in \Omega\}\ .$$

(Here $\mathbb{X}^*$ is the dual space of $\mathbb{X}$). If $\mathbb{X} = \mathbb{R}^l$ and $\partial\Omega$ is a C^1 (continuously differentiable) manifold at x, then $\mathcal{NC}(\Omega, x)$ is a ray with base point at the origin that points in the direction normal to $\partial\Omega$ at x. Even if $\partial\Omega$ is not C^1 at x, a standard result of differential geometry is that

$$\dim \mathcal{NC}(\Omega, x) = l - j\,,$$

where j is the dimension of the largest C^1 submanifold embedded in Ω that contains x.

An important function of cones is to expand the concept of scalar inequalities to general vector spaces. Given x and y in a vector space $\mathbb{X}$, we say that $x \leq y$, or

$y \geq x$, (with respect to C) if and only if $y - x \in C$. The *dual (or polar) cone* C^* consists of all elements $x^* \in \mathbb{X}^*$ such that the pairing $x^*(x) \geq 0$ for all $x \in C$. Given x^* and y^* in $\mathbb{X}^*$, we say that $x^* \leq y^*$, or $y^* \geq x^*$, (with respect to C^*) if and only if $y^* - x^* \in C^*$. Consider, for example, the convex cone

$$A \equiv \left\{ \boldsymbol{\alpha} \in \mathbb{R}^l : \boldsymbol{\alpha}^T \mathbf{m} \geq 0 \right\}$$

and its dual

$$A^* \equiv \left\{ \boldsymbol{\sigma} \in \mathbb{R}^l : \boldsymbol{\alpha}^T \boldsymbol{\sigma} \geq 0 \quad \forall\, \boldsymbol{\alpha} \in A \right\},$$

both of which depend on the vector $\mathbf{m}$. Given a vector $\boldsymbol{\alpha}, \boldsymbol{\sigma} \in \mathbb{R}^l$, we say that $\boldsymbol{\alpha} \geq 0$ (or $0 \leq \boldsymbol{\alpha}$) if and only if $\boldsymbol{\alpha} \in A$. and $\boldsymbol{\sigma} \geq^* 0$, or $0 \leq^* \boldsymbol{\sigma}$, if and only if $\boldsymbol{\sigma} \in A^*$. Similar convex cones, corresponding to each vector $\mathbf{m}_j$ of even degree polynomials, are given by

$$A_j \equiv \left\{ \boldsymbol{\alpha}_j \in \mathbb{R}^{l_j} : \boldsymbol{\alpha}_j^T \mathbf{m}_j(\omega) \geq 0 \quad \forall\, \omega \in \mathbb{S}^{d-1} \right\}, \qquad j \text{ even}, j \leq N \tag{4.13a}$$

$$A_j^* \equiv \left\{ \boldsymbol{\sigma}_j \in \mathbb{R}^{l_j} : \boldsymbol{\alpha}_j^T \boldsymbol{\sigma}_j \geq 0 \quad \forall\, \boldsymbol{\alpha}_j \in A_j \right\}, \qquad j \text{ even}, j \leq N \tag{4.13b}$$

In the following subsections, we will discuss several other important cones.

4.1.4.1 Realizable Densities When solving (4.4), we are only interested in those vectors $\boldsymbol{\rho}$ that are physically realizable—that is, they are moments of a function $f \in \mathbb{F}_{\mathbf{m}}$ with respect to $\mathbf{m}$. The image of $\mathbb{F}_{\mathbf{m}}$ under the *moment mapping* $g \mapsto \langle \mathbf{m} g \rangle$

is called the *set of realizable densities*:

$$\mathcal{R}_{\mathbf{m}} \equiv \left\{ \boldsymbol{\rho} \in \mathbb{R}^l : \boldsymbol{\rho} = \langle \mathbf{m} g \rangle \, , \, g \in \mathbb{F}_{\mathbf{m}} \right\} \, .$$

A density $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ has a natural decomposition based on the decomposition of $\mathbf{m}$ in (4.10):

$$\boldsymbol{\rho} = \left(\boldsymbol{\rho}_0, \boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_N \right)^T \, ,$$

where $\boldsymbol{\rho}_j = \langle \mathbf{m}_j g \rangle$ for some $g \in \mathbb{F}_{\mathbf{m}}$. The set $\mathcal{R}_{\mathbf{m}}$ has some very nice properties.

Theorem 12 *The set $\mathcal{R}_{\mathbf{m}}$ is an open, pointed, convex cone in $\mathbb{R}^l$. In fact, $\mathcal{R}_{\mathbf{m}}$ is the dual cone A^*, and every vector in $\mathcal{R}_{\mathbf{m}}$ can be realized by a non-negative function supported on a compact set.*

Proof. The fact that $\mathcal{R}_{\mathbf{m}}$ is a pointed, convex cone follows directly from those same properties of $\mathbb{F}_{\mathbf{m}}$. To show that it is open, choose any $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, and let $g \in \mathbb{F}_{\mathbf{m}}$ be such that $\langle \mathbf{m} g \rangle = \boldsymbol{\rho}$. Then there exists a compact set $E \subset \mathbb{R}^d$ with positive measure and a constant $c > 0$ such that $g \geq c$ on E. Define the linear map $\mathbf{p} : \mathbb{R}^N \to \mathbb{R}^N$ by

$$\boldsymbol{\beta} \mapsto \mathbf{p}(\boldsymbol{\beta}) \equiv \textstyle\int_E (\boldsymbol{\beta}^T \mathbf{m}) \mathbf{m}. \tag{4.14}$$

Then $\mathbf{p}(0) = 0$, and

$$\frac{\partial \mathbf{p}}{\partial \boldsymbol{\beta}} = \textstyle\int_E \mathbf{m}^T \mathbf{m}$$

is a constant, positive-definite matrix, which implies that $\mathbf{p}$ has a continuous inverse defined on all of $\mathbb{R}^l$. Hence, for each $\delta > 0$, the image of the set $\mathcal{B}_\delta \equiv$

$\left\{\boldsymbol{\beta} \in \mathbb{R}^l : |\boldsymbol{\beta}| < \delta\right\}$ under $\mathbf{p}$ is open.

Define $g_{\boldsymbol{\beta}} \in L_1(\mathbb{R}^d)$ by

$$g_{\boldsymbol{\beta}}(v) \equiv \begin{cases} \boldsymbol{\beta}^T \mathbf{m}(v), & v \in E, \\ 0, & \text{otherwise}. \end{cases} \tag{4.15}$$

Then $\mathbf{p}(\boldsymbol{\beta})$ is realized by $g_{\boldsymbol{\beta}}$, and

$$g + g_{\boldsymbol{\beta}} \geq 0$$

whenever $|\boldsymbol{\beta}^T \mathbf{m}| < c$. This means that $g + g_{\boldsymbol{\beta}} \in \mathbb{F}_{\mathbf{m}}$ whenever $|\boldsymbol{\beta}^T \mathbf{m}| < c$, in which case

$$\left\{\boldsymbol{\sigma} \in \mathbb{R}^l : \boldsymbol{\sigma} = \langle \mathbf{m}(g + g_{\boldsymbol{\beta}}) \rangle, \, |\boldsymbol{\beta}| < \frac{c}{\sup_{v \in E}(|\mathbf{m}|)}\right\}$$

is an open subset of $\mathcal{R}_{\mathbf{m}}$ that contains $\boldsymbol{\rho}$. Thus $\mathcal{R}_{\mathbf{m}}$ is open.

We now prove $\mathcal{R}_{\mathbf{m}} = A^*$. Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ be realized by a function $f \in \mathbb{F}_{\mathbf{m}}$. Then for any $\boldsymbol{\alpha} \in A$,

$$\boldsymbol{\alpha}^T \boldsymbol{\rho} = \left\langle \boldsymbol{\alpha}^T \mathbf{m} f \right\rangle \geq 0.$$

Therefore $\mathcal{R}_{\mathbf{m}} \subset A^*$. Conversely, for any $\boldsymbol{\rho} \in A^*$, define the function $z : \mathbb{R}^N \to \mathbb{R}$ by

$$\boldsymbol{\alpha} \mapsto z(\boldsymbol{\alpha}; \boldsymbol{\rho}, E) \equiv \boldsymbol{\alpha}^T \boldsymbol{\rho} - \textstyle\int_E \exp(\boldsymbol{\alpha}^T \mathbf{m})$$

where E is now *any* compact subset of $\mathbb{R}^d$. Following the arguments found in the appendix of [42], it is straight-forward to show that z is convex and that $\lim_{|\boldsymbol{\alpha}| \to \infty} z(\boldsymbol{\alpha}) =$

$-\infty$. (The proof utilizes the fact that, since $\boldsymbol{\rho} \in A^*$, $\boldsymbol{\alpha}^T\boldsymbol{\rho} \geq 0$ whenever $\boldsymbol{\alpha} \in A$). These properties imply that z has a maximum at some point $\bar{\boldsymbol{\alpha}} \in \mathbb{R}^l$, and first-order optimality conditions imply that

$$\boldsymbol{\rho} = \int_E \exp(\bar{\boldsymbol{\alpha}}^T \mathbf{m}) \, .$$

Therefore $\boldsymbol{\rho}$ is realized by a non-negative function in $\mathbb{F}_{\mathbf{m}}$ with compact support and $A^* \subset \mathcal{R}_{\mathbf{m}}$. ■

4.1.4.2 Exponentially Realizable Densities An important subset of $\mathcal{R}_{\mathbf{m}}$ consists of those vectors $\boldsymbol{\rho}$ that can be realized by functions of the form

$$G_{\boldsymbol{\alpha}} \equiv \exp(\boldsymbol{\alpha}^T \mathbf{m}) \, . \tag{4.16}$$

Define the set

$$\mathcal{A}_{\mathbf{m}} \equiv \left\{ \boldsymbol{\alpha} \in \mathbb{R}^l : \mathbf{m} G_{\boldsymbol{\alpha}} \in L_1\left(\mathbb{R}^d\right) \right\} \tag{4.17}$$

and the function $\mathbf{r} : \mathcal{A}_{\mathbf{m}} \to \mathbb{R}^l$ given by

$$\mathbf{r}(\boldsymbol{\alpha}) \equiv \langle \mathbf{m} G_{\boldsymbol{\alpha}} \rangle \, . \tag{4.18}$$

The image of $\mathcal{A}_{\mathbf{m}}$ under $\mathbf{r}$ is the set of *exponentially realizable densities*:

$$\mathcal{R}_{\mathbf{m}}^{\exp} \equiv \left\{ \boldsymbol{\rho} \in \mathbb{R}^l : \boldsymbol{\rho} = \mathbf{r}(\boldsymbol{\alpha}) \, , \; \boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \right\} \, .$$

This set is a cone and clearly $\mathcal{R}_{\mathbf{m}}^{\exp} \subset \mathcal{R}_{\mathbf{m}}$.

Characterizing the set $\mathcal{A}_{\mathbf{m}}$ turns out to be very important. Like $\mathcal{R}_{\mathbf{m}}$, it is a pointed, convex cone in $\mathbb{R}^l$. However, generally speaking, $\mathcal{A}_{\mathbf{m}}$ is not open. Its interior is simple:

$$\begin{aligned} \operatorname{int} \mathcal{A}_{\mathbf{m}} &= \left\{ \boldsymbol{\alpha} \in \mathbb{R}^l : \boldsymbol{\alpha}_N^T \mathbf{m}_N(\omega) < 0 \text{ for all } \omega \in \mathbb{S}^{d-1} \right\} \\ &= \left\{ \boldsymbol{\alpha} \in \mathbb{R}^l : \boldsymbol{\alpha}_N \in -\operatorname{int} A_N \right\} . \end{aligned} \tag{4.19}$$

Notice that Condition III is equivalent to $\operatorname{int} \mathcal{A}_{\mathbf{m}}$ being non-empty. If $\boldsymbol{\alpha} \in \operatorname{int} \mathcal{A}_{\mathbf{m}}$, then the behavior of $p = \boldsymbol{\alpha}^T \mathbf{m}$ will be dominated for large $|v|$ by the homogeneous component $p_N = \boldsymbol{\alpha}_N^T \mathbf{m}_N$, and

$$\lim_{|v| \to \infty} p(v) = \lim_{|v| \to \infty} p_N(v) = \lim_{|v| \to \infty} |v|^N p_N\left(v/|v|\right) = -\infty .$$

Therefore $G_{\boldsymbol{\alpha}}$ will decay exponentially, in which case all of its moments will be finite.

The boundary component $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ is much more complicated. If $\boldsymbol{\alpha} \in \partial \mathcal{A}_{\mathbf{m}}$, then $p_N(\omega) = 0$ for at least one $\omega \in \mathbb{S}^{d-1}$, and it may be that there are unbounded sequences $\{v_i\}_{i=1}^{\infty}$ such that $\lim_{i \to \infty} p(v_i) > -\infty$. Whether $G_{\boldsymbol{\alpha}}$ has finite moments in these cases is not entirely clear. We therefore introduce the first of two conditions.

Condition 13 *The set $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ can be decomposed into a finite union of disjoint smooth manifolds of codimension two or greater in $\mathbb{R}^l$. If s is one such manifold, then the projection $\boldsymbol{\alpha} \mapsto (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{N-1})$ maps s onto a manifold of codimension one or greater in $\mathbb{R}^{l - l_N}$.*

This decomposition of $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ is called a *stratification*. The manifolds that are its elements are called *strata*.

Let us discuss the dimensional restrictions of this condition. Clearly $\partial\mathcal{A}_{\mathbf{m}} \subset \{\boldsymbol{\alpha} \in \mathbb{R}^l : \boldsymbol{\alpha}_N \in \partial(-A_N)\}$, the latter of which has codimension one in $\mathbb{R}^l$. Then, in order to maintain integrability condition (4.17) that defines $\mathcal{A}_{\mathbf{m}}$, we expect further restrictions on the components $\boldsymbol{\alpha}_j$, $j < N$ that reduce the dimension of $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ by at least one degree.

4.1.5 Semi-algebraic Sets

In this subsection, we briefly discuss semi-algebraic sets, which will serve as a tool for later results.

Definition 14 *The class of semi-algebraic subsets of $\mathbb{R}^l$ is the smallest Boolean algebra of subsets of $\mathbb{R}^l$ which contains sets of the form*

$$\{x \in \mathbb{R}^l : p(x) > 0\}$$

for any polynomial function $p : \mathbb{R}^l \to \mathbb{R}$.

By definition, the class of semi-algebraic sets is closed under finite unions, intersections, complements, and Cartesian products. The reader is referred to [10, 31, 60] for a thorough discussion that includes the following facts.

Lemma 15 *Let $\Omega \subset \mathbb{R}^l$ be a semi-algebraic set. Then the following hold.*

1. *If $p : \mathbb{R}^j \to \mathbb{R}^l$ is a polynomial mapping, then $p^{-1}(\Omega)$ is semi-algebraic*

2. *(Tarski-Seidenberg Theorem) If* $p : \mathbb{R}^l \to \mathbb{R}^j$ *is a polynomial mapping, then* $p(\Omega)$ *is semi-algebraic. This holds in particular when* $j < l$ *and* p *projects elements in* $\mathbb{R}^l$ *onto any set of* j *coordinates.*

3. *The closure, interior, and boundary of a semi-algebraic sets are semi-algebraic.*

4. Ω *can be written as a finite disjoint union of smooth manifolds of codimension one or greater.*

We use several of these facts to prove another lemma.

Lemma 16 *Let* A_j *be given by (4.13a). Then the following hold.*

1. *The cone* A_j *and its boundary are semi-algebraic for* j *even,* $2 \leq j \leq N$.

2. *The sets* $\operatorname{int} \mathcal{A}_{\mathbf{m}}$, $\operatorname{cl} \mathcal{A}_{\mathbf{m}}$, *and* $\partial \mathcal{A}_{\mathbf{m}}$ *are all semi-algebraic.*

Proof. To prove the first statement, let $\mathbf{m}$ be given, and let j be an even integer. Define the set

$$S_j = \left\{ (\boldsymbol{\alpha}_j, \omega) \in \mathbb{R}^{l_j} \times \mathbb{S}^{d-1} : \boldsymbol{\alpha}_j^T \mathbf{m}_j(\omega) < 0 \right\} .$$

Clearly S_j is algebraic. Furthermore, the cone A_j is the complement of the projection of S_j onto its first l_j components:

$$A_j = \left\{ \boldsymbol{\alpha}_j \in \mathbb{R}^{l_j} : (\boldsymbol{\alpha}_j, \omega) \notin S_j \right\} .$$

Lemma 15.2 implies that A_j is semi-algebraic, and Lemma 15.3 implies that $\partial(A_j)$ is semi-algebraic as well. The second statement then follows immediately from (4.19) and Lemma 15.3. ■

It should be noted that one way to show Condition 13 holds is to prove that $\partial\mathcal{A}_{\mathbf{m}} \cap \mathcal{A}_{\mathbf{m}}$, or equivalently $\mathcal{A}_{\mathbf{m}}$ itself, is semi-algebraic. The simple form of $G_{\boldsymbol{\alpha}}$ leads us to believe that this is a plausible result, but we have not been able to prove or disprove it.

4.2 Entropy Minimization

4.2.1 Formulation

Given $\boldsymbol{\rho} = (\boldsymbol{\rho}_0, \ldots, \boldsymbol{\rho}_N) \in \mathcal{R}_{\mathbf{m}}$, we seek a solution of (4.4), where the relation $\langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}$ (or, equivalently, $\boldsymbol{\rho} \succeq^* \langle \mathbf{m} g \rangle$) is a shorthand for

$$\langle \mathbf{m}_j g \rangle = \boldsymbol{\rho}_j \,, \qquad 0 \le j \le N-1 \,, \tag{4.20a}$$

$$\langle \mathbf{m}_N g \rangle \preceq^* \boldsymbol{\rho}_N \,, \tag{4.20b}$$

and the inequality in (4.20b) is understood in the sense of the dual cone A_N^*. The components of $\langle \mathbf{m}_j g \rangle$, $0 \le j < N$, will be referred to as *lower-order moments*, and the components of $\langle \mathbf{m}_N g \rangle$ will be referred to as *higher-order moments*.

The main result in [73] concerning (4.4) is the following theorem.

Theorem 17 (Schneider) *Problem (4.4) possesses a unique minimizer of the form* $G_{\mathbf{a}(\boldsymbol{\rho})}$, *where* $G_{\boldsymbol{\alpha}}$ *is given by (4.16) and* $\mathbf{a}(\boldsymbol{\rho}) \in \mathbb{R}^l$ *is a vector of Lagrange multipliers.*

We briefly sketch the existence proof below. See [73] for details. Uniqueness of the minimizer follows immediately from the strict convexity of $\mathcal{H}$, and we will re-prove the form of the minimizer during the course of the discussion that follows.

Sketch of Proof (Existence). Let $C_{\mathbf{m}} = \{g \in \mathbb{F}_{\mathbf{m}} : \langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}\}$. Since $\mathcal{H}(g)$ is bounded below on $C_{\mathbf{m}}$ (see equation (4.12)), there exists a minimizing sequence $\{g_i\}_{i=1}^{\infty} \subset C_{\mathbf{m}}$ such that $\mathcal{H}(g_i) \to h(\boldsymbol{\rho})$. The fact that the entropy sequence $\mathcal{H}(g_i)$ and the sequence of moments $\langle \mathbf{m}_N g_i \rangle$ are bounded implies, via the Dunford-Pettis Lemma, that g_i converges weakly in L^1 through a subsequence. Let $\hat{g}_{\boldsymbol{\rho}}$ be the limit of that subsequence. Then Fatou's Lemma implies $\langle \mathbf{m}_j \hat{g}_{\boldsymbol{\rho}} \rangle \leq^* \boldsymbol{\rho}_j$, $0 \leq j \leq N$. Using the fact that g_i converges weakly in L^1 and that $\langle \mathbf{m}_N g_i \rangle \leq^* \boldsymbol{\rho}_N$, it can be shown that $\langle \mathbf{m}_j \hat{g}_{\boldsymbol{\rho}} \rangle = \boldsymbol{\rho}_j$ for $0 \leq j < N$. Thus $\hat{g}_{\boldsymbol{\rho}}$ is feasible and solves (4.4). ■

Note that if one were to take $\{g_i\}_{i=1}^{\infty} \subset C_{\mathbf{m}}^0 \equiv \{g \in \mathbb{F}_{\mathbf{m}} : \langle \mathbf{m} g \rangle = \boldsymbol{\rho}\}$ rather than in $C_{\mathbf{m}}$, then $\{g_i\}_{i=1}^{\infty}$ would still converge with $\langle \mathbf{m}_j \hat{g}_{\boldsymbol{\rho}} \rangle = \boldsymbol{\rho}_j$ for $j < N$. However, Fatou's Lemma implies only that $\langle \mathbf{m}_N \hat{g}_{\boldsymbol{\rho}} \rangle \leq^* \boldsymbol{\rho}_N$, and there is no way to ensure that $\langle \mathbf{m}_N \hat{g}_{\boldsymbol{\rho}} \rangle = \boldsymbol{\rho}_N$. This is precisely why (4.20b) is an inequality constraint: $C_{\mathbf{m}}$ is closed in the weak-L^1 topology whereas $C_{\mathbf{m}}^0$ is not.

Such behavior begs the following question: For what values of $\boldsymbol{\rho}$ does the sequence $\{g_i\}_{i=1}^{\infty}$ *not* converge inside $C_{\mathbf{m}}^0$? In other words, what does the set $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ look like? In [73], the author attempts to address this issue in the following theorem.

Theorem 18 (Schneider) *Problem (4.3) has a minimum if and only if there exists no function of the form $G_{\boldsymbol{\alpha}}$ in $C_{\mathbf{m}} \backslash C_{\mathbf{m}}^0$.*

The proof of this theorem follows immediately from uniqueness of the minimizer and the remarks following the proof of Theorem 17. The result, however, is of little practical use. In particular, it provides no insight into the geometry of $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$.

On the other hand, a geometric interpretation of $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ is given in [42, 43] for the special case when $\mathbf{m}_N = |v|^N$. Our goal here is to describe the geometry of $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ in the general setting. We do so by formulating the dual problem for (4.3) and then analyzing the relationship between $\boldsymbol{\rho}$ and $\mathbf{a}(\boldsymbol{\rho})$ given by the complementary slackness conditions. In the process, we will recover results from [42, 43] and also [73].

4.2.2 The Dual Function

Because $\mathcal{H}$ is convex on $\mathbb{F}_{\mathbf{m}}$ and the constraints are linear, (4.4) can be solved via the dual formulation [56]. Define the Lagrangian function $\mathcal{L} : \left(\mathbb{F}_{\mathbf{m}} \times \mathbb{R}^N \times \mathcal{R}_{\mathbf{m}}\right) \mapsto \mathbb{R}$ by

$$\mathcal{L}\left(g, \boldsymbol{\alpha}, \boldsymbol{\rho}\right) \equiv \mathcal{H}(g) + \boldsymbol{\alpha}^T \left(\boldsymbol{\rho} - \langle \mathbf{m} g \rangle\right) \tag{4.21}$$

and the dual function $\psi : \mathbb{R}^N \times \mathcal{R}_{\mathbf{m}} \mapsto \mathbb{R}$ by

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \equiv \inf_{g \in \mathbb{F}_{\mathbf{m}}} \mathcal{L}\left(g, \boldsymbol{\alpha}, \boldsymbol{\rho}\right). \tag{4.22}$$

We can compute $\psi(\boldsymbol{\alpha}, \boldsymbol{\rho})$ explicitly.

Theorem 19 *For all* $\boldsymbol{\alpha} \in \mathbb{R}^N$ *and* $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$,

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) = \begin{cases} \boldsymbol{\alpha}^T \boldsymbol{\rho} - \langle G_{\boldsymbol{\alpha}} \rangle, & G_{\boldsymbol{\alpha}} \in L_1(\mathbb{R}^d) \\ -\infty, & G_{\boldsymbol{\alpha}} \notin L_1(\mathbb{R}^d) \end{cases}. \tag{4.23}$$

Proof. Suppose first that $G_{\boldsymbol{\alpha}} \in L_1(\mathbb{R}^d)$. Then refer to the inequality (4.11b) and make the identification $z = g(v)$ and $y = G_{\boldsymbol{\alpha}}(v)$ to derive the point-wise inequality

$$(g \log g - g) \geq g \log G_{\boldsymbol{\alpha}} - G_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^T \mathbf{m} g - G_{\boldsymbol{\alpha}}$$

which implies that

$$(g \log g - g) - \boldsymbol{\alpha}^T \mathbf{m} g \geq -G_{\boldsymbol{\alpha}}. \tag{4.24}$$

Integration of (4.24) over $\mathbb{R}^d$ and addition of $\boldsymbol{\alpha}^T \boldsymbol{\rho}$ to both sides gives a lower bound on $\mathcal{L}$ (and hence ψ):

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \geq \boldsymbol{\alpha}^T \boldsymbol{\rho} - \langle G_{\boldsymbol{\alpha}} \rangle \ . \tag{4.25}$$

However, if $G_{\boldsymbol{\alpha}} \in L_1\left(\mathbb{R}^d\right)$, then $\boldsymbol{\mathcal{L}}(G_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = \boldsymbol{\alpha}^T \boldsymbol{\rho} - \langle G_{\boldsymbol{\alpha}} \rangle$ and (4.25) becomes an equality.

If $G_{\boldsymbol{\alpha}} \notin L_1\left(\mathbb{R}^d\right)$, then define for any set measurable $K \subset \mathbb{R}^d$, the function

$$G_{\boldsymbol{\alpha}}^K(v) = \begin{cases} G_{\boldsymbol{\alpha}}(v), & v \in K \\ 0, & \text{otherwise} \end{cases} .$$

Then $G_{\boldsymbol{\alpha}}^K \in \mathbb{F}_{\mathbf{m}}$ whenever K is bounded, and

$$\boldsymbol{\mathcal{L}}\left(G_{\boldsymbol{\alpha}}^K, \boldsymbol{\alpha}, \boldsymbol{\rho}\right) = \boldsymbol{\alpha}^T \boldsymbol{\rho} - \int_K G_{\boldsymbol{\alpha}} \, dv.$$

If $\{K_i\}$ is a sequence of compact sets with $K_1 \subset K_2 \subset K_3 \subset \cdots$ and $\cup_i K_i = \mathbb{R}^d$, then

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \leq \lim_{i \to \infty} \mathcal{L}\left(G_{\boldsymbol{\alpha}}^{K_i}, \boldsymbol{\alpha}, \boldsymbol{\rho}\right) = -\infty.$$

■

It should be noted that ψ differs only by a linear term from the density potential h^* that was introduced in Chapter 3:

$$h^*(\boldsymbol{\alpha}) = \langle G_{\boldsymbol{\alpha}} \rangle = \boldsymbol{\alpha}^T \boldsymbol{\rho} - \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}).$$

Because optimality conditions are frequently expressed in terms of first and second derivatives, the smoothness properties of ψ are important. This is true both when trying to identify analytical solutions and when developing numerical algorithms.

Theorem 20 *For any $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, the following hold*

1. *For any $\boldsymbol{\alpha}, \boldsymbol{\alpha} + \boldsymbol{\delta} \in \mathcal{A}_{\mathbf{m}}$, the function*

$$\phi(\tau) \equiv \psi(\boldsymbol{\alpha} + \tau\boldsymbol{\delta}, \boldsymbol{\rho})$$

is a twice differentiable function with

$$\phi'(\tau) = \boldsymbol{\delta}^T \mathbf{r}(\boldsymbol{\alpha} + \tau\boldsymbol{\delta}) - \boldsymbol{\delta}^T \boldsymbol{\rho}\,, \tag{4.26a}$$

$$\phi''(\tau) = \left\langle \left(\boldsymbol{\delta}^T \mathbf{m}\right)^2 G_{\boldsymbol{\alpha}+\tau\boldsymbol{\delta}} \right\rangle\ . \tag{4.26b}$$

Its first derivative is an increasing function of τ.

2. $\psi(\cdot, \boldsymbol{\rho})$ *is strictly convex on* $\mathcal{A}_{\mathbf{m}}$ *and infinitely Fréchet differentiable on* $\operatorname{int} \mathcal{A}_{\mathbf{m}}$, *with derivatives*

$$\frac{\partial \psi}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\rho}) = \mathbf{r}(\boldsymbol{\alpha}) - \boldsymbol{\rho}\,, \tag{4.26c}$$

$$\frac{\partial^{(i)} \psi}{\partial \boldsymbol{\alpha}^{(i)}}(\boldsymbol{\alpha}, \boldsymbol{\rho}) = \left\langle \mathbf{m}^{\vee(i)} G_{\boldsymbol{\alpha}} \right\rangle\,, \qquad i > 1\,. \tag{4.26d}$$

.

Proof. For the proofs of these statements, we refer the reader to Lemmas 5.1 and 5.2 in [43] along with a few comments. First, the lemmas in [43] refer to h^* rather than $\psi(\cdot, \boldsymbol{\rho})$. This makes little difference since they differ only by a linear factor. Also, the proofs in [43] are constructed specifically for the special case when $m_N = |v|^N$; however, modifications to the general setting are straight-forward. ■

In spite of the smoothness properties given by Theorem 20, the dual function is not even continuous at the boundary of $\mathcal{A}_{\mathbf{m}}$. Indeed, given a sequence $\{\boldsymbol{\alpha}_i\}_{i=1}^{\infty} \in \mathcal{A}_{\mathbf{m}}$ with limit $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$, it is possible that

$$h^*(\boldsymbol{\alpha}) < \lim_{i \to \infty} h^*(\boldsymbol{\alpha}_i).$$

As an example, consider the one-dimensional case when $\mathbf{m} = (1, v, v^2, v^3, v^4)^T$, which has been studied in detail in [42]). Given the following five points in the (v, w) plane:

$$(v_0, w_0) = (0, 0)\ , \qquad (v_1, w_1) = (1, 0)\ , \qquad (v_2, w_2) = \left(i, -i^2\right)\ ,$$

$$(v_3, w_3) = (2i, i)\ , \qquad (v_4, w_4) = (2i + 1, 0)\ ,$$

the unique degree four polynomial interpolating these points is

$$p_i(v) = (\boldsymbol{\alpha})_i^T \mathbf{m}(v) = \sum_{j=0}^{4} (\alpha_j)_i v^j\ ,$$

where

$$(\alpha_0)_i = 0\,, \qquad (\alpha_1)_i = \frac{2i+1}{4i-2} + \frac{4i^2+2i}{i^2-1}\,, \qquad (\alpha_2)_i = -\frac{4i^2+6i+1}{i^2-1} - \frac{2i^2+4i+1}{4i^2-2i}$$

$$(\alpha_3)_i = \frac{4i+2}{i^2-1} + \frac{3i+2}{4i^2-2i}\,, \qquad (\alpha_4)_i = -\frac{1}{i^2-1} - \frac{1}{4i^2-2i}$$

(The notation $(\boldsymbol{\alpha})_i$ denotes a sequence of vectors rather than the usual notation $\boldsymbol{\alpha}_i$, which denotes the components of a single vector $\boldsymbol{\alpha}$ corresponding to polynomials of degree i). As $i \to \infty$, $(\boldsymbol{\alpha})_i \to \boldsymbol{\alpha} = (0, 3/4, -9/2, 0, 0)^T$; hence $G_{\boldsymbol{\alpha}}$ is integrable. However, one may readily check that p is positive and concave on the interval $[2i, 2i + 1]$,

in which case

$$\begin{aligned} h^*(\boldsymbol{\alpha}_i) &= \langle G_{\boldsymbol{\alpha}_i} \rangle \\ &> \int_{2i}^{2i+1} e^{p_i(v)}\, dv \\ &> \int_{2i}^{2i+1} (1 + p_i(v))\, dv \\ &> 1 + \frac{i}{2} \to \infty \qquad \text{as } i \to \infty\,. \end{aligned}$$

Note that positivity of p_i on $[2i, 2i+1]$ gives the second inequality above since $e^x > 1 + x$ for $x > 0$, and concavity implies that the graph of p_i lies above the line segment ℓ joining the points $(2i, i)$ and $(2i+1, 0)$ in the (v, w) plane. Therefore the integral of p_i over $[2i, 2i+1]$ is bounded below by the area of the triangle formed by ℓ, the v-axis, and the line $\{v = 2i\}$. The area of this triangle is $i/2$. A similar argument shows that, for any $j \geq 0$,

$$\left\langle |v|^j G_{\boldsymbol{\alpha}_i} \right\rangle \to \infty \qquad \text{as } i \to \infty$$

while $\langle v^j G_{\boldsymbol{\alpha}} \rangle$ is finite.

The reason that $\psi(\cdot, \boldsymbol{\rho})$ is discontinuous at the boundary of $\mathcal{A}_{\mathbf{m}}$ is the same reason that the minimization problem (4.2) with equality constraints fails: because mass at the tails of the functions escapes as $i \to \infty$. In the example above, this is precisely what happens to the mass of $G_{\boldsymbol{\alpha}_i}$ that is supported on the interval $[2i, 2i+1]$. A similar effect occurs with the minimizing sequence $\{g_i\}_{i=1}^{\infty}$ in the proof of Theorem 17. The difference is that only the highest moments fail to converge in the minimizing

sequence, whereas none of moments in this example converge. The reason for this difference is that the moments $\langle \mathbf{m} g_i \rangle$ from the minimizing sequence are all bounded. In particular, the bound on $|\langle \mathbf{m}_N f_i \rangle|$ ensures that, for each $j < N$ and for all i,

$$\begin{aligned}
\lim_{R\to\infty} \int_{|v|<R} |\mathbf{m}_j f_i| \, dv &= \lim_{R\to\infty} \int_{|v|<R} \left| \frac{\mathbf{m}_j}{\mathbf{m}_N} \mathbf{m}_N f_i \right| \, dv \\
&\leq \lim_{R\to\infty} \frac{1}{R} \int_{|v|<R} |\mathbf{m}_N f_i| \, dv \\
&\leq \lim_{R\to\infty} \frac{\text{Const.}}{R} \\
&= 0 \, .
\end{aligned}$$

Thus not enough mass is lost in the limit to make a difference for the lower-order moments. A similar result would hold for the sequence $\{G_{\boldsymbol{\alpha}_i}\}_{i=1}^{\infty}$ if the moments of $G_{\boldsymbol{\alpha}_i}$ were controlled in some way. Controlling the moments is, in effect, the same as requiring $\boldsymbol{\alpha}_i \to \boldsymbol{\alpha}$ is along a specified path. In fact, we see later that the function $\psi\left(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}\right)$, where $\mathbf{a}(\boldsymbol{\rho})$ is the Lagrange multiplier associated with $\boldsymbol{\rho}$, is a continuous function of $\boldsymbol{\rho}$.

Unfortunately, the smoothness properties given by Theorem 20 are not enough for our purposes, and we will need to assume another condition concerning the behavior of the dual function when restricted to $\partial\mathcal{A}_{\mathbf{m}} \cap \mathcal{A}_{\mathbf{m}}$.

Condition 21 *Let s be an element of the stratification of $\partial\mathcal{A}_{\mathbf{m}} \cap \mathcal{A}_{\mathbf{m}}$ as described in Condition 13, and let $\psi_s(\cdot, \rho)$ be the restriction of ψ to s. Then $\psi_s(\cdot, \rho)$ is infinitely Fréchet differentiable on s with derivatives given by (2).*

We remark that this condition holds in one dimension and that, like the one dimensional example, the lack of smoothness in $\psi(\cdot, \boldsymbol{\rho})$ at the boundary is due to the loss of mass at the tails of the integrand. However, the behavior of polynomials in multiple dimensions in much more complex and we have, this in case, no conditions by which to define the elements in the stratification of $\partial\mathcal{A}_{\mathbf{m}} \cap \mathcal{A}_{\mathbf{m}}$. Clearly one must first prove Condition 13 before the validity of Condition 21 can really be addressed.

4.2.3 Duality Theorems

We present two duality theorems which, in conjunction with the explicit expression for ψ, provide a solution to (4.4). In addition, we establish a *complementary slackness condition* which will be later used to describe the geometry of the set $\mathcal{R}_{\mathbf{m}}^{\exp}$.

Theorem 22 *Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, and let h and ψ be given by (4.4) and (4.23), respectively. Then*

$$h(\boldsymbol{\rho}) = \max_{\boldsymbol{\alpha}_N \leq \mathbf{0}} \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \,, \tag{4.27}$$

where the maximum on the right is achieved by a unique $\hat{\boldsymbol{\alpha}} \in \left\{\boldsymbol{\alpha} \in \mathbb{R}^N : \boldsymbol{\alpha}_N \leq \mathbf{0}\right\}$. Furthermore if $\hat{g}_{\boldsymbol{\rho}}$ solves (4.4), then $\hat{g}_{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\alpha}}$ satisfy the complementary slackness condition,

$$\hat{\boldsymbol{\alpha}}^T \left(\boldsymbol{\rho} - \langle \mathbf{m} \hat{g}_{\boldsymbol{\rho}} \rangle\right) = 0 \,, \tag{4.28}$$

and $\hat{g}_{\boldsymbol{\rho}}$ minimizes $\mathcal{L}\left(g, \hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}\right)$ over $\mathbb{F}_{\mathbf{m}}$, i.e.,

$$\psi(\hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}) = \mathcal{L}\left(\hat{g}_{\boldsymbol{\rho}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}\right) . \tag{4.29}$$

The proof of this theorem is a bit technical and therefore left to the appendix. The argument makes no assumptions about the differentiability of $\mathcal{H}$; rather, it is based purely on convex analysis.

With respect to (4.27), we may assume $\hat{\boldsymbol{\alpha}}_N$ is such that $G_{\hat{\boldsymbol{\alpha}}} \in L^1(\mathbb{R}^d)$, since otherwise $\psi(\hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}) = -\infty$. Knowing $G_{\hat{\boldsymbol{\alpha}}} \in L^1(\mathbb{R}^d)$ allows us to compute

$$\mathcal{L}\left(G_{\hat{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}\right) = \hat{\boldsymbol{\alpha}}^T \boldsymbol{\rho} - \langle G_{\hat{\boldsymbol{\alpha}}} \rangle = \psi(\hat{\boldsymbol{\alpha}}, \boldsymbol{\rho})\,. \tag{4.30}$$

Because $\mathcal{L}$ is strictly convex in first argument, its minimizer is unique. Consequently, (4.29) and (4.30) imply that $\hat{g}_{\boldsymbol{\rho}} = G_{\hat{\boldsymbol{\alpha}}}$, where $\hat{\boldsymbol{\alpha}}$ solves (4.27). In order to satisfy the primary feasibility conditions, $\hat{\boldsymbol{\alpha}} \in \mathcal{A}_{\mathbf{m}}$, and (4.27) becomes

$$\hat{\boldsymbol{\alpha}}^T \left(\boldsymbol{\rho} - \langle \mathbf{m} G_{\hat{\boldsymbol{\alpha}}} \rangle\right) = 0\,. \tag{4.31}$$

Since $\boldsymbol{\rho}_j = \langle \mathbf{m}_j G_{\hat{\boldsymbol{\alpha}}} \rangle$ for $j < N$, the really important part of (4.31) is that

$$\hat{\boldsymbol{\alpha}}_N^T \left(\boldsymbol{\rho}_N - \langle \mathbf{m}_N G_{\hat{\boldsymbol{\alpha}}} \rangle\right) = 0\,. \tag{4.32}$$

The following useful result is an immediate consequence of the complementary slackness condition.

Corollary 23 *Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ and let $\hat{\boldsymbol{\alpha}}$ solve (4.27). Then*

$$h(\boldsymbol{\rho}) = \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) : \hat{\boldsymbol{\alpha}}^T \langle \mathbf{m} g \rangle = \hat{\boldsymbol{\alpha}}^T \boldsymbol{\rho} \right\}\,.$$

A duality theorem similar to Theorem 22 exists for the minimization problem

$$\tilde{h}(\boldsymbol{\rho}) = \inf_{g \in \mathbb{F}_{\mathbf{m}}} \{\mathcal{H}(g) : \langle \mathbf{m} g \rangle = \boldsymbol{\rho}\} \ . \tag{4.33}$$

Theorem 24 *Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, and let $\tilde{h}(\boldsymbol{\rho})$ and ψ be given by (4.33) and (4.23), respectively. Then*

$$\tilde{h}(\boldsymbol{\rho}) = \max_{\boldsymbol{\alpha}} \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \tag{4.34}$$

where the maximum on the right is achieved a unique $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^N$. Furthermore, if the infimum in (4.33) is attained by some function $\tilde{g}_{\boldsymbol{\rho}} \in \mathbb{F}_{\mathbf{m}}$ satisfying the equality constraints, then $\tilde{g}_{\boldsymbol{\rho}}$ minimizes $\mathcal{L}(g, \tilde{\boldsymbol{\alpha}}, \boldsymbol{\rho})$, i.e.,

$$\psi(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\rho}) = \mathcal{L}(\tilde{g}_{\boldsymbol{\rho}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}) \ .$$

The proof of this theorem is analogous to that of Theorem 22. See the appendix for additional comments. As with Theorem 22, we may assume in (4.34) that $\tilde{\boldsymbol{\alpha}} \in \mathcal{A}_{\mathbf{m}}$. Therefore the infimum in (4.4) is the same as in (4.33)—that is,

$$h(\boldsymbol{\rho}) = \tilde{h}(\boldsymbol{\rho}) = \max_{\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}} \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \ ,$$

and $\hat{g}_{\boldsymbol{\rho}} = \tilde{g}_{\boldsymbol{\rho}}$ whenever the latter exists. The equivalence of (4.4) and (4.33) shows that (4.2) has a solution if and only if $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$.

Corollary 25 *The infimum in (4.33) is a minimum if and only if $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$. Thus (4.2) has a solution if and only if $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$.*

Proof. If a minimum $\tilde{g}_{\boldsymbol{\rho}}$ exists, then it is given by $G_{\hat{\boldsymbol{\alpha}}}$ where $\hat{\boldsymbol{\alpha}}$ solves (4.34). Thus $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$. Conversely, if there exists a function $G_{\hat{\boldsymbol{\alpha}}}$ such that $\langle \mathbf{m} G_{\hat{\boldsymbol{\alpha}}} \rangle = \boldsymbol{\rho}$, then

$$\mathcal{H}(G_{\hat{\boldsymbol{\alpha}}}) \leq \mathcal{H}(g) - \boldsymbol{\alpha}^T(\langle \mathbf{m} g \rangle - \langle \mathbf{m} G_{\hat{\boldsymbol{\alpha}}} \rangle) = \mathcal{H}(g)$$

for all functions g in the constraint set of (4.33). Hence $G_{\hat{\boldsymbol{\alpha}}}$ is a minimum of (4.33), i.e., $\tilde{h}(\boldsymbol{\rho}) = \mathcal{H}(G_{\hat{\boldsymbol{\alpha}}})$. ■

4.3 The Relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$

The motivation for studying (4.4) is its application to an evolution equation for $\boldsymbol{\rho}$. It is therefore important to understand the relationship between $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$ as $\boldsymbol{\rho}$ varies over $\mathcal{R}_{\mathbf{m}}$. We should note that a similar analysis to what follows can be found in [43] for the special case when $\mathbf{m}_N = |v|^N$.

4.3.1 Justification of the Formal Legendre Duality

Let $\mathbf{a} : \mathcal{R}_{\mathbf{m}} \to \mathbb{R}^l$ be the function that maps each $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ to the multiplier $\hat{\boldsymbol{\alpha}} \in \mathcal{A}_{\mathbf{m}}$ that solves (4.4). Because $\psi(\cdot, \boldsymbol{\rho})$ is strictly convex on $\mathcal{A}_{\mathbf{m}}$, $\mathbf{a}(\boldsymbol{\rho})$ is uniquely defined for each $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ so that

$$\hat{g}_{\boldsymbol{\rho}} = G_{\mathbf{a}(\boldsymbol{\rho})} \qquad \text{and} \qquad h(\boldsymbol{\rho}) = \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) \,. \tag{4.35}$$

It turns out that $\mathbf{a}$ and the function $\mathbf{r}$ defined in (4.18) are inverses of one another.

Theorem 26 *The function* $\mathbf{r}$ *is one-to-one from* $\mathcal{A}_{\mathbf{m}}$ *onto* $\mathcal{R}_{\mathbf{m}}^{\exp}$ *with inverse* $\mathbf{a}$*, and*

it is a is diffeomorphism between $\operatorname{int} \mathcal{A}_{\mathbf{m}}$ *and* $\operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$.

Proof. We first identify $\mathbf{a}$ as the inverse of $\mathbf{r}$. Since $\mathbf{r}$ is (by definition) onto $\mathcal{R}_{\mathbf{m}}^{\exp}$, we need only to show that $\mathbf{a}(\mathbf{r}(\boldsymbol{\alpha})) = \boldsymbol{\alpha}$ for each $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$; and since the solution of the dual problem is unique, it is sufficient to show that

$$\psi(\boldsymbol{\alpha}, \mathbf{r}(\boldsymbol{\alpha})) = \psi(\mathbf{a}(\mathbf{r}(\boldsymbol{\alpha})), \mathbf{r}(\boldsymbol{\alpha})) \ \ \forall\, \boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}. \tag{4.36}$$

Because $\mathcal{H}$ is convex,

$$\mathcal{H}(G_{\boldsymbol{\alpha}}) \geq \mathcal{H}(G_{\boldsymbol{\alpha}_*}) + \langle \log(G_{\boldsymbol{\alpha}_*})\,(G_{\boldsymbol{\alpha}} - G_{\boldsymbol{\alpha}_*}) \rangle \quad \forall\, \boldsymbol{\alpha}, \boldsymbol{\alpha}_* \in \mathcal{A}_{\mathbf{m}}\,,$$

which, from the definitions of $\mathcal{H}$ and $G_{\boldsymbol{\alpha}}$, gives

$$\psi(\boldsymbol{\alpha}, \mathbf{r}(\boldsymbol{\alpha})) \geq \psi(\boldsymbol{\alpha}_*, \mathbf{r}(\boldsymbol{\alpha})) \ \ \forall\, \boldsymbol{\alpha}, \boldsymbol{\alpha}_* \in \mathcal{A}_{\mathbf{m}}\,.$$

This proves (4.36).

We now show that $\mathbf{r}$ is a diffeomorphism. For $\boldsymbol{\alpha} \in \operatorname{int} \mathcal{A}_{\mathbf{m}}$, $\mathbf{r}$ is the derivative of the density potential h^*:

$$\mathbf{r}(\boldsymbol{\alpha}) = \frac{\partial h^*}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha})\,.$$

Therefore $\mathbf{r}$ inherits smoothness properties from h (see Theorem 20). In particular, its Jacobian is

$$\frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \frac{\partial^2 h}{\partial^2 \boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\rho}) = \left\langle \mathbf{m}^T \mathbf{m} G_{\boldsymbol{\alpha}} \right\rangle\,,$$

is a positive-definite matrix. Thus, by the inverse function theorem, $\mathbf{r}$ is a diffeomorphism from $\operatorname{int}\mathcal{A}_{\mathbf{m}}$ onto $\operatorname{int}\mathcal{R}^{\exp}_{\mathbf{m}}$. This means that $\mathbf{a} = \mathbf{r}^{-1}$ is smooth on $\operatorname{int}\mathcal{R}^{\exp}_{\mathbf{m}}$ with Jacobian

$$\frac{\partial \mathbf{a}}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}) = \left[\frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}}\left(\mathbf{a}(\boldsymbol{\rho})\right)\right]^{-1}.$$

■

Remark 27 *It should be noted that if Conditions 13 and 21 hold, then* $\mathbf{r}$ *is a smooth diffeomorphism when restricted to any manifold in the stratification of* $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$.

This theorem rigorously establishes the formal Legendre duality relations used in Sections 4 and 5 of Chapter 3 when $\mathcal{A}_{\mathbf{m}} = \operatorname{int}\mathcal{A}_{\mathbf{m}}$ and $\mathcal{R}_{\mathbf{m}} = \mathcal{R}^{\exp}_{\mathbf{m}}$. From the definitions of h, h^*, and ψ, we deduce that

$$h(\boldsymbol{\rho}) + h^*(\mathbf{a}(\boldsymbol{\rho})) = \mathbf{a}(\boldsymbol{\rho})^T\boldsymbol{\rho}\,, \tag{4.37}$$

where

$$\left[\frac{\partial h^*}{\partial \boldsymbol{\alpha}}\left(\mathbf{a}(\boldsymbol{\rho})\right)\right]^T = \mathbf{r}\left(\mathbf{a}(\boldsymbol{\rho})\right) = \boldsymbol{\rho}\,, \qquad \boldsymbol{\rho} \in \mathcal{R}^{\exp}_{\mathbf{m}}\,. \tag{4.38}$$

Moreover, differentiating (4.38) with respect to $\boldsymbol{\rho}$ gives

$$\left[\frac{\partial h}{\partial \boldsymbol{\rho}}\left(\mathbf{r}(\boldsymbol{\alpha})\right)\right]^T = \mathbf{a}\left(\mathbf{r}(\boldsymbol{\alpha})\right) = \boldsymbol{\alpha}\,, \qquad \boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \tag{4.39}$$

(Note that the calculation in (4.39) requires that $\mathbf{a}$ be differentiable). Finally,

$$\frac{\partial^2 h}{\partial \boldsymbol{\rho}^2}(\boldsymbol{\rho}) = \frac{\partial^2 h^*}{\partial^2 \boldsymbol{\alpha}}\left(\mathbf{a}(\boldsymbol{\rho})\right)$$

is positive-definite so that h is strictly convex. Recall from Chapter 3 that strict convexity of h is required in order for entropy based closures to be symmetric hyperbolic systems.

4.3.2 Examples

For $N = 2$, $\mathcal{A}_{\mathbf{m}} = \operatorname{int} \mathcal{A}_{\mathbf{m}}$ and $\mathcal{R}_{\mathbf{m}} = \mathcal{R}_{\mathbf{m}}^{\exp}$. We recall specifically the following cases from Chapter 3.

1. **Maxwellian closure**. If $\mathbf{m} = (1, v, \frac{1}{2}|v|^2)^T$, the minimizer of (4.2) is a Maxwellian distribution:

$$\mathcal{M}_{n,u,\theta}(v) = \frac{n}{(2\pi\theta)^{3/2}} \exp\left(-\frac{|v-u|^2}{2\theta}\right),$$

where the fluid variables (n, u, θ) are related to the densities $\boldsymbol{\rho}_i$ by

$$\boldsymbol{\rho}_0 = n\,, \qquad \boldsymbol{\rho}_1 = nu\,, \qquad \boldsymbol{\rho}_2 = \frac{1}{2}nu^2 + \frac{3}{2}n\theta$$

and to the Lagrange multipliers $\hat{\boldsymbol{\alpha}}_i$ by

$$\hat{\boldsymbol{\alpha}}_0 = \log\left(\frac{n}{(2\pi\theta)^{3/2}}\right) - \frac{|u|^2}{2\theta}\,, \qquad \hat{\boldsymbol{\alpha}}_1 = \frac{u}{\theta}\,, \qquad \hat{\boldsymbol{\alpha}}_2 = -\frac{1}{\theta}\,.$$

2. **Gaussian closure**. If $\mathbf{m} = (1, v, v \vee v)^T$, the minimizer of (4.2) is a Gaussian

distribution:

$$\mathcal{G}_{n,u,\Theta}(v) = \frac{n}{\sqrt{\det(2\pi\Theta)}} \exp\left(-\frac{1}{2}(v-u)\cdot\Theta^{-1}\cdot(v-u)\right) ,$$

where the fluid variables (n, u, Θ) are related to the densities $\boldsymbol{\rho}_i$ by

$$\boldsymbol{\rho}_0 = n\,, \qquad \boldsymbol{\rho}_1 = nu\,, \qquad \boldsymbol{\rho}_2 = nu \vee u + n\Theta$$

and to the Lagrange multipliers $\hat{\boldsymbol{\alpha}}_i$ by

$$\hat{\boldsymbol{\alpha}}_0 = \log\left(\frac{n}{\sqrt{\det(2\pi\Theta)}}\right) - \frac{1}{2}u\cdot\Theta^{-1}\cdot u\,, \qquad \hat{\boldsymbol{\alpha}}_1 = \Theta^{-1}\cdot u\,, \qquad \hat{\boldsymbol{\alpha}}_2 = -\frac{1}{2}\Theta^{-1}\,.$$

In both of these examples, the expressions for $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\rho}$ can be used to find $\mathbf{a}(\boldsymbol{\rho})$ explicitly.

4.3.3 Degenerate Densities

If $\mathcal{R}_{\mathbf{m}}\backslash\mathcal{R}_{\mathbf{m}}^{\exp}$ is non-empty, then there are densities $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ such that $\boldsymbol{\rho} \neq \mathbf{r}(\boldsymbol{\alpha})$ for *any* $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$. In such cases (4.2) has no solution, and the Legendre duality between h and h^* is no longer valid. In particular, h is no longer strictly convex, $\mathbf{r}\left(\mathbf{a}(\boldsymbol{\rho})\right) \neq \boldsymbol{\rho}$, and (4.38) no longer holds. We call such densities *dengenerate.* Unfortunately, it turns out in most cases that $\mathcal{R}_{\mathbf{m}}\backslash\mathcal{R}_{\mathbf{m}}^{\exp}$ is non-empty.

Theorem 28 *The set $\mathcal{R}_{\mathbf{m}}\backslash\mathcal{R}_{\mathbf{m}}^{\exp}$ is empty if and only if $\mathcal{A}_{\mathbf{m}}$ is open.*

Proof. Suppose that $\mathcal{A}_{\mathbf{m}}$ is open. Then for each $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, $\psi(\cdot, \boldsymbol{\rho})$ is smooth on all of $\mathcal{A}_{\mathbf{m}}$. First order optimality conditions for the dual problem imply that

$$\frac{\partial \psi}{\partial \boldsymbol{\alpha}} \left(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}\right) = 0 \implies \boldsymbol{\rho} = \left\langle \mathbf{m} G_{\mathbf{a}(\boldsymbol{\rho})} \right\rangle \ .$$

Therefore $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$.

Now suppose that $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$ is non-empty and $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$. Choose a nonzero element $\boldsymbol{\sigma} \succeq^* 0$ such that $\boldsymbol{\alpha}^T \boldsymbol{\sigma} = 0$. Then $\boldsymbol{\alpha}^T \boldsymbol{\sigma} \leq 0$ for any $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}}$ and Theorem 12 implies that $\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma} \in \mathcal{R}_{\mathbf{m}}$. Therefore

$$\begin{aligned}
\psi(\boldsymbol{\alpha}, \mathbf{r}(\boldsymbol{\alpha})) &= \psi(\boldsymbol{\alpha}, \mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma}) \\
&\leq \psi(\mathbf{a}\left(\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma}\right), \mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma}) \\
&\leq \psi(\mathbf{a}\left(\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma}\right), \mathbf{r}(\boldsymbol{\alpha})) \\
&\leq \psi(\boldsymbol{\alpha}, \mathbf{r}(\boldsymbol{\alpha}))
\end{aligned}$$

so that

$$\psi(\boldsymbol{\alpha}, \mathbf{r}(\boldsymbol{\alpha})) = \psi(\mathbf{a}\left(\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma}\right), \mathbf{r}(\boldsymbol{\alpha})).$$

Uniqueness of the dual solution implies that $\mathbf{a}\left(\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma}\right) = \boldsymbol{\alpha}$. *If* $\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma} \in \mathcal{R}_{\mathbf{m}}^{\exp}$, then

$$\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma} = \mathbf{r}(\boldsymbol{\alpha}) \ ,$$

which contradicts the fact that $\boldsymbol{\sigma}$ is nonzero. Thus $\mathbf{r}(\boldsymbol{\alpha}) + \boldsymbol{\sigma} \in \mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$. ■

For $N > 2$, $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ is non-empty and thus Theorem 28 shows that the well-posed examples of the last subsection are the exception rather than the rule. However, in spite of the difficulties encountered for $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$, (4.37) still holds for all $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, and the following theorem states that (4.39) does as well.

Theorem 29 *The function h has a continuous Fréchet derivative everywhere on $\mathcal{R}_{\mathbf{m}}$ that is given by*

$$\frac{\partial h}{\partial \boldsymbol{\rho}}(\boldsymbol{\rho}) = \mathbf{a}(\boldsymbol{\rho})\,.$$

Proof. Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$. Because h minimizes the dual function ψ,

$$\begin{aligned} h(\boldsymbol{\rho}+\boldsymbol{\delta}) &= \psi\left(\mathbf{a}(\boldsymbol{\rho}+\boldsymbol{\delta}), \boldsymbol{\rho}+\boldsymbol{\delta}\right) \\ &\geq \psi\left(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}+\boldsymbol{\delta}\right) \\ &= \psi\left(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}\right) + \mathbf{a}(\boldsymbol{\rho})^T\boldsymbol{\delta} \\ &= h(\boldsymbol{\rho}) + \mathbf{a}(\boldsymbol{\rho})^T\boldsymbol{\delta} \end{aligned} \tag{4.40}$$

and, similarly,

$$h(\boldsymbol{\rho}+\boldsymbol{\delta}) \leq h(\boldsymbol{\rho}) + \mathbf{a}(\boldsymbol{\rho}+\boldsymbol{\delta})^T\boldsymbol{\delta}\,. \tag{4.41}$$

Together (4.40) and (4.41) imply that

$$\frac{\left|h(\boldsymbol{\rho}+\boldsymbol{\delta}) - h(\boldsymbol{\rho}) - \mathbf{a}(\boldsymbol{\rho})^T\boldsymbol{\delta}\right|}{\boldsymbol{\delta}} \leq \left|\mathbf{a}(\boldsymbol{\rho}+\boldsymbol{\delta}) - \mathbf{a}(\boldsymbol{\rho})\right|\,.$$

Thus we need to show that $\mathbf{a}$ is continuous.

Equation (4.40) implies also that $\mathbf{a}(\boldsymbol{\rho})$ is a *subgradient* of h at $\boldsymbol{\rho}$ [11]. The set of all subgradients is called the *subdifferential* of h at $\boldsymbol{\rho}$ and is denoted by $\partial h(\boldsymbol{\rho})$. It is a general result from convex analysis [71] that the set

$$\partial h(S) \equiv \{\partial h(\boldsymbol{\rho}) : \boldsymbol{\rho} \in S\}$$

is bounded whenever $S \subset \mathbb{R}^l$ is bounded. In particular, if $\left\{\boldsymbol{\rho}_j\right\}_{j=1}^{\infty} \subset \mathcal{R}_{\mathbf{m}}$ converges to $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$, then $\{\mathbf{a}(\boldsymbol{\rho}_j)\}_{j=1}^{\infty}$ is a bounded sequence. Let $\boldsymbol{\alpha}_*$ be any subsequential limit for this sequence. Then

$$\psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) = \lim_{i\to\infty} \psi\left(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}_{j_i}\right) \leq \lim_{i\to\infty} \psi\left(\mathbf{a}(\boldsymbol{\rho}_{j_i}), \boldsymbol{\rho}_{j_i}\right) \leq \psi(\boldsymbol{\alpha}^*, \boldsymbol{\rho}) \leq \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) \,, \tag{4.42}$$

where $\{j_i\}_{i=1}^{\infty}$ is any sequence of integers such that $\boldsymbol{\alpha}_* = \lim_{i\to\infty} \mathbf{a}(\boldsymbol{\rho}_{j_i})$. Note that the first and last inequalities in (4.42) follow because $\psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho})$ maximizes $\psi(\cdot, \boldsymbol{\rho})$, whereas the middle inequality is a consequence of Fatou's Lemma.

From (4.42), we deduce that

$$\psi(\boldsymbol{\alpha}_*, \boldsymbol{\rho}) = \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) \,,$$

and since $\mathbf{a}(\boldsymbol{\rho})$ is the *unique* minimizer of $\psi(\cdot, \boldsymbol{\rho})$, it follows that $\boldsymbol{\alpha}_* = \mathbf{a}(\boldsymbol{\rho})$. Finally, because this result holds for any subsequential limit of $\{\mathbf{a}(\boldsymbol{\rho}_j)\}_{j=1}^{\infty}$, $\mathbf{a}$ is continuous and thus h is continuously differentiable. ■

One important facet of this result in that $h(\boldsymbol{\rho}) = \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho})$ is a differentiable function of $\boldsymbol{\rho}$ even though $\psi(\cdot, \boldsymbol{\rho})$ may not be continuous for $\boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$.

4.3.4 Geometry of $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$

Even if $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ is non-empty, there is evidence to suggest that the dynamics of entropy closure is such that vectors in this set might never be attained—that is, if $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$ initially, then it will continue to be so for all later times. Consider, for example, the following.

Proposition 30 *Suppose that the function*

$$\chi(\boldsymbol{\rho}) \equiv \left\langle |v\mathbf{m}|\, G_{\mathbf{a}(\boldsymbol{\rho})} \right\rangle \tag{4.43}$$

is bounded on an open set O containing $\boldsymbol{\rho}_* \in \mathcal{R}_{\mathbf{m}}$. Then $\mathbf{r}$ is continuous at $\boldsymbol{\rho}_*$.

Proof. Let $\{\boldsymbol{\rho}_j\}_{j=1}^{\infty} \subset O$ be any sequence such that $\boldsymbol{\rho}_j \to \boldsymbol{\rho}$. Then for any constant $R > 0$,

$$0 < \left| \int_{|v|>R} \mathbf{m} G_{\mathbf{a}(\boldsymbol{\rho})}\, dv \right| \leq \frac{1}{R} \int_{\mathbb{R}^d} |v\mathbf{m}|\, G_{\mathbf{a}(\boldsymbol{\rho})}\, dv \to 0 \text{ as } \; R \to \infty\,. \tag{4.44}$$

Since $\mathbf{a}$ is continuous, $\mathbf{a}(\boldsymbol{\rho}_j) \to \mathbf{a}(\boldsymbol{\rho})$ and the sequence $G_{\mathbf{a}(\boldsymbol{\rho}_j)}$ is uniformly bounded on $\{v \in \mathbb{R}^d : |v| \leq R\}$. Hence the Lebesgue Bounded Convergence Theorem implies that

$$\lim_{j\to\infty} \int_{|v|<R} \mathbf{m} G_{\mathbf{a}(\boldsymbol{\rho}_j)}\, dv = \int_{|v|<R} \mathbf{m} G_{\mathbf{a}(\boldsymbol{\rho})}\, dv\,. \tag{4.45}$$

Together (4.44) and (4.45) give the result. ■

In light of Proposition 30, let us suppose that $\boldsymbol{\rho}_* \in \mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$, $\{\boldsymbol{\rho}_j\}_{j=1}^{\infty} \subset \mathcal{R}_{\mathbf{m}}^{\exp}$, and $\boldsymbol{\rho}_j \to \boldsymbol{\rho}_*$. Then $\mathbf{a}(\boldsymbol{\rho}_j) \to \mathbf{a}(\boldsymbol{\rho}_*)$, and *if* $\mathbf{r}$ is continuous, then

$$\boldsymbol{\rho}_* = \lim_{j \to \infty} \boldsymbol{\rho}_j = \lim_{j \to \infty} \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_j)) = \mathbf{r}(\mathbf{a}(\boldsymbol{\rho}_*)) \,,$$

which contradicts the fact that $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$. We conclude that $\mathbf{r}$ is not continuous, whereby the function χ given in (4.43) cannot be bounded near $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$. Such behavior was first observed for the one dimensional example in [42]. In particular it was found that $\left\langle v \mathbf{m}_N G_{\mathbf{a}(\boldsymbol{\rho})} \right\rangle$ diverged to positive or negative infinity as $\boldsymbol{\rho}_j \to \boldsymbol{\rho}_* \in \mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$, depending on the direction of approach. Note that $\left\langle v \mathbf{m}_N G_{\mathbf{a}(\boldsymbol{\rho})} \right\rangle$ is the flux associated with the moment $\boldsymbol{\rho}_N$ in the entropy based moment closure, and as pointed out in [42] the divergent behavior of this flux raises the possibility that $\operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$ is invariant under the dynamics of the closure.

Now suppose it can be proven that vectors $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ will never be attained during the dynamics of an entropy closure. Then if $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}^{\exp}$ initially, (4.2) will always have a solution and the formal properties of the closure based on the Legendre duality between h and h^* will be maintained. However, in order for such closures to be physically relevant, it must be shown—at a minimum—that $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ is small is some sense. This is our current objective.

Define the projection $\boldsymbol{\pi} : \mathcal{R}_{\mathbf{m}} \to \mathcal{R}_{\mathbf{m}}^{\exp}$ by

$$\boldsymbol{\pi}(\boldsymbol{\rho}) \equiv \mathbf{r}(\mathbf{a}(\boldsymbol{\rho})) \,.$$

Then $\boldsymbol{\pi}(\boldsymbol{\rho})$ is the spatial density that is realized by the minimizer of (4.4). Theorem 26 implies that

$$(i)\, \boldsymbol{\pi} \text{ is the identity on } \mathcal{R}_{\mathbf{m}}^{\exp}\,; \tag{4.46a}$$

$$(ii)\, \boldsymbol{\pi}(\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}) = \mathbf{r}\left(\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}\right) = \mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial \mathcal{R}_{\mathbf{m}}^{\exp}\,; \tag{4.46b}$$

$$(iii)\, \mathbf{a}(\boldsymbol{\pi}(\boldsymbol{\rho})) = \mathbf{a}(\boldsymbol{\rho}))\,. \tag{4.46c}$$

Like $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$, the functions $\mathbf{r}$, $\mathbf{a}$, and $\boldsymbol{\pi}$ all have a natural decomposition based on the decomposition of $\mathbf{m}$ in (4.10):

$$\mathbf{r} = (\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_N)^T\,, \qquad \mathbf{a} = (\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_N)^T\,, \qquad \boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N)^T\,,$$

so that

$$\begin{aligned} G_{\mathbf{a}(\boldsymbol{\rho})} &= \exp\left(\sum_{j=1}^{N} \mathbf{a}_j(\boldsymbol{\rho})^T \mathbf{m}_j\right)\,, \\ \mathbf{r}_j(\boldsymbol{\alpha}) &= \langle \mathbf{m}_j G_{\boldsymbol{\alpha}} \rangle\,, \\ \boldsymbol{\pi}_j(\boldsymbol{\rho}) &= \mathbf{r}_j(\mathbf{a}(\boldsymbol{\rho}))\,. \end{aligned}$$

With this decomposition, $\boldsymbol{\pi}_j(\boldsymbol{\rho}) = \boldsymbol{\rho}_j$ for all $j < N$. Roughly speaking, the following theorem says that $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ is constructed by attaching a cone to each point in $\mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial \mathcal{R}_{\mathbf{m}}^{\exp}$.

Theorem 31 *The vector $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}} \backslash \operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$ if and only if $\boldsymbol{\rho}_j = \bar{\boldsymbol{\rho}}_j$ for all $j < N$ and*

$\boldsymbol{\rho}_N \in \bar{\boldsymbol{\rho}}_N + \mathcal{NC}(-A_N, \mathbf{a}_N(\bar{\boldsymbol{\rho}}))$ *for some vector* $\bar{\boldsymbol{\rho}} \in \mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$.

Proof. Begin with the "only if" part. Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}} \backslash \operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$ and set $\bar{\boldsymbol{\rho}} = \boldsymbol{\pi}(\boldsymbol{\rho})$. According to (4.46), $\bar{\boldsymbol{\rho}} \in \mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$ and $\mathbf{a}(\boldsymbol{\rho}) = \mathbf{a}(\bar{\boldsymbol{\rho}})$. The complementary slackness condition (4.32) implies that

$$\mathbf{a}_N(\bar{\boldsymbol{\rho}})^T \boldsymbol{\rho}_N = \mathbf{a}_N(\bar{\boldsymbol{\rho}})^T \bar{\boldsymbol{\rho}}_N \,,$$

where $\mathbf{a}_N(\bar{\boldsymbol{\rho}})^T \in \partial(-A_N)$; and the feasibility condition in (4.4) implies that

$$\boldsymbol{\alpha}_N^T \boldsymbol{\rho}_N \leq \boldsymbol{\alpha}_N^T \bar{\boldsymbol{\rho}}_N$$

for all $\boldsymbol{\alpha}_N \in -A_N$. Therefore,

$$\left(\boldsymbol{\alpha}_N - \mathbf{a}_N(\bar{\boldsymbol{\rho}})\right)^T \left(\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N\right) \leq 0$$

for all $\boldsymbol{\alpha}_N \in -A_N$, which means that $\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N$ is in the normal cone of $-A_N$ at $\mathbf{a}_N(\bar{\boldsymbol{\rho}})$:

$$\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N \in \mathcal{NC}(-A_N, \alpha).$$

Now prove the "if" part. Suppose that there exists $\boldsymbol{\rho} \in \mathbb{R}^l$ and $\bar{\boldsymbol{\rho}} \in \mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$ such that $\boldsymbol{\rho}_j = \bar{\boldsymbol{\rho}}_j$ for all $j < N$ and $\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N \in \mathcal{NC}(-A_N, \mathbf{a}_N(\bar{\boldsymbol{\rho}}))$. Then

$$\left(\boldsymbol{\alpha}_N - \mathbf{a}_N(\bar{\boldsymbol{\rho}})\right)^T \left(\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N\right) \leq 0 \;\; \forall\, \boldsymbol{\alpha}_N \in -A_N \,. \tag{4.47}$$

Setting $\boldsymbol{\alpha}_N = 0$ and then $\boldsymbol{\alpha}_N = 2\mathbf{a}_N(\bar{\boldsymbol{\rho}})$ in (4.47) gives

$$\mathbf{a}_N(\bar{\boldsymbol{\rho}})^T (\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N) \geq 0 \qquad \text{and} \qquad \mathbf{a}_N(\bar{\boldsymbol{\rho}})^T (\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N) \leq 0\,,$$

respectively. Therefore

$$\mathbf{a}_N(\bar{\boldsymbol{\rho}})^T (\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N) = 0\,. \tag{4.48}$$

Next, setting $\boldsymbol{\alpha}_N = \mathbf{a}_N(\bar{\boldsymbol{\rho}}) + \mathbf{a}_N(\boldsymbol{\rho})$ in (4.47) gives

$$\mathbf{a}_N(\boldsymbol{\rho})^T (\boldsymbol{\rho}_N - \bar{\boldsymbol{\rho}}_N) \leq 0\,. \tag{4.49}$$

Since $\boldsymbol{\rho}_j = \bar{\boldsymbol{\rho}}_j$ for all $j < N$, (4.48) and (4.49) imply that

$$\psi(\mathbf{a}(\bar{\boldsymbol{\rho}}), \bar{\boldsymbol{\rho}}) \overset{(4.48)}{=} \psi(\mathbf{a}(\bar{\boldsymbol{\rho}}), \boldsymbol{\rho}) \leq \psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho}) \overset{(4.49)}{\leq} \psi(\mathbf{a}(\boldsymbol{\rho}), \bar{\boldsymbol{\rho}}) \leq \psi(\mathbf{a}(\bar{\boldsymbol{\rho}}), \bar{\boldsymbol{\rho}})\,. \tag{4.50}$$

Here, the first and last inequalities follow because $\psi(\mathbf{a}(\boldsymbol{\rho}), \boldsymbol{\rho})$ maximizes $\psi(\cdot, \boldsymbol{\rho})$. Furthermore, since this maximizer is unique, (4.50) shows that $\mathbf{a}(\bar{\boldsymbol{\rho}}) = \mathbf{a}(\boldsymbol{\rho})$, which means that either $\boldsymbol{\rho} = \bar{\boldsymbol{\rho}} \in \mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$ or $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$. In either case, the claim is proven.

■

Theorem 31 provides a nice description of the degenerate values of $\boldsymbol{\rho}$ associated with each $\bar{\boldsymbol{\rho}} \in \mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$. However, a clean description of $\mathcal{R}_{\mathbf{m}} \backslash \operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$ requires also that $\mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$ itself possess some nice structure, and the purpose of Conditions 13 and 21 it to ensure that this is the case. Condition 13 ensures that the set $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ is a "nice" in a well-defined sense while Condition 21 ensures its image

under $\mathbf{r}$, $\mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$, is also nice.

A (continuous) fiber bundle [36] $\mathcal{B}$ consists of topological spaces B, called the base space, and F, called a fiber space, along with a projection $P : \mathcal{B} \to B$ such that $P^{-1}(B)$ is locally homeomorphic to the cross product of B and F. Roughly speaking, $\mathcal{B}$ is constructed by attached a (topological equivalent) copy of F to each point in B.

Theorem 32 *Suppose Conditions 13 and 21 hold. Then* $\mathcal{R}_{\mathbf{m}} \setminus \operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$ *is the finite union of smooth bundles of codimension one or greater in* $\mathbb{R}^l$. *The base space of each bundle is a smooth manifold in* $\mathcal{R}_{\mathbf{m}}^{\exp} \cap \partial\mathcal{R}_{\mathbf{m}}^{\exp}$. *The fiber attached to each point* $\bar{\boldsymbol{\rho}}$ *in the base space is the cone* $\mathcal{NC}(-A_N, \mathbf{a}_N(\bar{\boldsymbol{\rho}}))$.

Before proving this theorem, we need to address a technical point. Let $\mathcal{S}$ be a stratification of $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ and $\mathcal{T}$ be a stratification of $\partial(-A_N)$, the latter of which was proven to exist in Lemma 16. The projection $\boldsymbol{\alpha} \mapsto \boldsymbol{\alpha}_N$ of an element $s \in \mathcal{S}$ onto $\mathbb{R}^{l_N}$, which we denote by s_N, is a subset of $\partial(-A_N)$. Since each stratification is a *finite* union of smooth manifolds, s_N can be further decomposed, if necessary, into a finite union of smooth manifolds, each of which is a subset of a unique element of $\mathcal{T}$. We summarize with following lemma.

Lemma 33 *Let* $\mathcal{S}$ *be a stratification of* $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ *and* $\mathcal{T}$ *be a stratification of* $\partial(-A_N)$, *and suppose that Condition 13 holds. Then, without loss of generality, we may assume that the projection* $\boldsymbol{\alpha} \mapsto \boldsymbol{\alpha}_N$ *applied to any element of* $\mathcal{S}$ *is a subset of an element of* $\mathcal{T}$.

For simplicity of exposition, we maintain the assumption given in the proceeding lemma for the proof of Theorem 32.

Proof of Theorem 32. Let $\mathcal{S}$ be a stratification of $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ and $\mathcal{T}$ be a stratification of $\partial(-A_N)$. Let $s \in \mathcal{S}$ and $s_N \subset t \in \mathcal{T}$. Then, for any $\boldsymbol{\alpha} \in s$, $\boldsymbol{\alpha}_N \in t$ and

$$\dim(\mathcal{N}(-A_N, \boldsymbol{\alpha}_N)) = l_N - \dim(t) \leq l_N - \dim(s_N)\,. \tag{4.51}$$

According to Condition 13,

$$l_N - \dim(s_N) < l - \dim(s)\,, \tag{4.52}$$

while by Condition 21, $\mathbf{r}$ is diffeomorphic when restricted to s so that set $\mathbf{r}(s)$ is also a smooth manifold with the same dimension as s. Attached to each point $\mathbf{r}(\boldsymbol{\alpha}) \in \mathbf{r}(s)$ is the cone $\mathcal{N}(-A_N, \boldsymbol{\alpha}_N)$. The complete structure is a fiber bundle with base $\mathbf{r}(s)$ and fibers $\mathcal{N}(-A_N, \boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in s$. If we denote the bundle by $\mathcal{B}(s)$, then (4.51) and (4.52) imply that

$$\dim(\mathcal{B}(s)) = \dim(s) + \dim(\mathcal{N}(-A_N, \boldsymbol{\alpha}_N)) \leq \dim(s) + (l_N - \dim(s_N)) < l\,.$$

Finally $\mathcal{R}_{\mathbf{m}} \backslash \operatorname{int} \mathcal{R}_{\mathbf{m}}^{\exp}$ is the union of all sets $\mathcal{B}$ that are generated by strata in $\mathcal{S}$, of which it is assumed there are a finite number. ■

As a consequence of this theorem, we see that if Conditions 13 and 21 hold, then $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ is a set with no interior and zero measure. Determining cases for which these (or appropriate similar) conditions hold is therefore an important open question.

4.3.5 Examples

We will assume that Conditions 13 and 21 hold in the following examples.

4.3.5.1 Junk's example The example $m_N = |v|^N$ has been studied in [42, 43] and [73], particularly when $N = 4$. For general N,

$$A_N = \{\boldsymbol{\alpha}_N \in \mathbb{R} : \boldsymbol{\alpha}_N \geq 0\} \qquad \text{and} \qquad \partial(-A_N) = \{0\}\,.$$

If $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ and $\mathbf{a}_N(\rho) = 0$, then $\mathbf{a}_{N-1}(\boldsymbol{\rho}) = 0$ as well; otherwise, $G_{\mathbf{a}(\rho)} \notin \mathbb{F}_{\mathbf{m}}$. With this fact in mind, we recall from Corollary 23 that

$$\mathcal{H}(G_{\mathbf{a}(\boldsymbol{\rho})}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) : \mathbf{a}(\boldsymbol{\rho})^T \langle \mathbf{m} g \rangle = \mathbf{a}(\boldsymbol{\rho})^T \boldsymbol{\rho} \right\} .$$

Therefore $G_{\mathbf{a}(\rho)}$ is the actually the minimizer of $\mathcal{H}$ subject to fewer constraints:

$$\mathcal{H}(G_{\mathbf{a}(\rho)}) = \min_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) : \langle \mathbf{m}_j g \rangle = \boldsymbol{\rho}_j \,,\ j \leq N-2 \right\}. \tag{4.53}$$

Let $\bar{\mathbf{m}}$ contain the components of $\mathbf{m}$ of degree $\bar{N} \equiv N-2$ and less:

$$\bar{\mathbf{m}} = (\mathbf{m}_0, \mathbf{m}_1, \ldots, \mathbf{m}_{N-2}) \,,$$

and let the variables $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\rho}}$ and the functions $\bar{\mathbf{r}}$, $\bar{\mathbf{a}}$, and $\bar{\boldsymbol{\pi}}$ be defined similarly. For this example,

$$\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}} \subset \mathcal{A}_{\bar{\mathbf{m}}} \times \{\boldsymbol{\alpha}_{N-1} = 0\} \times \{\boldsymbol{\alpha}_N = 0\} \ , \tag{4.54}$$

but these two sets are not necessarily equal, since that latter may include $\boldsymbol{\alpha}$ for which $G_{\boldsymbol{\alpha}} \in \mathbb{F}_{\bar{\mathbf{m}}}$, but $G_{\boldsymbol{\alpha}} \notin \mathbb{F}_{\mathbf{m}}$. However, one may readily conclude that $G_{\boldsymbol{\alpha}} \in \mathbb{F}_{\mathbf{m}}$ for all $\bar{\boldsymbol{\alpha}} \in \operatorname{int} \mathcal{A}_{\bar{\mathbf{m}}}$. Hence,

$$\operatorname{int} \mathcal{A}_{\bar{\mathbf{m}}} \times \{\boldsymbol{\alpha}_{N-1} = 0\} \times \{\boldsymbol{\alpha}_N = 0\} \subset \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}} \,.$$

Let $\mathcal{S}$ be a stratification of $\mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}$. The projection of any $s \in \mathcal{S}$ onto $\partial\,(-A_N)$ is the point $\{\boldsymbol{\alpha}_N = 0\}$, so the normal cone attached to $\boldsymbol{\alpha} \in s$ is just a ray:

$$\mathcal{NC}(-A_N, \alpha_N) = (-A_N)^* = A_N = \{\boldsymbol{\alpha}_N : \boldsymbol{\alpha}_N \geq 0\} \,.$$

Therefore

$$\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp} = \{\boldsymbol{\rho} : \boldsymbol{\rho}_N > \mathbf{r}_N(\boldsymbol{\alpha}),\ \boldsymbol{\alpha} \in \mathcal{A}_{\mathbf{m}} \cap \partial \mathcal{A}_{\mathbf{m}}\} \,, \tag{4.55}$$

Because A_N is one-dimensional, the inequality in (4.55) is scalar.

If $N = 4$, the situation simplifies further, because $\operatorname{int} \mathcal{A}_{\bar{\mathbf{m}}} = \mathcal{A}_{\bar{\mathbf{m}}}$ and the inclusion in (4.54) becomes an equality. In addition, $\mathcal{R}_{\bar{\mathbf{m}}} = \mathcal{R}_{\bar{\mathbf{m}}}^{\exp}$ and $\bar{\mathbf{r}}$ is a diffeomorphism on all of $\mathcal{A}_{\bar{\mathbf{m}}}$. Therefore

$$\begin{aligned} \mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp} &= \{\boldsymbol{\rho} : \boldsymbol{\rho}_N > \mathbf{r}_N(\boldsymbol{\alpha}), \qquad \bar{\boldsymbol{\alpha}} \in \mathcal{A}_{\mathbf{m}},\ \boldsymbol{\alpha}_N = \boldsymbol{\alpha}_{N-1} = 0\} \,. \\ &= \left\{\boldsymbol{\rho} : \boldsymbol{\rho}_N > \boldsymbol{\pi}_N(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})),\ \boldsymbol{\rho}_{N-1} = \boldsymbol{\pi}_{N-1}(0, 0, \bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})),\ \bar{\boldsymbol{\rho}} \in \mathcal{R}_{\bar{\mathbf{m}}}\right\} \,. \end{aligned}$$

The components $\boldsymbol{\pi}_N(0,0,\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}}))$ and $\boldsymbol{\pi}_{N-1}(0,0,\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}}))$ are simple to compute since

$$\boldsymbol{\pi}(0,0,\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})) = \left\langle \mathbf{m} G_{\bar{\mathbf{a}}(\bar{\rho})} \right\rangle ,$$

and $\bar{\mathbf{a}}(\bar{\rho})$ has an explicit formula when $\bar{N} = 2$. (See the examples in Section 4.3.2.)

4.3.5.2 A Non-Junkian Example The situation becomes more complicated when $\mathbf{m}_N$ includes polynomials other than $|v|^N$ because the inequality constraints in (4.4) are no longer scalar. The simplest example of this type occurs when

$$\mathbf{m}_N = (v \vee v)\,|v|^{N-2}\,.$$

We examine in detail the two-dimensional case ($d = 2$) and write $\boldsymbol{\alpha}_N$ in the form of a symmetric matrix:

$$\boldsymbol{\alpha}_N = \begin{pmatrix} (\boldsymbol{\alpha}_N)_{11} & (\boldsymbol{\alpha}_N)_{12} \\ (\boldsymbol{\alpha}_N)_{21} & (\boldsymbol{\alpha}_N)_{22} \end{pmatrix} = \begin{pmatrix} a+b & c \\ c & a-b \end{pmatrix} . \tag{4.56}$$

With respect to the (a,b,c) coordinates, the set A_N is a cone in $\mathbb{R}^3$ that can be found in a high school geometry text:

$$A_N = \left\{ (a,b,c) \in \mathbb{R}^3 : a^2 \geq \sqrt{b^2+c^2} \right\} ,$$

and the boundary of $-A_N$ is

$$\partial(-A_N) = \left\{(a,b,c) \in \mathbb{R}^3 : a = -\sqrt{b^2+c^2}\right\}. \tag{4.57}$$

Let $\mathcal{S}$ be the stratification of $\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}}$ and let $s \in \mathcal{S}$ so that $s_N \in \partial(-A_N)$. The set $\partial(-A_N)$ has a stratification $\mathcal{T}$ consisting of two manifolds: t_1 is the origin in $\mathbb{R}^3$ and t_2 is the remainder of the cone. We consider each manifold separately.

1. $\boldsymbol{\alpha}_N \in t_1$. In this case, $a = b = c = 0$ and

$$\mathcal{NC}(-A_N, \boldsymbol{\alpha}_N) = (-A_N)^* = A_N$$

The situation essentially reduces to the Junkian case, and the bundle associated with $s \subset \{\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}} : \boldsymbol{\alpha}_N = 0\}$ is

$$\mathcal{B}(s) = \{\boldsymbol{\rho} : \boldsymbol{\rho}_N >^* \mathbf{r}_N(\boldsymbol{\alpha}),\ \boldsymbol{\alpha} \in s\ \}\ , \tag{4.58}$$

and if $N = 4$,

$$\mathcal{B}(s) = \left\{\boldsymbol{\rho} : \boldsymbol{\rho}_N >^* \boldsymbol{\pi}_N(0,0,\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})),\ \boldsymbol{\rho}_{N-1} = \boldsymbol{\pi}_{N-1}(0,0,\bar{\mathbf{a}}(\bar{\boldsymbol{\rho}})),\ \bar{\boldsymbol{\rho}} \in \mathcal{R}_{\bar{\mathbf{m}}}\right\} \tag{4.59}$$

However, unlike the Junkian case, the inequality in (4.58) and (4.59) is no longer scalar. Rather, it must be understand in terms of the dual cone A_N^*.

2. $\boldsymbol{\alpha}_N \in t_2$. In this case $a \geq |b| > 0$. In the (a,b,c) coordinates $\mathcal{NC}(-A_N, \boldsymbol{\alpha}_N)$ is

a ray given by

$$\left\{\lambda\left(\sqrt{b^2+c^2}, b, c,\right) : \lambda \geq 0\right\} .$$

Therefore

$$\mathcal{NC}(-A_N, \boldsymbol{\alpha}_N) = \left\{ \lambda \begin{pmatrix} \sqrt{b^2+c^2}+b & c \\ +c & \sqrt{b^2+c^2}-b \end{pmatrix} : \lambda > 0 \right\} , \qquad (4.60)$$

which can expressed in terms of the components of $\boldsymbol{\alpha}_N$ by inverting (4.56). The bundle associated with any $s \subset \{\mathcal{A}_{\mathbf{m}} \cap \partial\mathcal{A}_{\mathbf{m}} : \boldsymbol{\alpha}_N \neq 0\}$ is

$$\mathcal{B}(s) = \left\{\boldsymbol{\rho} : \boldsymbol{\rho}_N = \mathbf{r}_N(\boldsymbol{\alpha}) + \mathcal{NC}(-A_N, \boldsymbol{\alpha}_N), \boldsymbol{\rho}_j = \mathbf{r}_j(\boldsymbol{\alpha}) , \;\; j < N , \boldsymbol{\alpha} \in s\right\} . \qquad (4.61)$$

The set $\mathcal{R}_{\mathbf{m}} \backslash \mathcal{R}_{\mathbf{m}}^{\exp}$ is the union of bundles of the type given in (4.58) and (4.61).

4.4 Appendix: Duality Theorems

Proof of Theorem 22. The form of the constraints in (4.4) requires that $\mathbf{m}$ be separated into lower-order and higher-order polynomials. Define the polynomial vector of lower degree polynomials

$$\mathbf{m}_{\mathrm{L}} = (\mathbf{m}_1, \mathbf{m}_2, \ldots \mathbf{m}_{N-1}) \qquad (4.62)$$

and let $l_{\mathrm{L}} = l_0 + \ldots + l_{N-1}$ be the number of components in $\mathbf{m}_{\mathrm{L}}$. Let $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ and introduce the sets

$$\mathbb{A} \equiv \big\{(\eta, \boldsymbol{\sigma}_{\mathrm{L}}, \boldsymbol{\sigma}_N) \in \mathbb{R} \times \mathbb{R}^{l_{\mathrm{L}}} \times \mathbb{R}^{l_N} : \\ \eta \geq \mathcal{H}(g),\, \boldsymbol{\sigma}_{\mathrm{L}} = \langle \mathbf{m}_{\mathrm{L}} g \rangle\,,\, \boldsymbol{\sigma}_N \geq^* \langle \mathbf{m}_N g \rangle \text{ for some } g \in \mathbb{F}_{\mathbf{m}}\big\}\,,$$

$$\mathbb{B} \equiv \big\{(\eta, \boldsymbol{\sigma}_{\mathrm{L}}, \boldsymbol{\sigma}_N) \in \mathbb{R} \times \mathbb{R}^{l_{\mathrm{L}}} \times \mathbb{R}^{l_N} : \eta \leq h(\boldsymbol{\rho}),\, \boldsymbol{\sigma}_{\mathrm{L}} = \boldsymbol{\rho}_{\mathrm{L}},\, \boldsymbol{\sigma}_N \leq^* \boldsymbol{\rho}_N\big\}\,,$$

$$\tilde{\mathbb{B}} \equiv \big\{(\eta, \boldsymbol{\sigma}_{\mathrm{L}}, \boldsymbol{\sigma}_N) \in \mathbb{R} \times \mathbb{R}^{l_{\mathrm{L}}} \times \mathbb{R}^{l_N} : \eta \leq h(\boldsymbol{\rho}),\, \boldsymbol{\sigma}_{\mathrm{L}} \leq^* \boldsymbol{\rho}_{\mathrm{L}},\, \boldsymbol{\sigma}_N \leq^* \boldsymbol{\rho}_N\big\}\,.$$

Here the vectors $\boldsymbol{\rho}_{\mathrm{L}}$, $\boldsymbol{\sigma}_{\mathrm{L}}$, and $\boldsymbol{\alpha}_{\mathrm{L}}$ are defined in a manner analogous to 4.62. Using the convexity of $\mathcal{H}$, it is fairly easy to show that $\mathbb{A}$ and $\tilde{\mathbb{B}}$ are convex and that $\tilde{\mathbb{B}}$ has a non-empty interior that is disjoint from $\mathbb{A}$. Therefore the Eidelheit Separation Theorem (see Theorem 3 in Section 5.12 of [56]) implies that $\mathbb{A}$ and $\tilde{\mathbb{B}}$ are separated by a hyperplane in $\mathbb{R} \times \mathbb{R}^{l_{\mathrm{L}}} \times \mathbb{R}^{l_N}$. Since $\mathbb{B} \subset \tilde{\mathbb{B}}$, this hyperplane separates $\mathbb{A}$ and $\mathbb{B}$ as well. This means that there exists $(\hat{\eta}, \hat{\boldsymbol{\alpha}}_{\mathrm{L}}, \hat{\boldsymbol{\alpha}}_N) \in \mathbb{R} \times \mathbb{R}^{l_{\mathrm{L}}} \times \mathbb{R}^{l_N}$, not all zero, such that

$$\hat{\eta}\eta^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{A}} \geq \hat{\eta}\eta^{\mathbb{B}} + \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{B}} + \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{B}} \tag{4.63}$$

for all $\left(\eta^{\mathbb{A}}, \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}}, \boldsymbol{\sigma}_N^{\mathbb{A}}\right) \in \mathbb{A}$ and $\left(\eta^{\mathbb{B}}, \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{B}}, \boldsymbol{\sigma}_N^{\mathbb{B}}\right) \in \mathbb{B}$, or since $\boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{B}} = \boldsymbol{\rho}_{\mathrm{L}}$,

$$\hat{\eta}\eta^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{A}} \geq \hat{\eta}\eta^{\mathbb{B}} + \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \boldsymbol{\rho}_{\mathrm{L}} + \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{B}}\,. \tag{4.64}$$

The relation in (4.63) can be written more compactly as

$$\hat{\eta}\eta^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}^T \boldsymbol{\sigma}^{\mathbb{A}} \geq \hat{\eta}\eta^{\mathbb{B}} + \hat{\boldsymbol{\alpha}}^T \boldsymbol{\sigma}^{\mathbb{B}}\,.$$

The nature of $\mathbb{A}$ and $\mathbb{B}$ now leads to conclusions about the elements $\hat{\eta}$ and $\hat{\boldsymbol{\alpha}}_N$. For example, letting $\boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} = \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{B}} = \boldsymbol{\rho}_{\mathrm{L}}$ and $\boldsymbol{\sigma}_N^{\mathbb{A}} = \boldsymbol{\sigma}_N^{\mathbb{B}} = \boldsymbol{\rho}_N$ in (4.64) yields

$$\hat{\eta}\eta^{\mathbb{A}} \geq \hat{\eta}\eta^{\mathbb{B}}$$

for all $\eta^{\mathbb{B}} \leq h(\boldsymbol{\rho})$ and all $\eta^{\mathbb{A}} \geq \mathcal{H}(g)$ with $g \in \mathbb{F}_{\mathbf{m}}$. Thus $\hat{\eta} \geq 0$. (Note that the choice of $\boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}}$ and $\boldsymbol{\sigma}_N^{\mathbb{A}}$ is possible since $\boldsymbol{\rho}$ is assumed to be in $\mathcal{R}_{\mathbf{m}}$.) Also letting $\boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} = \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{B}} = \boldsymbol{\rho}_{\mathrm{L}}$, $\eta^{\mathbb{B}} = h(\boldsymbol{\rho})$, $\boldsymbol{\sigma}_N^{\mathbb{A}} = \boldsymbol{\rho}_N$, and $\eta^{\mathbb{A}} \to h(\boldsymbol{\rho})$ yields

$$\hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\rho}_N \geq \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{B}}$$

for all $\boldsymbol{\sigma}_N^{\mathbb{B}} \leq^* \boldsymbol{\rho}_N$ and therefore $\hat{\boldsymbol{\alpha}}_N \geq 0$. (Recall that inequalities between vectors are interpreted in the sense of cones as described in Section 2.4).

We now prove by contradiction that $\hat{\eta}$ is positive. If $\hat{\eta} = 0$, then letting $\boldsymbol{\sigma}_N^{\mathbb{B}} = \boldsymbol{\rho}_N$ in (4.64) gives,

$$\hat{\boldsymbol{\alpha}}^T \left(\boldsymbol{\sigma}^{\mathbb{A}} - \boldsymbol{\rho}\right) = \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \left(\boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} - \boldsymbol{\rho}_{\mathrm{L}}\right) + \hat{\boldsymbol{\alpha}}_N^T \left(\boldsymbol{\sigma}_N^{\mathbb{A}} - \boldsymbol{\rho}_N\right) \geq 0\,. \tag{4.65}$$

for all $\boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} = \langle \mathbf{m}_{\mathrm{L}} g \rangle$ and $\boldsymbol{\sigma}_N^{\mathbb{A}} \geq \langle \mathbf{m}_N g \rangle$ with $g \in \mathbb{F}_{\mathbf{m}}$. In particular (4.65) holds for $\boldsymbol{\sigma} = \boldsymbol{\rho}$. Thus, since $\mathcal{R}_{\mathbf{m}}$ is open, there exist $\boldsymbol{\rho} \in \mathcal{R}_{\mathbf{m}}$ such that $\hat{\boldsymbol{\alpha}}^T \left(\boldsymbol{\sigma}^{\mathbb{A}} - \boldsymbol{\rho}\right) < 0$ unless $\hat{\boldsymbol{\alpha}} = 0$ which, assuming that $\hat{\eta} = 0$, contradicts the Eidelheit Separation Theorem. We conclude that $\hat{\eta} > 0$, and by multiplying (4.64) by an appropriate constant, we may assume, without loss of generality, that $\hat{\eta} = 1$. In this case, (4.64)

becomes

$$\eta^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{A}} \geq \eta^{\mathbb{B}} + \hat{\boldsymbol{\alpha}}_{\mathrm{L}}^T \boldsymbol{\rho}_{\mathrm{L}} + \hat{\boldsymbol{\alpha}}_N^T \boldsymbol{\sigma}_N^{\mathbb{B}} \tag{4.66}$$

for all $\left(\eta^{\mathbb{A}}, \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{A}}, \boldsymbol{\sigma}_N^{\mathbb{A}}\right) \in \mathbb{A}$ and $\left(\eta^{\mathbb{B}}, \boldsymbol{\sigma}_{\mathrm{L}}^{\mathbb{B}}, \boldsymbol{\sigma}_N^{\mathbb{B}}\right) \in \mathbb{B}$.

We next utilize (4.66) understand the relationship between $\hat{\boldsymbol{\alpha}}$ and $h(\boldsymbol{\rho})$. If $\left(\eta^{\mathbb{B}}, \boldsymbol{\sigma}_N^{\mathbb{B}}\right) = (h(\boldsymbol{\rho}), \boldsymbol{\rho}_N)$, then (4.66) gives

$$h(\boldsymbol{\rho}) \leq \eta^{\mathbb{A}} + \hat{\boldsymbol{\alpha}}^T \left(\boldsymbol{\sigma}^{\mathbb{A}} - \boldsymbol{\rho}\right) \tag{4.67}$$

for all $\left(\eta^{\mathbb{A}}, \boldsymbol{\sigma}^{\mathbb{A}}\right) \in \mathbb{A}$. Considering that $(h(\boldsymbol{\rho}), \boldsymbol{\rho}) \in \mathbb{A}$, it follows then that

$$h(\boldsymbol{\rho}) = \inf_{(\eta, \boldsymbol{\sigma}) \in \mathbb{A}} \left\{\eta + \hat{\boldsymbol{\alpha}}^T (\boldsymbol{\sigma} - \boldsymbol{\rho})\right\} . \tag{4.68}$$

In addition, by letting $\eta^{\mathbb{A}} = \mathcal{H}(g)$ and $\boldsymbol{\sigma}^{\mathbb{A}} = \langle \mathbf{m} g \rangle$ for any $g \in \mathbb{F}_{\mathbf{m}}$, (4.67) gives

$$h(\boldsymbol{\rho}) \leq \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{\mathcal{H}(g) + \hat{\boldsymbol{\alpha}}^T (\langle \mathbf{m} g \rangle - \boldsymbol{\rho})\right\} . \tag{4.69}$$

Therefore,

$$\begin{aligned} h(\boldsymbol{\rho}) &\leq \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{\mathcal{H}(g) + \hat{\boldsymbol{\alpha}}^T (\langle \mathbf{m} g \rangle - \boldsymbol{\rho})\right\} \\ &\leq \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{\mathcal{H}(g) + \hat{\boldsymbol{\alpha}}^T (\langle \mathbf{m} g \rangle - \boldsymbol{\rho}) : \langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}\right\} \\ &\leq \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{\mathcal{H}(g) : \langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}\right\} \\ &= h(\boldsymbol{\rho}) . \end{aligned} \tag{4.70}$$

The first inequality in (4.70) is just (4.69); the second follows from the fact that the set of feasible functions $g \in \mathbb{F}_{\mathbf{m}}$ has been restricted; the third follows because $\hat{\boldsymbol{\alpha}}^T \left(\langle \mathbf{m} g \rangle - \boldsymbol{\rho} \right) \leq 0$ whenever $\langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}$.

From (4.70), it follows that

$$h(\boldsymbol{\rho}) = \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \hat{\boldsymbol{\alpha}}^T \left(\langle \mathbf{m} g \rangle - \boldsymbol{\rho} \right) \right\} , \tag{4.71}$$

and since $\mathcal{H}(\hat{g}_{\boldsymbol{\rho}}) = h(\boldsymbol{\rho})$, (4.71) implies that

$$\mathcal{H}(\hat{g}_{\boldsymbol{\rho}}) \leq \mathcal{H}(\hat{g}_{\boldsymbol{\rho}}) + \hat{\boldsymbol{\alpha}}^T \left(\langle \mathbf{m} \hat{g}_{\boldsymbol{\rho}} \rangle - \boldsymbol{\rho} \right) \leq \mathcal{H}(\hat{g}_{\boldsymbol{\rho}}) \, .$$

This proves the complementary slackness condition (4.28) and also that

$$h(\boldsymbol{\rho}) = \mathcal{L} \left(\hat{g}_{\boldsymbol{\rho}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\rho} \right) \, . \tag{4.72}$$

Furthermore, given *any* $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}_N \geq \mathbf{0}$,

$$\begin{aligned} \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \boldsymbol{\alpha}^T (\langle \mathbf{m} g \rangle - \boldsymbol{\rho}) \right\} &\leq \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \boldsymbol{\alpha}^T (\langle \mathbf{m} g \rangle - \boldsymbol{\rho}) : \langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho} \right\} \\ &\leq \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) : \langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho} \right\} \\ &= h(\boldsymbol{\rho}) \, . \end{aligned} \tag{4.73}$$

The first inequality above holds because the set of possible functions $g \in \mathbb{F}_{\mathbf{m}}$ has been restricted, and the second inequality holds because $\boldsymbol{\alpha}^T \left(\langle \mathbf{m} g \rangle - \boldsymbol{\rho} \right) \leq 0$ whenever $\langle \mathbf{m} g \rangle \preceq^* \boldsymbol{\rho}$. In addition, (4.71) shows that both inequalities become equalities when

$\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$. This means that

$$\begin{aligned} h(\boldsymbol{\rho}) &= \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \hat{\boldsymbol{\alpha}}^T \left(\langle \mathbf{m} g \rangle - \boldsymbol{\rho} \right) \right\} \\ &= \max_{\boldsymbol{\alpha}_N \geq \mathbf{0}} \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \boldsymbol{\alpha}^T (\langle \mathbf{m} g \rangle - \boldsymbol{\rho}) \right\} \\ &= \max_{\boldsymbol{\alpha}_N \leq \mathbf{0}} \inf_{g \in \mathbb{F}_{\mathbf{m}}} \left\{ \mathcal{H}(g) + \boldsymbol{\alpha}^T (\boldsymbol{\rho} - \langle \mathbf{m} g \rangle) \right\} \\ &= \max_{\boldsymbol{\alpha}_N \leq \mathbf{0}} \inf_{g \in \mathbb{F}_{\mathbf{m}}} \mathcal{L} \left(g, \boldsymbol{\alpha}, \boldsymbol{\rho} \right) \\ &= \max_{\boldsymbol{\alpha}_N \leq \mathbf{0}} \psi(\boldsymbol{\alpha}, \boldsymbol{\rho}) \, , \end{aligned} \tag{4.74}$$

which proves (4.27), and moreover, that the maximum in (4.27) is attained by $\hat{\boldsymbol{\alpha}}$, i.e.,

$$h(\boldsymbol{\rho}) = \psi(\hat{\boldsymbol{\alpha}}, \boldsymbol{\rho}) \, . \tag{4.75}$$

Together (4.75) and (4.72) give (4.29). ■

The proof of Theorem 24 is very similar to that of Theorem 22, and the differences are fairly transparent. First, the sign of $\boldsymbol{\alpha}_N$ is not determined, although it is determined later. (See the remarks following Theorem 24). Even so, one may still deduce that $\hat{\eta} > 0$ using the fact that $\mathcal{R}_{\mathbf{m}}$ is open. Also, because the constraints in (4.33) are all equalities, the condition

$$\boldsymbol{\alpha}^T \left(\langle \mathbf{m} g \rangle - \boldsymbol{\rho} \right) \leq 0$$

holds trivially. It is this fact that is key to the arguments in (4.70) and (4.71).

It should be noted that this proof is based on arguments found in [56]. In

particular, one should consult Theorem 1 in section 8.3, Theorem 1 in section 8.6, and Exercise 7 at the end of Chapter 8. Many other texts discuss duality theory in a variety of contexts, but most assume that the argument of the objective ($\mathcal{H}$ in this case) lives in a linear vector space, which $\mathbb{F}_{\mathbf{m}}$ is not.

Chapter 5

Simulation of an n^+-n-n^+ Diode

In this chapter, we compute numerical solutions for several second-order models generated by closures from Chapter 3. An n^+-n-n^+ diode [41] acts as a benchmark problem for comparing and contrasting various aspects of these models. It is assumed that the diode is endowed with a slab symmetry, which means that the distribution of electrons is constant when restricted to planes perpendicular to a given axis. This assumption is often employed when the length scale of a device along such an axis is much smaller than the length scales perpendicular to the axis.

We confirm several previously known facts. The first is that the use of Monte Carlo relaxation coefficients improves numerical results. The second is that heat flux is a necessary component of an accurate model. However, most expressions for the heat flux in the models studied here are not sufficient to accurately describe the behavior of the diode. This includes convective corrections derived in [4] which have little effect in the diode drain where velocity overshoot [33] is prevalent.

We also make some new observations. Most important among these is the fact that anisotropic stress plays a major role in the velocity and temperature profiles of the Anile-Penisi (AP) and perturbed entropy-based (PEB) models. When treated

as a diffusive pertubation, the anisotropic stress removes velocity overshoot effect in the diode drain at the cost of smearing the temperature profile. When treated as an independent variable in the Gaussian closure, it has less effect on velocity overshoot but also less smearing in the temperature profile.

We now lay out the organization of the chapter. In Section 5.1, we reduce moment systems to one spatial dimension by slab symmetry along the x_1 axis. In Section 5.2, we give a complete list of models which we will study. In Section 5.3, we discuss the central-upwind scheme [45] that is the basis for our computations. Finally, in Section 5.4, we present results and provide comments.

5.1 Reduction to One Dimension

In this section, we invoke the slab symmetry of the diode to reduce all second-order models to a description in one spatial dimension. All of these models of are derived using moments of the polynomial vector

$$\mathbf{m} = \begin{pmatrix} 1 \\ v \\ \frac{1}{2}|v|^2 \end{pmatrix} \qquad \text{or} \qquad \mathbf{m} = \begin{pmatrix} 1 \\ v \\ v \vee v \end{pmatrix}$$

and are supplemented by a Poisson equation for Φ:

$$-(\epsilon \Phi_x)_x = q_e(D - n)\,. \tag{5.1}$$

5.1.1 The Case $\mathbf{m} = (1, v, \frac{1}{2}|v|^2)^T$

Models based on the polynomial vector $\mathbf{m} = (1, v, \frac{1}{2}|v|^2)^T$ all have the form

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{5.2a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu^2 + n\theta I + \Sigma) - n\nabla_x \Phi = -\frac{1}{\tau_p} nu\,, \tag{5.2b}$$

$$\partial_t \left(\frac{nu^2}{2} + \frac{3n\theta}{2} \right) + \nabla_x \cdot \left(\frac{nu^3}{2} + \frac{5}{2} nu\theta + \Sigma \cdot u + q \right) - nu \cdot \nabla_x \Phi = C_{\frac{1}{2}|v|^2}\,, \tag{5.2c}$$

where

$$C_{\frac{1}{2}|v|^2} - \frac{1}{\tau_w} \left(\frac{n\,|u|^2}{2} + \frac{3n\,(\theta - \theta_\ell)}{2} \right).$$

The fact that the diode has slab symmetry means that

$$\nabla_x = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ 0 \\ 0 \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ 0 \\ 0 \end{pmatrix}, \qquad q = \begin{pmatrix} q_1 \\ 0 \\ 0 \end{pmatrix},$$

and that

$$\Theta = \mathrm{diag}(\theta_L, \theta_T, \theta_T)\,,$$

$$\Sigma = n\,\mathrm{diag}(\theta_L - \theta, \theta_T - \theta, \theta_T - \theta)\,,$$

$$\theta = \frac{1}{3}\,(\theta_L + 2\theta_T)\,.$$

For convenience, we abuse notation by dropping the subscript from the components x_1, u_1, and q_1. Since the remaining components of these vectors play no role in

what follows, their should be no chance of confusion. With this notation, the balance equations for concentration, momentum, and energy are

$$\partial_t n + \partial_x (nu) = 0 \tag{5.4a}$$

$$\partial_t (nu) + \partial_x (nu^2 + n\theta + \sigma) - \frac{q_e}{m_e^*} n \partial_x \Phi = -\frac{1}{\tau_p} nu \tag{5.4b}$$

$$\partial_t \left(\frac{nu^2 + 3n\theta}{2} \right) + \partial_x \left(\frac{nu^3 + nu(3\theta + 2\theta_L)}{2} + q \right) - \frac{q_e}{m_e^*} nu \partial_x \Phi = -\frac{1}{\tau_w} \left(\frac{nu^2 + 3n(\theta - \theta)_\ell}{2} \right) , \tag{5.4c}$$

where $\sigma = n(\theta_L - \theta)$ is the *anisotropy*. A closure can then be specified by giving θ_L and q in terms of n, u, and θ.

5.1.2 The Case $\mathbf{m} = (1, v, v \vee v)^T$

Models based on the polynomial vector $\mathbf{m} = (1, v, v \vee v)^T$ all have the form

$$\partial_t n + \nabla_x \cdot (nu) = 0 \tag{5.5a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\Theta) - \frac{q_e}{m_e^*} n \nabla_x \Phi = -\frac{1}{\tau_p} nu \tag{5.5b}$$

$$\partial_t (nu \vee u + n\Theta) + \nabla_x \cdot \left(nu^{\vee 3} + 3n\Theta \vee u + Q \right) - \frac{q_e}{m_e^*} nu \cdot \nabla_x \Phi = C_{v \vee v} , \tag{5.5c}$$

where

$$C_{v \vee v} = -\frac{1}{\tau_\sigma} (n\Theta - n\theta I) - \frac{1}{\tau_p} \left(nu \vee u - \frac{1}{3} n|u|^2 \right) - \frac{1}{\tau_w} \left(\frac{1}{3} n|u|^2 I + n(\theta - \theta)_\ell I \right) .$$

The assumption of slab symmetry implies that the only nonzero components of Q are Q_{111} and all permutations of $Q_{122} = Q_{133}$. Equations (5.5a) and (5.5b) and the one-half of the trace of (5.5c) give provide a description for the evolution for the concentration, momentum, and energy that is analogous to (5.4):

$$\partial_t n + \partial_x (nu) = 0 \tag{5.6a}$$

$$\partial_t (nu) + \partial_x (nu^2 + n\theta_L) - \frac{q_e}{m_e^*} n \partial_x \Phi = -\frac{1}{\tau_p} nu \tag{5.6b}$$

$$\begin{aligned} \partial_t \left(\frac{nu^2 + 3n\theta}{2} \right) + \partial_x \left(\frac{nu^3 + nu(3\theta + 2\theta_L)}{2} + q \right) \\ - \frac{q_e}{m_e^*} nu \partial_x \Phi = -\frac{1}{\tau_w} \left(\frac{nu^2 + 3n(\theta - \theta)_\ell}{2} \right) . \end{aligned} \tag{5.6c}$$

(Note the same abuse of notation with the subscripts from x_1, u_1, and q_1 all dropped).

There is one more independent scalar equation that may be extracted from (5.5). By taking the $(1,1)$ component of (5.5c) and subtracting one-third of the trace of (5.5c), one finds that

$$\begin{aligned} \partial_t \left(\frac{2}{3} nu^2 + n(\theta_L - \theta) \right) \\ + \partial_x \left(\frac{2}{3} nu^3 + nu \left(\frac{8}{3} \theta_L - \theta \right) + \tilde{q} \right) - \frac{4}{3} \frac{q_e}{m_e^*} nu \partial_x \Phi + \tilde{q} \\ = -\frac{1}{\tau_\sigma} \left(\frac{2}{3} nu^2 + n(\theta_L - \theta) \right) + \frac{2}{3} \left(\frac{1}{\tau_p} - \frac{1}{\tau_\sigma} \right) nu^2 , \end{aligned} \tag{5.6d}$$

where

$$\tilde{q} = \left(Q_{111} - \frac{1}{3} (Q_{111} + Q_{122} + Q_{133}) \right) .$$

This additional equation is to track the anisotropy of Θ which is known once θ_L , θ_T, or σ is determined. A closure for (5.6) is specified by giving q and $\tilde{q}$ in terms of n, u, θ, and θ_L.

5.2 The Models

5.2.1 Bløtekjær-Type Models.

Several variations of the Bløtekjær model, all of which have the form (5.2), are listed below. As in Chapter 3, q and σ are separated into diffusive components , $\sigma^{(1)}$ and $q^{(1)}$, and their convective components, $\sigma^{(2)}$ and $q^{(2)}$.

- **Maxwellian Baccarani-Wordeman (MBW).** Maxwellian closure with Baccarani-Wordeman formulas for relaxation times:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = q^{(2)} = 0\,.$$

- **Maxwellian Monte Carlo (MMC).** Maxwellian closure with Monte-Carlo relaxation times:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = q^{(2)} = 0\,.$$

- **Bløtekjær Baccarani-Wordeman 1 (BBW1).** Bløtekjær, Baccarani, Wordemann model with $\gamma = -1.0$:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,.$$

$$q^{(1)} = -\frac{3}{2} n\theta \tau_p^{\mathrm{BW}} \partial_x \theta\,, \qquad q^{(2)} = 0\,.$$

- **Bløtekjær Baccarani-Wordeman 2 (BBW2).** Bløtekjær, Baccarani, Wordemann model with $\gamma = -2.1$:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -0.4 n\theta \tau_p^{\mathrm{BW}} \partial_x \theta\,, \qquad q^{(2)} = 0\,.$$

- **Bløtekjær Monte Carlo 1 (BMC1).** Same as BBW1, except relaxation times are Monte Carlo:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -\frac{3}{2} \rho\theta \tau_p^{\mathrm{MC}} \partial_x \theta\,, \qquad q^{(2)} = 0\,.$$

- **Bløtekjær Monte Carlo 2 (BMC2).** Same as BBW2, except relaxation

times are Monte Carlo:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -0.4n\theta\tau_p^{\mathrm{MC}}\partial_x\theta,\ q^{(2)} = 0\,.$$

The BBW1 and BBW2 models can be found in various places in the literature. (See [25, 28, 29, 41, 72].).We include the models MBW and MMC to see the effects of the heat flux and the models BMC1 and BMC2 to see the improved results provided by Monte Carlo relaxation times when contrasted with the analytical expressions of Baccarani and Wordemann. The Monte Carlo relaxation times are modeled as functions of the electron energy and can be found in [63].

5.2.1.1 Anile-Pennisi Models The next group of models are variants of the Anile-Pennisi closure [4] that is based on extended thermodynamics. Each of them has the form of (5.2), except for APPV1 and APPV2. These models make inconsistent use of the anisotropy σ, which is included in the energy equation but not the momentum equation. Several of these models include an important aspect of the Anile-Pennisi closure, which is a convective contribution to the heat flux.

- **Anile-Pennisi No Viscosity 1 (APNV1)**. Anile-Pennisi model with Monte Carlo relaxation times. No anisotropy:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$
$$q^{(1)} = -\frac{5}{2} n\theta\tau_q^{\mathrm{MC}}\partial_x\theta\,, \qquad q^{(2)} = \frac{5}{2} n\theta u\left(\frac{\tau_q^{\mathrm{MC}}}{\tau_p^{\mathrm{MC}}} - 1\right).$$

- **Anile-Pennisi No Viscosity 2 (APNV2).** Anile-Pennisi model with Monte Carlo relaxation times. No anisotropy and no convective heat correction:

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_q^{\mathrm{MC}}\partial_x\theta\,, \qquad q^{(2)} = 0\,.$$

- **Anile-Pennisi Partial Viscosity 1 (APPV1).** Anile-Pennisi model with Monte Carlo relaxation times. Anisotropy applied only in the energy equation:

$$\sigma^{(1)} = \frac{4}{3} n\theta\tau_\sigma^{\mathrm{MC}}\partial_x u = 0\,, \qquad \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_q^{\mathrm{MC}}\partial_x\theta\,, \qquad q^{(2)} = \frac{5}{2} n\theta u\left(\frac{\tau_q^{\mathrm{MC}}}{\tau_p^{\mathrm{MC}}} - 1\right).$$

- **Anile-Pennisi Partial Viscosity 2 (APPV1).** Anile-Pennisi model with Monte Carlo relaxation times. Anisotropy applied only in energy equation.

No convective corrections to the heat flux:

$$\sigma^{(1)} = \frac{4}{3} n\theta\tau_{\sigma}^{\mathrm{MC}} \partial_x u = 0\,, \qquad \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_{q}^{\mathrm{MC}} \partial_x \theta\,, \qquad q^{(2)} = 0\,.$$

- **Anile-Pennisi Full Viscosity 1 (APFV1).** Anile-Pennisi model with Monte Carlo relaxation times. Viscous anisotropy applied in the momentum and energy equations:

$$\sigma^{(1)} = \frac{4}{3} n\theta\tau_{\sigma}^{\mathrm{MC}} \partial_x u = 0\,, \qquad \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_{q}^{\mathrm{MC}} \partial_x \theta\,, \qquad q^{(2)} = \frac{5}{2} n\theta u \left(\frac{\tau_{q}^{\mathrm{MC}}}{\tau_{p}^{\mathrm{MC}}} - 1 \right)\,.$$

- **Anile-Pennisi Full Viscosity 2 (APFV2).** Anile-Pennisi model with Monte Carlo relaxation times. Viscous anisotropy applied in the momentum and energy equations. No convective corrections to heat flux:

$$\sigma^{(1)} = -\frac{4}{3} n\theta\tau_{\sigma}^{\mathrm{MC}} \partial_x u = 0\,, \qquad \sigma^{(2)} = 0\,,$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_{q}^{\mathrm{MC}} \partial_x \theta\,, \qquad q^{(2)} = 0\,.$$

Computational results for APNV1 and APPV1 can be found in [72] and [63], respectively. We believe that the inclusion of anisotropic effects is an important calculation, which is why APFV1 and APFV2 have been included. In addition, we believe it is important to examine the real effects of the convective heat flux corrections that are the main advance of the Anile-Pennisi closure. Many times, difficulties with accurate modeling occur at places in the physical domain of the problem where the spatial gradients of macroscopic variables are large, in which case diffusive corrections will likely dominate convective corrections. The three models APNV2, APPV2, and APFV2 provide comparisons to determine if this in the case.

5.2.1.2 Perturbed Entropy-Based Closures The final group of models are the entropy based models discussed in Chapter 3. The first three of these are based on the Maxwellian Closure and have the form of (5.2).

- **Maxwellian Monte-Carlo (MMC).** Maxwellian closure with Monte-Carlo relaxation times (already considered with the Bløtekjær models):

$$\sigma^{(1)} = \sigma^{(2)} = 0\,,$$

$$q^{(1)} = q^{(2)} = 0\,.$$

- **Perturbed Maxwellian Monte Carlo 1 (PMMC1).** Perturbed Maxwellian

closure with Monte-Carlo relaxation times:

$$\sigma^{(1)} = -\frac{4}{3} n\theta\tau_p^{\mathrm{MC}} \partial_x u \,, \qquad \sigma^{(2)} = \frac{2}{3} n|u|^2 \,.$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_p^{\mathrm{MC}} \partial_x \theta \,, \qquad q^{(2)} = -\frac{4}{3} nu^3 + \frac{\tau_p^{\mathrm{MC}}}{\tau_w^{\mathrm{MC}}} \left(\frac{5}{6} nu^3 + \frac{5}{2} n \left(\theta - \theta_\ell\right) u \right) .$$

- **Perturbed Maxwellian Monte Carlo 2 (PMMC2).** Perturbed Maxwellian closure with Monte-Carlo relaxation times. Convective corrections are left out:

$$\sigma^{(1)} = -\frac{4}{3} n\theta\tau_p^{\mathrm{MC}} \partial_x u \,, \qquad \sigma^{(2)} = 0 \,.$$

$$q^{(1)} = -\frac{5}{2} n\theta\tau_p^{\mathrm{MC}} \partial_x \theta \,, \qquad q^{(2)} = 0 \,.$$

The remaining cases come from the Gaussian closure and take the form (5.6). In all of the previous models σ is expressed as a function of the variables n, u, and θ. Now, however, the evolution of σ is determined by an additional equation, and it is q and $\tilde{q}$ must be specified via the closure. The term $\tilde{q}$ can be written as the sum of a diffusive part $\tilde{q}^{(1)}$ and a convective part $\tilde{q}^{(2)}$. There are three models to consider.

- **Gaussian Monte Carlo (GMC).** Gaussian closure with Monte-Carlo relaxation times:

$$q = \tilde{q} = 0 \,.$$

- **Perturbed Gaussian Monte Carlo 1 (PGMC1).** Perturbed Gaussian

closure with Monte-Carlo relaxation times but no convective corrections:

$$
\begin{aligned}
q^{(1)} &= -\frac{1}{2} n \tau_\sigma^{\mathrm{MC}} \theta_L \partial_x \left(3\theta_L + 2\theta_T\right) , \\
q^{(2)} &= -\frac{\tau_\sigma^{\mathrm{MC}}}{\tau_p^{\mathrm{MC}}} \left(\frac{4}{3} n u^3 + n(\theta - \theta_L) u\right) + \frac{\tau_\sigma^{\mathrm{MC}}}{\tau_p^{\mathrm{MC}}} \left(\frac{5}{6} n u^3 + \frac{5}{2} n(\theta - \theta_\ell) u\right) , \\
\tilde{q}^{(1)} &= -\frac{1}{2} n \tau_\sigma^{\mathrm{MC}} \theta_L \partial_x \left(\frac{8}{3}\theta_L - 2\theta\right) , \\
\tilde{q}^{(2)} &= -\frac{\tau_\sigma^{\mathrm{MC}}}{\tau_p^{\mathrm{MC}}} \left(\frac{8}{9} n u^3 + \frac{7}{3} n(\theta - \theta_L) u\right) + \frac{\tau_\sigma^{\mathrm{MC}}}{\tau_p^{\mathrm{MC}}} \left(\frac{4}{9} n u^3 + \frac{4}{3} n(\theta - \theta_\ell) u\right) .
\end{aligned}
$$

- **Perturbed Gaussian Monte Carlo 2 (PGMC2).** Perturbed Gaussian closure with Monte-Carlo relaxation times but no convective corrections:

$$
q^{(1)} = -\frac{1}{2} n \tau_\sigma^{\mathrm{MC}} \theta_L \partial_x \left(3\theta_L + 2\theta_T\right) , \qquad q^{(2)} = 0 ,
$$

$$
\tilde{q}^{(1)} = -\frac{1}{2} n \tau_\sigma^{\mathrm{MC}} \theta_L \partial_x \left(\frac{8}{3}\theta_L - 2\theta\right) , \qquad q^{(2)} = 0 .
$$

To the author's knowledge, these are the first computations of their kind in the context of semiconductor models. We will compare and contrast these models with the AP models. We will also investigate whether the additional equation provided by the Gaussian closure improves the accuracy of results, especially with respect to the velocity and temperature profiles near the diode drain. Finally, as with the AP models, we will examine the effects of the convective corrections.

5.2.2 The Benchmark Device

The benchmark device is an n^+-n-n^+ diode that is used to simulate the channel in MOSFET and MESFET devices [84]. We assume it made of silicon with electric permittivity $\epsilon = 1.04 \times 10^{-16}\,\mathrm{C}/\,\mu\mathrm{m}$ and effective mass $m_e^* = 0.32 m_e$, where $m_e = 9.109 \times 10^{-31}\,\mathrm{kg}$ is the free electron mass [84]. Because of slab symmetry, the diode can be represent by an interval of length $L = 0.6$ microns with a doping profile

$$D(x) = \begin{cases} 1.0 \times 10^{18}\,\mathrm{cm}^{-3}, & 0.0\,\mu\mathrm{m} < x < 0.1\,\mu\mathrm{m} \\ 1.0 \times 10^{16}\,\mathrm{cm}^{-3}, & 0.1\,\mu\mathrm{m} < x < 0.5\,\mu\mathrm{m} \\ 1.0 \times 10^{18}\,\mathrm{cm}^{-3}, & 0.5\,\mu\mathrm{m} < x < 0.6\,\mu\mathrm{m} \end{cases} .$$

The left end is called the source, the right end is called the drain, and the center portion is the channel. An external battery with a potential $V_{\mathrm{bias}} = 1\,\mathrm{V}$ is attached to the device. The temperature of the device is $T = 300\,\mathrm{K} = 0.0259\,\mathrm{eV}$.

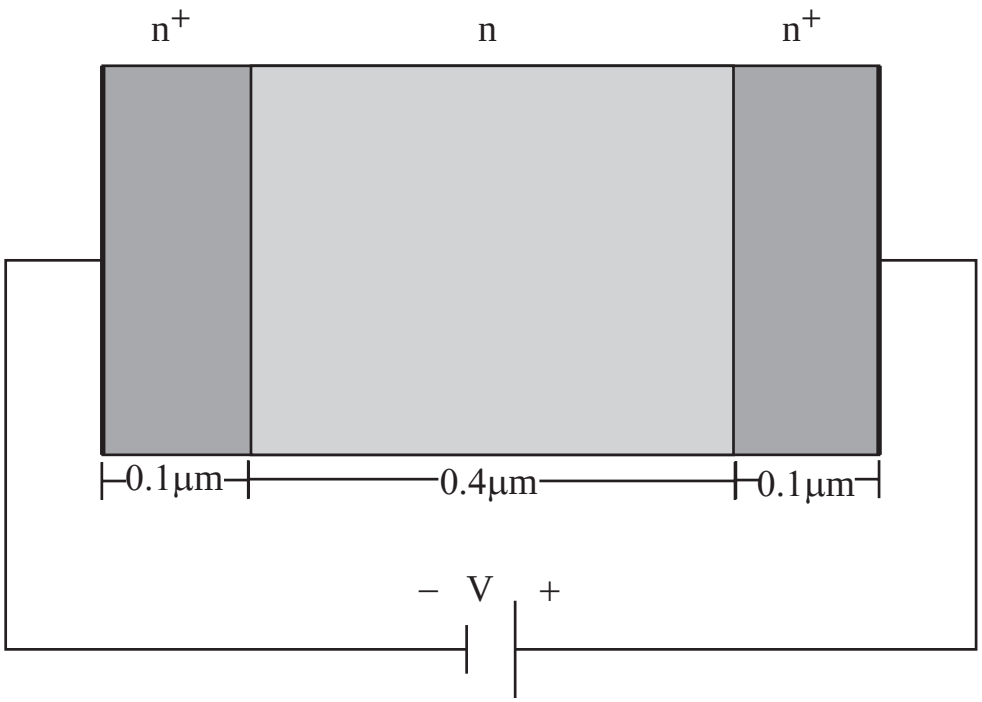

Figure 5.1: The n-n^+-n diode.

5.2.3 Boundary Conditions

Boundary conditions for (5.4) and (5.6) have not yet been given. Depending on the form of Σ, q, and $\tilde{q}$, boundary conditions can be of hyperbolic or mixed parabolic-hyperbolic type [82] [7]. However, as in [25], we find that our numerical solutions are not at all sensitive to over-specification of the boundary conditions. We therefore apply the following boundary conditions, which are consistent with a boundary layer

in thermal equilibrium:

$$n(0) = D(0)\,, \qquad n(L) = D(L)\,, \tag{5.7a}$$

$$\partial_x u(0) = \partial_x u(L) = 0\,, \tag{5.7b}$$

$$\partial_x \theta(0) = \partial_x \theta(L) = 0\,. \tag{5.7c}$$

Equation (5.6d) for the Gaussian based models requires the additional boundary condition

$$\partial_x \theta_L(0) = \partial_x \theta_L(L) = 0\,.$$

The boundary condition for the Poisson equation is

$$\Phi(L) = \Phi(0) - \log\left(\frac{n(L)}{n(0)}\right) + V_{\text{bias}}\,. \tag{5.8}$$

Since Φ is a relative quantity, the specification of $\Phi(0)$ can be arbitrary and has no effect on the numerics. We ignore traditional convention [59], and simply set $\Phi(0) = 0$. Given, (5.2a), this means that $\Phi(L) = V_{\text{bias}}$.

5.3 The Numerical Scheme

The models presented in the last section will be computed using central-upwind schemes. The schemes adapt central schemes into a traditional semi-discrete framework [47,50]. They maintain the key feature of central schemes—simplicity, but with less dissipation and without the cumbersome problems involved with staggering. For completeness we give a brief description below. Following traditional notation [47,50],

we let $\mathbf{u}$ (rather than $\boldsymbol{\rho}$) be the vector of spatial densities.

5.3.1 Finite Volume Formulation

All of the models in the last section has the form

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{l}(\mathbf{u})\Phi_x + (\mathbf{D}(\mathbf{u}) \cdot \mathbf{g}(\mathbf{u})_x)_x + \mathbf{r}(\mathbf{u}) \,, \tag{5.9}$$

where $\mathbf{f}(\mathbf{u})$ is the vector of fluxes, $\mathbf{l}(\mathbf{u})$ is the vector of field terms, $\mathbf{r}(\mathbf{u})$ is a vector of collision terms and $\mathbf{D}(\mathbf{u}) \cdot \mathbf{g}(\mathbf{u})_x$ is a vector of diffusive terms. The matrix $\mathbf{D}(\mathbf{u})$ is called the diffusion matrix. We assume that (5.9) is hyperbolic—that is, its homogeneous version

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0 \tag{5.10}$$

is hyperbolic—and let $\lambda_1 < \lambda_2 < \ldots < \lambda_r$ be the eigenvalues of the linearized flux matrix

$$\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \,.$$

The domain $[0, L]$ is divided into N uniform cells $I_i \equiv \left[x_{i-1/2}, x_{i+1/2}\right]$ with centers $\{x_i\}_{i=1}^N$. A semi-discrete, finite-volume formulation is obtained by integrating (5.9)

in space over each of these cells:

$$\begin{aligned}\frac{d}{dt}\bar{\mathbf{u}}_i(t) = &- \frac{\mathbf{f}(\mathbf{u}\left(x_{i+1/2},t\right)) - \mathbf{f}(\mathbf{u}\left(x_{i-1/2},t\right))}{\Delta x} \\ &+ \frac{1}{\Delta x}\int_{I_i} \mathbf{l}(\mathbf{u}(x,t))\Phi_x(x)\,dx \\ &+ \frac{\mathbf{D}(\mathbf{u}(x_{i+1/2},t))\cdot\mathbf{g}(\mathbf{u})_{\mathbf{x}}(x_{i+1/2},t) - \mathbf{D}(\mathbf{u}(x_{i-1/2},t))\cdot\mathbf{g}(\mathbf{u})_{\mathbf{x}}(x_{i-1/2},t)}{\Delta x} \\ &+ \frac{1}{\Delta x}\int_{I_i} \mathbf{r}(\mathbf{u}(x,t))\,dx\,.\end{aligned} \tag{5.11}$$

Here, the cell average $\bar{\mathbf{u}}_i$ is given by

$$\bar{\mathbf{u}}_i(t) \equiv \frac{1}{\Delta x}\int_{I_i} \mathbf{u}(x,t)\,dx\,. \tag{5.12}$$

Any algorithm for updating the evolution of $\bar{\mathbf{u}}$ requires that all of the terms on the right hand side of (5.11) be evaluated, at least approximately.

5.3.2 Flux Evaluation

In light of the fact that $\mathbf{u}$ may be discontinuous, the main issue in developing (5.11) is how to evaluate the fluxes at cell edges. This is done with a reconstruction procedure coupled with an efficient Riemann solver. We will use a Riemann solver found with central-upwind schemes [45].

5.3.2.1 Reconstruction The first step in evaluating the flux at an interface is to (approximately) reconstruct $\mathbf{u}$ with a function

$$\mathbf{p}(x) = \sum_{i=1}^{N} \mathbf{p}_i(x)\chi_i(x)\,, \tag{5.13}$$

where χ_i is the indicator function on the interior of I_i and $\mathbf{p}_i$ is a polynomial that satisfies

$$\frac{1}{\Delta x}\int_{I_k} \mathbf{p}_i(x) = \bar{\mathbf{u}}_k \tag{5.14}$$

for all k in the stencil of x_i. This stencil, $S(i; s_1, s_2)$, is a collection of mesh points

$$S(i; s_1, s_2) = \{x_{i-s_1}, \ldots, x_i, \ldots, x_{i+s_2-1}\}\,,$$

where the integers s_1 and s_2 are chosen based on two factors. The first of these is formal accuracy. If $\mathbf{u}$ is smooth, then a stencil with $s = s_1 + s_2$ points gives an order s approximation of $\mathbf{u}$ on I_i:

$$\mathbf{p}_i(x) = \mathbf{u}(x,t) + \mathcal{O}(\Delta x)^s\,, \qquad x \in I_i$$

Because $\mathbf{u}$ may be discontinuous, the other major consideration when choosing a stencil is that $\mathbf{p}$ reproduce the discontinuities in $\mathbf{u}$ without producing spurious oscillations. For second-order spatial accuracy, the reconstruction of $\mathbf{u}$ is a simple linear interpolation:

$$\mathbf{u}(x,t) = \bar{\mathbf{u}}_i(t) + \mathbf{u}_i'(t)(x - x_i)\,.$$

Here

$$\mathbf{u}_i' = SL(\bar{\mathbf{u}}_{i-1}, \bar{\mathbf{u}}_i, \bar{\mathbf{u}}_{i+1}),$$

where SL can be any appropriate slope limiter [47, 50]. For our calculations,

$$SL(\bar{\mathbf{u}}_{i-1}, \bar{\mathbf{u}}_i, \bar{\mathbf{u}}_{i+1}) = \text{minmod}\left(\frac{\bar{\mathbf{u}}_{i+1} - \bar{\mathbf{u}}_i}{\Delta x}, \frac{\bar{\mathbf{u}}_{i+1} - \bar{\mathbf{u}}_{i-1}}{2\Delta x}, \frac{\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i-1}}{\Delta x}\right),$$

where the minmod function is applied to a vector component-wise.

5.3.2.2 Riemann Solver Given reconstructions $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, the (approximate) value of $\mathbf{f}(\mathbf{u}\left(x_{i+1/2}\right))$ must be determined. Because $\mathbf{p}$ may be discontinuous at $x_{i+1/2}$—that is, beyond the smooth order of accuracy error between $\mathbf{p}_i\left(x_{i+1/2}\right)$ and $\mathbf{p}_{i+1}\left(x_{i+1/2}\right)$—an (approximate) Riemann solver [47] [50] must be employed. Given the Riemann problem

$$\begin{aligned} &\mathbf{v}_t + \mathbf{f}(\mathbf{v})_x = 0 \\ &\mathbf{v}(x,0) = \begin{cases} \mathbf{p}_i\left(x_{i+1/2}\right), & x < x_{i+1/2} \\ \mathbf{p}_{i+1}\left(x_{i+1/2}\right), & x > x_{i+1/2} \end{cases}, \end{aligned}$$

an (approximate) Riemann solver R gives an (approximate) solution

$$\mathbf{v}(x_{i+1/2}, t+\tau) = R(\mathbf{p}_i\left(x_{i+1/2}\right), \mathbf{p}_{i+1}\left(x_{i+1/2}\right), t+\tau)$$

for all τ sufficiently small. This solution is then used to approximate $\mathbf{f}(\mathbf{u}\left(x_{i+1/2}\right), t+\tau)$ by the numerical flux

$$\mathbf{F}_{i+1/2}(t+\tau) = \mathbf{f}\left(\mathbf{v}(x_{i+1/2}, t+\tau)\right) .$$

5.3.2.3 Central Schemes A Riemann solver computes $\mathbf{v}(x_{i+1/2}, t+\tau)$ by determining how waves emanate from discontinuities of $\mathbf{p}$ across adjacent cells. The computation cost associated with R can be large because it involves diagonalizing the matrix $\mathbf{A}(\mathbf{u})$ at each cell interface in order to analyze the local wave structure. Moreover, exact solvers may require iterative methods. Motivating by the desire to avoid Riemann solvers, the authors in [65] introduced *central schemes*, which use an integral formulation of (5.10) over the *staggered* cell $I_{i+1/2} = (x_i, x_{i+1})$.

A fully discrete central scheme for the homogeneous equation with step size Δt is

$$\frac{\bar{\mathbf{u}}_{i+1/2}(t+\Delta t) - \bar{\mathbf{u}}_{i+1/2}(t)}{\Delta t} + \int_t^{t+\Delta t} \frac{\mathbf{f}(\mathbf{u}\left(x_{i+1}, \tau\right)) - \mathbf{f}(\mathbf{u}\left(x_i, \tau\right))}{\Delta x}\, d\tau = 0 . \qquad (5.15)$$

The staggered average $\bar{\mathbf{u}}_{i+1/2}$ is updated by calculating $\bar{\mathbf{u}}_{i+1/2}(t)$ and the time integral in (5.15) using a reconstruction that interpolates the unstaggered averages $\bar{\mathbf{u}}_i(t)$. If $\mathbf{p}$ is an order s reconstruction of $\mathbf{u}$ that satisfies (5.14), then

$$\bar{\mathbf{u}}_{i+1/2}(t) = \int_{x_i}^{x_{i+1/2}} \mathbf{p}_i\left(x, t\right)\, dx + \int_{x_{i+1/2}}^{x_{i+1}} \mathbf{p}_{i+1}\left(x, t\right)\, dx + \mathcal{O}(\Delta x)^s . \qquad (5.16)$$

However, the key to central schemes is in the evaluation of the time integrals in (5.15). Since the reconstruction of $\mathbf{u}$ occurs on the unstaggered cells I_i, $\mathbf{p}$ will be

smooth at each x_i as long as discontinuities do not propagate there from neighboring cell interfaces. This will be the case as long as $\Delta t \leq \frac{1}{2}\lambda \Delta x$, where

$$\lambda = \max_{0 \leq x \leq L} \max_{1 \leq a \leq r} \{|\lambda_a(\mathbf{u}(x))|\} . \tag{5.17}$$

Given the restriction on Δt, the time integral in (5.15) can be evaluated by replacing $\mathbf{u}$ with $\mathbf{p}$ and applying standard quadrature formulas. For an in-depth review see [83].

5.3.2.4 Central-Upwind Schemes The benefit of central schemes is that rather than try to resolve the wave structure at cell-interfaces with a Riemann solver, one may simply integrate over any discontinuities that occur between adjacent cells, as prescribed in (5.16). The result is a very simple algorithm for solving (5.15), but there is a price for this simplicity. Central schemes do not possess a semi-discrete formulation, and the numerical dissipation associated with (5.15) for an order s reconstruction will be $\mathcal{O}(\Delta x^{2s-1}/\Delta t)$ [46]. For steady state problems with small steps sizes, the cumulative effects of the numerical dissipation can degrade the final accuracy of the solution.

It is possible to recover a semi-discrete scheme by using non-uniform staggered cells and then projecting solutions from these cells onto the original unstaggered grid. This projection idea was first proposed in [38] as a way to remove the staggering which, as a practical matter, can be tedious to implement. Then in [45, 46], a non-uniform staggering was introduced to address the dissipation issue specifically. The authors found that the $\mathcal{O}(1/\Delta t)$ dissipation could be removed by introducing cells with widths just wide enough to capture any discontinuities emanating from cell

interfaces. Because point-wise interpolation with non-uniform cells is not possible [76], the finite volume formulation is a requirement here. These schemes, which we describe below, are called *central-upwind schemes*.

Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r$ be eigenvalues of the linearized flux matrix $\mathbf{A}$ and set

$$a^+_{i+1/2} = \max\left\{\lambda_r\left(\mathbf{p}_i(x_{i+1/2})\right),\ \lambda_r\left(\mathbf{p}_{i+1}(x_{i-1/2}\right),\ 0\right\}\,, \tag{5.18a}$$

$$a^-_{i+1/2} = \min\left\{\lambda_1\left(\mathbf{p}_i(x_{i+1/2}\right),\ \lambda_1\left(\mathbf{p}_{i+1}(x_{i-1/2}\right),\ 0\right\}\,. \tag{5.18b}$$

It can be shown that any discontinuity propagating from the cell interface at $x_{i+1/2}$ between I_i and I_{i+1} is contained in the the interval.

$$\tilde{I}_{i+1/2} = (x_{i+1/2,l}\,,\ x_{i+1/2,r}) \equiv (x_{i+1/2} + a^-_{i+1/2}\Delta t\,,\ x_{i+1/2} + a^+_{i+1/2}\Delta t)$$

for $\tau \in (t, t+\Delta t)$. The integral formulation of (5.10) on $\tilde{I}_{i+1/2}$ is

$$\frac{\bar{\mathbf{w}}_{i+1/2}(t+\Delta t) - \bar{\mathbf{w}}_{i+1/2}(t)}{\Delta t}\,dx + \int_t^{t+\Delta t} \frac{\mathbf{f}(\mathbf{u}\left(x_{i+1/2,r},\tau\right)) - \mathbf{f}(\mathbf{u}\left(x_{i+1/2,l},\tau\right))}{\Delta x} d\tau = 0\,, \tag{5.19}$$

where

$$\bar{\mathbf{w}}_{i+1/2}(t) = \frac{1}{x_{i+1/2,r} - x_{i+1/2,l}} \int_{x_{i+1/2,l}}^{x_{i+1/2,r}} \mathbf{u}\,(x,t)\ dx\,. \tag{5.20}$$

The remainder of the spatial domain is composed of cells

$$\tilde{I}_i = (x_{i-1/2,r}\,,\ x_{i+1/2,l})\,,$$

and the integral formulation of (5.10) on $\tilde{I}_i$ is

$$\frac{\bar{\mathbf{w}}_i(t+\Delta t)-\bar{\mathbf{w}}_i(t)}{\Delta t}\,dx+\int_t^{t+\Delta t}\frac{\mathbf{f}(\mathbf{u}\left(x_{i+1/2,l},\tau\right))-\mathbf{f}(\mathbf{u}\left(x_{i-1/2,r},\tau\right))}{\Delta x}d\tau=0\,, \quad (5.21)$$

where

$$\bar{\mathbf{w}}_i(t)=\frac{1}{x_{i+1/2,l}-x_{i-1/2,r}}\int_{x_{i-1/2,r}}^{x_{i+1/2,l}}\mathbf{u}\,(x,t)\;\,dx\,. \quad (5.22)$$

We now outline the steps of the semi-discrete scheme.

1. Given cell averages $\bar{\mathbf{u}}_i(t)$, construct an approximation $\mathbf{p}$ of $\mathbf{u}$ of desired spatial accuracy.

2. Use the reconstruction $\mathbf{p}$ to compute $\bar{\mathbf{w}}_i(t)$ via (5.22) and $\bar{\mathbf{w}}_{i+1/2}(t)$ via (5.20).

3. Replace $\mathbf{u}$ with $\mathbf{p}$ in the flux integrals in (5.19) and (5.21). Then use standard quadrature formulas to evaluate the integrals.

4. Update $\bar{\mathbf{w}}_i$ and $\bar{\mathbf{w}}_{i+1/2}$ using (5.19) and (5.21) and the calculations from steps 2 and 3.

5. Find a polynomial reconstruction $\mathbf{q}$ that interpolates the averages $\bar{\mathbf{w}}_i(t+\Delta t)$ and $\bar{\mathbf{w}}_{i+1/2}(t+\Delta t)$.

6. Use $\mathbf{q}$ to compute $\bar{\mathbf{u}}_i(t+\Delta t)$ with (5.12).

It is shown in [46] that in the limit $\Delta t \to 0$, that fully discrete formulation recovers

the semi-discrete homogeneous form of (5.11) with numerical flux

$$\begin{aligned}\mathbf{F}_{i+1/2} =& \frac{a^+_{i+1/2}\mathbf{f}_{i+1/2}(\mathbf{p}_{i+1}(x_{i+1/2})) - a^-_{i+1/2}\mathbf{f}_{i+1/2}(\mathbf{p}_i(x_{i+1/2}))}{a^+_{i+1/2} - a^-_{i+1/2}} \\ &+ \frac{a^+_{i+1/2}a^-_{i+1/2}}{a^+_{i+1/2} - a^-_{i+1/2}}(\mathbf{p}_{i+1}(x_{i+1/2}) - \mathbf{p}_i(x_{i+1/2})\,.\end{aligned} \tag{5.23}$$

Notice that there is no explicit dependence on the intermediate staggered averages $\bar{\mathbf{w}}_i$ and $\bar{\mathbf{w}}_{i+1/2}$ or the reconstruction $\mathbf{q}$.

5.3.3 Remaining Discretization

We discretize the fluxes in (5.11) using the central-upwind flux. Then a spatial discretization of the diffusive, field, and collision terms must be given, and a temporal discretization must be specified.

To discretize the remaining terms in (5.11), we use the fact that for any smooth function ψ

$$\frac{1}{\Delta x}\int_{I_i}\psi(\mathbf{u}(x,t))\,dx = \psi(\bar{\mathbf{u}}_i(t)) + \mathcal{O}(\Delta x^2)\,. \tag{5.24}$$

The diffusive terms and the electric potential can be safely discretized with center

differences, which in conjunction with (5.24) gives

$$\frac{1}{\Delta x}\int_{I_i}\mathbf{l}(\mathbf{u}(x,t))\Phi_x(x)\,dx = \mathbf{l}(\bar{\mathbf{u}}_i(t))\left[\frac{\Phi_{i+1}-\Phi_{i-1}}{2\Delta x}\right] + \mathcal{O}(\Delta x^2)\,, \tag{5.25a}$$

$$\mathbf{D}(\mathbf{u}(x_{i+1/2},t)) = \left[\frac{\mathbf{D}(\bar{\mathbf{u}}_i(t)) + \mathbf{D}(\bar{\mathbf{u}}_{i+1}(t))}{2}\right] + \mathcal{O}(\Delta x^2)\,, \tag{5.25b}$$

$$\mathbf{g}(\mathbf{u})_{\mathbf{x}}(x_{i+1/2},t) = \left[\frac{\mathbf{g}(\bar{\mathbf{u}}_{i+1}(t)) - \mathbf{g}(\bar{\mathbf{u}}_{i-1}(t))}{2\Delta x}\right] + \mathcal{O}(\Delta x^2) \tag{5.25c}$$

$$\frac{1}{\Delta x}\int_{I_i}\mathbf{r}(\mathbf{u}(x,t)) = \mathbf{r}(\bar{\mathbf{u}}_i(t)) + \mathcal{O}(\Delta x^2)\,. \tag{5.25d}$$

The Poisson equation is discretized with standard central differences. Since ϵ is constant in the benchmark problem,

$$\frac{-\Phi_{i+1}+2\Phi_i-\Phi_{i+1}}{\Delta x^2} = \frac{q}{\epsilon}\left(\bar{D}_i - \bar{n}_i\right) + \mathcal{O}(\Delta x)$$

where the bar denotes averages over a given cell. Finally, the averages $\bar{\mathbf{u}}_i$ are evolved with a simple forward Euler method:

$$\frac{d}{dt}\bar{\mathbf{u}}_i = \frac{\bar{\mathbf{u}}_i(t+\Delta t) - \bar{\mathbf{u}}_i(t)}{\Delta t} + \mathcal{O}(\Delta t)\ ,$$

which should be sufficient for steady-state calculations. Higher-order methods are easily implementable for studying transient behavior [76].

5.3.4 Remarks

Some remarks about the choices for our scheme are in order. We freely acknowledge that the scheme outlined above is not the most efficient or well-behaved from a computational viewpoint. However, our primary interest at this point is to compare the qualitative features of the models. We briefly mention a few issues here so that the reader may be aware of them. Some of these issues will be addressed in the next chapter in the development of better schemes.

1. **Temporal accuracy and stiffness**. Because we are interested primarily in steady-state calculations, the use of a first-order in time method should suffice. However, there is also a problem with efficiency. The system (5.9) in its non-dimensional form may take on the scaling

$$\mathbf{u}_t + \frac{1}{\varepsilon}\mathbf{f}(\mathbf{u})_x = \frac{1}{\varepsilon}\mathbf{l}(\mathbf{u})\Phi_x + \left(\mathbf{D}(\mathbf{u}) \cdot \mathbf{g}(\mathbf{u})_x\right)_x + \frac{1}{\varepsilon^2}\mathbf{r}(\mathbf{u})$$

in the drift-diffusion limit or

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \frac{1}{\varepsilon}\mathbf{l}(\mathbf{u})\Phi_x + \left(\mathbf{D}(\mathbf{u}) \cdot \mathbf{g}(\mathbf{u})_x\right)_x + \frac{1}{\varepsilon}\mathbf{r}(\mathbf{u})$$

in the drift-collision limit. In cases where ε is small, these equations become stiff, in which case the use of implicit schemes is in order. This can be particularly difficult for the first scaling since the fluxes are approximated in a highly nonlinear fashion. Furthermore, if the diffusive terms are nonzero, than an explicit scheme forces an additional restriction in the time step $\Delta t \sim \Delta x^2$. These

issues are addressed in the next chapter.

2. **Spatial accuracy**. Increasing the order of spatial accuracy in (5.24) with a finite volume method requires a reconstruction method that incorporates function averages from adjacent cells. The best option in this case is to change to a finite difference formulation. However, this means abandoning the central-upwind approach, and for this reason, we limit the spatial accuracy to second-order.

3. **Well-balanced schemes**. The use of central differences in (5.25) means that the scheme will not be well-balanced. *Well-balanced* schemes are numerical schemes that formulate non-conservative terms into a conservative framework. This means that the semi-discrete scheme for the balance law (5.9) would have the form

$$\frac{d}{dt}\bar{\mathbf{u}}_i(t) = -\frac{\hat{\mathbf{F}}_{j+1/2} - \hat{\mathbf{F}}_{j-1/2}}{\Delta x},$$

in analogy with the semi-discrete formulation of the homogeneous equation (5.10). This type of formulation is used to preserve certain properties of a numerical solution such a positivity or a particular steady state. Usually, the form of $\hat{\mathbf{F}}$ depends heavily known information about the solution. Examples where well-balanced schemes have proven fruitful can be found in [6] and [24]. The problems studied in these cases have a form similar to (5.9) but are generally much simpler than the models we are interested in here.

The lack of a well-balanced scheme will be evident in our benchmark computations at the source and drain junctions of the diode, most noticeably is the

results for the current $J = -qnu$. (Current and momentum differ only by a constant). For most of the diode, the computed steady-state current will be constant, as expected. However, large oscillations will appear at the junctions. Such behavior can be found in similar computations [3, 9, 16, 25].

4. **Hyperbolicity**. The reader should be reminded that it is not known whether the schemes APNV1, APPV1, APFV1, PMMC1, and PGMC1 are really hyperbolic. Because they are all based on pertubations and/or reductions of hyperbolic systems, it is reasonable to believe they are hyperbolic in some non-trivial subset the state space of densities. Whether or not such a subset incorporates all physically realized values is unknown. For computational purposes, we ignore this fact and use the wave structure of their hyperbolic counterparts to compute λ given by (5.17) and the values $a^{\pm}_{i+1/2}$ given by (5.18). This means we use the wave structure of APNV2 for computing APNV1 and the wave structure of APPV1 for computing APPV2 an so on. Experience has shown that the same time step restrictions are required for stable computations.

5.4 Numerical Results

Below we present the results of calculations. Each simulation is allowed to run until the following stop criterion is reached

$$\frac{\sum_{i=1}^{N} [n_i(t_k) - n_i(t_k + \Delta t_k)]}{\sum_{i=1}^{N} n_i(t_k)} \leq tol \cdot \Delta t_k$$

Here we set the tolerance $tol = 10^{-4}$.

5.4.1 Bløtekjær-Type Models

Figures 5.2-5.13 contain results for the six Bløtekjær-type models. Most figures consists of six subplots shadowed by the corresponding data taken from Monte-Carlo experiments. The exception is Figure 5.9, which shows the electric fields. No Monte Carlo data was available for this figure.

Figures 5.2 and 5.3 are the electron concentration. Figure 5.2 shows that each model has the same basic behavior. One must refer to Figure 5.3 to see the differences. From these rescaled pictures, several things are clear. First is the need to add corrections to the straight-forward Maxwellian closure (MBW, MMC). Second is effect of a non-zero heat flux, which has the greatest effect at the drain junction even most of the models do not necessary give improved results. Finally, the accuracy gained by using Monte-Carlo relaxation times is very noticeable. Quite surprisingly the model BMC2 gives among the best results of all the models presented in the chapter. It would be interesting to see if this accuracy is robust by varying the device parameters.

In Figure 5.4, the Monte Carlo relaxation times give slightly better results. However, the most noticeable features in Figure 5.4 is the presence of large oscillations at the junctions $x = 0.1$ and $x = 0.5$. This will be the case for all current figures presented here.

Velocity results can be seen in Figure 5.5 and with a zoomed view in Figure 5.6. As with most of the models in this chapter, the majority of Bløtekjær-type models underestimate the velocity in the channel region yet display a velocity overshoot

effect at the drain junction that is characteristic of hydrodynamic models. Changes in the heat diffusion coefficient κ make a significant difference in both areas. Note again the accuracy of models BMC1 and BMC2 which use Monte-Carlo relaxations times as compared to the BBW1 and BBW2 models which use Baccarani-Wordeman relaxation times.

Temperature results are given in Figure 5.7 in units of thermal energy. In addition to the remarkable accuracy of model BMC2, we note the small spike just before the drain junction in the MBW and MMC models. This spike is not a numerical defect, but rather a small shock in the temperature profile which is smoothed away by the heat dissipation in the other models. The energy profile in Figure 5.8 shows the same behavior.

The electric field results are displayed in Figure 5.9. We have no Monte Carlo data to compare here and thus use the MBW model as a reference. Note the large spike in the electric field at the drain junction which is made sharper by the addition of a heat flux

Heat flux and energy flux results are given in Figures 5.10-5.13. None of the models correctly predict the heat flux, although the models BBW1 and BMC1 are closer than the BBW2 and BMC2 models. Even so, the energy flux for BBW2 and BMC2 is much closer to Monte Carlo data than BBW1 and BMC1 at the drain junction. Once again, the BMC2 model is surprisingly accurate.

5.4.2 Anile-Pennisi Models

The results for the six AP models are given in Figures 5.14-5.24. The basic features are the same as the Bløtekjær models. Among the AP models, variation in results depends mostly on the presence of diffusive terms in momentum equations. This fact is apparent in Figures 5.15, 5.17 and 5.18, where the full viscosity models APFV1 and APFV2 gives much improved results near the drain. However, there is a noticeable degradation in accuracy for the temperature results in Figure 5.19.

There seems to be very little difference between models that do or do not include the convective heat flux correction in the energy equations. The most noticeable difference is seen in the heat flux results themselves, given in Figure 5.22. Here the models APNV1, APPNV2, and APNV3 are much more accurate than there counterparts APNV2, APNV3, and APNV4 in the channel of the diode. However, they show no substantial improvement at the drain junction, which is always the most problematic area. This is consistent with our earlier stance that convective corrections may be dominated by diffusive corrections at places what gradients are large—such as at the drain junction. Also, the increased accuracy in the heat flux actually translates to decreased accuracy in the energy flux, which is the flux that actually drives the energy equation. This last fact should not be considered a weakness of the AP model; rather one should that there is still important dynamics that has not yet been given a thorough accounting.

5.4.3 Entropy-Based Models

Results for the entropy based models are given in Figures 5.25-5.35. These figures include the same type of results as the Anile-Pennisi and Bløtekjær-type models with the same general behavior. In addition, Figure 5.34 compares the anisotropy σ of several of the models and Figure 5.35 compares the longitudinal and transverse temperature components of the perturbed Gaussian models.

From Figures 5.25-5.30, we make several observations. First, the electron concentration in the MMC an GMC models is very similar—both cases show the need for a non-zero heat flux near the drain junction. The perturbed Maxwellian closures are more accurate than their Gaussian counterparts near the drain but slightly less accurate near the source and along the center of the channel. The current results from the perturbed Gaussian models are also slightly more accurate.

The perturbed Maxwellian closures have a significantly smaller velocity overshoot than their Gaussian counterparts. This is due to the fact that the anisotropy in the Gaussian model is not a diffusive correction. However, as with the different AP models, there is a trade-off that comes in the form of degraded temperature results. Although the velocities in PGMC1 and PGMC2 show a significant overshoot at the drain junction, this overshoot is much smaller than most of the other models that do not include diffusive effects in the momentum equation. Again, the notable exception is BMC2. A marked similarity of these two models in a smaller diffusive heat flux. The heat diffusivity for PGMC1 and PGMC2 is significantly greater than that of BMC2, but it is less than for all other models.

The anisotropy for APFV1, APFV2, PMMC1, PMMC2, PGMC1, and PGMC2 is plotted in Figure 5.34. Here one can see the similarities in the diffusive models APFV1, APFV2, PMMC1, PMMC2 and also how they differ from the Gaussian models PGMC1 and PGMC2.

For the most part, the differences between PMMC1 and PMMC2 and between PGMC1 and PGMC2 are small. The model PGMC1 is slightly better predictor of the temperature than is PGMC2, but the inclusion of convective corrections is most notable in the heat flux data, although none of the entropy based models is as accurate in this respect as the APNV1, APPV1, and APFV1 models are. Similar to the AP results, the increased accuracy in the heat flux translates to a loss of accuracy in the energy flux, and moreover, the heat flux results at the drain junction are still quite inaccurate for all of the models considered.

Finally, Figure 5.35 gives a comparison of the temperature components for the perturbed Gaussian models. The values θ_L and θ_T do not suffer any significant changes when moving from PGMC1 to PGMC2. In both cases, $\theta_L > \theta_T$. This is expected since the kinetic distribution is stretched in the longitudinal direction by the electric field.

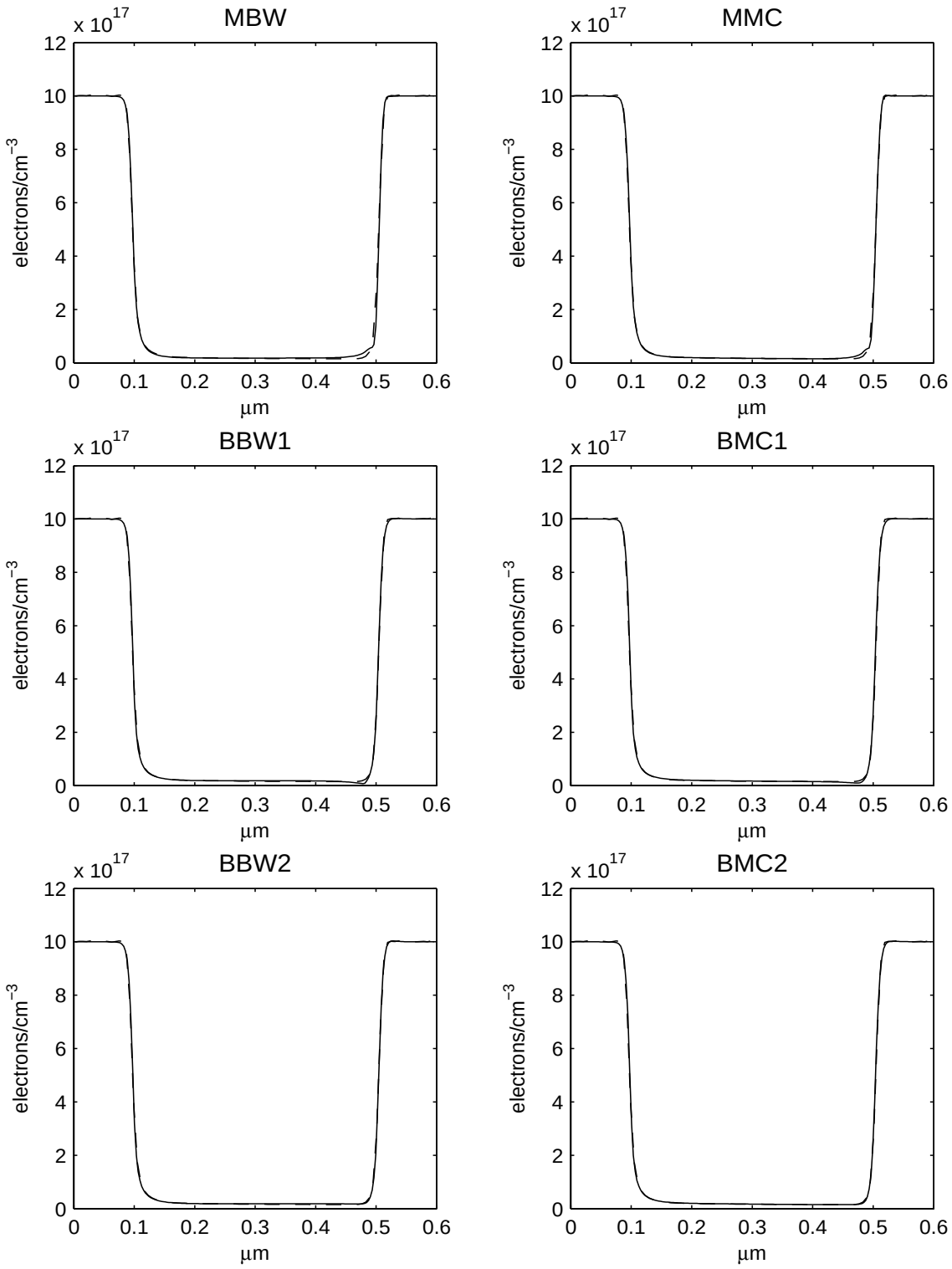

Figure 5.2: Electron concentration n for Bløtekjær-type models. Dashed line is Monte Carlo data.

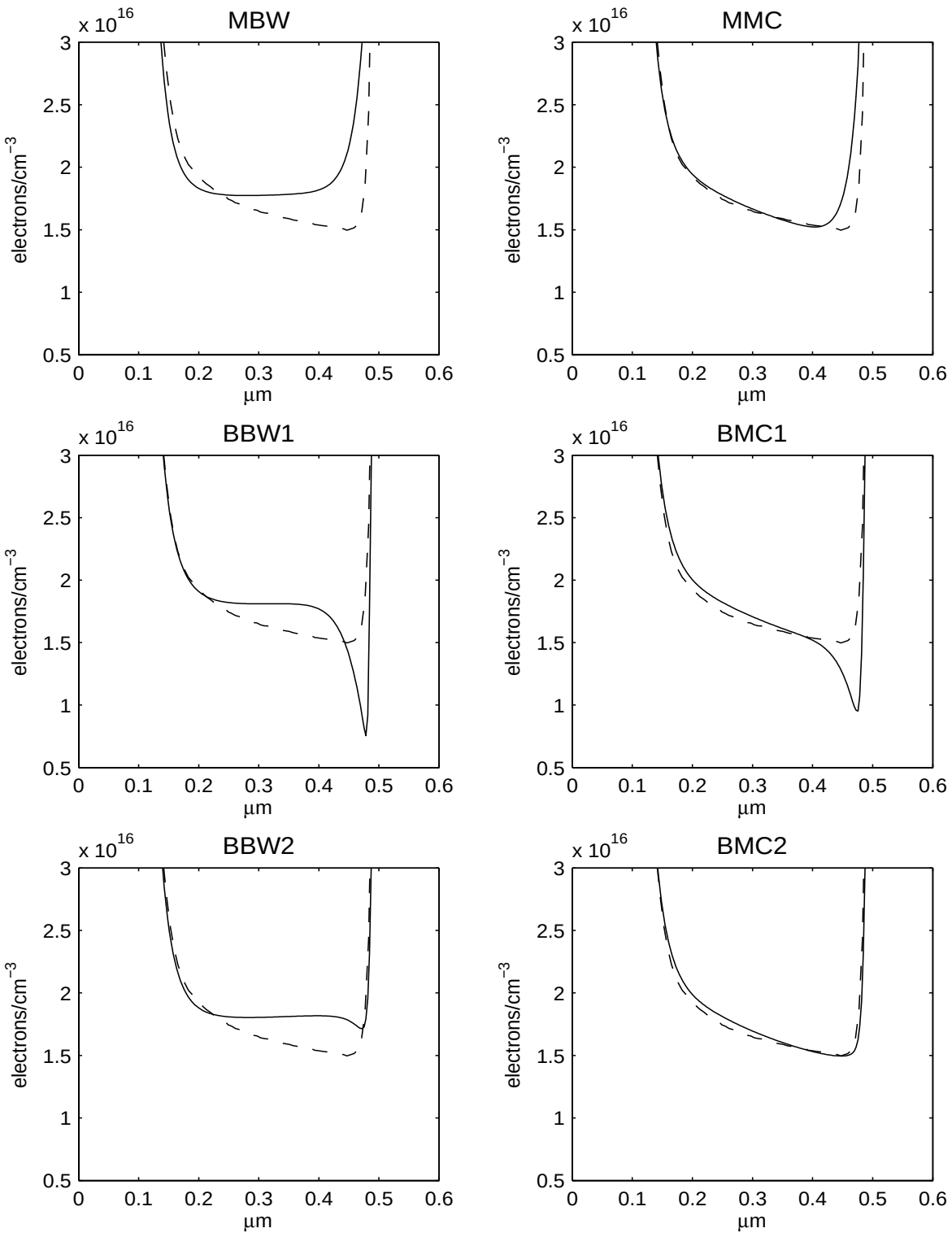

Figure 5.3: Electron concentration n for Bløtekjær-type models, magnified view. Dashed line is Monte Carlo data.

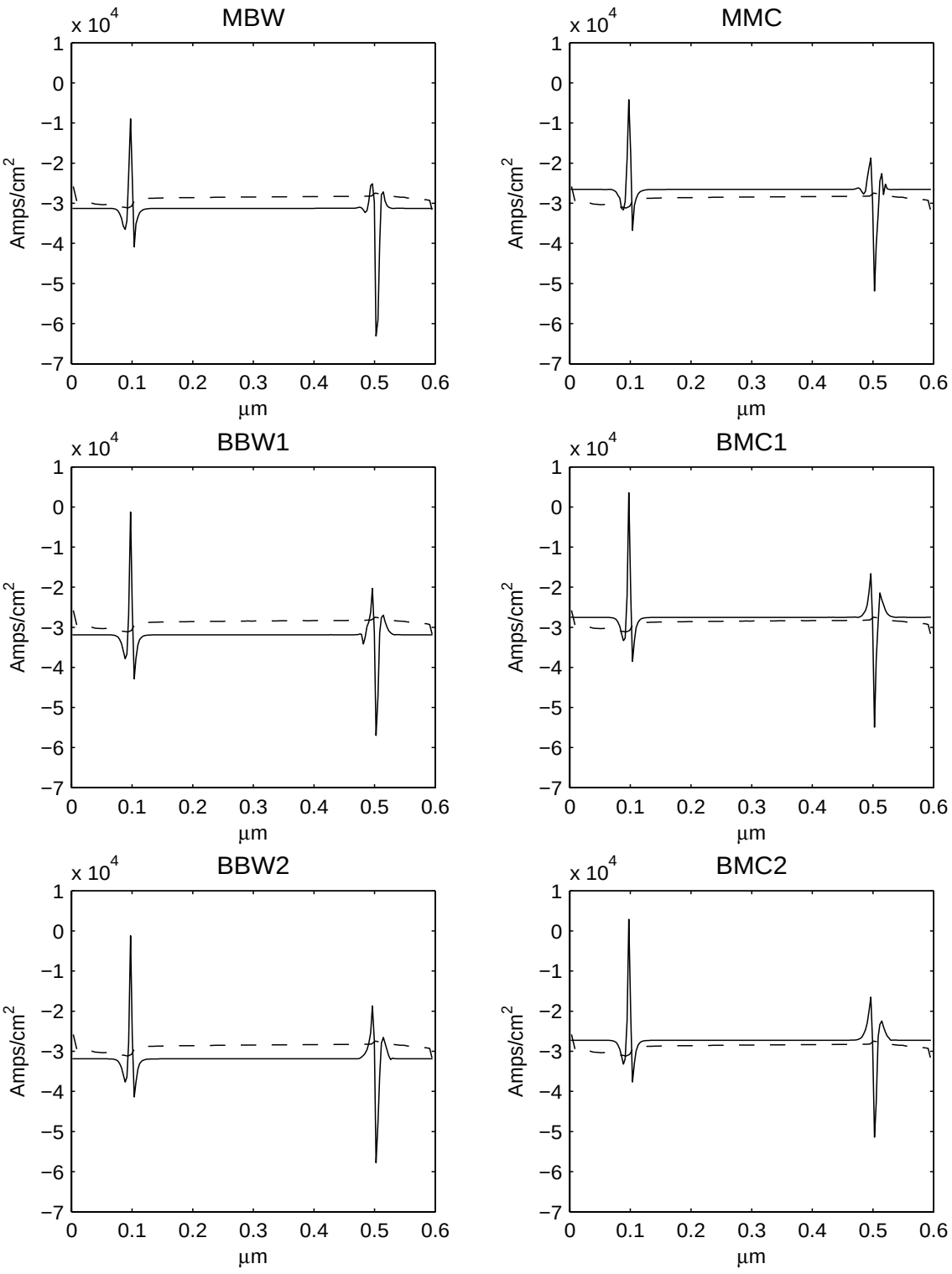

Figure 5.4: Electron current $J = -qnu$ for Bløtekjær-type models. Dashed line is Monte Carlo data.

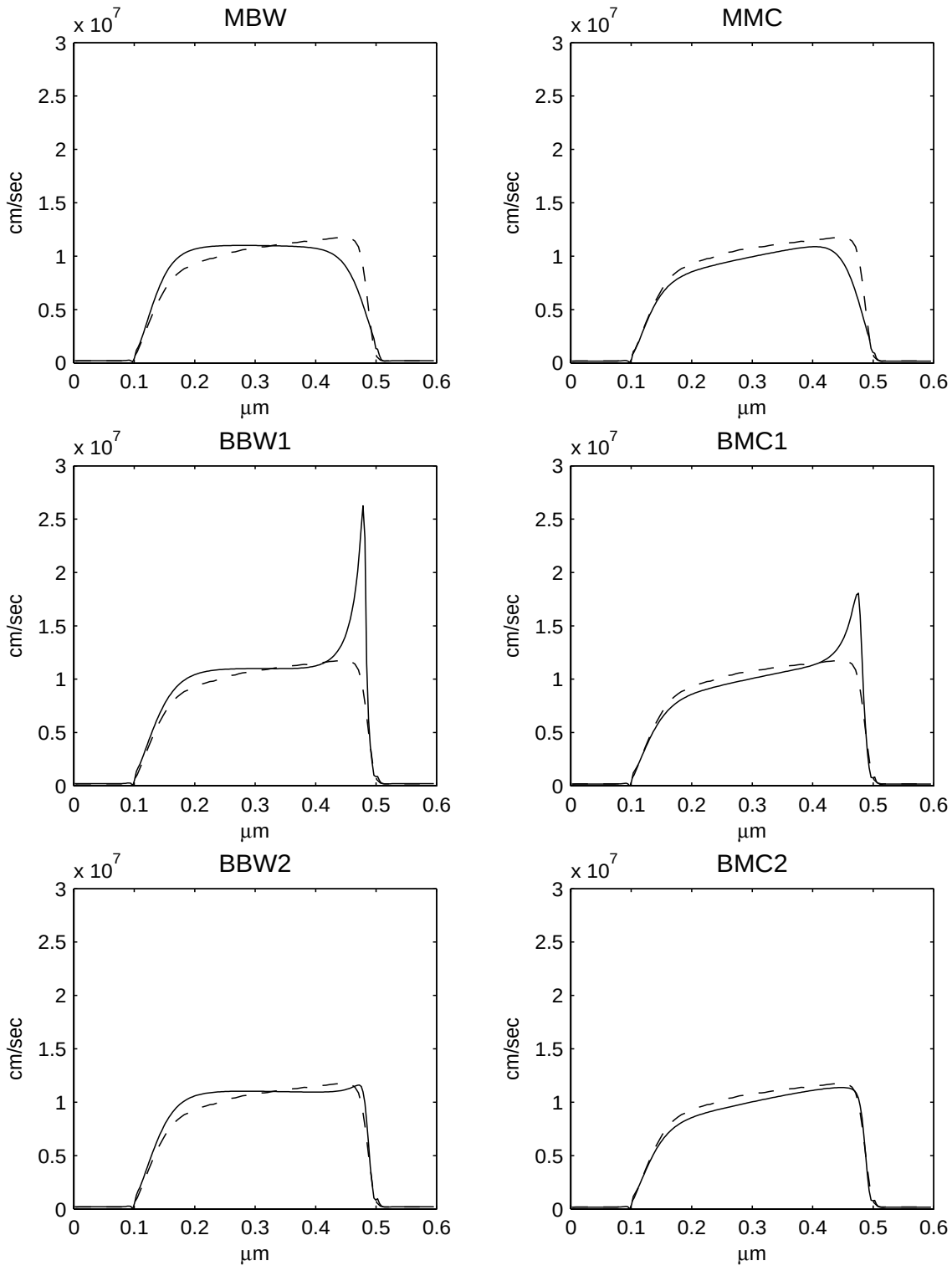

Figure 5.5: Electron velocity u for Bløtekjær-type models. Dashed line is Monte Carlo data.

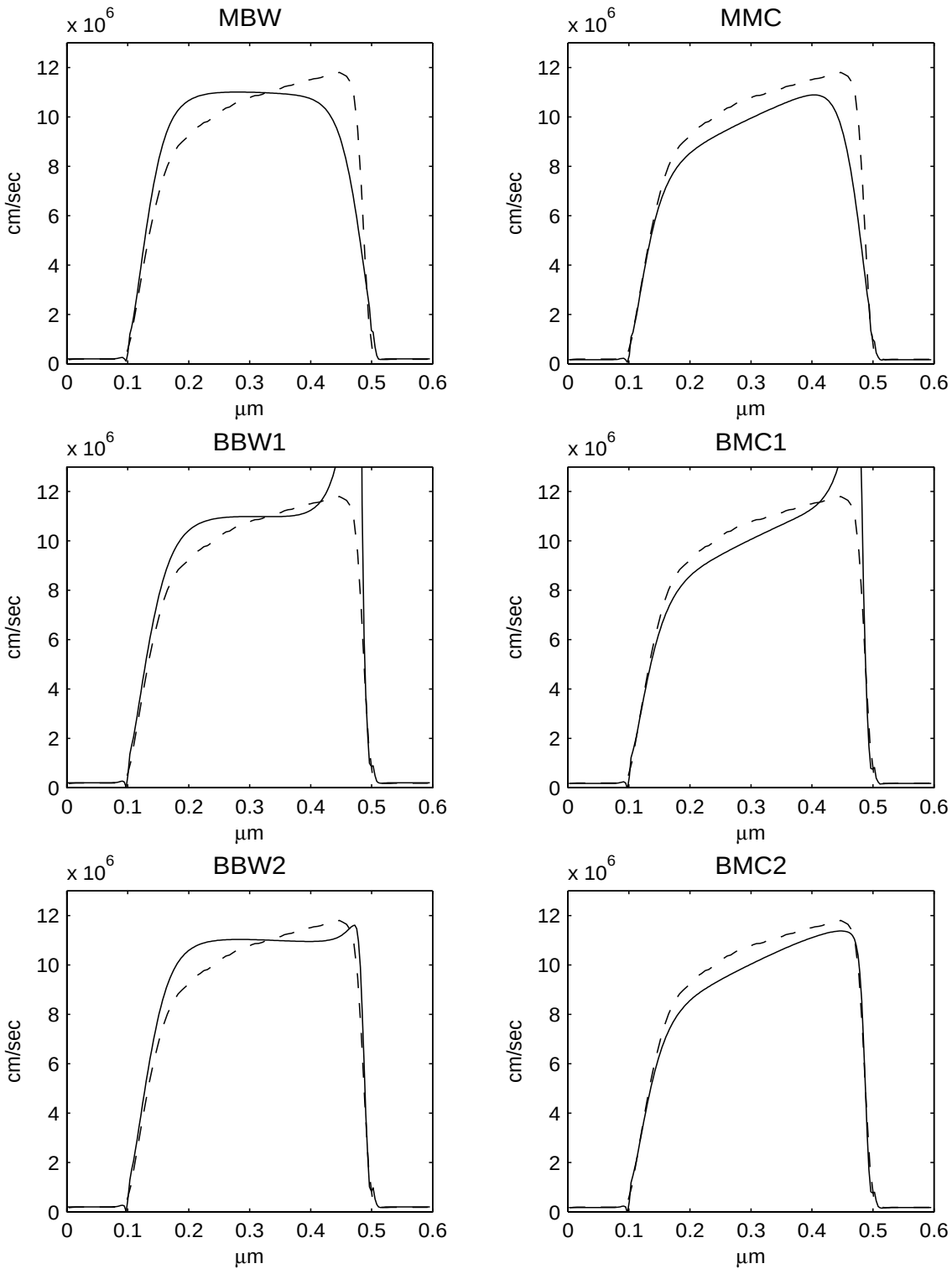

Figure 5.6: Electron velocity u for Bløtekjær-type models, magnified view. Dashed line is Monte Carlo data.

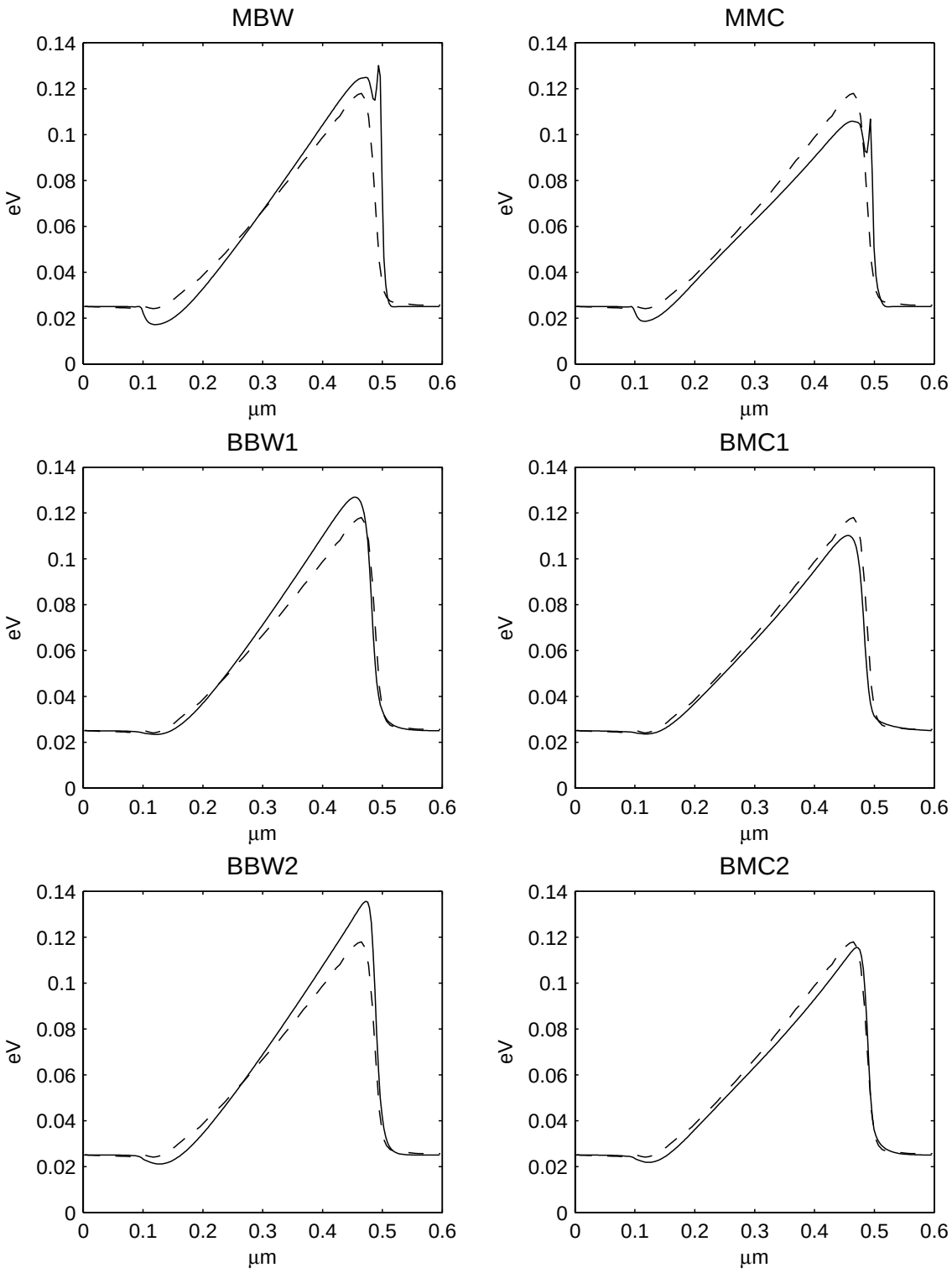

Figure 5.7: Thermal energy $m_e\theta$ for Bløtekjær-type models. Dashed line is Monte Carlo data.

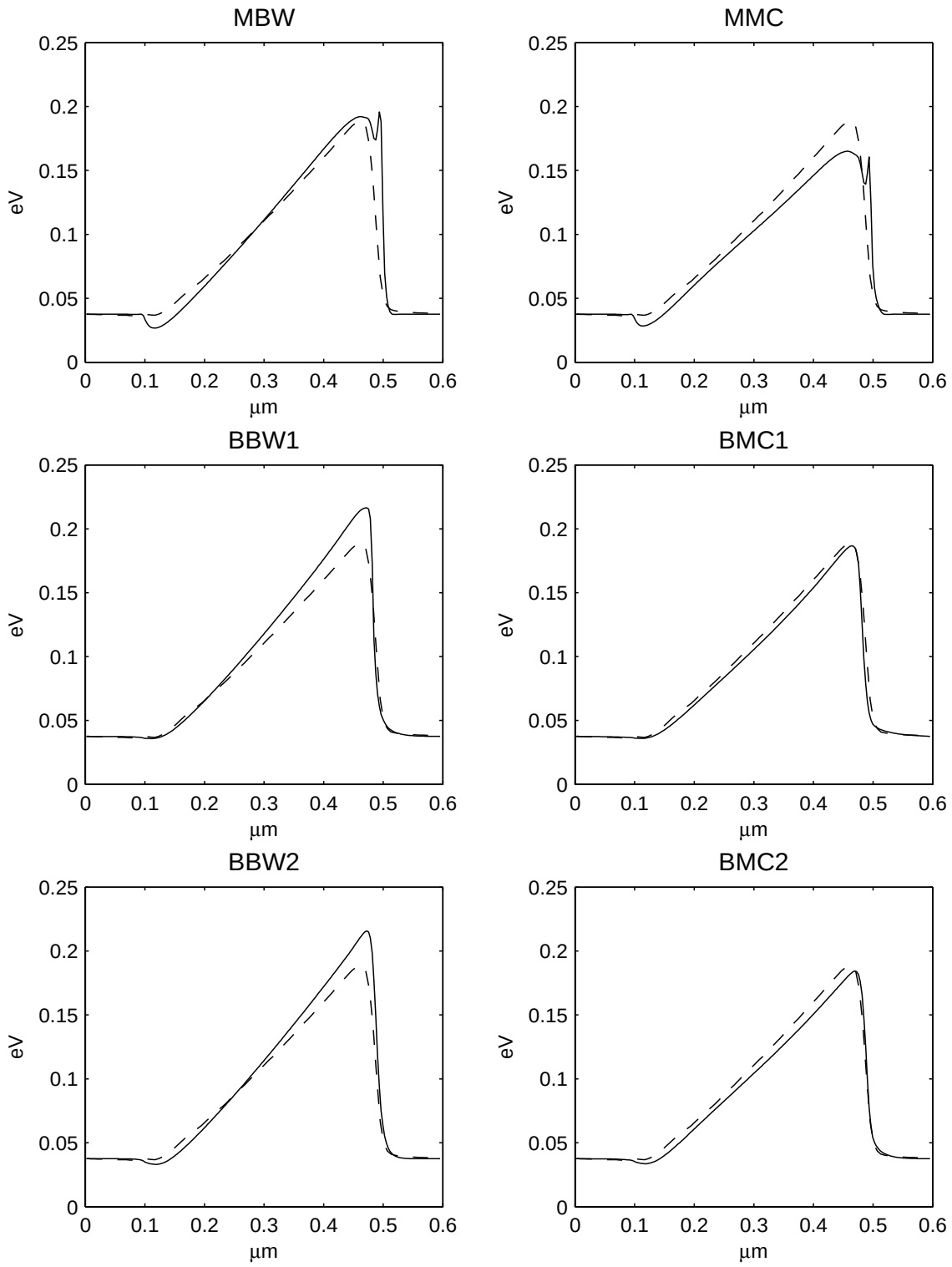

Figure 5.8: Electron energy $\frac{1}{2}m_e u^2 + \frac{3}{2}m_e\theta$ for Bløtekjær-type models. Dashed line is Monte Carlo data.

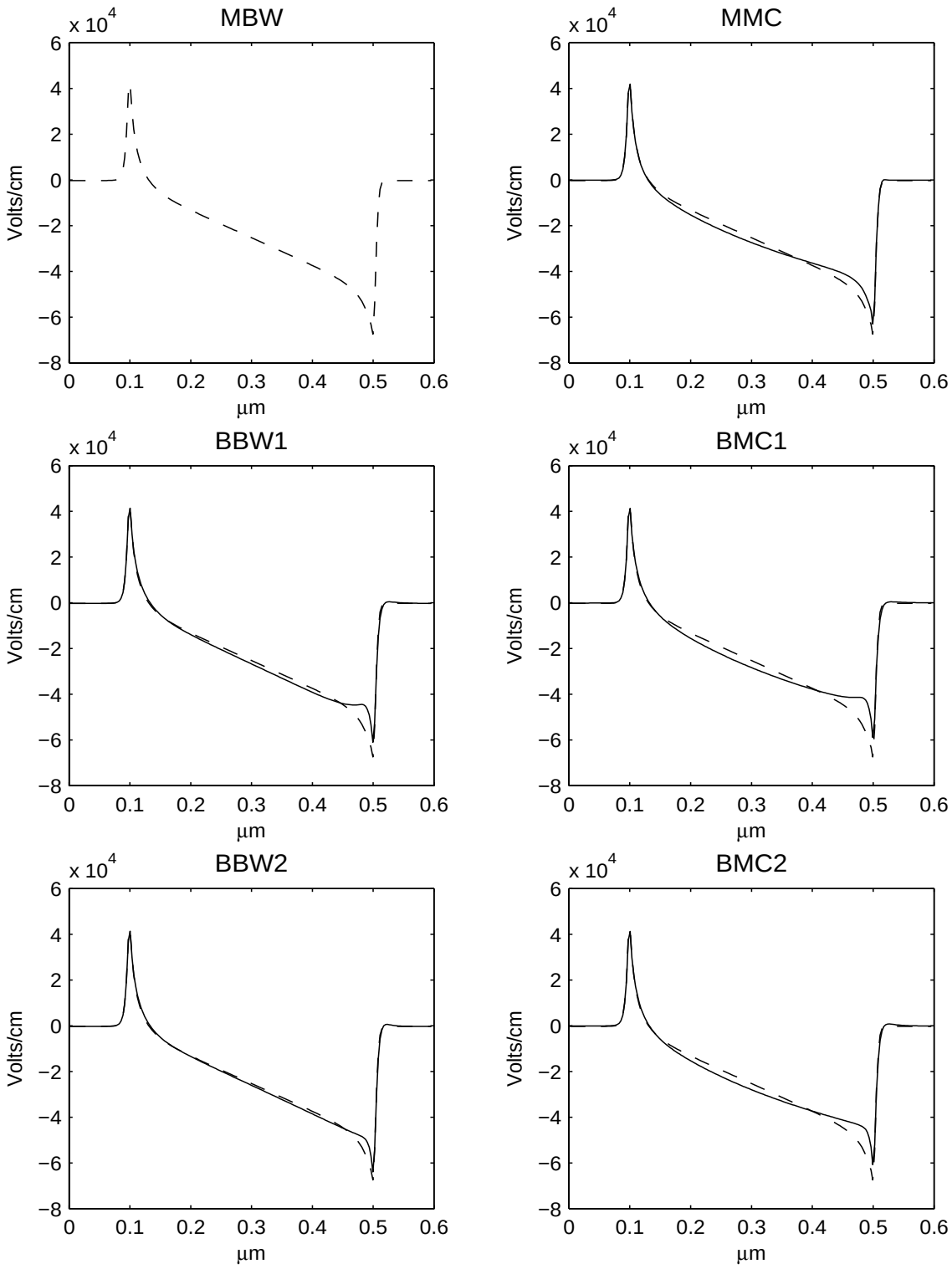

Figure 5.9: Electric field $E = -\partial_x \Phi$ for Bløtekjær-type models. Dashed line is MBW model.

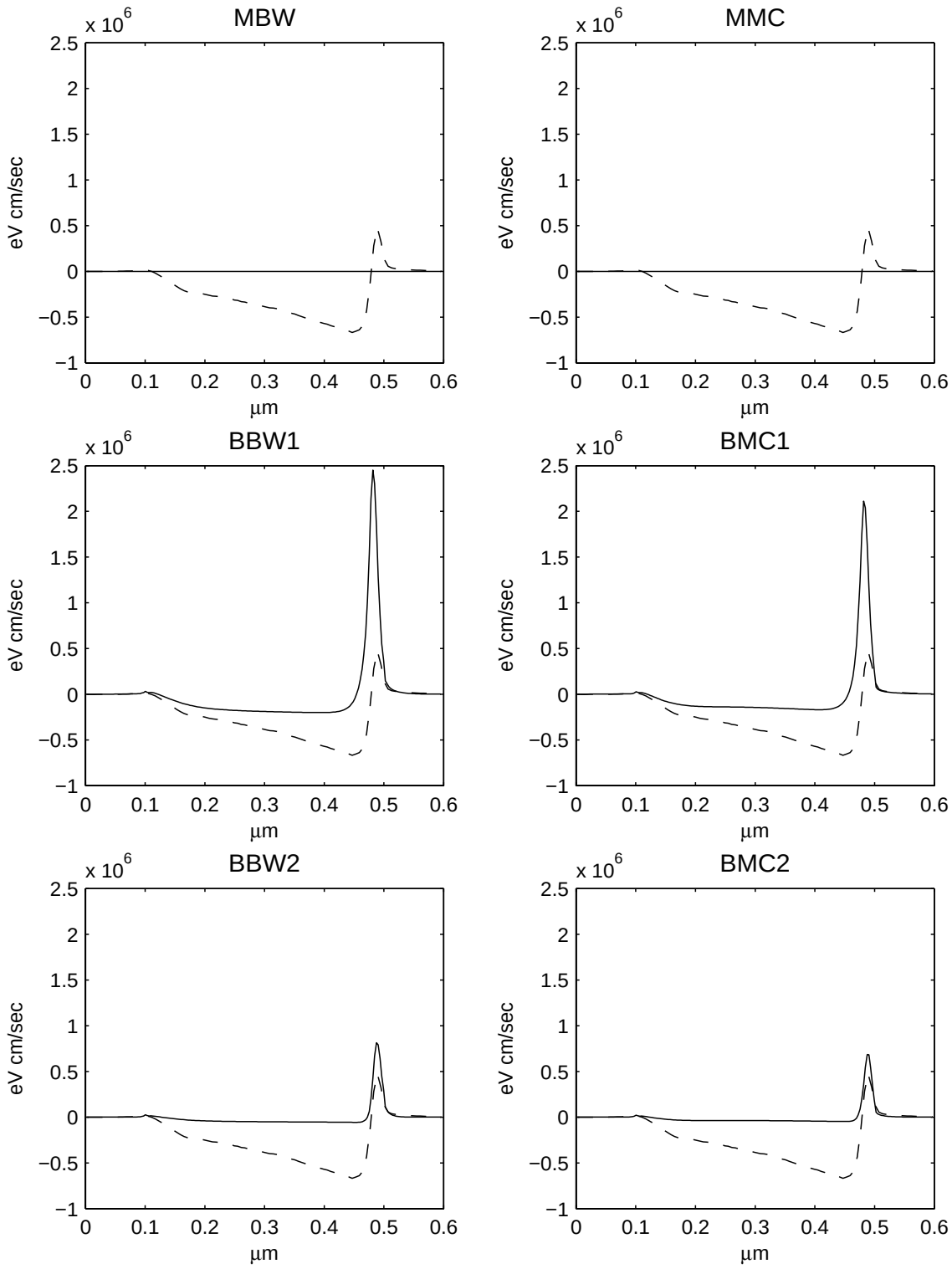

Figure 5.10: Heat flux $\frac{m_e}{n}q$ for Bløtekjær-type models. Dashed line is Monte Carlo data.

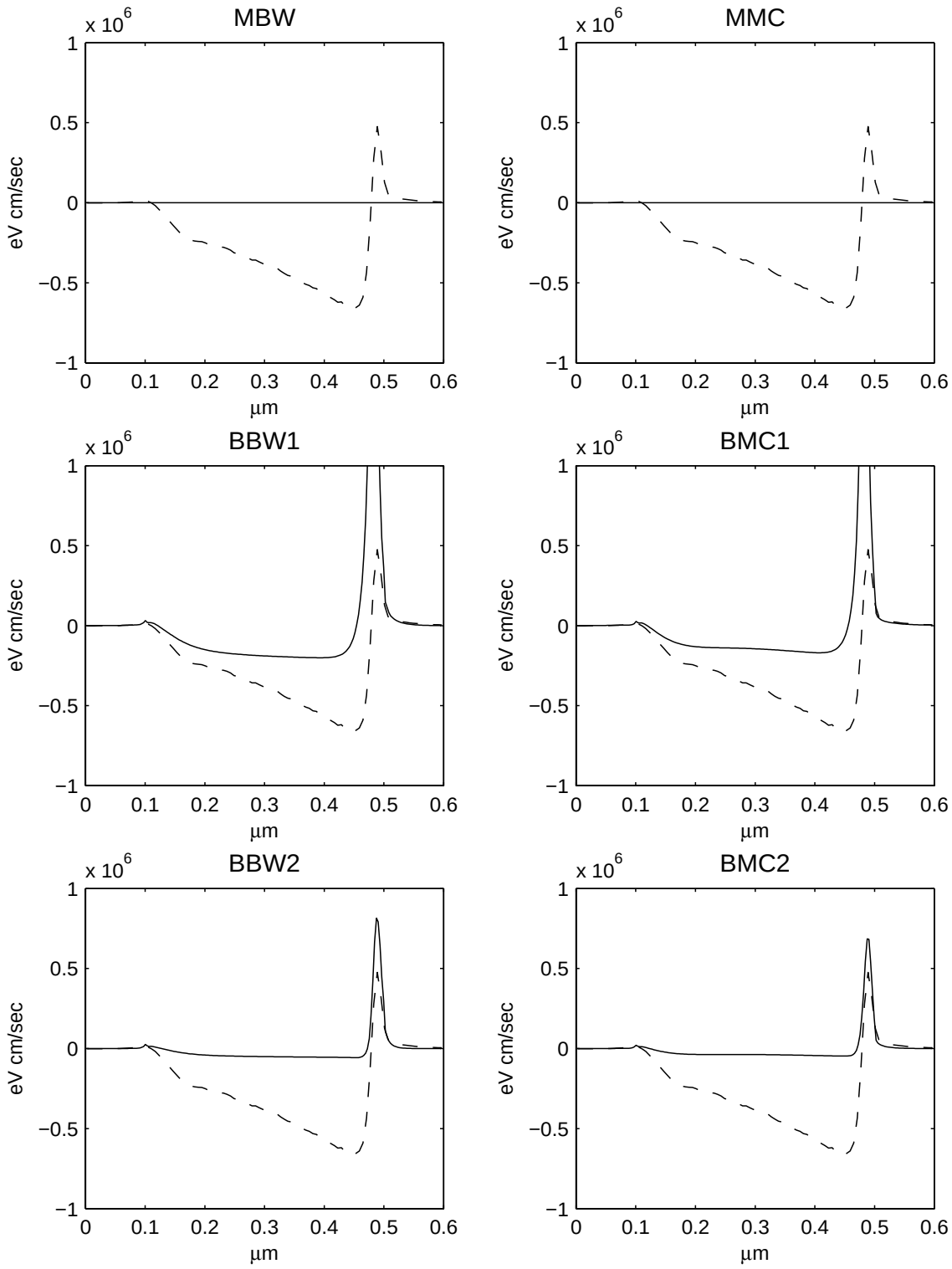

Figure 5.11: Heat flux $\frac{m_e}{n}q$ for Bløtekjær-type models, magnified view. Dashed line is Monte Carlo data.

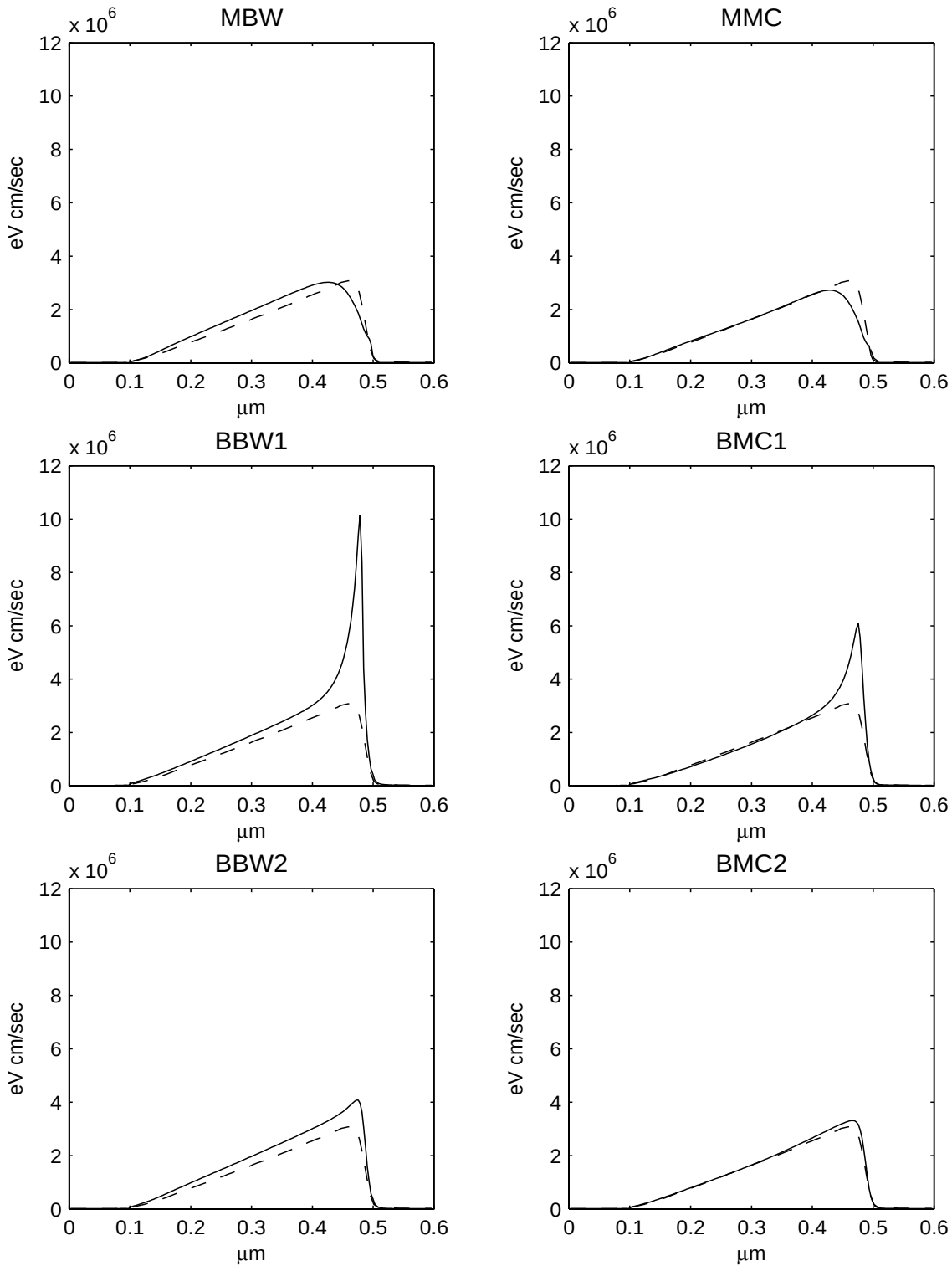

Figure 5.12: Electron energy flux $\frac{1}{2}m_e u^3 + \frac{5}{2}m_e\theta + m_e\frac{q}{n}$ for Bløtekjær-type models. Dashed line is Monte Carlo data.

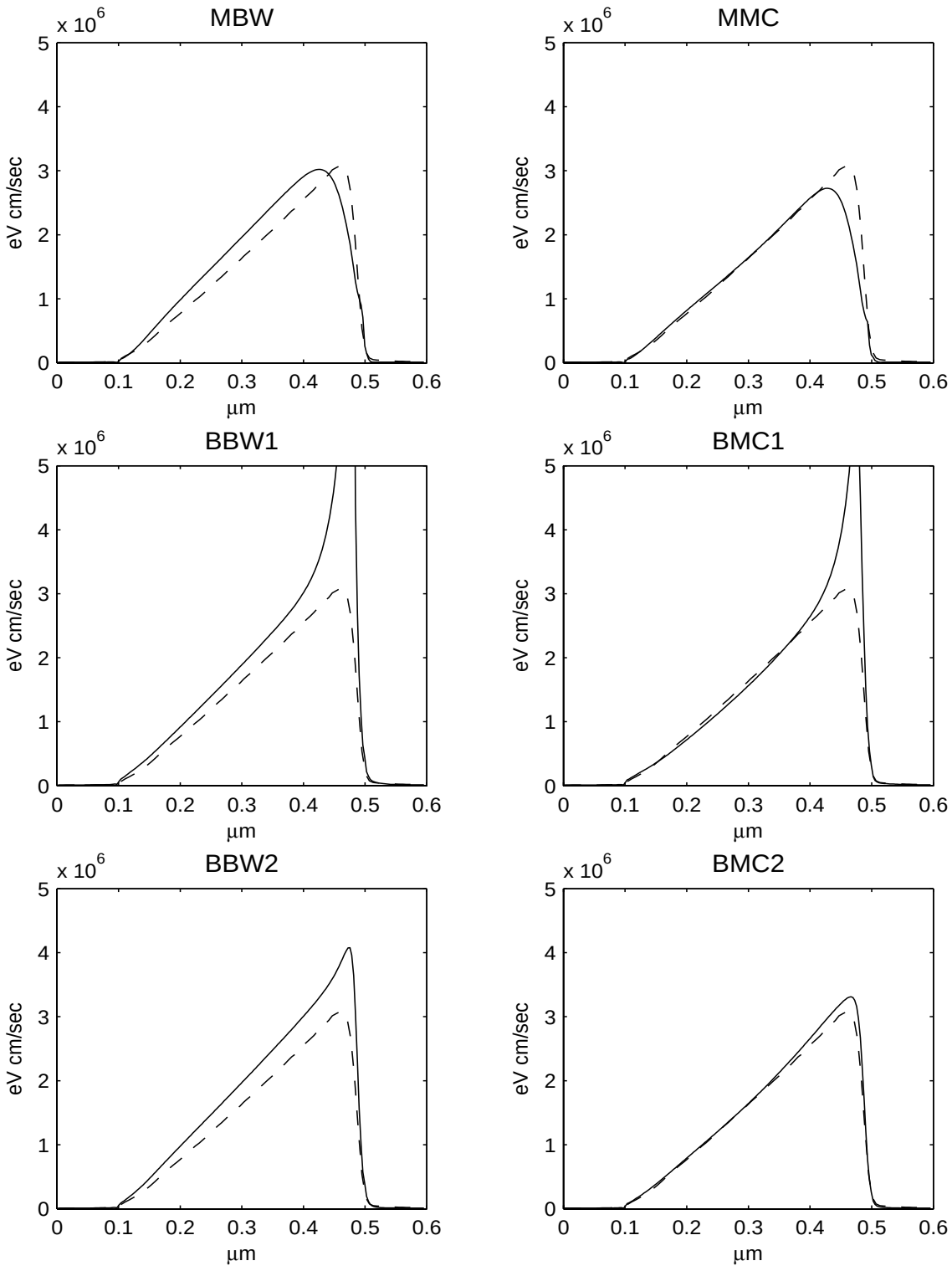

Figure 5.13: Electron energy flux $\frac{1}{2}m_e u^3 + \frac{5}{2}m_e\theta + m_e\frac{q}{n}$ for Bløtekjær-type models, magnified view. Dashed line is Monte Carlo data.

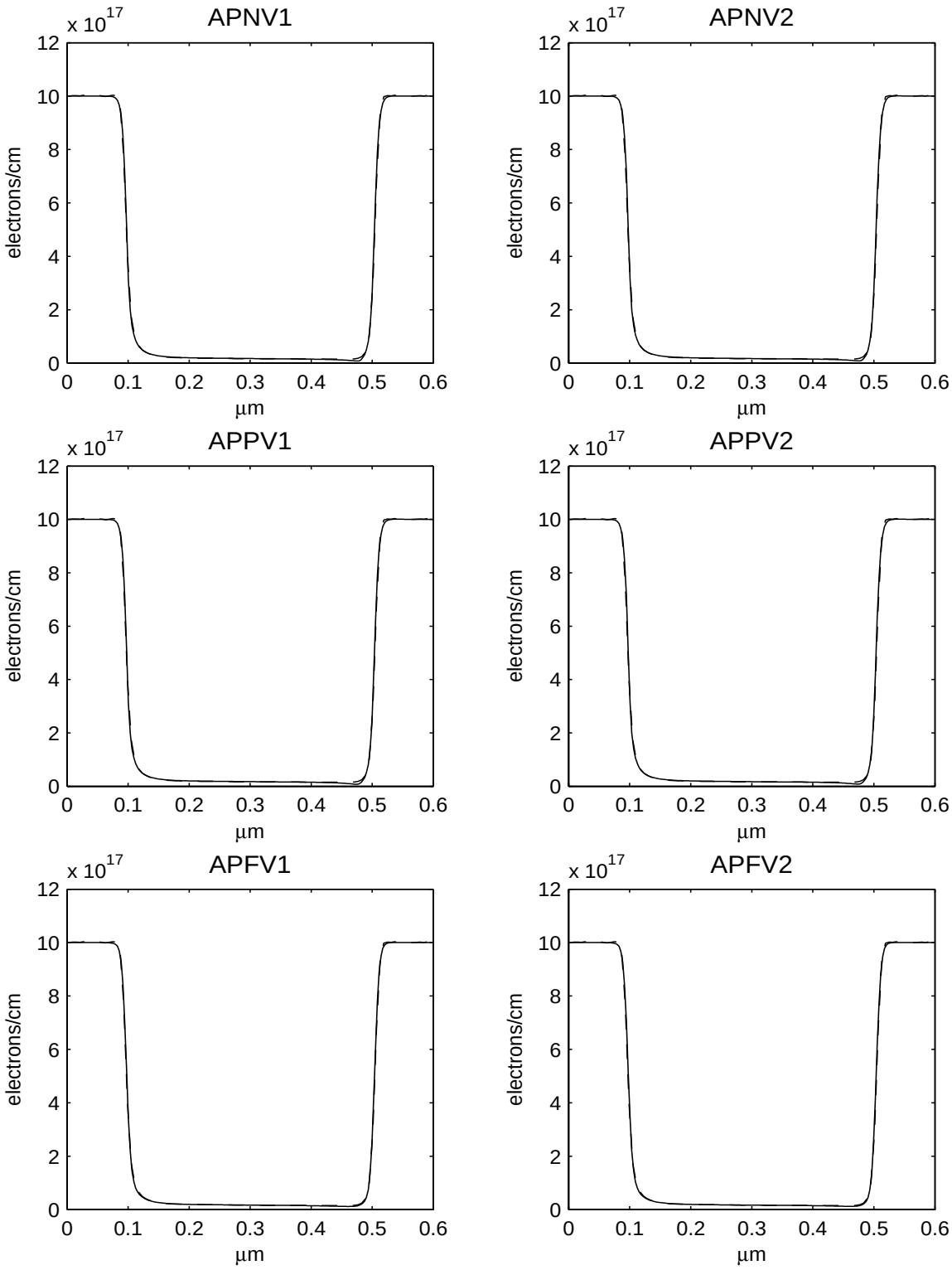

Figure 5.14: Electron concentration n for AP models. Dashed line is Monte Carlo data.

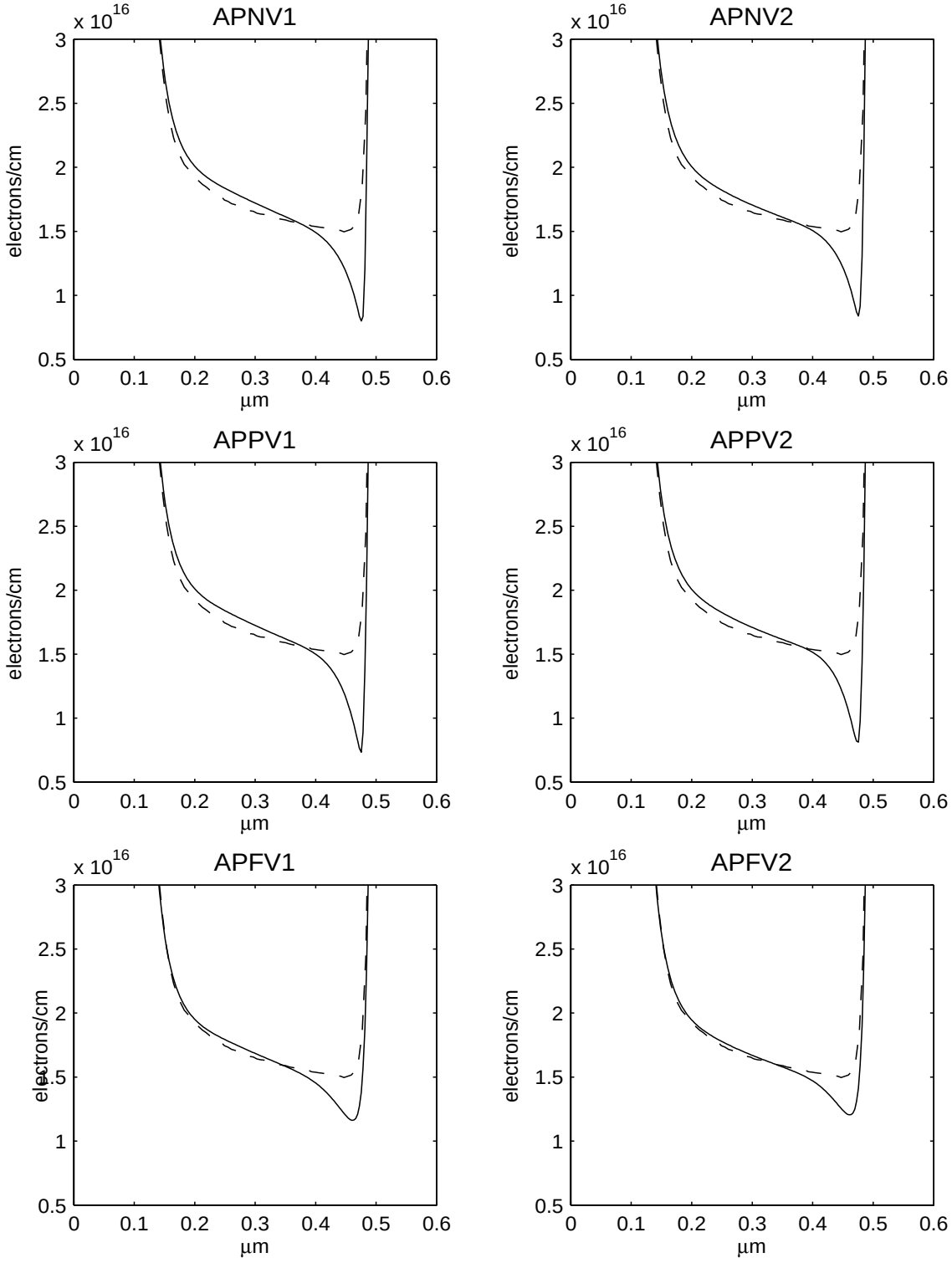

Figure 5.15: Electron concentration n for AP models, magnified view. Dashed line is Monte Carlo data.

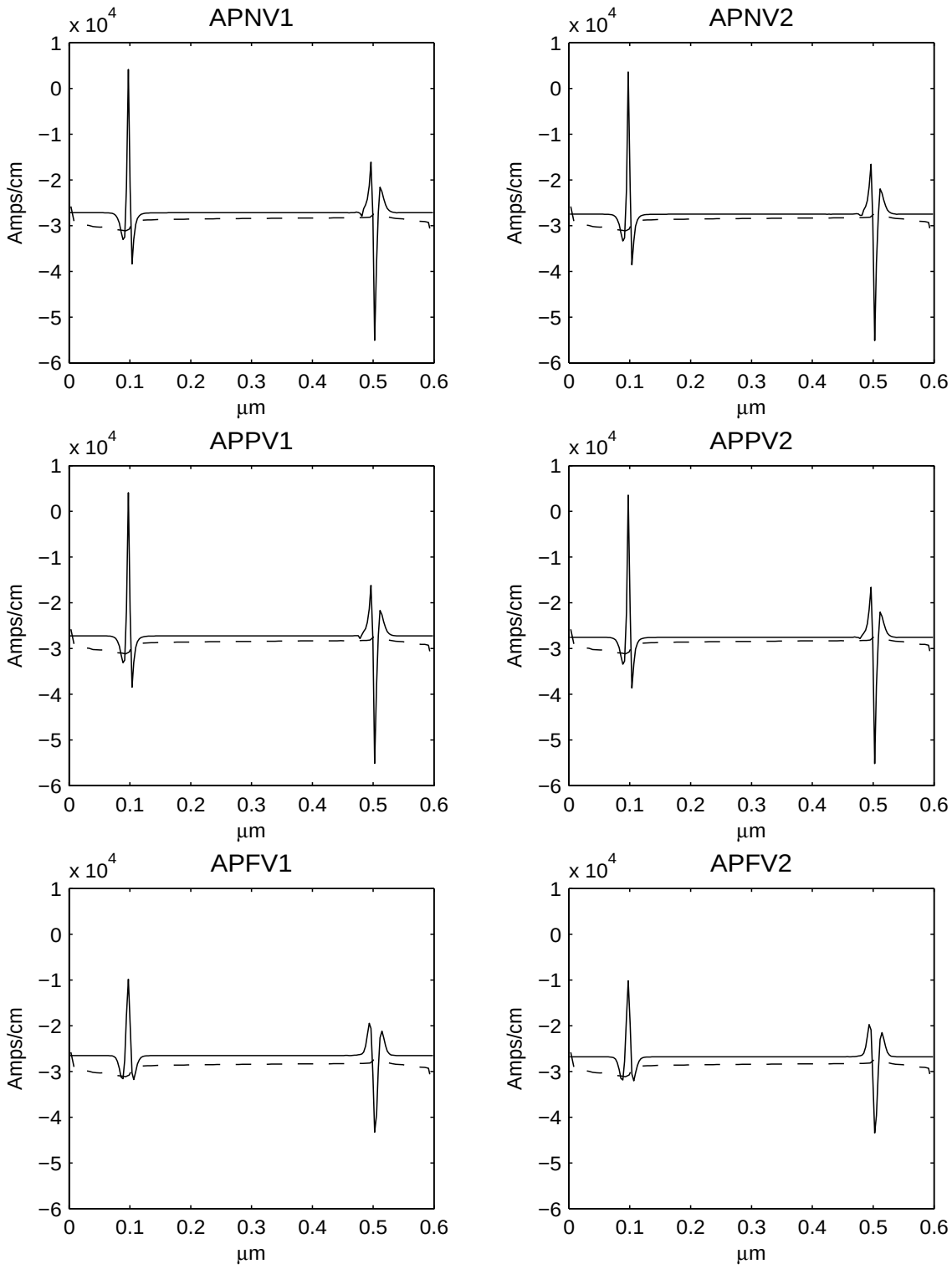

Figure 5.16: Electron current $J = -qnu$ for AP models. Dashed line is Monte Carlo data.

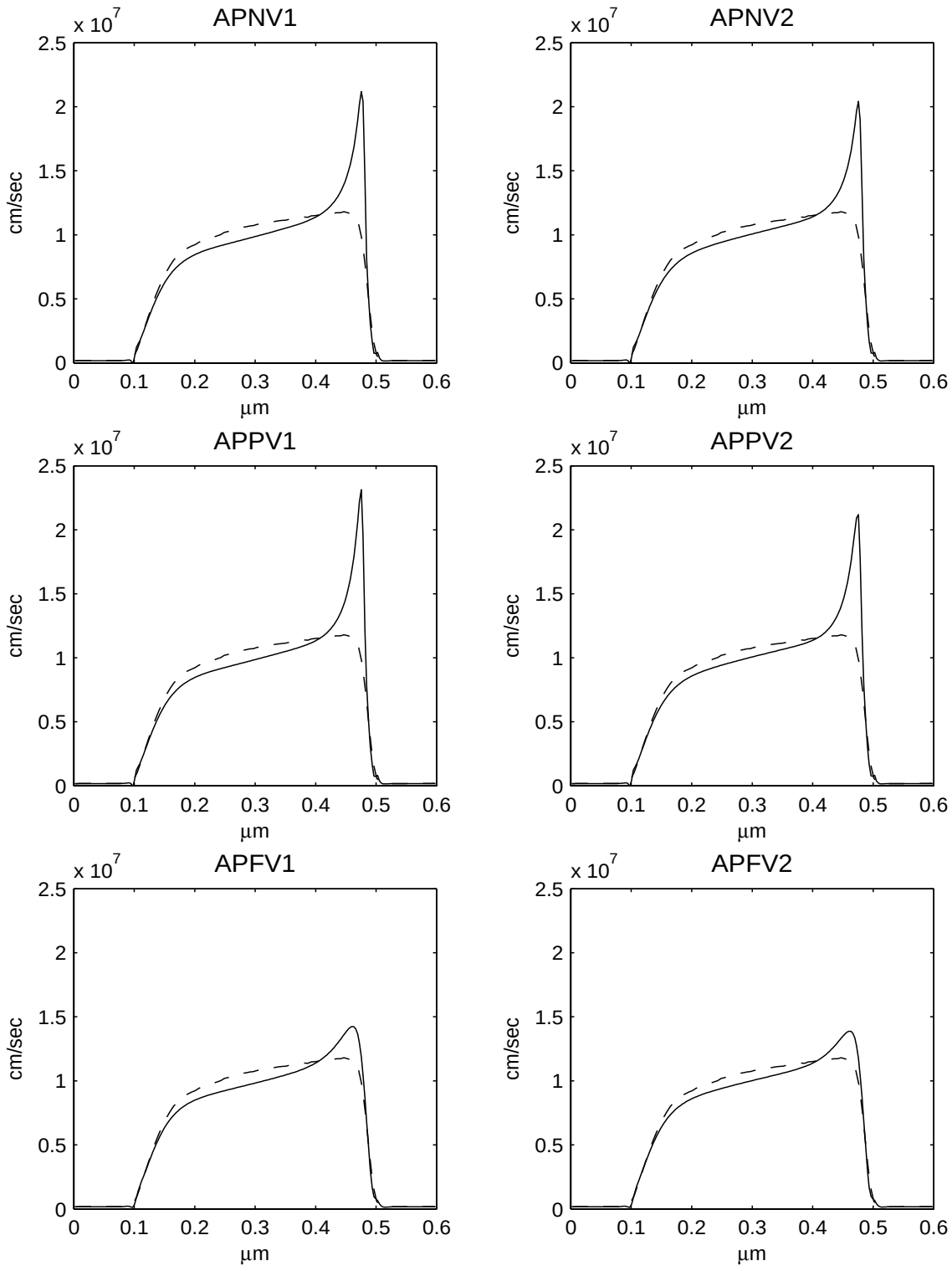

Figure 5.17: Electron velocity u for AP models. Dashed line is Monte Carlo data.

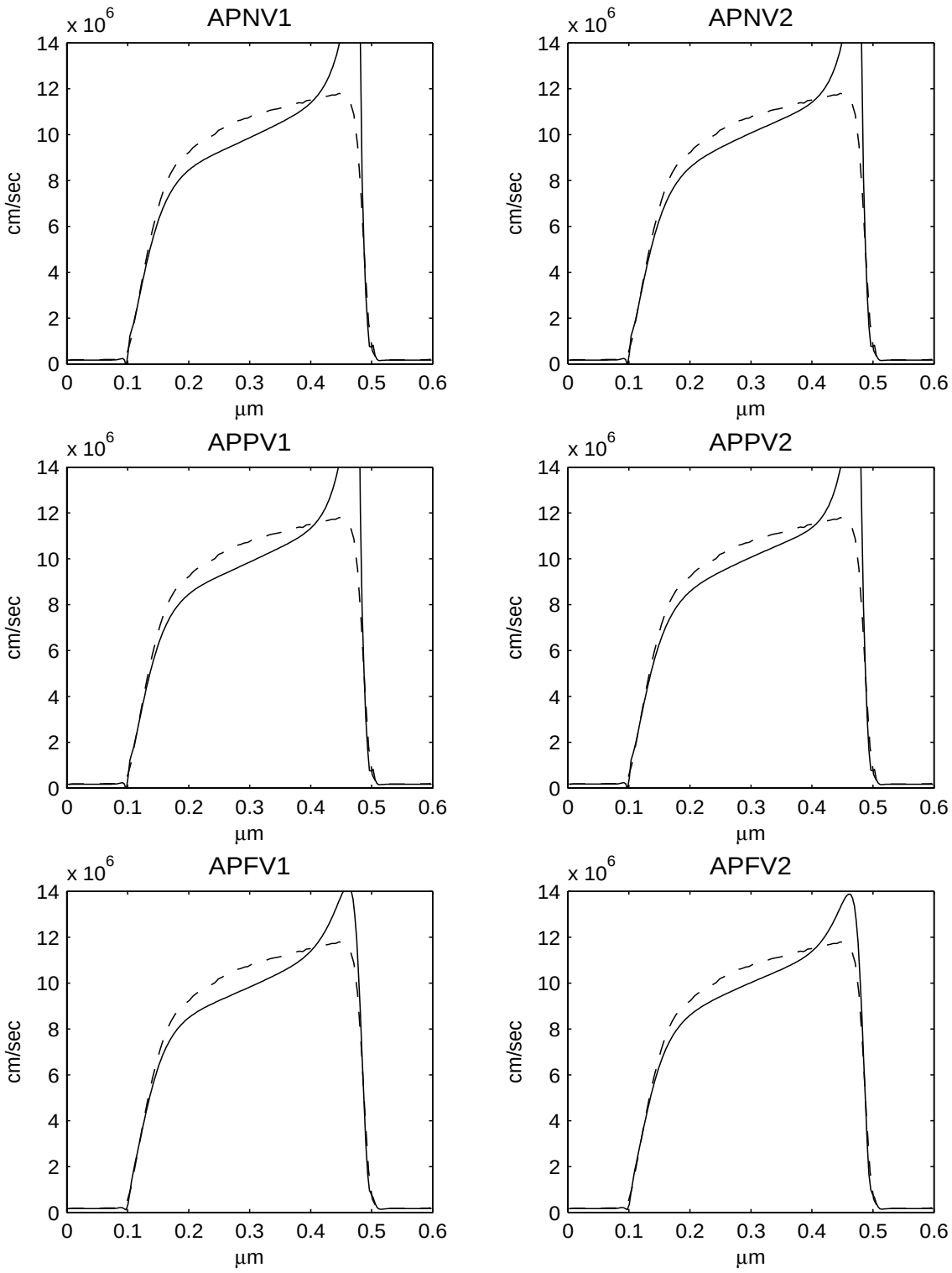

Figure 5.18: Electron velocity u for AP models, magnified view. Dashed line is Monte Carlo data.

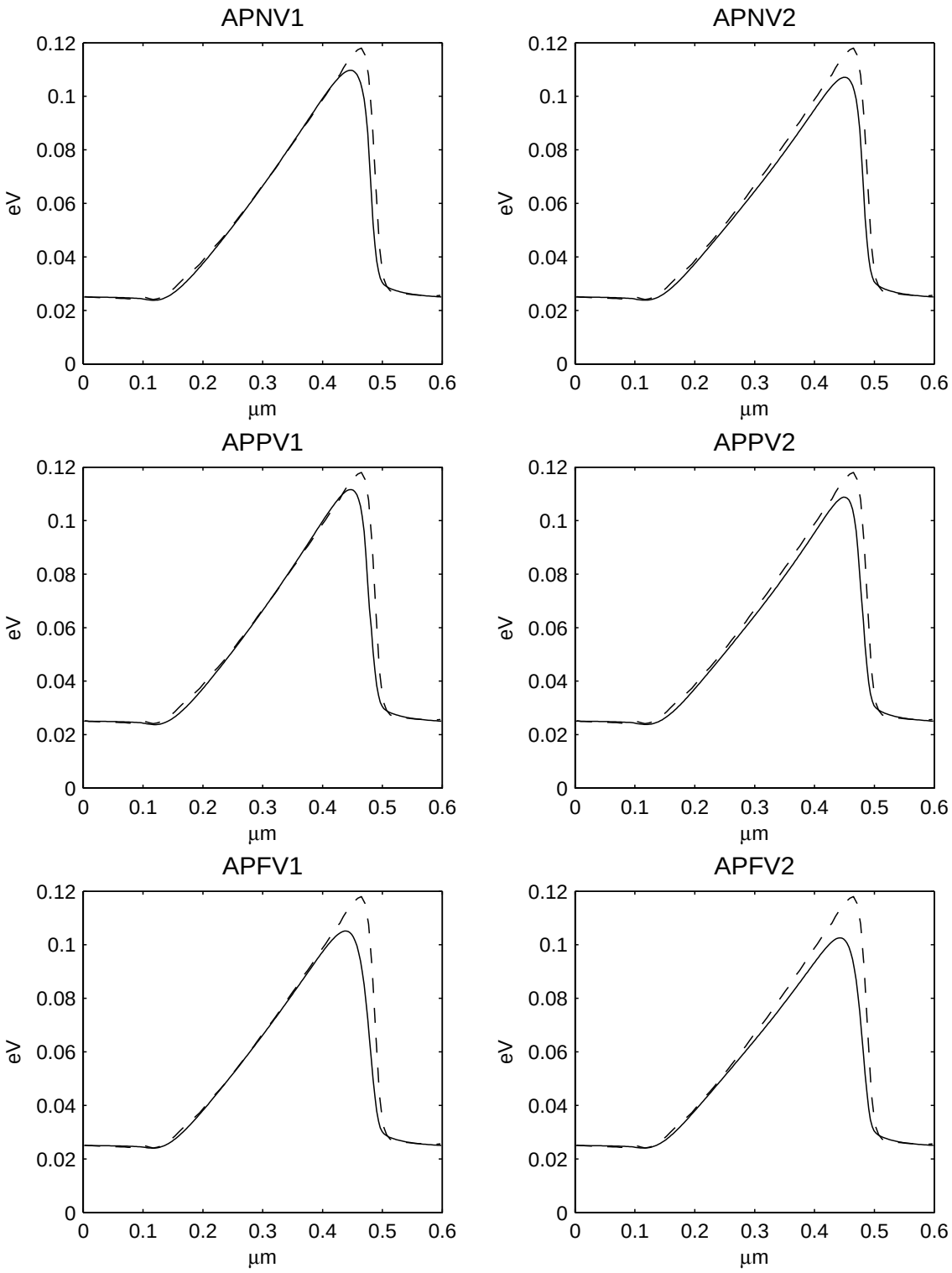

Figure 5.19: Electron thermal energy $m_e\theta$ for AP models. Dashed line is Monte Carlo data.

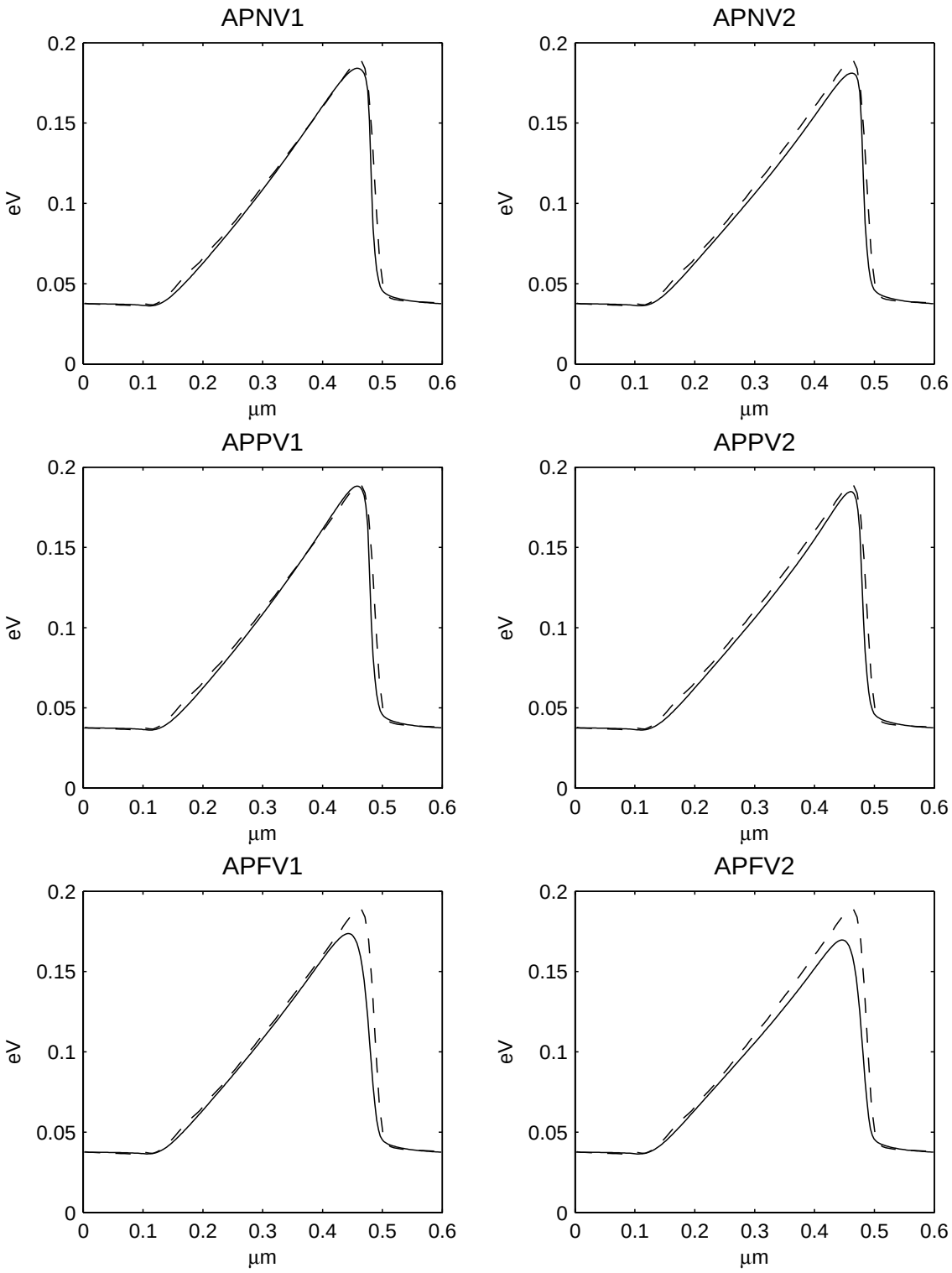

Figure 5.20: Electron energy $\frac{1}{2}m_e u^2 + \frac{3}{2}m_e\theta$ for AP models. Dashed line is Monte Carlo data.

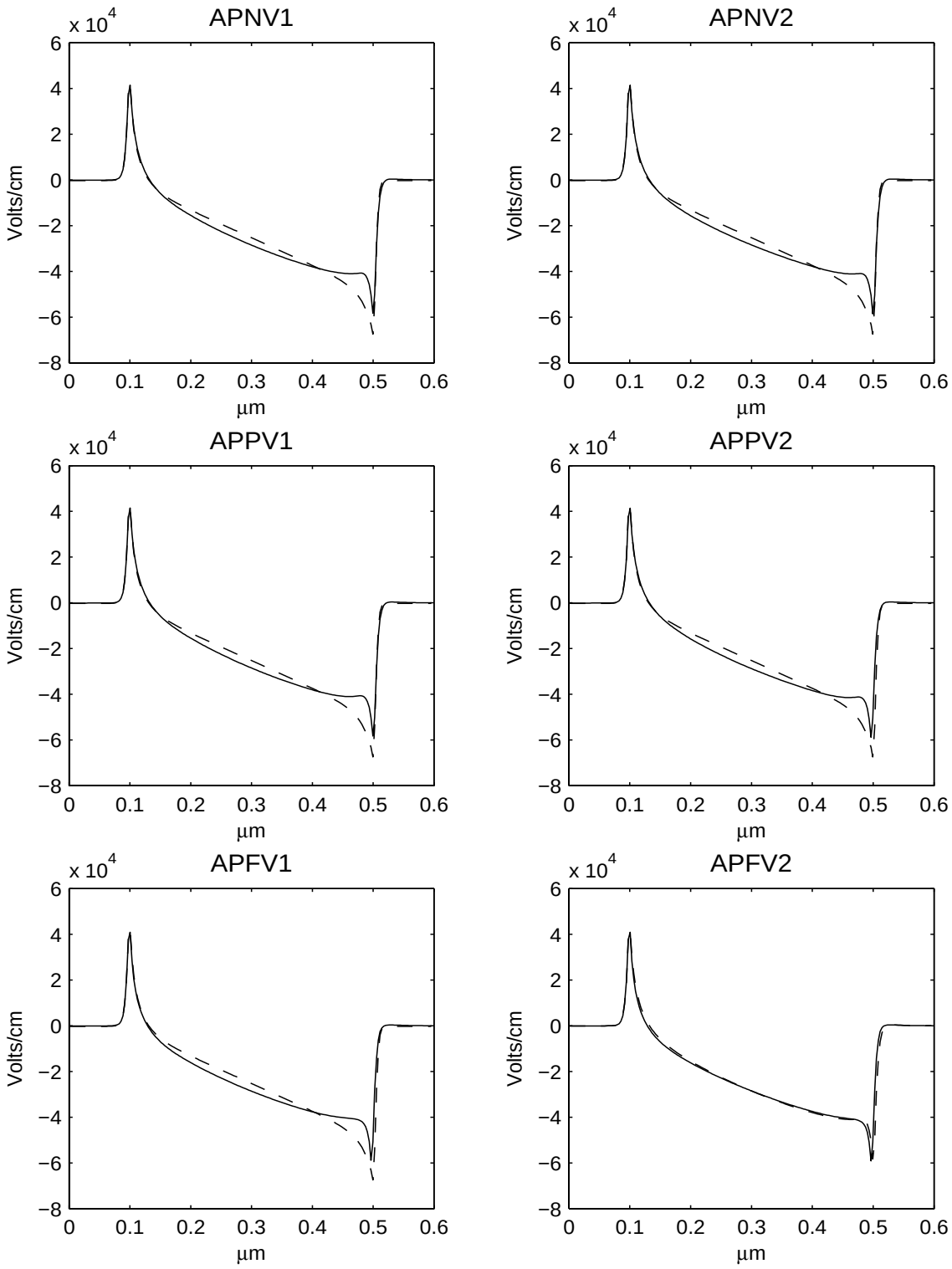

Figure 5.21: Electric field $E = -\partial_x \Phi$ for AP models. Dashed line is MBW model.

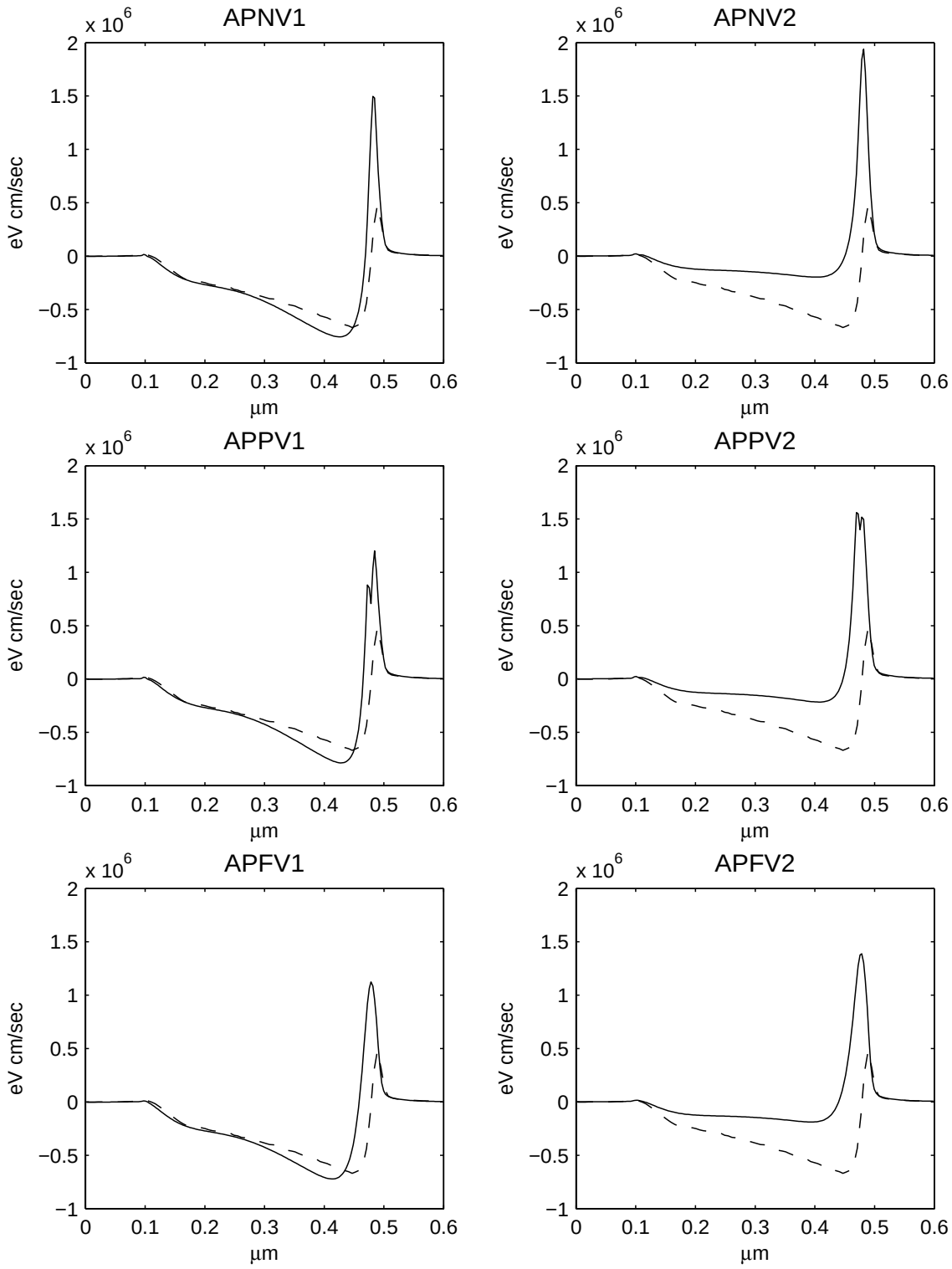

Figure 5.22: Heat flux $\frac{m_e}{n}q$ for AP models. Dashed line is Monte Carlo data.

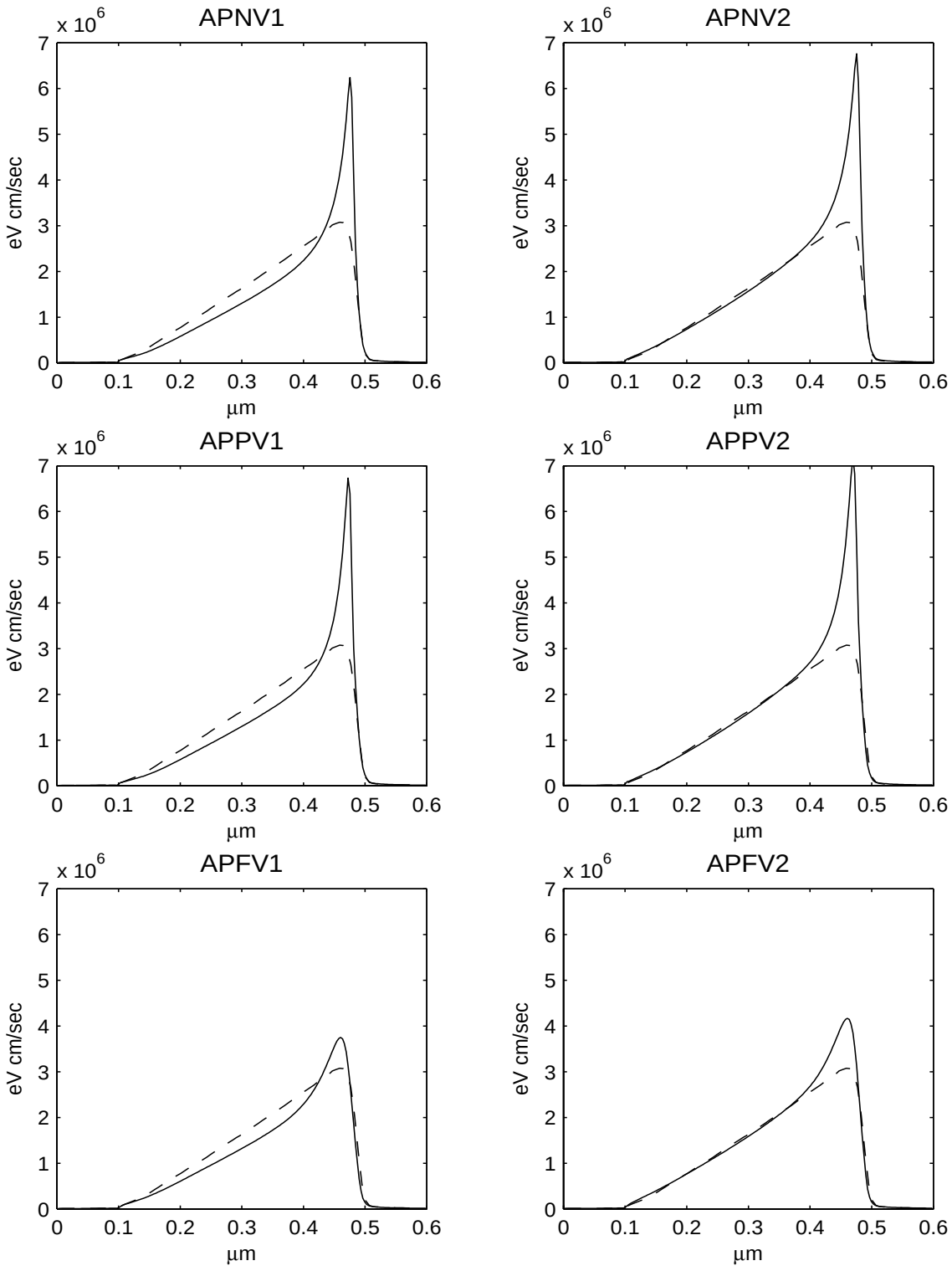

Figure 5.23: Electron energy flux $\frac{1}{2}m_e u^3 + m_e \frac{5}{2}\theta + m_e \frac{q}{n}$ for AP models. Dashed line is Monte Carlo data.

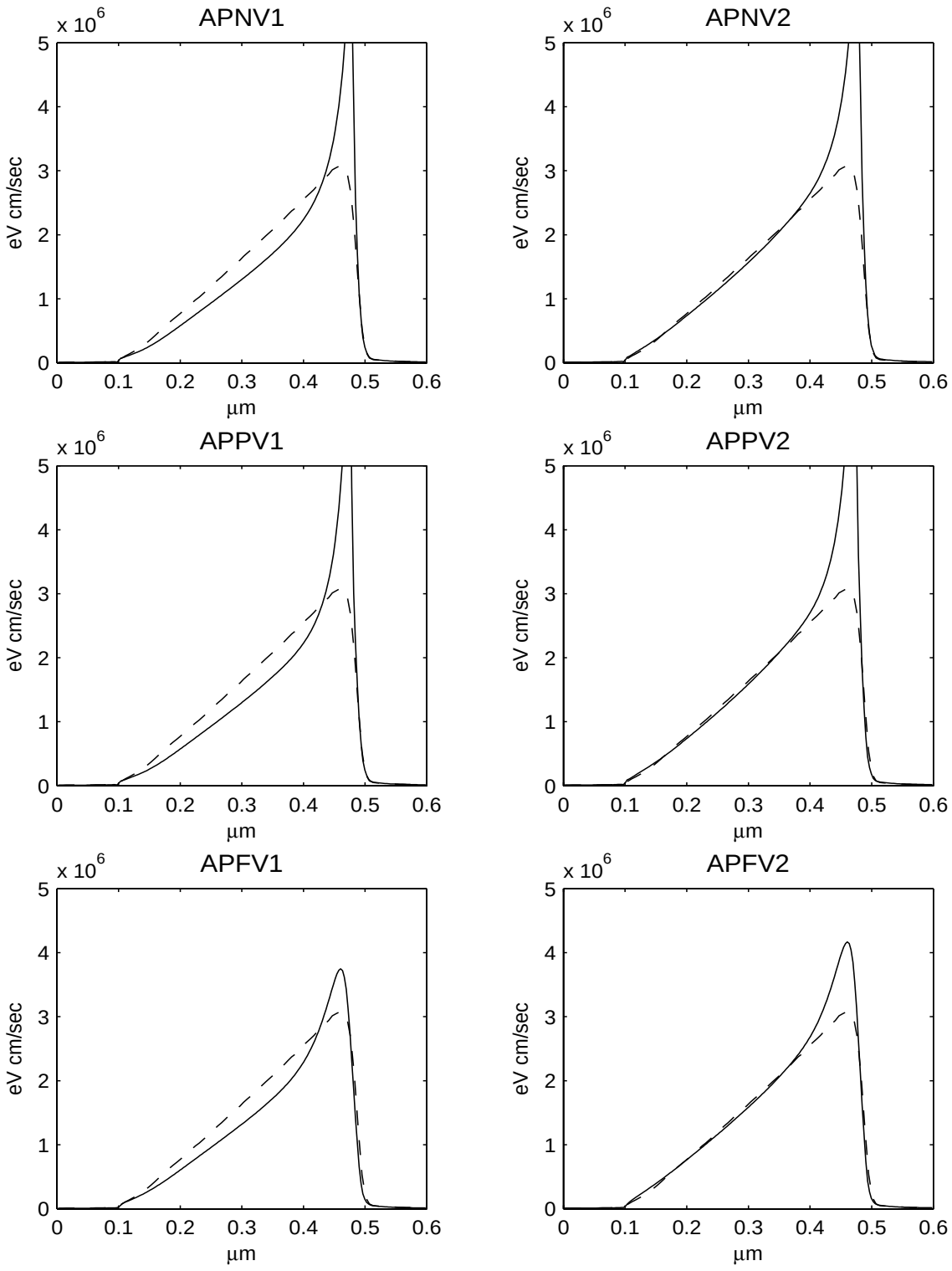

Figure 5.24: Electron energy flux $\frac{1}{2}m_e u^3 + m_e \frac{5}{2}\theta + m_e \frac{q}{n}$ for AP models, magnified view. Dashed line is Monte Carlo data.

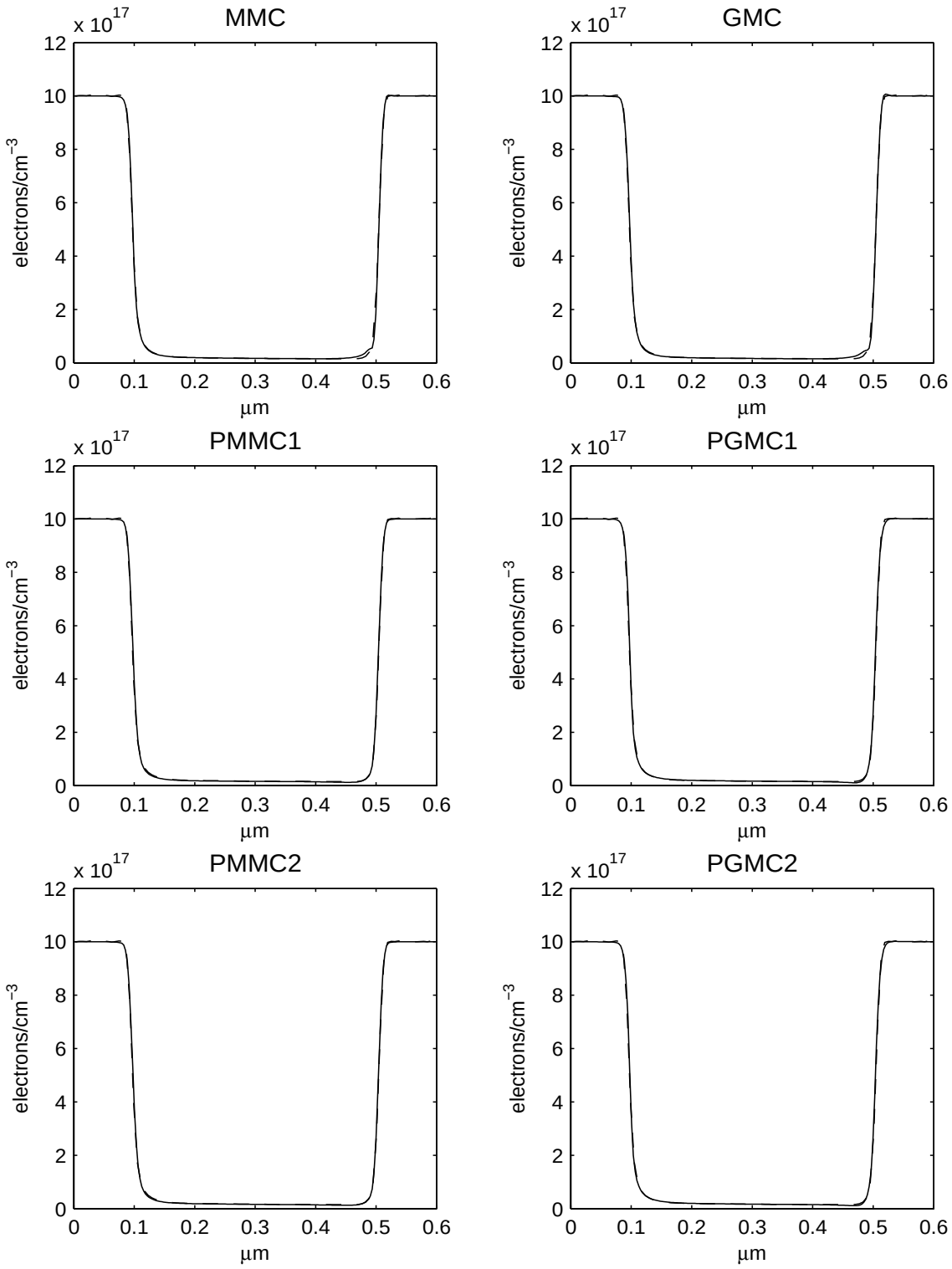

Figure 5.25: Electron concentration n for entropy-based models. Dashed line is Monte Carlo data.

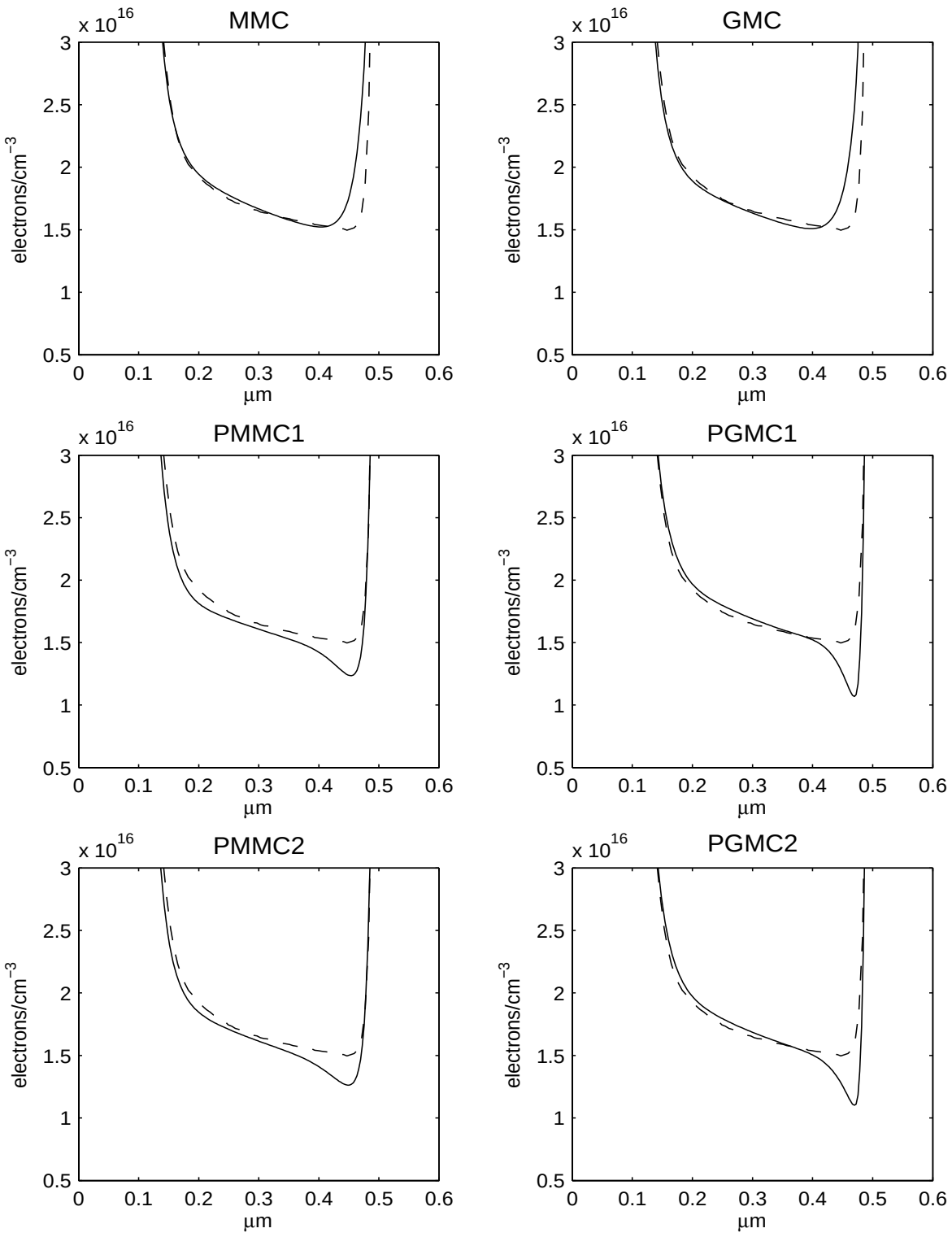

Figure 5.26: Electron concentration n for entropy-based models, magnified view. Dashed line is Monte Carlo data.

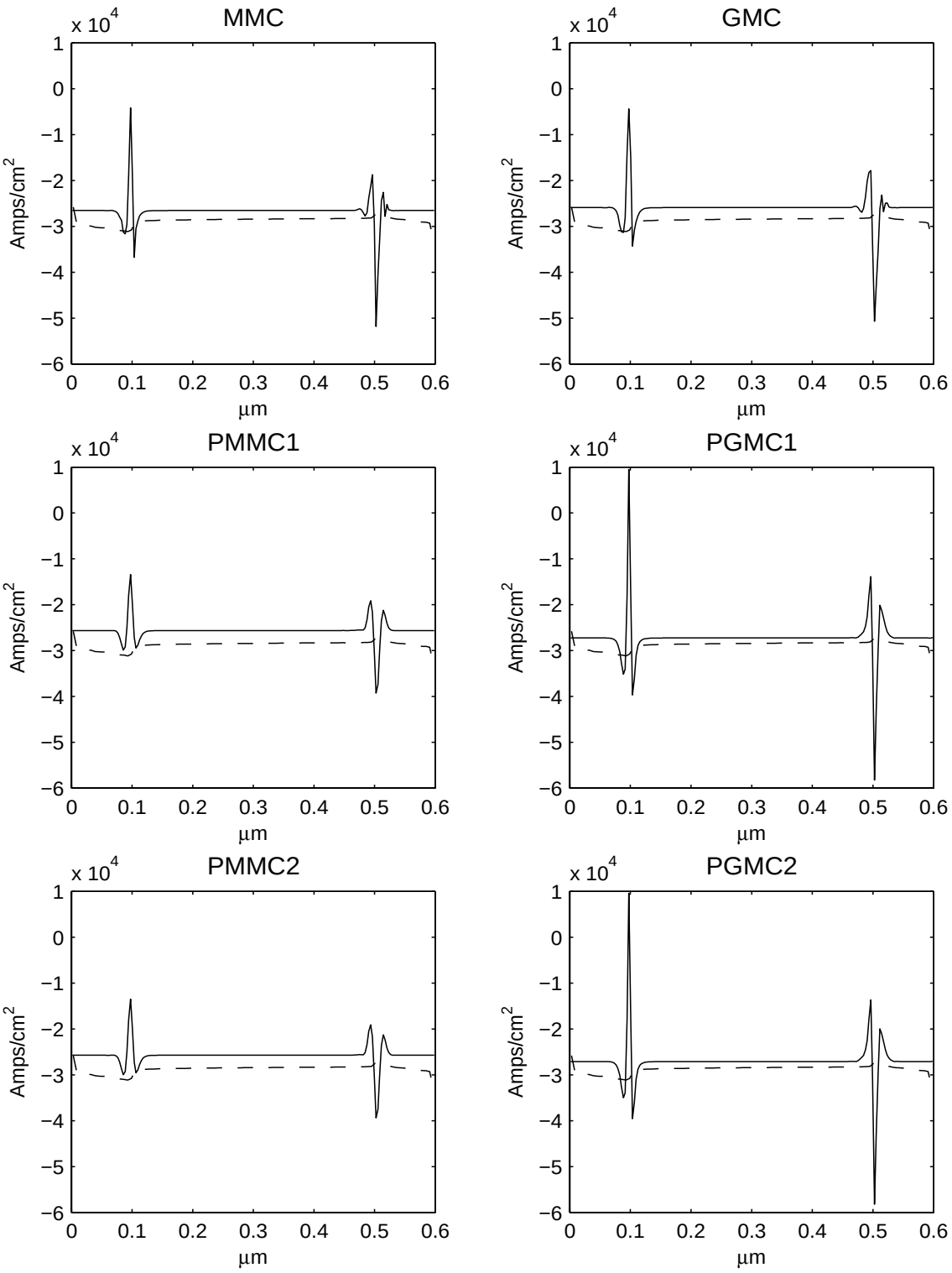

Figure 5.27: Electron current $J = -qnu$ for entropy based models. Dashed line is Monte Carlo data.

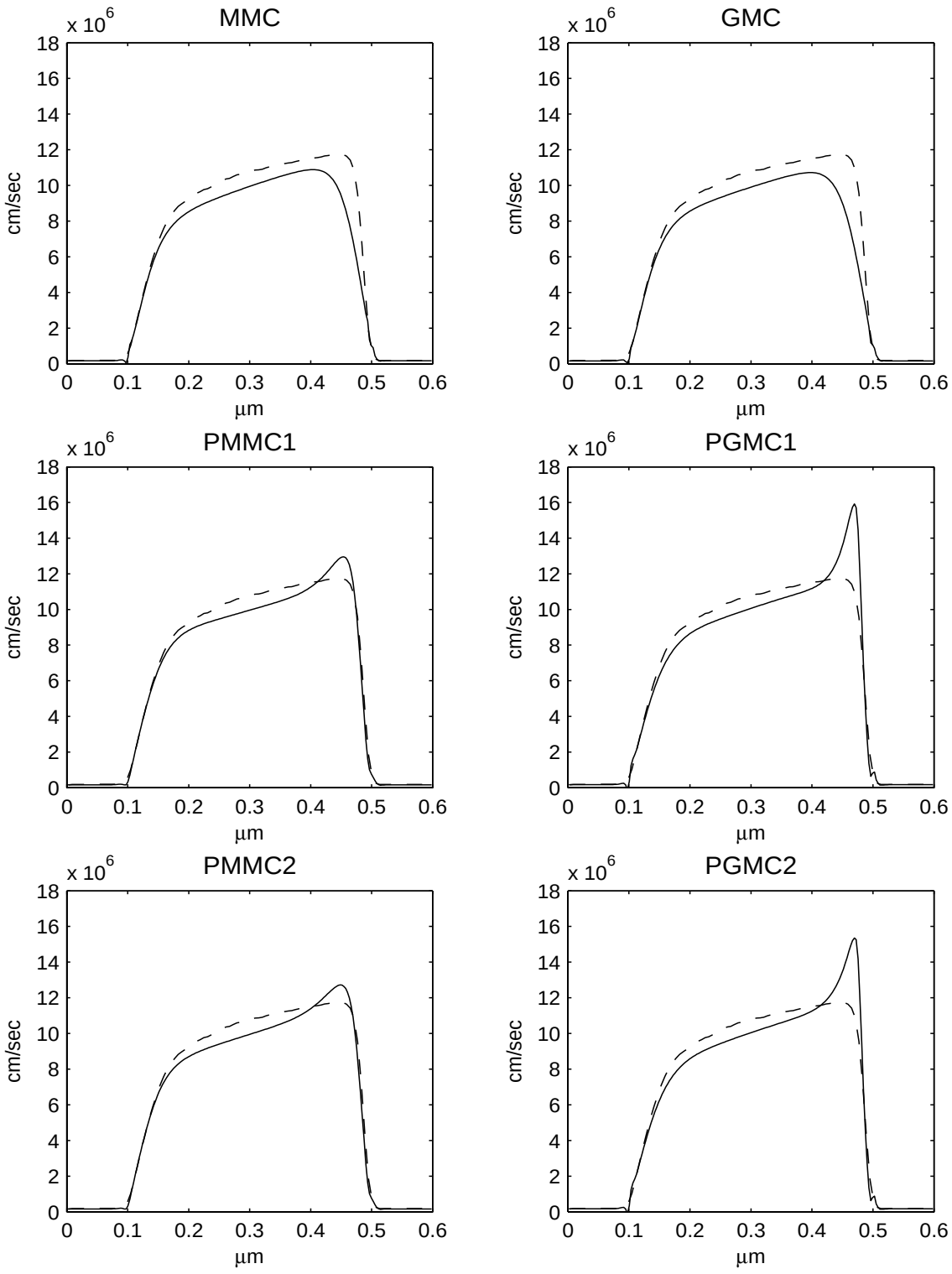

Figure 5.28: Electron velocity u for entropy-based models. Dashed line is Monte Carlo data.

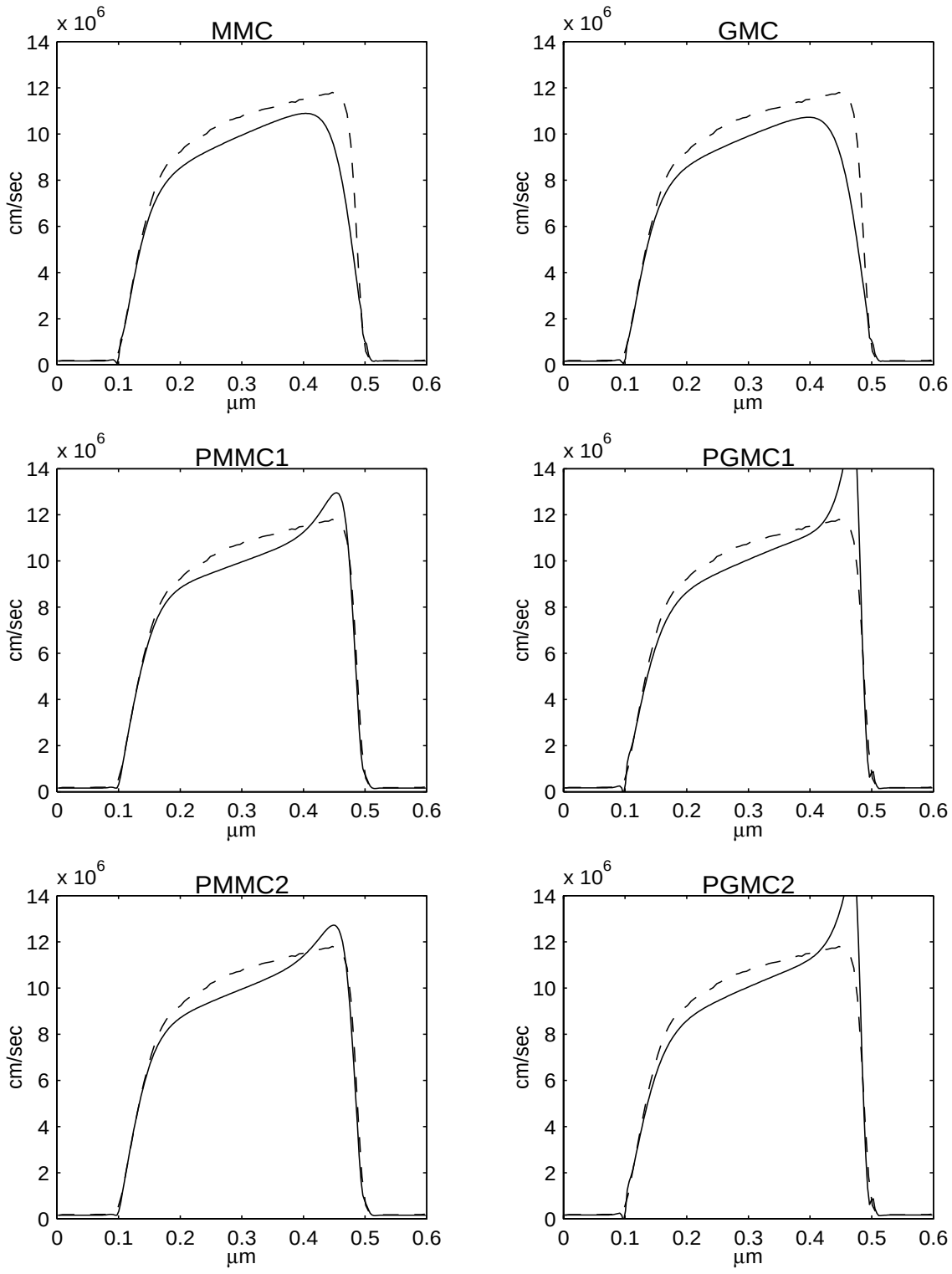

Figure 5.29: Electron velocity u for entropy-based models, magnified view. Dashed line is Monte Carlo data.

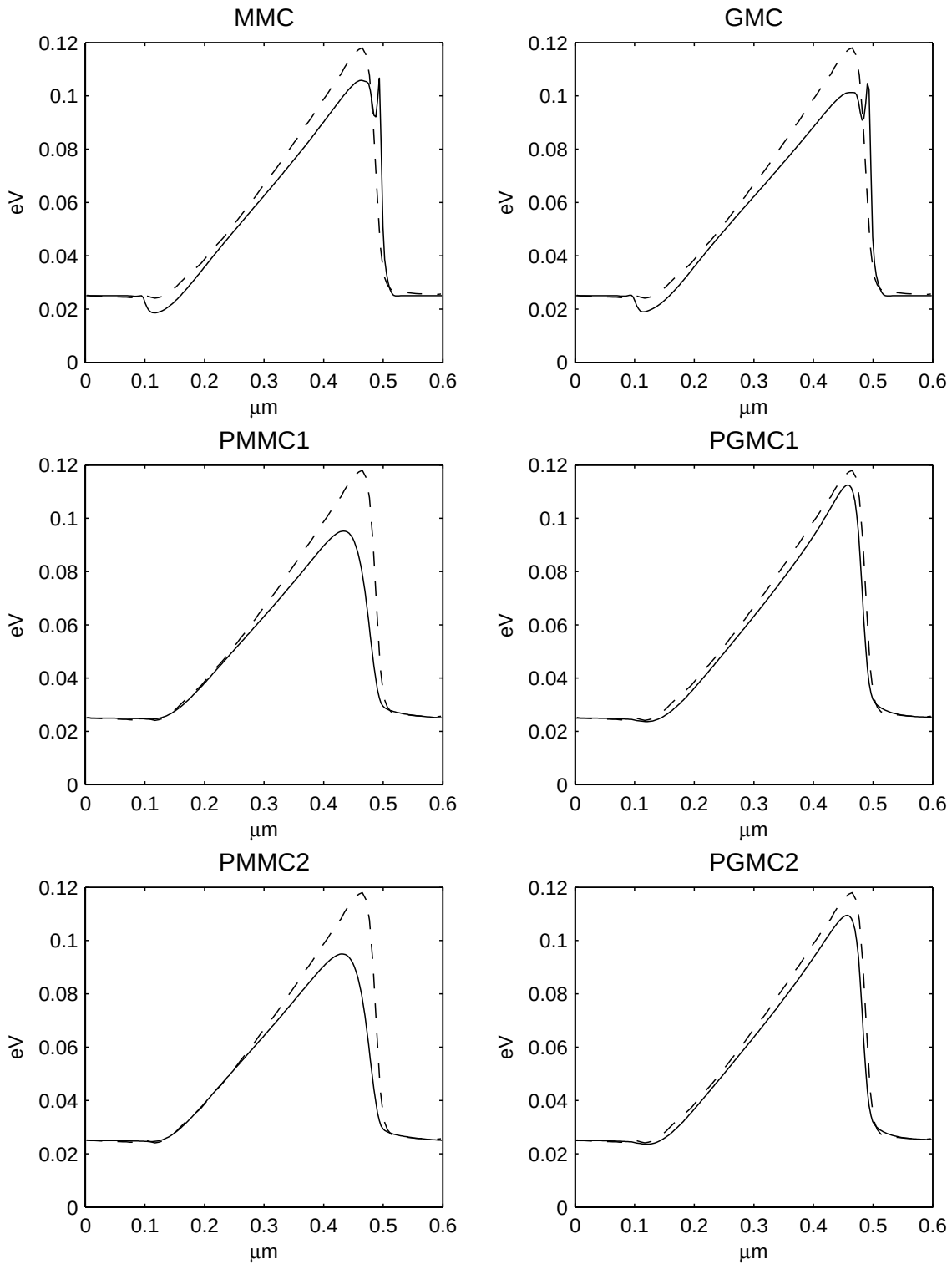

Figure 5.30: Electron thermal energy $m_e\theta$ for entropy based models. Dashed line is Monte Carlo data.

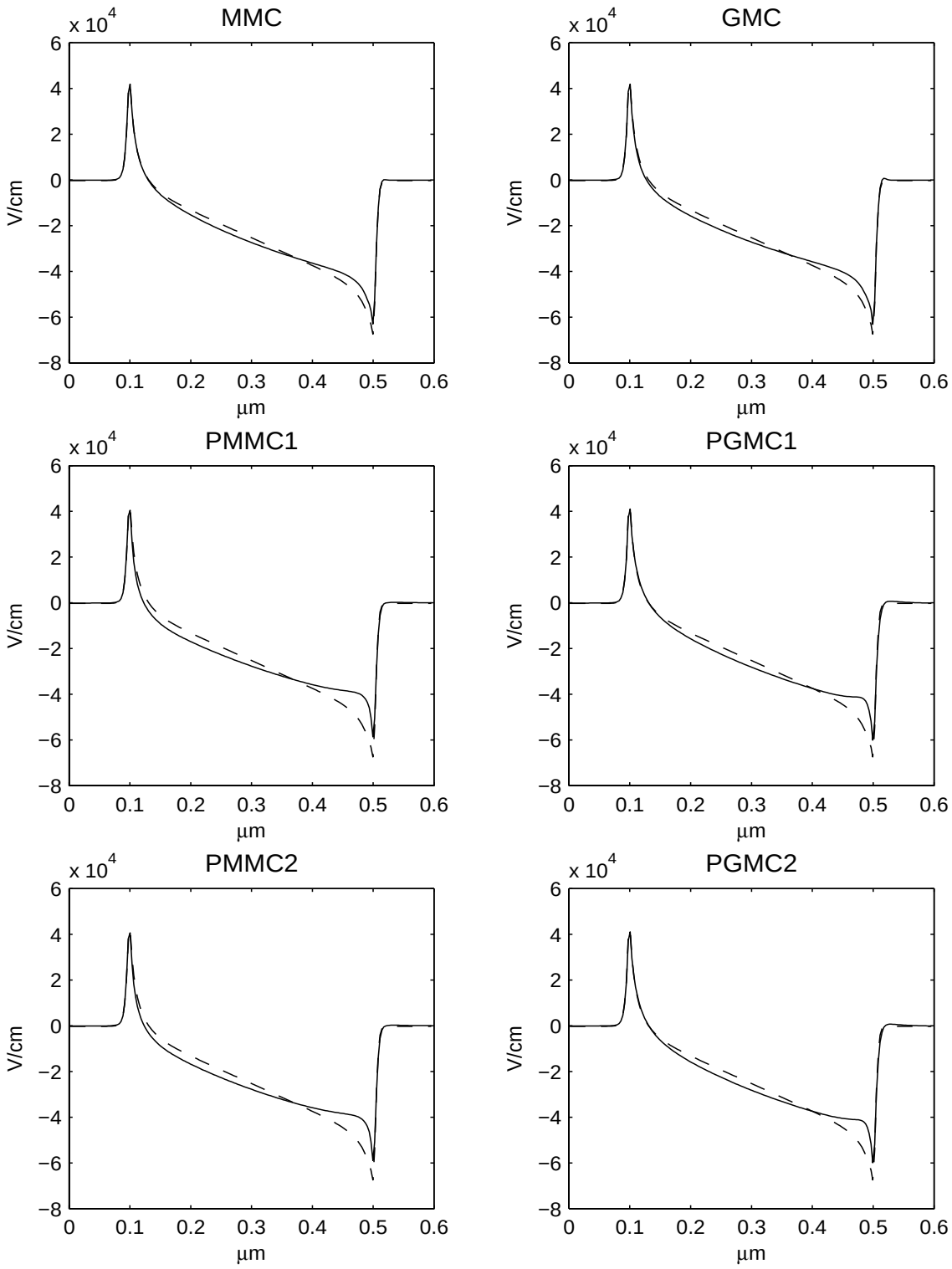

Figure 5.31: Electric field $E = -\partial_x \Phi$ for entropy based models. Dashed line is MBW model.

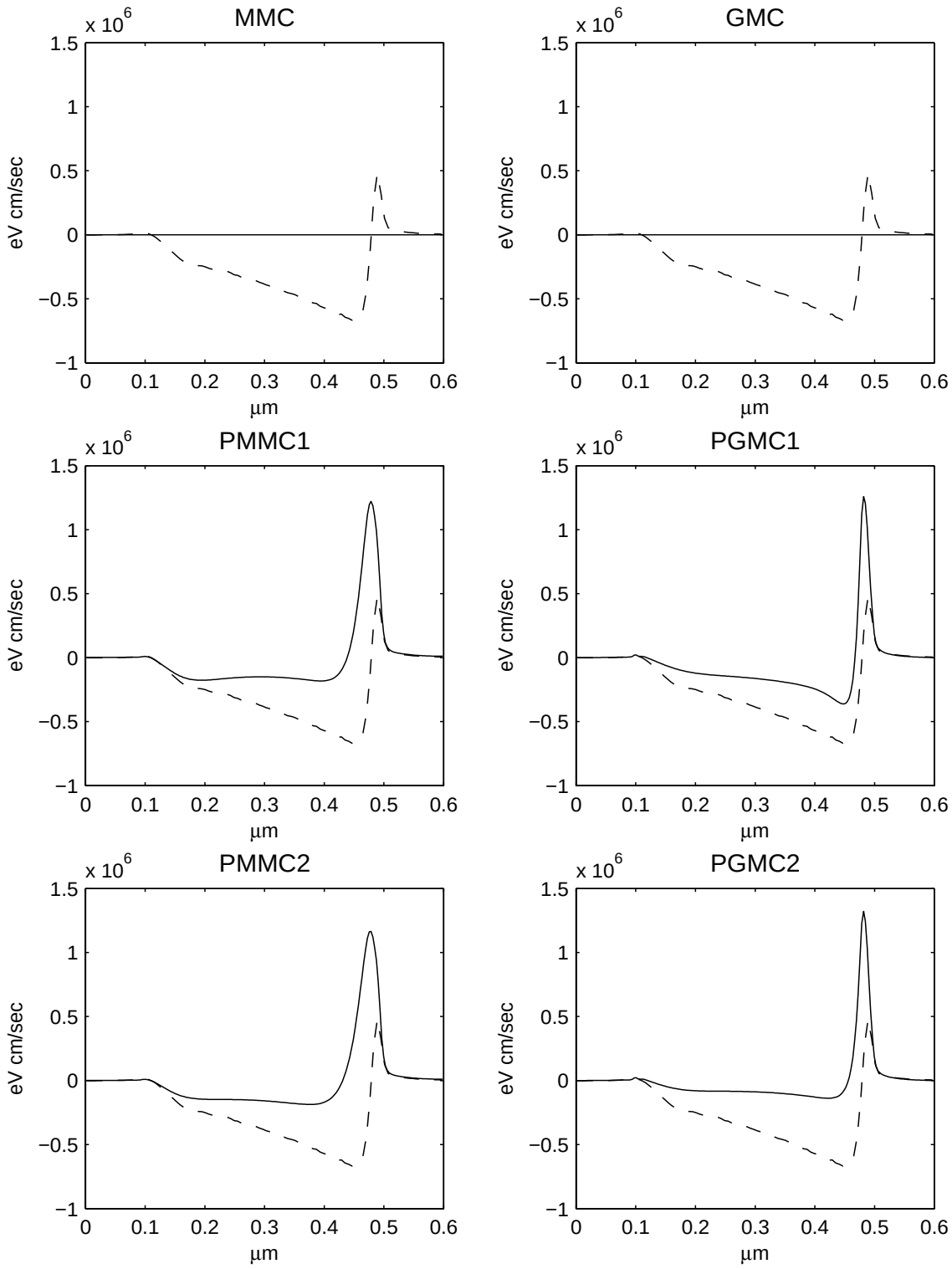

Figure 5.32: Electron heat flux $\frac{m_e}{n}q$ for entropy-based models. Dashed line is Monte Carlo data.

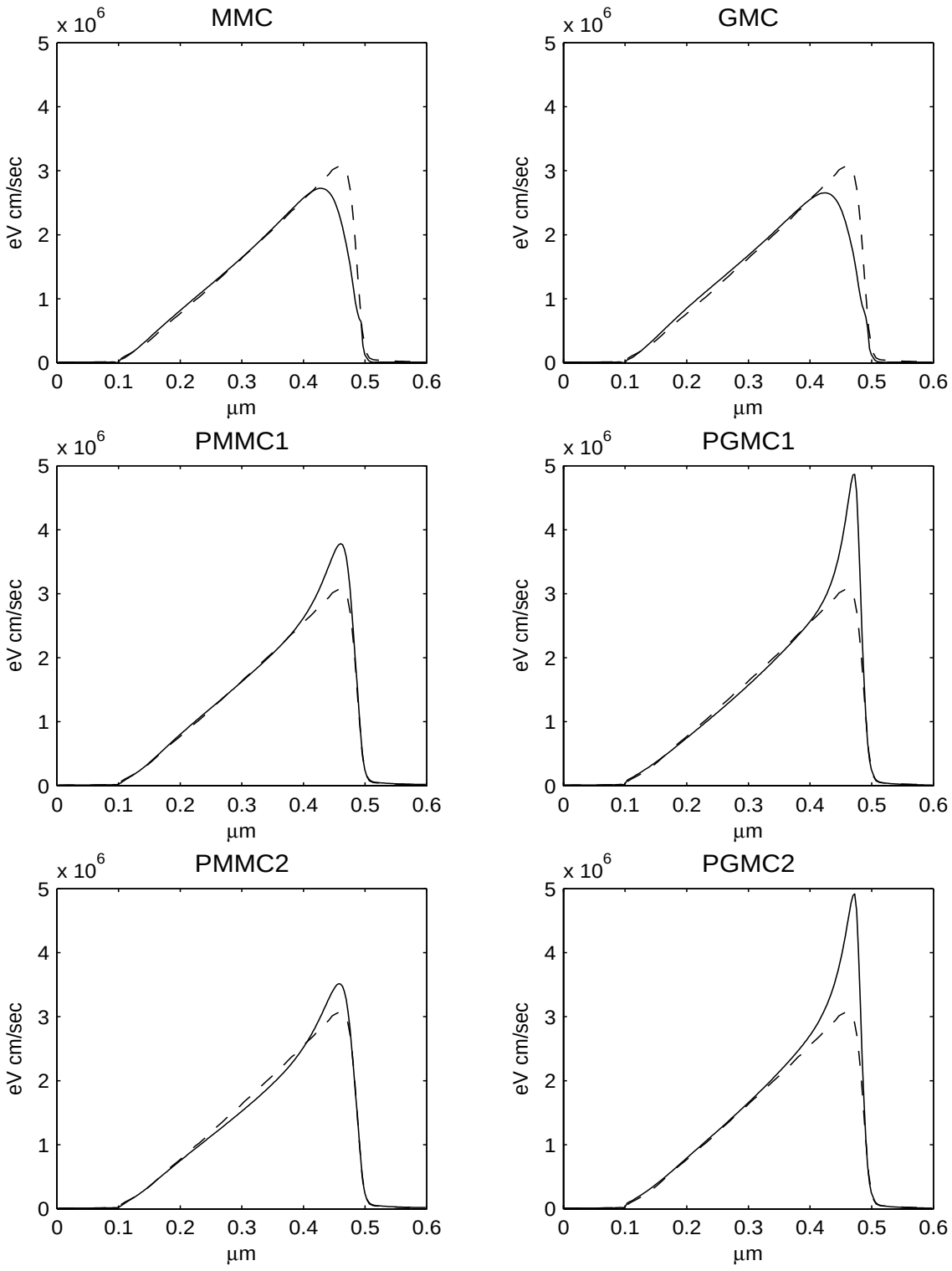

Figure 5.33: Electron energy flux $\frac{1}{2}m_e u^3 + m_e\frac{5}{2}\theta + m_e\frac{q}{n}$ for entropy based models. Dashed line is Monte Carlo data.

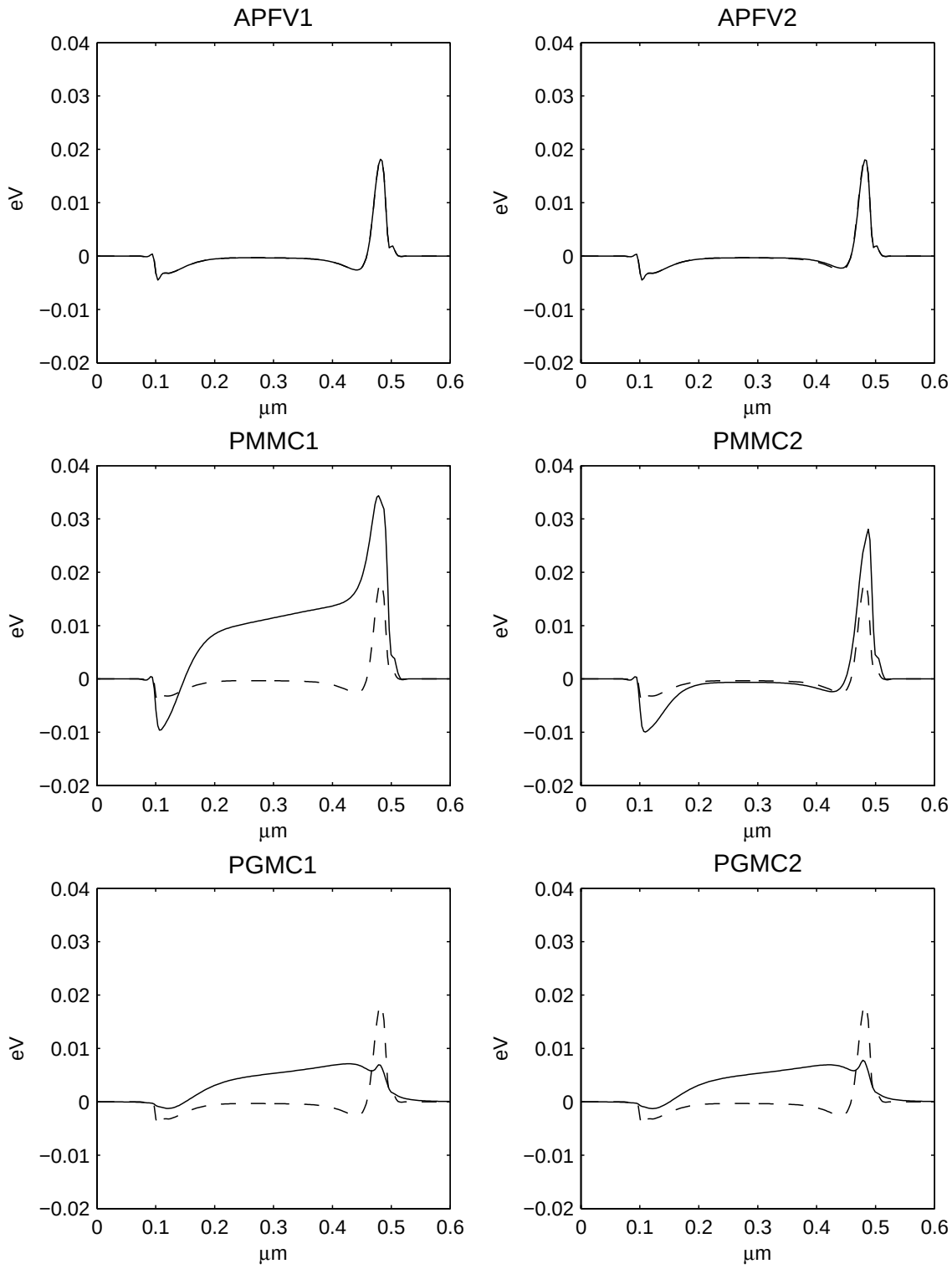

Figure 5.34: Anistropic stress $\frac{m_e \sigma}{n}$ for AP and PEB models. Dashed line is APFV1 model.

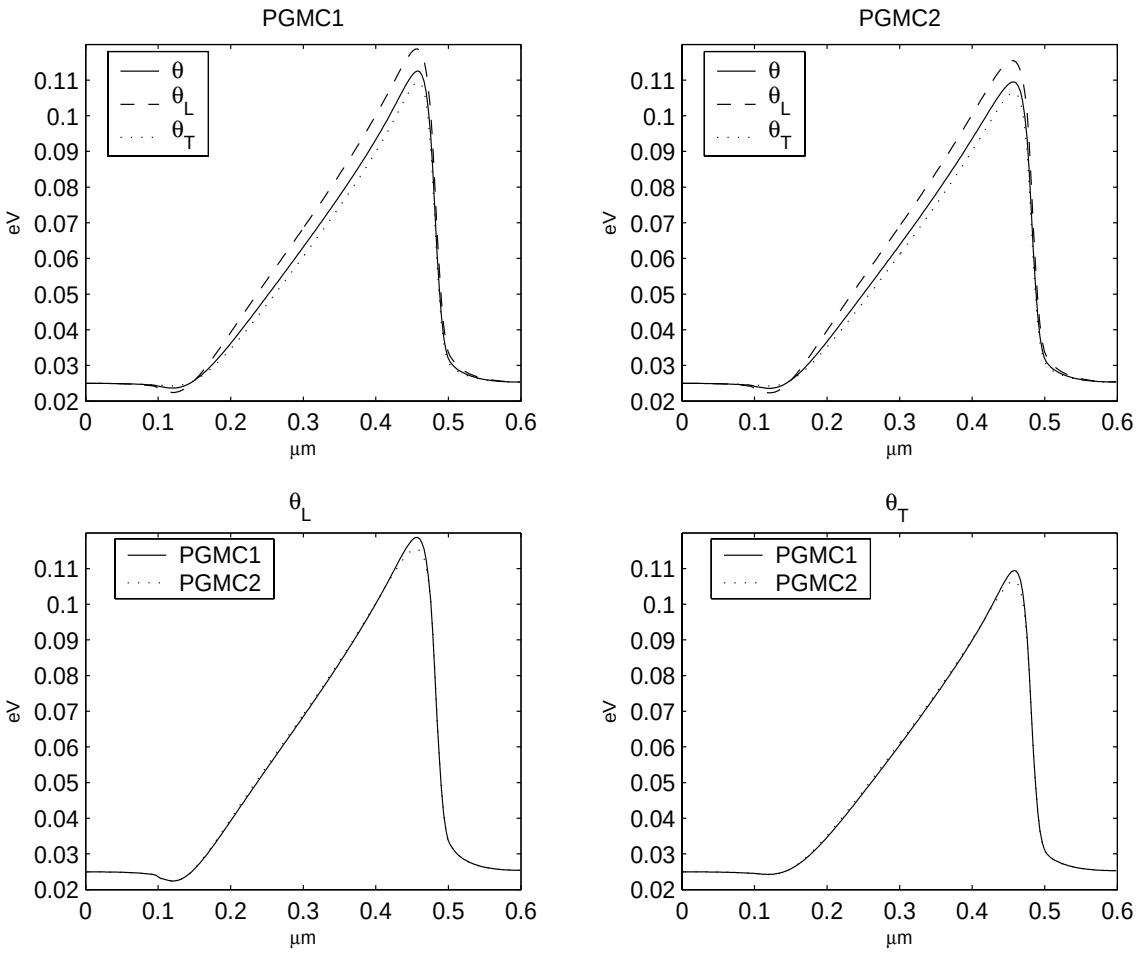

Figure 5.35: Components of thermal energy for perturbed Gaussian closures.

Chapter 6

Computational Issues: Stiffness and Balance

In Chapter 5, we computed solutions for several hydrodynamic models without much concern for the computational issues involved. In the current chapter, we examine some of the important numerical aspects of these closures which are commonly found in hyperbolic systems with relaxation and source terms. We then present a new scheme based on a splitting method that was first introduced in [40]. This splitting is based on the balance of forces in hydrodynamic models that, for regimes with small electric field, recover the drift-diffusion system in the asymptotic limit of small mean-free-path. The advantage of the scheme is that it removes stiffness and excessive dissipation which is often found with standard discretizations of hydrodynamic models in the drift-diffusion regime. In addition, the scheme also significantly reduces the size of numerical current oscillations found at material junctions in a n^+-n-n^+ diode with slab symmetry.

For the purposes of this chapter, the term hydrodynamic model will refer specifi-

cally to the following system:

$$\partial_t n + \partial_x (nu) = 0 \,, \tag{6.1a}$$

$$\partial_t (nu) + \partial_x (nu^2 + n\theta + \sigma) - \frac{q_e}{m_e^*} n \partial_x \Phi = -\frac{1}{\tau_p} nu \,, \tag{6.1b}$$

$$\partial_t \left(\frac{nu^2}{2} + \frac{3n\theta}{2} \right) + \partial_x \left(\frac{nu^3}{2} + \frac{5nu\theta}{2} + \sigma u + q \right) - \frac{q_e}{m_e^*} nu \partial_x \Phi = -\frac{1}{\tau_w} \left(\frac{nu^2}{2} + \frac{3n(\theta - \theta_\ell)}{2} \right) . \tag{6.1c}$$

with appropriate boundary conditions. Here the fluid variables are the electron concentration n, the bulk velocity u, and the temperature θ. The constants q_e and m_e^* are the electron charge and effective mass, and θ_ℓ is the temperature of the semiconductor lattice. The quantities τ_p and τ_w are relaxation times for momentum and energy that are fit with Monte Carlo calculations as functions of electron energy. Their exact form can be found in [63]. The anisotropy σ and the heat flux q are given by the constitutive relations

$$\sigma = -\frac{4}{3} \tau_p n \theta \partial_x \theta \,, \qquad q = -\frac{5}{2} \tau_p n \theta \partial_x \theta \,. \tag{6.2}$$

Finally, Φ is the electric potential that satisfies Poisson equation,

$$-\partial_x (\epsilon \partial_x \Phi) = q_e (D - n) \,, \tag{6.3}$$

with appropriate boundary conditions. The derivative $-\partial_x \Phi$ is the electric field. The variable $\epsilon = \epsilon(x)$ is the electric permittivity, and $D = D(x)$ is the doping profile

that is created by the ionization of atoms in the crystal lattice of the semiconductor material. The shape of the doping profile is set during the fabrication process of a device. Together, (6.1) and (6.2) make up the drift-diffusion-Poisson system.

The behavior of solutions to (6.1) is heavily dependent on the relative sizes of the different forces that act free electrons in the semiconductor. For regimes in which the potential energy and thermal energy balance and the mean free path of electrons is small compared to the device length, the electron concentration is formally approximated by the drift-diffusion equation

$$\partial_t n + \partial_x \left(\mu n \partial_x \Phi - a \partial_x n \right) = 0 \,, \tag{6.4}$$

where there mobility and diffusivity are given by

$$\mu = \frac{q_e \tau_p}{m_e^*} \qquad \text{and} \qquad a = \tau_p \theta_\ell$$

respectively, and Φ still satisfies (6.3). In such cases, $\theta = \theta_\ell$ and u is determined by a balance between electrical and diffusive forces:

$$u = \mu \partial_x \Phi - a \partial_x (\log(n)) \,.$$

Rigorous results connecting the hydrodynamic and drift-diffusion models can be found in [22, 23, 30].

When solving (6.1) numerically, two issued must be addressed. The first issue is numerical stiffness for systems in the drift-diffusion regime, when the evolution of

n given in (6.1) can be accurately approximated by (6.4). In this situation, stiff flux terms in (6.1) can create excessive numerical dissipation that leads to a distorted approximation of the diffusive term in (6.4). In addition, stiff flux terms imply large wave speeds and a hyperbolic CFL condition that is much more restrictive than the natural diffusive CFL condition, $\Delta t \sim \Delta x^2$ associated with (6.4). Therefore, explicit schemes for (6.1) will be highly inefficient in the drift-diffusion regime. On the other hand, a standard implicit approach is impractical for modern shock capturing methods which approximate fluxes in a highly nonlinear fashion.

The second numerical issue is that of *balance*. At steady state, the current profile for (6.1) is constant in space. However, the existence of non-conservative electric field terms in (6.1) often leads to non-physical oscillations. This is because standard discretization for (6.1) fail to mimic the balance of forces at the continuum level that give rise to steady-state solutions. In places where the fluid variables and the potential vary rapidly, the effects can be quite dramatic. Results such as these are found in other applications such as shallow water models and chemotaxis [6, 24]. In fact almost any hyperbolic system with a non-conservative force term is subject to this behavior. In some cases, so-called *well-balanced schemes* have been developed that the preserve the balance of forces in the steady state at the discrete level. Although frequently successful, these schemes require explicit information from the steady-state equations that is typically not available with the hydrodynamic model.

In this chapter, we adapt a splitting method that was originally put forth to fix the problems with efficiency and numerical dissipation associated with stiff flux terms in a simple 2×2 hyperbolic system [40, 64, 78]. Our new splitting is based on the

balance of dominant forces in the drift-diffusion regime. We find that, in addition to removing stiff fluxes, this split scheme significantly reduces the presence of non-physical oscillations in the steady-state current profile. We are hopeful that these results will lead to more general concepts of well-balanced schemes that are more robust and applicable to transient as well as steady-state problems.

The outline of the chapter is as follows. In Section 2, we describe the benchmark problem for testing our scheme. In Section 3, we formally derive the drift-diffusion scaling of (6.1), the drift-diffusion limit, and discuss the numerical issues that arise. In Section 3, we present previous work and numerical results. In Section 4, a new scheme is introduced with details presented in Section 5. In Section 6, numerical results are given. Section 7 is for discussion and conclusions.

6.1 The Benchmark Problem

As in Chapter 5, we will be simulating electron transport for an n^+-n-n^+ diode (see Figure 6.1) of length L that is used to simulate the channel in MOSFET and MESFET devices [84]. The diode possesses slab symmetry and is therefore described in one spatial dimension. The left end of the diode is called the source; the right end is called the drain; and the center portion is the channel. The boundaries between these regions are called junctions. It is here that numerical oscillations in the current tend to appear.

We assume the diode is made of silicon with constant electric permittivity $\epsilon = 1.04\times10^{-16}\,\mathrm{C}/\,\mu\mathrm{m}$ and effective mass $m_e^* = 0.32m_e$, where $m_e = 9.109\times10^{-31}\,\mathrm{kg}$ is the free electron mass [84]. The device length is $L = 0.6x_0$, where x_0 is a representative

length scale, and the doping profile is

$$D(x) = \begin{cases} 10D_0, & 0 < x < L/6 \\ 0.1D_0, & L/6 < x < 5L/6 \\ 10D_0, & 5L/6 < x < L \end{cases}, \tag{6.5}$$

where D_0 is a representative concentration. In the Chapter 5, the value of x_0 was fixed at $1\,\mu$m and D_0 was fixed at $10^{17}\,\mathrm{cm}^{-3}$. However, we now consider the behavior of the device over a range of values for D_0 and x_0.

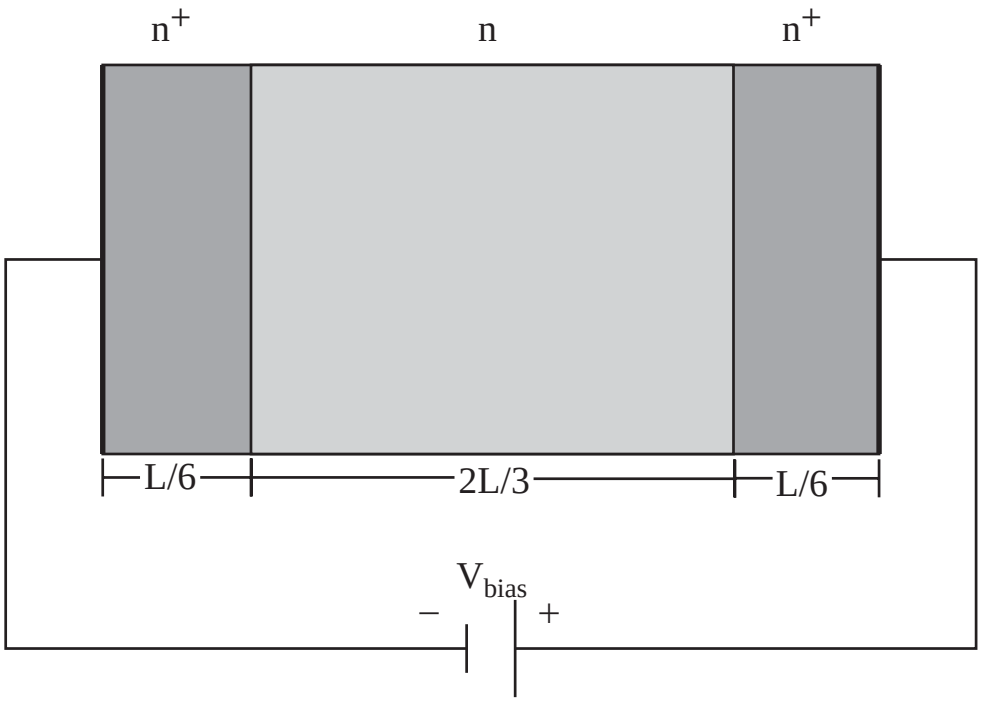

Figure 6.1: The one dimensional n-n^+-n diode of length L.

Current is driven by an external battery that creates a potential across the diode.

The temperature of the diode is

$$T_\ell \equiv \frac{m_e^* \theta_\ell}{k_B} = 300\,\mathrm{K}.$$

We apply the following boundary conditions, which are consistent with thermal equilibrium at the boundary:

$$n(0) = D(0)\,, \qquad n(L) = D(L)\,, \tag{6.6a}$$

$$\partial_x u(0) = \partial_x u(L) = 0\,, \tag{6.6b}$$

$$\partial_x \theta(0) = \partial_x \theta(L) = 0\,, \tag{6.6c}$$

$$\Phi(0) = 0\,, \qquad \Phi(L) = V_{\mathrm{bias}}\,. \tag{6.6d}$$

6.2 Drift-Diffusion

6.2.1 Non-Dimensionalization

We begin by recasting the hydrodynamic model (6.1) in a non-dimensional form. To this end, we define independent variables $\hat{x}$ and $\hat{t}$ by

$$x = x_0 \hat{x}\,, \qquad t = t_0 \hat{t}\,.$$

Here carats denote non-dimensional variables and "naught" subscripts denote reference values associated with each dimensional variable. We also define the non-

dimensional dependent variables that are functions of $\hat{x}$ and $\hat{t}$:

$$n = n_0 \hat{n}\,, \qquad u = u_0 \hat{u}\,, \qquad \theta = \theta_0 \hat{\theta}\,, \qquad \Phi = [\Phi_0]\,\Phi$$

$$\tau_p = \tau_0 \hat{\tau}_p\,, \qquad \tau_w = \tau_0 \hat{\tau}_w\,, \qquad D = D_0 \hat{D}\,.$$

Here the notation $[\Phi_0]$ denotes the potential drop across the device.

The hydrodynamic model is characterized by three energy scales: the kinetic energy u_0^2, the thermal energy θ_0, and the potential energy drop $\frac{q_e}{m_e^*}[\Phi_0]$. Associated with each of these energies scales is a reference velocity: the bulk velocity u_0, the thermal velocity $\theta_0^{1/2}$, and the drift velocity v_E. The thermal velocity is typically given by the lattice temperature, i.e., $\theta_0^{1/2} = \theta_\ell^{1/2}$. To find a value for the drift velocity, we consider a particle initially at rest at time zero that is accelerated by a constant electric field E_0. Just before a collision at time τ_0, the particle has velocity

$$v_E = \frac{qE_0}{m}\tau_0\,,$$

If $E_0 = -\frac{[\Phi_0]}{x_0}$, then the drift velocity is

$$v_E = \varepsilon \frac{q_e}{m_e^*}\frac{[\Phi_0]}{\theta_0^{1/2}}\,.$$

The bulk velocity is then given by

$$u_0 = \min(\theta_0^{1/2}, v_E)\,.$$

This will ensure that dynamics at both the thermal and drift velocities will be taken into account.

The hydrodynamic model has two important time scales. The reference time t_0 is defined as the time it takes a particle of speed u_0 to traverse the distance x_0:

$$x_0 = u_0 t_0 \,. \tag{6.7}$$

The time τ_0 is the mean collision time between an electron and the semiconductor lattice. If $\theta_0^{1/2}$ is the thermal velocity, then

$$\text{mean free path} = \tau_0 \theta_0^{1/2} \,. \tag{6.8}$$

We now introduce several dimensionless ratios and place (6.1) in a non-dimensional form. First is the scaled Knudsen number which is the ratio of the mean free path to the length scale of the device:

$$\varepsilon \equiv \frac{\theta_0^{1/2} \tau_0}{x_0} \,. \tag{6.9}$$

Together (6.7)-(6.9) relate the two times scales τ_0 and t_0 via the thermal and bulk velocities:

$$\frac{t_0}{\tau_0} = \varepsilon \frac{\theta_0^{1/2}}{u_0} \,. \tag{6.10}$$

Next are the dimensionless velocity ratios

$$\eta \equiv \frac{u_0}{\theta_0^{1/2}}\,, \qquad \delta \equiv \frac{v_E}{\theta_0^{1/2}} = \varepsilon \frac{q_e}{m_e^*}\frac{[\Phi_0]}{\theta_0}\,,$$

which measure the relative size of the bulk to thermal velocity and the drift to thermal velocity. With these ratios, the non-dimensional hydrodynamic model is (dropping hats)

$$\partial_t n + \partial_x (nu) = 0\,, \tag{6.11a}$$

$$\partial_t (nu) + \partial_x \left(nu^2 + \frac{1}{\eta^2} n\theta + \frac{\varepsilon}{\eta}\sigma \right) - \frac{\delta}{\varepsilon\eta^2} n\partial_x \Phi = -\frac{1}{\eta\varepsilon}\frac{1}{\tau_p} nu\,, \tag{6.11b}$$

$$\begin{aligned} &\partial_t \left(\frac{nu^2}{2} + \frac{1}{\eta^2}\frac{3n\theta}{2} \right) \\ &\qquad + \partial_x \left(\frac{nu^3}{2} + \frac{1}{\eta^2}\frac{5nu\theta}{2} + \frac{\varepsilon}{\eta}\sigma u + \frac{\varepsilon}{\eta^3} q \right) \\ &\qquad\qquad - \frac{\delta}{\varepsilon\eta^2} nu\partial_x \Phi = -\frac{1}{\eta\varepsilon}\frac{1}{\tau_w}\left(\frac{nu^2}{2} + \frac{1}{\eta^2}\frac{3n(\theta - 1)}{2} \right)\,. \end{aligned} \tag{6.11c}$$

where q and σ retain the form given in (6.2) but with rescaled variables.

6.2.2 The Drift-Diffusion Scaling

The behavior of solutions to (6.11) depends heavily on the relative sizes of ε, η, and δ. The *drift-diffusion scaling* assumes that the potential energy and thermal energy balance, in which case $\delta = \varepsilon$ and $v_E = \varepsilon\theta_0^{1/2}$. In light of (6.10), $u_0 = v_E$ and $\eta = \varepsilon$,

in which case (6.11) becomes

$$\partial_t n + \partial_x (nu) = 0\,, \tag{6.12a}$$

$$\partial_t (nu) + \partial_x \left(nu^2 + \frac{1}{\varepsilon^2} n\theta + \sigma \right) - \frac{1}{\varepsilon^2} n \partial_x \Phi = -\frac{1}{\varepsilon^2} \frac{1}{\tau_p} nu\,, \tag{6.12b}$$

$$\begin{aligned} \partial_t \left(\frac{nu^2}{2} + \frac{1}{\varepsilon^2} \frac{3n\theta}{2} \right) & \\ + \partial_x \left(\frac{nu^3}{2} + \frac{1}{\varepsilon^2} \frac{5nu\theta}{2} + \sigma u + \frac{1}{\varepsilon^2} q \right) & \\ - \frac{1}{\varepsilon^2} nu \partial_x \Phi = -\frac{1}{\varepsilon^2} \frac{1}{\tau_w} \left(\frac{nu^2}{2} + \frac{1}{\varepsilon^2} \frac{3n(\theta - 1)}{2} \right)\,. & \end{aligned} \tag{6.12c}$$

The Poisson equation must also be recast in non-dimensional form. It turns out that the electron concentration is dominated by the doping profile so that $D_0 = n_0$. Hence, the non-dimensional Poisson equation is (in scaled variables)

$$-\lambda^2 \partial_x (\epsilon \partial_x \Phi) = (D - n)\,, \tag{6.13}$$

where the scaled Debye length is given by

$$\lambda = \frac{\epsilon_0 [\Phi_0]}{q D_0 x_0^2}\,. \tag{6.14}$$

This parameter characterizes the distance over which the potential responds to variations in the charge distribution. In particular, is determines the effective thickness of the diode junctions. In practice, λ must be a small fraction of the device length in order to maintain well-defined source, channel, and drain regions. For our numerical

experiments, we would like maintain a constant value of λ and therefore impose the condition that $D_0 x_0^2$ be held constant as x_0 changes.

6.2.3 The Drift-Diffusion Limit

The drift-diffusion scaling gets its name from the limiting equations derived from (6.12) in the limit $\varepsilon \to 0$. Formally, this limit (in non-dimensional variables) that

$$\partial_t n + \partial_x(\tau_p n \partial_x \Phi) = \partial_x \left(\tau_p \partial_x n\right) \tag{6.15a}$$

$$nu = \tau_p n \partial_x \Phi - \tau_p \partial_x n \tag{6.15b}$$

$$\theta = 1 \tag{6.15c}$$

It is important to note that (6.15a) can be rewritten in conservative form

$$\partial_t n + \partial_x \left(\tau_p e^{\Phi} \partial_x (e^{-\Phi} n)\right) = 0\,.$$

The quantity $e^{-\Phi} n$ is called a *Slotboom variable*. It plays an important role in the behavior of (6.15) near steady-state solutions that satisfy $u = 0$.

6.2.4 Physical Validity

It is important to consider the size of ε, η, and δ is a realistic setting and to assess the validity of the drift-diffusion scaling. At room temperature,

$$\varepsilon \sim \frac{10^{-8}\,\mathrm{m}}{x_0}\,.$$

For devices 15 years ago, $x_0 \sim 10\,\mu$m and $\varepsilon \sim 10^{-3}$. In modern devices , $x_0 \sim 0.1 - 1\,\mu$m and $\varepsilon \sim 10^{-2} - 10^{-1}$. Unlike many other types of kinetic systems, the physics of the collision process requires that $\varepsilon < 1$. In fact, $\varepsilon \sim 10^{-1}$ is approaching the limit of semiconductor operation.

Although the local value of ε is roughly constant throughout the device, the local values of η and δ can vary by several orders of magnitude. In many modern devices, channel sizes are small enough that the local potential energy is larger than the thermal energy near diode junctions, even though ε is still relatively small ($\varepsilon \sim 10^{-2}$). Rather than apply the drift-diffusion scaling, it is more appropriate in these cases to set $\delta = \eta = 1$. The non-dimension form of (6.1) becomes

$$\begin{aligned}
\partial_t n + \partial_x (nu) =& 0\,, \\
\partial_t (nu) + \partial_x \left(nu^2 + n\theta + \varepsilon\sigma\right) - \frac{1}{\varepsilon} n \partial_x \Phi =& -\frac{1}{\varepsilon}\frac{1}{\tau_p} nu\,, \\
\partial_t \left(\frac{nu^2}{2} + \frac{3n\theta}{2}\right) + \partial_x \left(\frac{nu^3}{2} + \frac{5nu\theta}{2} + \varepsilon\sigma u + \varepsilon q\right) & \\
- \frac{1}{\varepsilon} nu \partial_x \Phi =& -\frac{1}{\varepsilon}\frac{1}{\tau_w}\left(\frac{nu^2}{2} + \frac{3n(\theta - 1)}{2}\right).
\end{aligned}$$

This scaling is known as the *drift-collision scaling* due to the leading-order balance between the collision terms on the right-hand side and the field terms on the left-hand side of the momentum and energy equations. Asymptotic limits for the drift-collision scaling have been studied in [21], and a numerical investigation of the $n^+ - n - n^+$ diode can be found in [18]. It should be noted that (6.1) may not be an appropriate model for this scaling. This is because the closure is derived based on the assumption

that forces due to the electric field are dominated by collisional effects. As mentioned in Chapter 3, the constitutive relations for σ and q must be re-examined when the electric field is large. High field effects in the context of hydrodynamic models will be the subject of future work.

6.2.5 Preview of Numerical Issues

A desirable property for any numerical scheme used to simulate (6.1) is that it recover the drift-diffusion behavior given by (6.4) when ε is small—that is when devices are large relative to the mean free path. It is clear from (6.12) that the relaxation, field, *and* flux terms are all stiff when ε is small, i.e., when the device size is large. In particular, the wave speeds of (6.12) are

$$\lambda = u,\ u \pm \frac{1}{\varepsilon}\sqrt{\frac{5}{3}\theta}\,.$$

Stiffness leads to two problems. The first, which is fairly obvious, is that when $\varepsilon \ll \Delta x$, the stiff terms in (6.12) imply a time step condition that is much more restrictive than the explicit diffusive condition for that is natural for (6.15c), i.e., $\Delta t \sim (\Delta x)^2$. For the relaxation and field terms these restrictions can be overcome by an implicit time discretization. However, for Godunov-type schemes that employ spatial reconstructions beyond first-order, the implicit evaluation of flux terms is not very practical. This is because the reconstructions—whether they be slope-fitting or ENO or WENO—are very nonlinear and often discontinuous functions of the cell data.

The second, more subtle problem is excessive numerical dissipation. Even in the semi-discrete case, Godunov type schemes introduce numerical dissipation in positive correlation with the size of the wave speeds of the linearized flux matrix. In particular, a semi-discrete differencing of (6.12a) will result in an approximation for the spatial derivative of the momentum that looks like

$$\frac{d}{dt}n_j = -\frac{(nu)_{j+1/2} - (nu)_{j-1/2}}{\Delta x} = -\partial_x (nu) \ + \text{higher-order terms}$$

Included in the higher-order terms is numerical dissipation. In the drift-diffusion approximation, u is given by (6.15b) in which case

$$\frac{d}{dt}n_j = -\partial_x (\tau_p n \partial_x \Phi - \tau_p \partial_x n) \ + \text{higher-order terms}$$

which appears to be a consistent discretization for (6.15b). However, when $\varepsilon \ll \Delta x$ the numerical dissipation can be quite large—comparable to or even greater that the physical diffusion term $\partial_x \tau_p \partial_x n$. A more detailed calculation of this phenomenon is given in the next section a simple 2×2 linear model.

Another issue is that of *balance*. Numerical solutions of (6.1) are often characterized by non-physical oscillations in the current profile. For steady-state solutions, the current should be constant in space. However, many time-dependent schemes evolve to a steady-state in which large oscillations appear at the diode junctions. The size of these oscillations depends on the sharpness of the junction. For the sharp doping profile given in (6.5), the size of these oscillations can be of the same order

as the current itself. For transient solutions, such behavior can even break down a numerical scheme [9].

6.3 Numerical Background

In this section, we present previous work on the numerical issues introduced at the end of the last section, and discuss what ideas may carry over the hydrodynamic model. We begin with a simple 2×2 stiff hyperbolic system with relaxation and then consider the presence of additional source terms. We also present some results based on previous computations of the hydrodynamic model to emphasize the problem with current oscillations at the diode junctions.

6.3.1 A Model Problem

The stiffness and numerical issues introduced in the last section can be understood through the study of the model problem

$$\begin{aligned} \partial_t n + \partial_x m &= 0\,, \\ \partial_t m + \frac{1}{\varepsilon^2}\partial_x n &= -\frac{1}{\varepsilon^2} m\,, \end{aligned} \tag{M1}$$

which as $\varepsilon \to 0$ is approximated by the diffusion equation

$$\partial_t n = \partial_x^2 n\,, \qquad m = -\partial_x n\,. \tag{6.16}$$

Here, it is understood that the momentum $m = nu$. The only contribution to the momentum flux here is the pressure $p = \varepsilon^{-2} n$. The temperature is constant and equal to one. Numerical studies of (M1) can be found in [40, 55, 64] and references therein. Like the hydrodynamic model, (M1) is stiff when $\varepsilon \ll 1$. In addition to the obvious time step restrictions, the stiff flux in the momentum equation of (M1) creates excessive numerical dissipation in the concentration equation of (M1), thereby reducing its accuracy when approximating for (6.16) near the diffusive limit.

6.3.1.1 Numerical Diffusion For a given mesh size Δx, most numerical schemes of order s compute exact solution to a modified equation [50] for (M1) of the form

$$\partial_t n + \partial_x m + \sum_{k=s}^{\infty} \left(a_k \, \partial_x^{(k)} m + b_k \, \partial_x^{(k)} n \right) (\Delta x)^k = 0 \,,$$
$$\partial_t m + \frac{1}{\varepsilon^2} \partial_x n + \sum_{k=s}^{\infty} \left(c_k \, \partial_x^{(k)} m + d_k \, \partial_x^{(k)} n \right) (\Delta x)^k = -\frac{1}{\varepsilon^2} m \,,$$

where the coefficients a_k, b_k, c_k, and d_k depend on ε. For example, a center-difference, second-order, upwind scheme for (M1) has a modified equation

$$\partial_t n + \partial_x m - \frac{1}{12} (\Delta x)^2 \partial_x^3 m + \frac{1}{8} \frac{(\Delta x)^3}{\varepsilon} \partial_x^4 n + O\left(\varepsilon, (\Delta x)^4\right) = 0 \,, \tag{6.17a}$$
$$\partial_t m + \frac{1}{\varepsilon^2} \partial_x n - \frac{1}{12} \frac{(\Delta x)^2}{\varepsilon^2} \partial_x^3 n + \frac{1}{8} \frac{(\Delta x)^3}{\varepsilon} \partial_x^4 m + O\left(\varepsilon, (\Delta x)^4\right) = -\frac{1}{\varepsilon^2} m \,. \tag{6.17b}$$

If terms in (6.17b) are balanced according to powers of ε, then

$$m = -\partial_x n + \frac{1}{12} (\Delta x)^2 \partial_x^3 n + O\left(\varepsilon, (\Delta x)^3\right) \,,$$

Substituting this expression for m into (6.17a) gives the following modified diffusion equation

$$\partial_t n - \partial_x^2 n - \frac{1}{12}(\Delta x)^2 \partial_x^3 m + \frac{1}{8}\frac{(\Delta x)^3}{\varepsilon}\partial_x^4 n + O(\varepsilon, (\Delta x)^3) = 0 \tag{6.18}$$

which is inconsistent with (6.16) in the limit

$$\varepsilon \to 0\,, \qquad \Delta x \to 0\,, \qquad \frac{(\Delta x)^3}{\varepsilon} = \text{const.}$$

In practice, solutions to (6.17) will be smeared by the numerical dissipation from the term $\partial_x^4 n$ in (6.18) whenever $(\Delta x)^3$ is a reasonable fraction of ε [55]. It should also be noted that slope-limiting does not correct this problem, and that a similar result holds for central schemes. However, discontinuous Galerkin methods can remove the numerical dissipation is some cases [55].

6.3.1.2 Simple Splitting Approaches In [40], a split scheme is introduced to address the problems associated with the stiff system. The scheme consists of two steps. The first is a relaxation step:

$$\partial_t n = 0\,, \tag{6.19a}$$

$$\partial_t m + \left(\frac{1}{\varepsilon^2} - 1\right)\partial_x n = -\frac{1}{\varepsilon^2} m\,, \tag{6.19b}$$

followed by a convection step:

$$\partial_t n + \partial_x m = 0\,, \tag{6.20a}$$

$$\partial_t m + \partial_x n = 0\,. \tag{6.20b}$$

When $\varepsilon \ll 1$, the (6.19b) projects the solution into the diffusion balance

$$m = -\partial_x n + O(\varepsilon^2)\,. \tag{6.21}$$

This property has been shown in [39] to be an important aspect of capturing the proper behavior described in (6.16) when $\varepsilon \ll 1$. Meanwhile, the convective step is a hyperbolic system with wave speeds that are independent of ε.

When $\varepsilon \ll \Delta x$, the splitting improves efficiency by relaxing the hyperbolic CFL condition of the original stiff system. Because $\partial_t n = 0$ in the relaxation step, implicit and explicit updating the stiff flux term in (6.19b) is the same. Given that the convective step (6.20) is updated explicitly, the natural CFL condition for the entire scheme in the diffusive regime is $\Delta t \sim (\Delta x)^2 \gg \varepsilon \Delta x$.

The splitting also removes excessive numerical dissipation when $\varepsilon \ll \Delta x$. The modified equations for (6.20) for a center-difference, second-order, upwind scheme are

$$\partial_t n + \partial_x m - \frac{1}{12}(\Delta x)^2 \partial_x^3 m + \frac{1}{8}(\Delta x)^3 \partial_x^4 n + O(\Delta x)^4 = 0\,, \tag{6.22a}$$

$$\partial_t m + \partial_x n - \frac{1}{12}(\Delta x)^2 \partial_x^3 n + \frac{1}{8}(\Delta x)^3 \partial_x^4 m + O(\Delta x)^4 = 0\,. \tag{6.22b}$$

Substituting (6.21) into (6.22a) gives

$$\partial_t n = \partial_x^2 n + O(\varepsilon^2, (\Delta x)^2)\,,$$

which lacks the numerical dissipation term $O((\Delta x)^3/\varepsilon)$ found in (6.18).

6.3.2 Systems of Balance Laws

Simulation of balance laws containing source terms with spatial derivatives in non-divergent form is a challenging task. Numerical schemes often fail to capture key physical features of a system because of the difficulty involved with capturing the delicate balance of forces found at the continuum level. Such is the case in the following system, which is obtained by adding a source term and a convective term to the momentum equation in (M1):

$$\begin{aligned} \partial_t n + \partial_x m &= 0\,, \\ \partial_t m + \partial_x \left(\frac{m^2}{n} + \frac{1}{\varepsilon^2} n \right) - \frac{1}{\varepsilon^2} n \partial_x z &= -\frac{\nu}{\varepsilon^2} m\,. \end{aligned} \tag{M2}$$

Here z is either given or solved self-consistently as a function of n. The quantity ν is a relaxation rate or a friction coefficient that depends on n and m. Note that (M2) is essentially the drifted-diffusion system from in Chapter 3. A study of stiff numerics in this context can be found in [64].

Recently, *well-balanced schemes* [6,15,24,88] have been developed for solving (M2) in several special cases. These schemes are devised in order preserves certain steady-

state solutions which formally satisfy

$$\partial_x m = 0\,, \tag{6.23a}$$

$$\partial_x \left(\frac{1}{2} u^2 + \frac{1}{\varepsilon^2} \log(n) - \frac{1}{\varepsilon^2} z \right) = -\frac{\nu}{\varepsilon^2} u\,. \tag{6.23b}$$

When $\nu = 0$, the steady-state solutions are

$$m = Const. \tag{6.24a}$$

$$\frac{1}{2} u^2 + \frac{1}{\varepsilon^2} \log(n) - \frac{1}{\varepsilon^2} z = Const. \tag{6.24b}$$

The idea of well-balanced schemes is to construct a conservative approximation of (M2) that incorporates the information from (6.24) in order to preserve the steady-state:

$$\frac{d}{dt}\mathbf{u}_i + \frac{\hat{\mathbf{F}}_{i+1/2} - \hat{\mathbf{F}}_{i-1/2}}{\Delta x} = 0\,,$$

where $\mathbf{u} = (n, m)$ and

$$\hat{\mathbf{F}}_{i+1/2} = F(\mathbf{u}_{i+1}, \mathbf{u}_i, z_i, z_{i+1})$$

for some smooth function F that satisfies an appropriate consistency condition. For example, when $u = 0$, the Slotboom variable $e^{-z}n$ is constant. This fact is utilized in [24] to develop well-balanced schemes for chemotaxis models when $\nu = 0$ near the steady-state $u = 0$. A more general approach is proposed in [15] to model systems near any subsonic steady state. The scheme is applied to the Saint-Venant sys-

tem of shallow water equations and the Euler-Poisson system for collisionless charge transport, which are similar to (M2) except for the right-hand side is zero.

The well-balanced approaches from these examples are not directly applicable to the hydrodynamic model for several reasons. First of all, both conditions $\nu = 0$ and $u = 0$ are far from being satisfied, and the Slotboom variable $e^{-\Phi}n$ may vary by 10 to 20 orders of magnitude over the length of the diode. Furthermore, the hydrodynamic model has a temperature dependent pressure given by $p = p(n, \theta) = \varepsilon^{-2} n\theta$, where the evolution of θ is derived from an independent equation (6.12c) for the energy. Even if $\sigma = q = 0$ in (6.12), the statement analogous to (6.23b) for the hydrodynamic model is

$$\partial_x \left(\frac{1}{2} u^2 + \theta - \frac{1}{\varepsilon^2} \Phi \right) + \frac{1}{\varepsilon^2} \theta \, \partial_x \log(n) = -\frac{1}{\varepsilon^2} \frac{1}{\tau_p} u$$

which has the form of (6.23b) if and only if $\theta = 1$.

6.3.3 Previous Computations of the Hydrodynamic Model

Even though there are several versions of the hydrodynamic model in the literature with different expressions for σ, q, τ_p, the general behavior of solutions and the major computational issues are essentially the same. Steady state computations based on iterative methods can be found in [28, 29, 63], but our focus is on time-evolution methods, particular Godunov-type schemes. Once such scheme can be found in [25], where the authors use a sixth-order ENO reconstruction method with explicit, first-order time steps. In cases when $\varepsilon \ll 1$ this scheme is stiff, and in the drift-diffusion regime, it is subject to all of the limitations discussed above. These limitations are

partially circumvented by the high-order accuracy of the scheme.

In [72], split scheme is devise central schemes. There, the splitting separates the system into a relaxation component:

$$\partial_t n = 0\,, \tag{6.25a}$$

$$\partial_t (nu) + \frac{\varepsilon}{\eta}\partial_x \sigma - \frac{\delta}{\varepsilon\eta^2} n\partial_x \Phi = -\frac{1}{\eta\varepsilon}\frac{1}{\tau_p} nu\,, \tag{6.25b}$$

$$\partial_t \left(\frac{nu^2}{2} + \frac{1}{\eta^2}\frac{3n\theta}{2}\right) + \partial_x \left(\frac{\varepsilon}{\eta}\sigma u + \frac{\varepsilon}{\eta^3} q\right) - \frac{\delta}{\varepsilon\eta^2} nu\partial_x \Phi = -\frac{1}{\eta\varepsilon}\frac{1}{\tau_w}\left(\frac{nu^2}{2} + \frac{1}{\eta^2}\frac{3n\,(\theta-1)}{2}\right)\,. \tag{6.25c}$$

and a convective component

$$\partial_t n + \partial_x (nu) = 0\,, \tag{6.26a}$$

$$\partial_t (nu) + \partial_x \left(nu^2 + \frac{1}{\eta^2} n\theta\right) = 0\,, \tag{6.26b}$$

$$\partial_t \left(\frac{nu^2}{2} + \frac{1}{\eta^2}\frac{3n\theta}{2}\right) + \partial_x \left(\frac{nu^3}{2} + \frac{1}{\eta^2}\frac{5nu\theta}{2}\right) = 0\,, \tag{6.26c}$$

that has the form of the Euler equations for a compressible neutral-particle gas. This splitting is compatible with the drift-collision scaling when $\eta = \delta = 1$ and $\varepsilon \ll 1$. In such cases, only (6.25) is stiff and implicit methods can be used. However, in the drift-diffusion regime when $\eta = \delta = \varepsilon \ll 1$, this splitting does not address the restrictive CFL condition or numerical dissipation associated with stiff flux terms. Moreover, central schemes are not recommended for computing steady-state solutions with small time steps due to the cumulative effects of numerical dissipation [46].

Another important point is that the numerical experiments in [25] and [72] use a smoothed doping profile. The result is that the gradients of n and Φ are much smaller at the junctions than for the stiff doping profile given in (6.5). Even with such smoothing, oscillations still exist. In fact, it is noted in [25] that, even will the smoothed doping profile, the accuracy of a third-order ENO scheme is not sufficient to remove oscillations from the steady-state current profile.

To get an idea of the oscillatory nature of the numerical current at the junctions, we compute steady-state results for an explicit non-split scheme and a scheme based on the splitting in (6.25)-(6.26). The former will be denoted $S1$ and the latter $S2$. Both schemes evaluate numerical fluxes using central-upwind techniques [45] that are second-order in space. The time step for $S1$ is

$$\Delta t = c \min\left(\frac{\Delta x}{2\, sp(\mathbf{A})}, \Delta x^2, \frac{1}{\eta\varepsilon}\right) \tag{6.27}$$

where $\mathbf{A}$ is the linearized flux matrix, $\rho(\mathbf{A})$ is its spectral radius, and $c < 1$ is an $O(1)$ constant. For $S2$, the fact that (6.25) can be updated implicitly allows for a less restrictive time step

$$\Delta t = c\frac{\Delta x}{2\, sp(\mathbf{A})}\,. \tag{6.28}$$

We note that $\rho(\mathbf{A}) = O(1/\eta)$.

Simulation results for $S1$ and $S2$ are given in Figure 6.2 and Figure 6.3, respectively. The schemes are tested with different mesh size, time step, and doping profiles. The oscillatory nature of the steady-state current is clear. For the non-split scheme,

these oscillations are not reduced by decreasing the time steps or by increasing the temporal-order. Increasing the spatial resolution does help as does the introduction of a smoothed doping profile

$$\tilde{D} = D - \frac{0.99}{2}\left(\tanh\left(\frac{x - L/6}{0.2}\right) - \tanh\left(\frac{x - 5L/6}{0.2}\right)\right), \tag{6.29}$$

where D is given in (6.5). However, a smoothed doping profile significantly alters the value of the steady-state current, which we would like to avoid.

For the split scheme, decreasing the size of the time steps does reduce current oscillations somewhat as does the implementation of the second-order time marching used in [72]. However, the oscillations are still quite large. Increasing the spatial resolution has mush less effect than it does for $S1$, and smoothing the doping profile does not help much either.

6.4 A New Splitting Approach to the Hydrodynamic Model

In this section, we present a new approach based on the split method in [40]. We find that splitting the hydrodynamic model in a way that respects the drift-diffusion balance yields a scheme that bypasses the strict CFL condition and removes the excessive numerical dissipation in the drift-diffusion regime. At the same time, it significantly reduces oscillations in the steady-state solution for a range of ε that extends well into the transition regime.

Our split scheme will be expressed in terms of n, the current $m = nu$ (called momentum in previous contexts), and the *relative energy*

$$r = \frac{nu^2}{2} + \frac{1}{\varepsilon^2}\frac{3n\,(\theta - 1)}{2}\,,$$

which is identically zero at equilibrium ($u = 0$, $\theta = 1$). Rewriting (6.12) in these new variables gives:

$$\begin{aligned}
\partial_t n + \partial_x\,(nu) &= 0\,, \\
\partial_t m + \partial_x\left(\frac{2}{3}\frac{m^2}{n} + \frac{2}{3}r + \frac{1}{\varepsilon^2}n + \sigma\right) - \frac{1}{\varepsilon^2}n\partial_x\Phi &= -\frac{1}{\varepsilon^2}\frac{1}{\tau_p}m\,, \qquad \text{(H)} \\
\partial_t r + \partial_x\left(\frac{5}{3}\frac{rm}{n} - \frac{1}{3}\frac{m^3}{n^2} + \frac{1}{\varepsilon^2}m + \sigma u + \frac{q}{\varepsilon^2}\right) - \frac{1}{\varepsilon^2}nu\partial_x\Phi &= -\frac{1}{\varepsilon^2}\frac{1}{\tau_w}r\,,
\end{aligned}$$

where the anisotropy and heat flux are

$$\sigma = \frac{4}{3}n\theta\tau_p\partial_x\left(\frac{m}{n}\right)\,, \qquad q = \varepsilon^2\frac{5}{2}n\theta\tau_p\partial_x\left(\frac{2}{3}\frac{r}{n} - \frac{1}{3}\frac{m^2}{n^2}\right)\,. \tag{6.30}$$

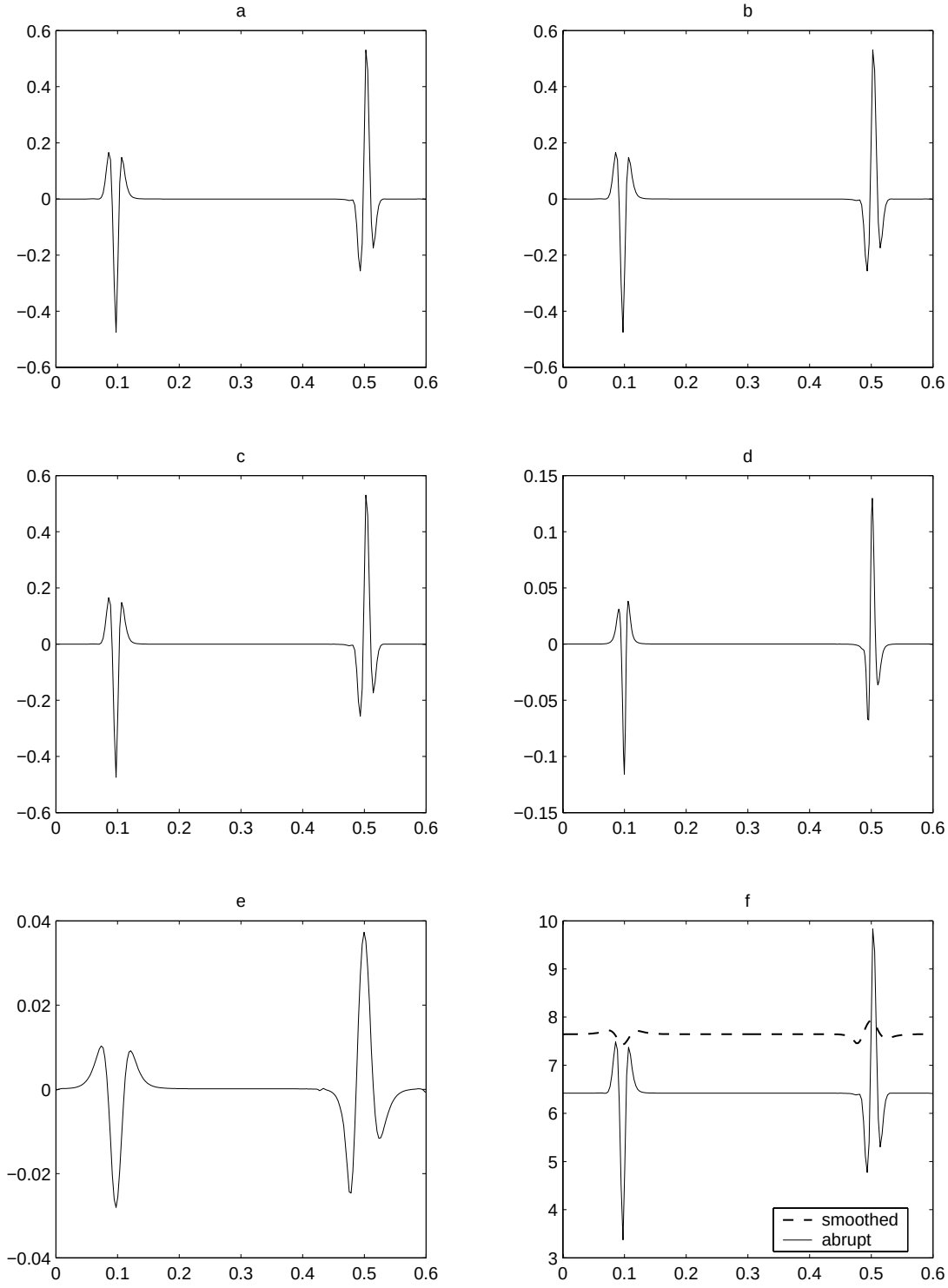

Figure 6.2: Steady state current oscillations around the spatial average for diode of length $L = 0.6 \times 10^{-6}$ m. The scheme is S1 with time step given in (6.27) (a) First order in time, 200 points, $c = 0.2$. (b) First order in time, 200 points, $c = 0.02$. (c) Second order in time, 200 points, $c = 0.2$. (d) First order in time, 400 points, $c = 0.2$. (e) First order in time, 200 points, $c = 0.02$, with smoothed doping (6.29). (f) Comparison of scaled current from (a) and (e).

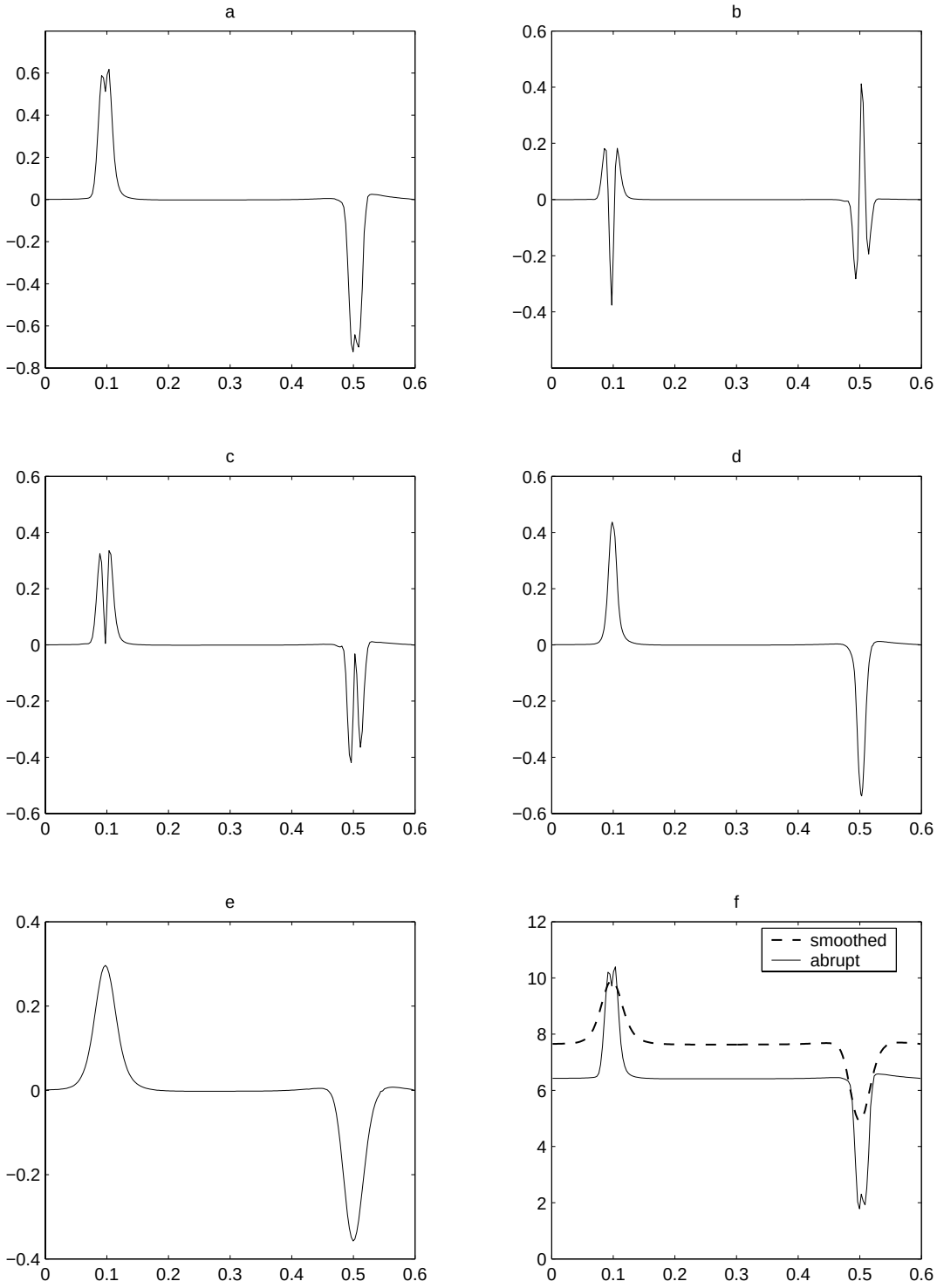

Figure 6.3: Steady state current oscillations around the spatial average for diode of length $L = 0.6 \times 10^{-6}$ m. The scheme is $S2$ with time step given in (6.28) (a) First order in time, 200 points, $c = 0.2$. (b) First order in time, 200 points, $c = 0.02$. (c) Second order in time, 200 points, $c = 0.2$. (d) First order in time, 400 points, $c = 0.2$. (e) First order in time, 200 points, $c = 0.02$, with smoothed doping (6.29). (f) Comparison of scaled current from (a) and (e).

6.4.1 Two Step Splitting

Following the splitting for the model problem (M1), we break (H) into a relaxation step:

$$\begin{aligned}
\partial_t n &= 0\,, \\
\partial_t m + \partial_x \left(\frac{1}{\varepsilon^2} n + \sigma \right) - \frac{1}{\varepsilon^2} n \partial_x \Phi &= -\frac{1}{\varepsilon^2} \frac{1}{\tau_p} m\,, \\
\partial_t r + \partial_x \left(\frac{1}{\varepsilon^2} m + \frac{\sigma m}{n} + \frac{q}{\varepsilon^2} \right) - \frac{1}{\varepsilon^2} m \partial_x \Phi &= -\frac{1}{\varepsilon^2} \frac{1}{\tau_p} r\,,
\end{aligned}$$

and a convection step:

$$\begin{aligned}
\partial_t n + \partial_x m &= 0\,, \\
\partial_t m + \partial_x \left(\frac{m^2}{n} + r \right) &= 0\,, \\
\partial_t r + \partial_x \left(\frac{5}{3} \frac{rm}{n} - \frac{1}{3} \frac{m^3}{n^2} \right) &= 0\,.
\end{aligned}$$

The relaxation step projects m and r into the correct drift-diffusion limit, and the convective step looks like the Euler equations. Because r is the *relative* energy, the convective step has wave speeds

$$u,\ u \pm \sqrt{\frac{5}{3}(\theta - 1)}\,.$$

Thus, although the convective step is not stiff, it fails to be hyperbolic when $\theta < 1$. It so happens that this condition occurs at the source junction of the diode where the electric field traps low-energy electrons. We therefore try a new approach.

6.4.2 Three Step Splitting

The problem with the two step splitting is essentially the nonlinear terms. We therefore remove the nonlinear convective components of (H),

$$\partial_t n = 0\,, \tag{6.31a}$$

$$\partial_t m + \partial_x \left(\frac{2}{3}\frac{m^2}{n} \right) = 0\,, \tag{6.31b}$$

$$\partial_t r + \partial_x \left(\frac{5}{3}\frac{rm}{n} - \frac{1}{3}\frac{m^3}{n^2} \right) = 0\,. \tag{6.31c}$$

which, by themselves, form a hyperbolic system with wave speeds

$$\lambda = 0,\ \frac{4}{3}u,\ \frac{5}{3}u\,.$$

The remainder of (H) is

$$\partial_t n + \partial_x (nu) = 0\,, \tag{6.32a}$$

$$\partial_t m + \partial_x \left(\frac{2}{3}r + \frac{1}{\varepsilon^2}n + \sigma \right) - \frac{1}{\varepsilon^2} n \partial_x \Phi = -\frac{1}{\varepsilon^2}\frac{1}{\tau_p} m\,, \tag{6.32b}$$

$$\partial_t r + \partial_x \left(\frac{1}{\varepsilon^2}m + \frac{\sigma m}{n} + \frac{q}{\varepsilon^2} \right) - \frac{1}{\varepsilon^2} nu \partial_x \Phi = -\frac{1}{\varepsilon^2}\frac{1}{\tau_w} r\,. \tag{6.32c}$$

It should be noted that if $\sigma = q = 0$ and we identify $\tau_p = \mu^{-1}$, then equations (6.32a)-(6.32b) recover (M2) exactly when $\theta = 1$.

The remainder (6.32) can now be broken into two parts: a relaxation step,

$$\partial_t n = 0\,, \tag{6.33a}$$

$$\partial_t m + \partial_x \left(\left(\frac{1}{\varepsilon^2} - 1 \right) n + \sigma \right) - \frac{1}{\varepsilon^2} n \partial_x \Phi = -\frac{1}{\varepsilon^2} \frac{1}{\tau_p} m\,, \tag{6.33b}$$

$$\partial_t r + \partial_x \left(\left(\frac{1}{\varepsilon^2} - 1 \right) m + \frac{\sigma m}{n} + \frac{q}{\varepsilon^2} \right) - \frac{1}{\varepsilon^2} n u \partial_x \Phi = -\frac{1}{\varepsilon^2} \frac{1}{\tau_w} r\,. \tag{6.33c}$$

and a convective step,

$$\partial_t n + \partial_x m = 0\,, \tag{6.34a}$$

$$\partial_t m + \frac{2}{3} \partial_x r + \partial_x n = 0\,, \tag{6.34b}$$

$$\partial_t r + \partial_x m = 0\,, \tag{6.34c}$$

that is hyperbolic with wave speeds

$$\lambda = 0, \pm \sqrt{\frac{5}{3}}\,.$$

The relaxation step (6.33) projects the variables into the drift-diffusion limit. By freezing τ_p and τ_w at the current time, this step can be updated implicitly. Since $\partial_t n = 0$ in (6.33a), an implicit evaluation of n is trivial and (6.33b) is updated using the current value of n. The new value of m is then used to update (6.33c). In this way, we obtain an easily implemented semi-implicit scheme. Note that σ and q may also be updated implicitly. Even though these terms are not stiff (recall from (6.30) that $q = O(\varepsilon^2)$) and do not play a role in the drift-diffusion balance, their explicit

evaluation will require that a diffusive step $\Delta t \sim (\Delta x)^2$ be enforced. Although we desire this type of time step for $\varepsilon \ll 1$, it becomes restrictive when $\varepsilon = O(1)$.

The relaxation step is followed by the linear convective step (6.34) and then the nonlinear convective step (6.32). Comparing powers of ε in (6.33b) gives

$$m = -\partial_x n + n\partial_x \Phi + O(\varepsilon^2)$$

which, when substituted into (6.34a), recovers (6.15c) in the limit $\varepsilon \to 0$. Moreover, since the wave speeds in both convective steps are independent of ε, excessive numerical dissipation is no longer an issue.

6.5 Details of the Scheme

In this section, we present the details of our scheme. The algorithm computes each step in the following order.

1. Relaxation

$$\partial_t n = 0\,, \tag{6.35a}$$

$$\partial_t m + \left(\frac{1}{\varepsilon^2} - 1\right)\partial_x n + \partial_x \sigma - \frac{1}{\varepsilon^2} n\partial_x \Phi = -\frac{1}{\varepsilon^2}\frac{1}{\tau_p} m\,, \tag{6.35b}$$

$$\partial_t r + \left(\frac{1}{\varepsilon^2} - 1\right)\partial_x m + \partial_x\left(\frac{\sigma m}{n} + \frac{q}{\varepsilon^2}\right) - \frac{1}{\varepsilon^2} m\partial_x \Phi = -\frac{1}{\varepsilon^2}\frac{1}{\tau_w} r\,. \tag{6.35c}$$

2. Linear Convection

$$\partial_t n + \partial_x m = 0\,, \tag{6.36a}$$

$$\partial_t m + \frac{2}{3}\partial_x r + \partial_x n = 0\,, \tag{6.36b}$$

$$\partial_t r + \partial_x m = 0\,. \tag{6.36c}$$

3. Nonlinear Convection

$$\partial_t n = 0\,, \tag{6.37a}$$

$$\partial_t m + \partial_x \left(\frac{2}{3}\frac{m^2}{n}\right) = 0\,, \tag{6.37b}$$

$$\partial_t r + \partial_x \left(\frac{5}{3}\frac{rm}{n} - \frac{1}{3}\frac{m^3}{n^2}\right) = 0\,. \tag{6.37c}$$

The electric field is updated after each iteration of these three steps using standard methods. The boundary conditions for the problem are given in (6.6). As in Chapter 5, we over-specify these conditions on both sides of the diode and enforce them after each step in the splitting. The initial condition is determined by setting $m = r = 0$ and solving the steady-state drift-diffusion-Poisson system,

$$\partial_x \left(\tau_p e^{\Phi} \partial_x (e^{-\Phi} n)\right) = 0,$$

$$-\lambda^2 \partial_x \left(\epsilon \partial_x \Phi\right) = (D - n),$$

with boundary conditions given in (6.6a) and (6.6b) and V_{bias} set to zero. The computation is performed with an iterative scheme based on Newton's method. Once this scheme converges, V_{bias} is set to 1.0 Volts, and the ensuing potential drop across the device drives the system to a steady state.

6.5.1 Spatial Discretization

If we let $\mathbf{u} = (n, m, r)^T$, then the convective steps have the form

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = 0$$

where

$$\mathbf{f}_{\text{linear}} = \left(m, \frac{2}{3} r + n, m \right)^T$$
$$\mathbf{f}_{\text{nonlinear}} = \left(0, \frac{2}{3} \frac{m^2}{n}, \frac{5}{3} \frac{rm}{n} - \frac{1}{3} \frac{m^3}{n^2} \right)^T$$

In either case let

$$\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}$$

be the linearized flux matrix with eigenvalues $\lambda_1 < \lambda_2 < \lambda_3$.

We use a finite-volume, central-upwind scheme [45] which updates $\mathbf{u}$ component-wise as

$$\frac{d}{dt} \bar{\mathbf{u}}_i + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{\Delta x} = 0$$

where

$$\bar{\mathbf{u}}_i(t) \equiv \frac{1}{\Delta x} \int_{I_i} \mathbf{u}(x,t)\, dx$$

is the cell average of $\mathbf{u}$ over the interval $I_i = (x_{i-1/2}, x_{i+1/2})$ centered at x_i and $\mathbf{F}_{i+1/2}$ is a numerical flux computed via a reconstruction of $\mathbf{u}$:

$$\mathbf{F}_{i+1/2} = \frac{a^+_{i+1/2}\mathbf{f}(\mathbf{p}_{i+1}(x_{i+1/2})) - a^-_{i+1/2}\mathbf{f}(\mathbf{p}_i(x_{i+1/2}))}{a^+_{i+1/2} - a^-_{i+1/2}} \tag{6.38}$$

$$+ \frac{a^+_{i+1/2}a^-_{i+1/2}}{a^+_{i+1/2} - a^-_{i+1/2}}(\mathbf{p}_{i+1}(x_{i+1/2}) - \mathbf{p}_i(x_{i+1/2}) \,. \tag{6.39}$$

The values $a^{\pm}_{i+1/2}$ depends on the local wave speeds and are given by

$$a^+_{i+1/2} = \max\left\{\lambda_3\left(\mathbf{p}_i(x_{i+1/2})\right),\ \lambda_3\left(\mathbf{p}_{i+1}(x_{i-1/2}\right),\ 0\right\}\ ,$$

$$a^-_{i+1/2} = \min\left\{\lambda_1\left(\mathbf{p}_i(x_{i+1/2}\right),\ \lambda_1\left(\mathbf{p}_{i+1}(x_{i-1/2}\right),\ 0\right\}\ ,$$

where $\mathbf{p}_i$ is a non-oscillatory reconstruction of $\mathbf{u}$ in cell i.

The benefit of the central-upwind scheme is its simplicity. It has a semi-discrete formulation but does not require wave decompositions or a Riemann solver. However, its usefulness here is limited to second-order. This is because evaluating the field and relaxation terms from the relaxation step in a finite volume setting requires a reconstruction process in order to achieve spatial accuracy beyond second-order. In light of the implicit time stepping, it is therefore more natural to work with finite differences when going to higher-order, which means abandoning the central-upwind method. We therefore use a second-order reconstruction with slope limiters to approximate

derivatives of $\mathbf{u}$:

$$\mathbf{p}_i(x) = \bar{\mathbf{u}}_i + \mathbf{u}'_i(x - x_i)\,, \qquad x \in I_i$$

where $\mathbf{u}'_i$ is a slope-limited numerical derivative

$$\mathbf{u}'_i = \text{minmod}\left(\frac{\bar{\mathbf{u}}_{i+1} - \bar{\mathbf{u}}_i}{\Delta x}, \frac{\bar{\mathbf{u}}_{i+1} - \bar{\mathbf{u}}_{i-1}}{2\Delta x}, \frac{\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i-1}}{\Delta x}\right).$$

The CFL condition for this scheme $\Delta t = 0.5 sp(\mathbf{A})\Delta x$ where $sp(\mathbf{A})$ is the spectral radius of $\mathbf{A}$.

The relaxation step can be updated in several ways. Because Φ satisfies a Poisson equation, it is natural to use center differences to compute its derivatives. However, it is not entirely clear how to compute the derivatives n and m. This is because the relaxation system is not hyperbolic. The convective flux in this case is

$$\mathbf{f}(\mathbf{u}) = \begin{pmatrix} 0 \\ n \\ m \end{pmatrix}.$$

(The viscosity and heat flux are diffusive terms and not included in $\mathbf{f}$). Therefore, the matrix $\mathbf{A}$ is degenerate with eigenvalues that are all zero. Even if one were to use a method for hyperbolic problems, it is unclear how to enforce any kind of upwinding at cell interfaces. The central upwind flux (6.38), for example, is not well-defined since

$a^+_{i+1/2} = a^-_{i+1/2} = 0$ for all i. We therefore consider the slightly perturbed system

$$\begin{aligned}
\partial_t n + \omega \partial_x m &= 0\,, \\
\partial_t m + \left(\frac{1}{\varepsilon^2} - 1\right)\partial_x n + \partial_x \sigma - \frac{1}{\varepsilon^2} n \partial_x \Phi &= -\frac{1}{\varepsilon^2}\frac{1}{\tau_p} m\,, \\
\partial_t r + \left(\frac{1}{\varepsilon^2} - 1\right)\partial_x m + \partial_x \left(\frac{\sigma m}{n} + q\right) - \frac{1}{\varepsilon^2} m \partial_x \Phi &= -\frac{1}{\varepsilon^2}\frac{1}{\tau_w} r\,.
\end{aligned}$$

where $\omega > 0$. Since $\varepsilon < 1$, this system *is* hyperbolic with eigenvalues

$$\lambda = 0, \pm\sqrt{\left(\frac{1}{\varepsilon^2} - 1\right)\omega}\,.$$

The numerical flux is

$$\begin{aligned}
\mathbf{F}_{i+1/2} &= \frac{\mathbf{f}(\mathbf{p}_{i+1}(x_{i+1/2})) + \mathbf{f}(\mathbf{p}_i(x_{i+1/2}))}{2} \\
&\quad + \sqrt{\left(\frac{1}{\varepsilon^2} - 1\right)\omega}\,\frac{\mathbf{p}_{i+1}(x_{i+1/2}) - \mathbf{p}_i(x_{i+1/2})}{2}.
\end{aligned}$$

and

$$\mathbf{F}_{i+1/2} = \frac{\mathbf{f}(\mathbf{p}_{i+1}(x_{i+1/2})) + \mathbf{f}(\mathbf{p}_i(x_{i+1/2}))}{2} \qquad \text{as} \qquad \omega \to 0. \tag{6.40}$$

For this reason, the flux terms at the interfaces are approximated by a simple average of the interpolated values from the two adjacent cells.

The remaining terms are discretized in the finite volume setting as follows:

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \partial_x \sigma \; dx = \frac{\sigma_{i+1/2} - \sigma_{i-1/2}}{\Delta x}$$

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \partial_x \left(\frac{\sigma m}{n}\right) \; dx = \frac{1}{\Delta x}\left[\frac{\sigma_{i+1/2} m_{i+1/2}}{n_{i+1/2}} - \frac{\sigma_{i-1/2} m_{i-1/2}}{n_{i-1/2}}\right] + O(\Delta x)^2$$

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \partial_x q \; dx = \frac{1}{\Delta x}\left[q_{i+1/2} - q_{i-1/2}\right]$$

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} n\partial_x \Phi \; dx = \frac{1}{\Delta x}\bar{n}_j \left(\Phi_{i+1} - \Phi_{i-1}\right) + O(\Delta x)^2$$

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} m\partial_x \Phi \; dx = \frac{1}{\Delta x}\bar{m}_j \left(\Phi_{i+1} - \Phi_{i-1}\right) + O(\Delta x)^2$$

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{1}{\tau_p} m \; dx = \frac{1}{\Delta x}\frac{1}{(\tau_p)_i}\bar{m}_i + O(\Delta x)^2$$

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{1}{\tau_p} r \; dx = \frac{1}{\Delta x}\frac{1}{(\tau_w)_i}\bar{r}_i + O(\Delta x)^2$$

Here $n_{i+1/2}$ and $m_{i+1/2}$ are computed by the average in (6.40), while

$$\sigma_{i+1/2} = \frac{\sigma_i + \sigma_{i+1}}{2}\,, \qquad q_{i+1/2} = \frac{q_i + q_{i+1}}{2}\,.$$

and

$$(\tau_p)_i = \tau_p(\bar{\mathbf{u}}_i)\,, \qquad (\tau_w)_i = \tau_w(\bar{\mathbf{u}}_i)\,.$$

The relaxation times are evaluated using a Monte-Carlo fitting [63] of the energy. It must be noted that, physically, these values are only accurate when $L = 0.6\,\mu$m. However, we continue to use these values for a range of length scales for the explicit purpose of computational comparisons.

Another way to discretize the relaxation step is to first use Slotboom variables to

place it in conservation form. Because n is constant in time during the relaxation step, so too is Φ. Therefore, (6.32) is formally equivalent to

$$\partial_t \left(e^{-\psi} n\right) = 0 \,, \tag{6.41a}$$

$$\partial_t \left(e^{-\psi} m\right) + \frac{1-\varepsilon^2}{\varepsilon^2} \partial_x \left(e^{-\psi} n\right) + e^{-\psi} \partial_x \sigma = -\frac{1}{\varepsilon^2} \frac{1}{\tau_p} \left(e^{-\psi} m\right) \,, \tag{6.41b}$$

$$\partial_t \left(e^{-\psi} r\right) + \frac{1-\varepsilon^2}{\varepsilon^2} \partial_x \left(e^{-\psi} m\right) + e^{-\psi} \partial_x \left(\frac{\sigma m}{n} + q\right) = -\frac{1}{\varepsilon^2} \frac{1}{\tau_w} \left(e^{-\psi} r\right) \,, \tag{6.41c}$$

where $\psi = \frac{\Phi}{1-\varepsilon^2}$. Note when $m = \sigma = q = 0$, (6.41b) recovers the expression

$$e^{-\psi} n = \text{const} \,,$$

which is reminiscent of the well balanced approach in [24]. The discretization proceeds as before, simply replacing n, m, and r by there respective Slotboom counterparts. The only difference of note is the approximation of the diffusive terms which are given by

$$\frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} e^{-\psi} \partial_x \sigma = \frac{e^{-\psi_i}}{\Delta x} \left(\sigma_{i+1/2} - \sigma_{i-1/2}\right) + O(\Delta x)^2 \,,$$

$$\frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} e^{-\psi} \partial_x \left(\frac{\sigma m}{n} + q\right) = \frac{e^{-\psi_i}}{\Delta x} \left(\frac{\sigma_{i+1/2} m_{i+1/2}}{n_{i+1/2}} - \frac{\sigma_{i-1/2} m_{i-1/2}}{n_{i-1/2}}\right) + O(\Delta x)^2 \,,$$

$$\frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} e^{-\psi} \partial_x \left(\frac{\sigma m}{n} + q\right) = \frac{e^{-\psi_i}}{\Delta x} \left(q_{i+1/2} - q_{i-1/2}\right) + O(\Delta x)^2 \,.$$

Finally, to transfer numerically between the original and the Slotboom variables, we

use

$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} e^{-\psi} n \; dx = e^{-\psi_i}\bar{n}_i + O(\Delta x)^2$$
$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} e^{-\psi} m \; dx = e^{-\psi_i}\bar{m}_i + O(\Delta x)^2$$
$$\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} e^{-\psi} r \; dx = e^{-\psi_i}\bar{r}_i + O(\Delta x)^2$$

We will see that using the Slotboom variables improves the behavior of the scheme in the drift-diffusion limit. (We have also observed this fact when computing steady-state solutions of the drift-diffusion-Poisson system.) The three step split scheme that uses the Slotboom variables in this way will be denoted $S3S$.

6.5.2 Time Discretization

Both of the convective steps, (6.36) and (6.37), are updated explicitly. The field terms in the relaxation step are updated implicitly, and the relaxation terms are updated semi-implicitly ("semi" only because the relaxation times are frozen at the current time step). The time step of our temporal first-order scheme is

$$\Delta t = \max\left(c(\Delta x)^2, \varepsilon\frac{\Delta x}{2}\right), \tag{6.42}$$

where the value of c depends on how σ and q are updated. For smaller devices, we find experimentally that an explicit update of these variables requires $c \leq 0.3$, whereas implicit updating allows $c \leq 1.0$. For larger devices, smaller values are required for stability. (See the numerical results in the next section).

We use Richardson extrapolation [5] to make the scheme second-order in time. Let T be the evolution operator for the numerical scheme with

$$T(U(t), \Delta t) = U(t + \Delta t) + O(\Delta t)$$

and set

$$\begin{aligned} U_1^{k+1} &= T(U^k, \Delta t)\,, \\ U_2^{k+1} &= T\left(T\left(U^k, \frac{\Delta t}{2}\right), \frac{\Delta t}{2}\right)\,, \\ U^{k+1} &= 2U_2^{k+1} - U_1^{k+1}\,. \end{aligned}$$

Then it is straightforward to show that $U^{k+1} = U(t + \Delta t) + O(\Delta t)^3$. Local third-order accuracy implies second-order global accuracy. The entire process requires three cycles of the three step scheme per time step.

6.6 Numerical Results.

In this section we present numerical results. In our discussion, we will refer to the following schemes:

- $S1$-$\boldsymbol{\tau}$: non-split scheme, order τ in time.
- $S2$-$\boldsymbol{\tau}$: split scheme in (6.25)-(6.26), order τ in time.
- **S3-E-$\boldsymbol{\tau}$** : three step scheme in (6.35)-(6.37), explicit update of σ and q, order τ in time.

- **S3-I-τ**: three step scheme in (6.35)-(6.37), implicit update of σ and q, order τ in time.

- **S3S-I-τ**: three step scheme in (6.35)-(6.37), discretization of Slotboom variables in the relaxation step, implicit update of σ and q, order τ in time.

6.6.1 The Transition Regime

Our initial computations are for the diode of Chapter 5 with length $L = 0.6\,\mu$m, in which case $\varepsilon = 2.0 \times 10^{-2}$. The time step in this case is $\Delta t \sim \varepsilon \Delta x$ where $\Delta x \sim \varepsilon$.

Figure (6.4) is a comparison of scheme $S3$ with the non-split scheme $S1$. It is clear that the two methods give nearly identical results. The only notable exception is the current oscillations at the junctions which are reduced by a factor of ten when using $S3$ as compared to $S1$.

Figure (6.5) gives current oscillations for several variations of $S3$. In the top left plot, the computation uses explicit updates of the diffusive terms σ and q and requires a value of $c = 0.2$ in (6.42) in order to maintain stability. In the top right plot, $c = 0.1$ and the oscillations decrease by a factor of 0.75. In the bottom left plot, σ and q are updated implicitly, thereby allowing a larger value of $c = 1.0$. However, the oscillations increase by a factor of five. This problem is resolved by going to a scheme that is second-order in time (bottom right).

6.6.2 The Drift-Diffusive Regime

To examine the diffusive regime, we consider a device of length $L = 0.6\,$cm, in which case $\varepsilon = 2 \times 10^{-6}$. Results for these computations are given in Figures 6.6-6.9. The

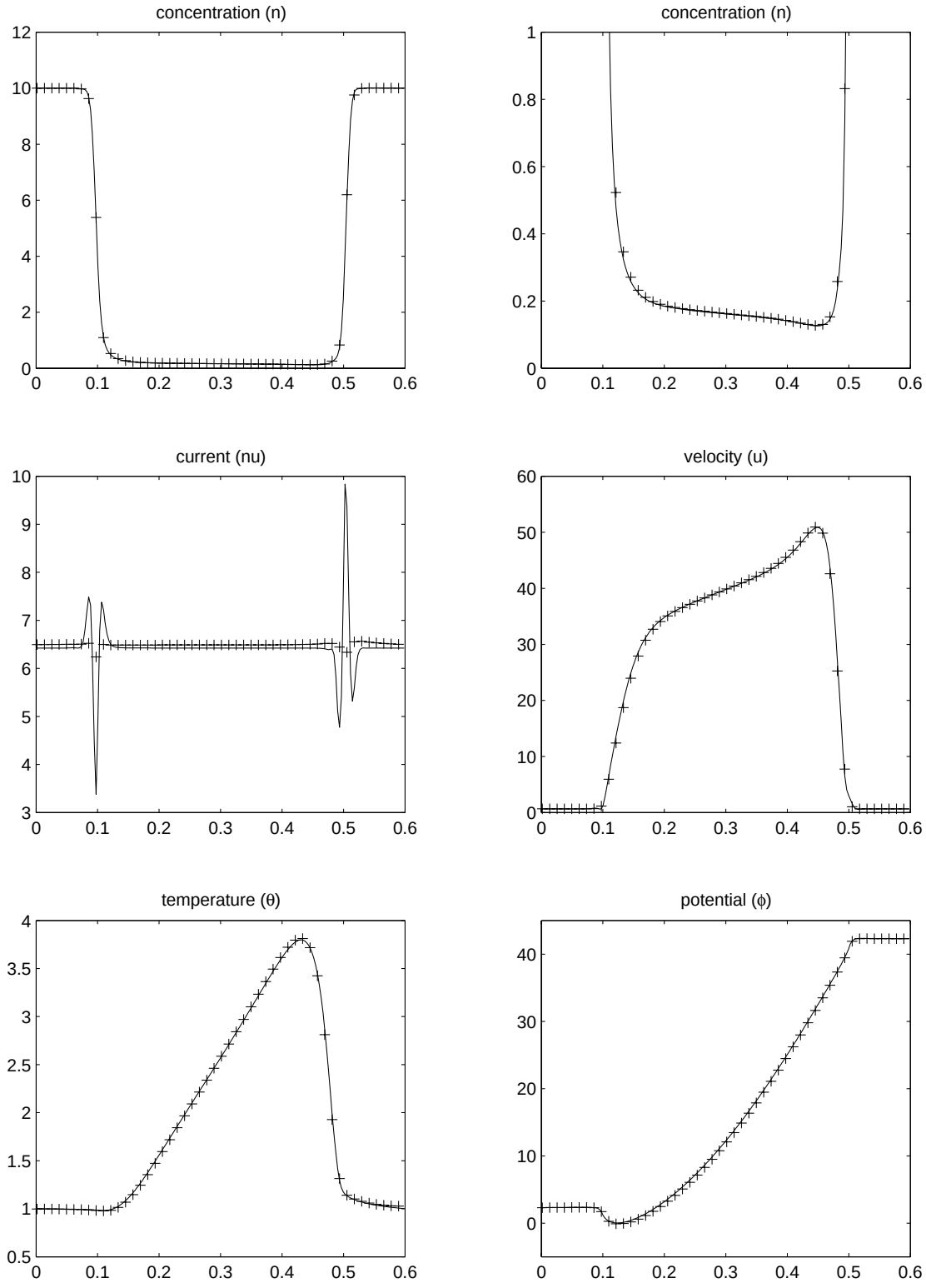

Figure 6.4: Steady state results for S-1 (solid line) and $S3E$-1 (pluses). Each scheme uses 200 meshpoints. Note that the top right plot is just a magnified version of top left in the diode channel.

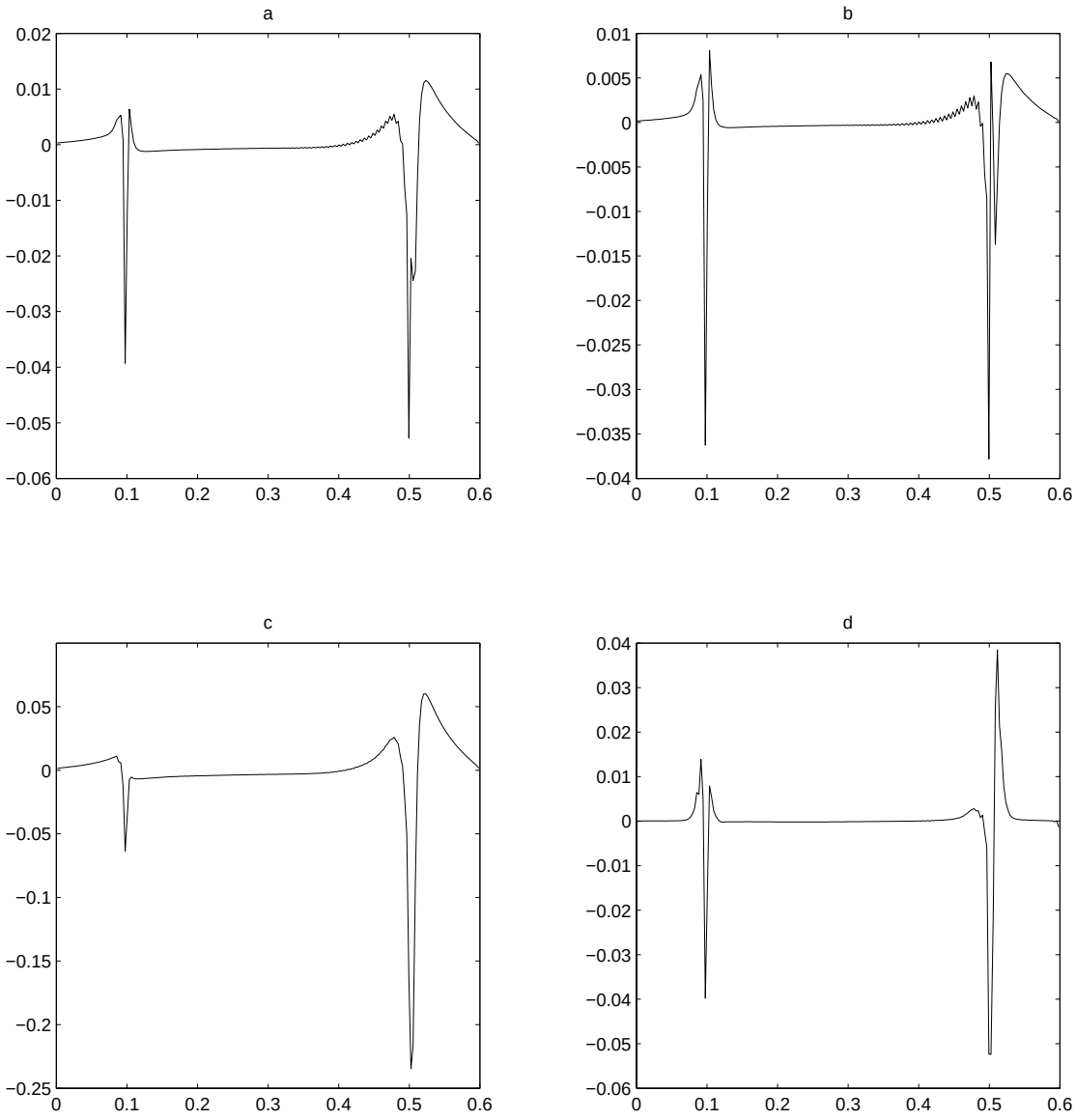

Figure 6.5: Steady state current oscillations around the spatial average using scheme S3. Each plot uses 200 points, with timestep given by (6.42). $L = 0.6 \times 10^{-6}\,\mathrm{m}$ and $\varepsilon = 2 \times 10^{-2}$. (a) S3E-1, $c = 0.2$. (b) S3E-1, $c = 0.1$. (c) S3I-1, $c = 1.0$ (c) S3E-2, $c = 1.0$.

time step is $\Delta t \sim \Delta x^2$ with $\Delta x > \varepsilon$. It is important to note that a device of this size is much too large to be physical. In practice, devices as small as $10\,\mu$m can be adequately described by the drift-diffusion equations. However, we would like to push the limits of the scheme much further.

Figures (6.6) and (6.7) compare the results from the hydrodynamic model with the drift-diffusion model. For simplicity, the scaled relaxation times τ_p and τ_w are set equal to one for these computations. As expected, the results of the two models are very similar. We recall the discussion of numerical dissipation from Section 6.3.1.1, in which the numerical dissipation of a simple non-split model was found to be proportional to $(\Delta x)^3/\varepsilon$. The computation in Figure (6.6) uses 1600 mesh points; therefore $(\Delta x)^3 \sim 2 \times 10^{-5}\varepsilon$. By contrast, the computation in Figure (6.7) uses 200 mesh points; therefore $(\Delta x)^3 \sim 1 \times 10^{-2}\varepsilon$. Even though the agreement in the current profile deteriorates slightly with the larger mesh, the concentration n shows none of the effects of numerical dissipation.

A serious problem encountered with the three step scheme in the diffusive regime is the onset of new current oscillations at the diode junctions that spread into the rest of the domain, as seen in the top two plots of Figure 6.8. We find that these oscillations can be reduced significantly by using the discretization based on the Slotboom variables. The bottom right plot of Figure 6.8 shows these results. Note, however, that the value of c in the time step (6.42) must be less than 0.1 in order for the computation to be stable. Otherwise, the current oscillations in the junction and the drain become completely unmanageable (bottom left plot). Plots of current oscillations for a selection of device lengths are given Figure 6.9.

The presence of these ringing type of oscillations is indeed curious. We find that they exist in the transition regime as well; they are just much smaller. However, it is not clear whether the source of the ringing is a balance problem, a problem with the scheme itself, or a combination of the two. There are at least two possible defects introduced by the splitting: boundary problems and large dispersive effects. Generally speaking, over-specification of boundary conditions is known to cause oscillations that can pollute the interior of a computational domain. However, no such problems have been observed with the hydrodynamic model. Even so, several alternative implementations of the boundary conditions have been tried with no positive results. One idea was to apply boundary conditions based on characteristics at each convective step in the scheme. Another was to enforce appropriate boundary conditions inherited from the original system. In all cases, it has been observed that oscillations originate at the junctions and spread to the boundary, and not from the boundary to the junctions. Therefore, an analysis of the dispersive terms in the modified equations for the split system will be our next step.

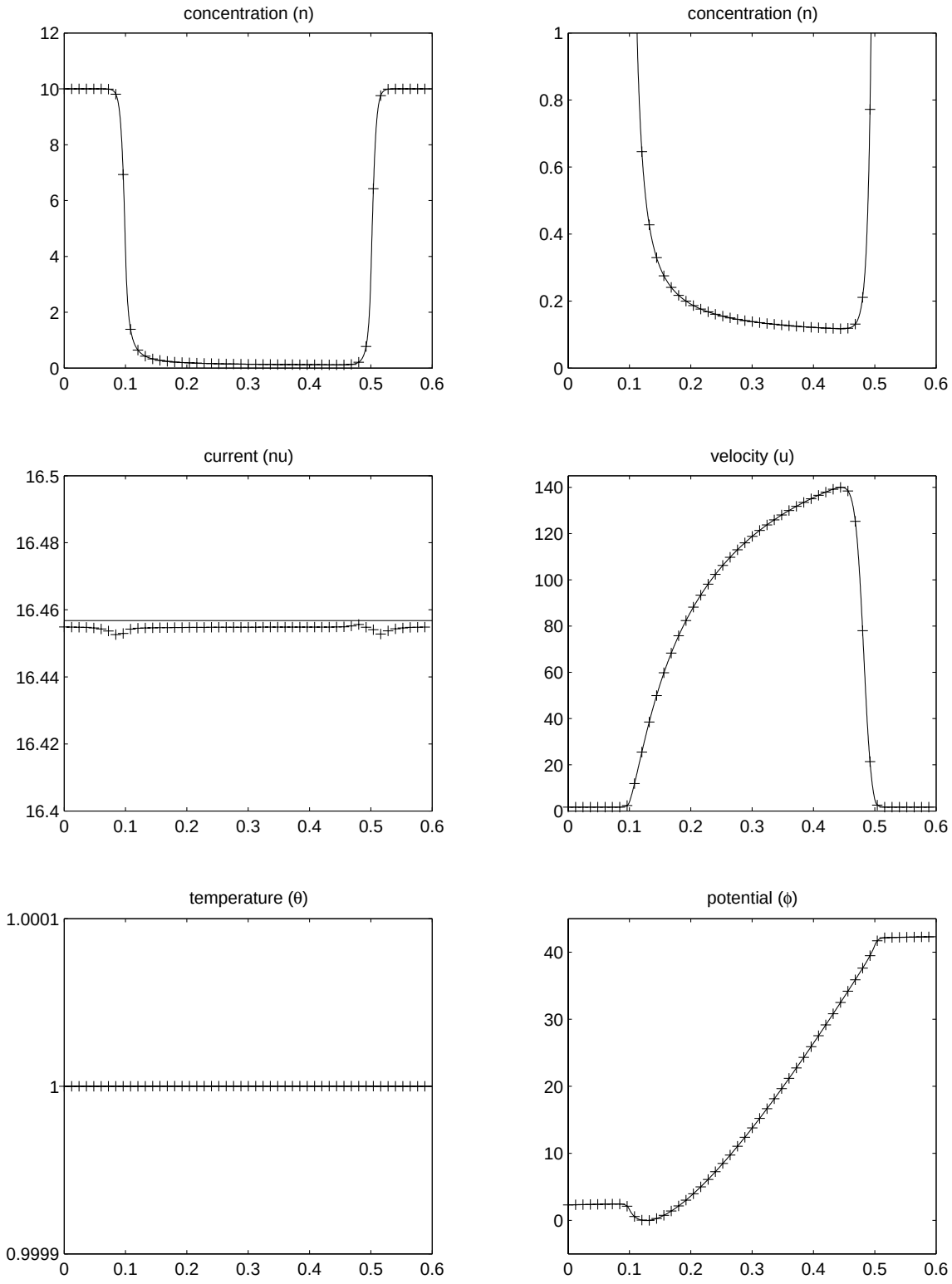

Figure 6.6: Steady state results for S3E-1 vs. drift-diffusion results. Channel length $L = 0.6 \times 10^{-2}$ m. Each scheme uses 1600 meshpoints. Note that the top right plot is just a magnified version of top left plot in the diode channel.

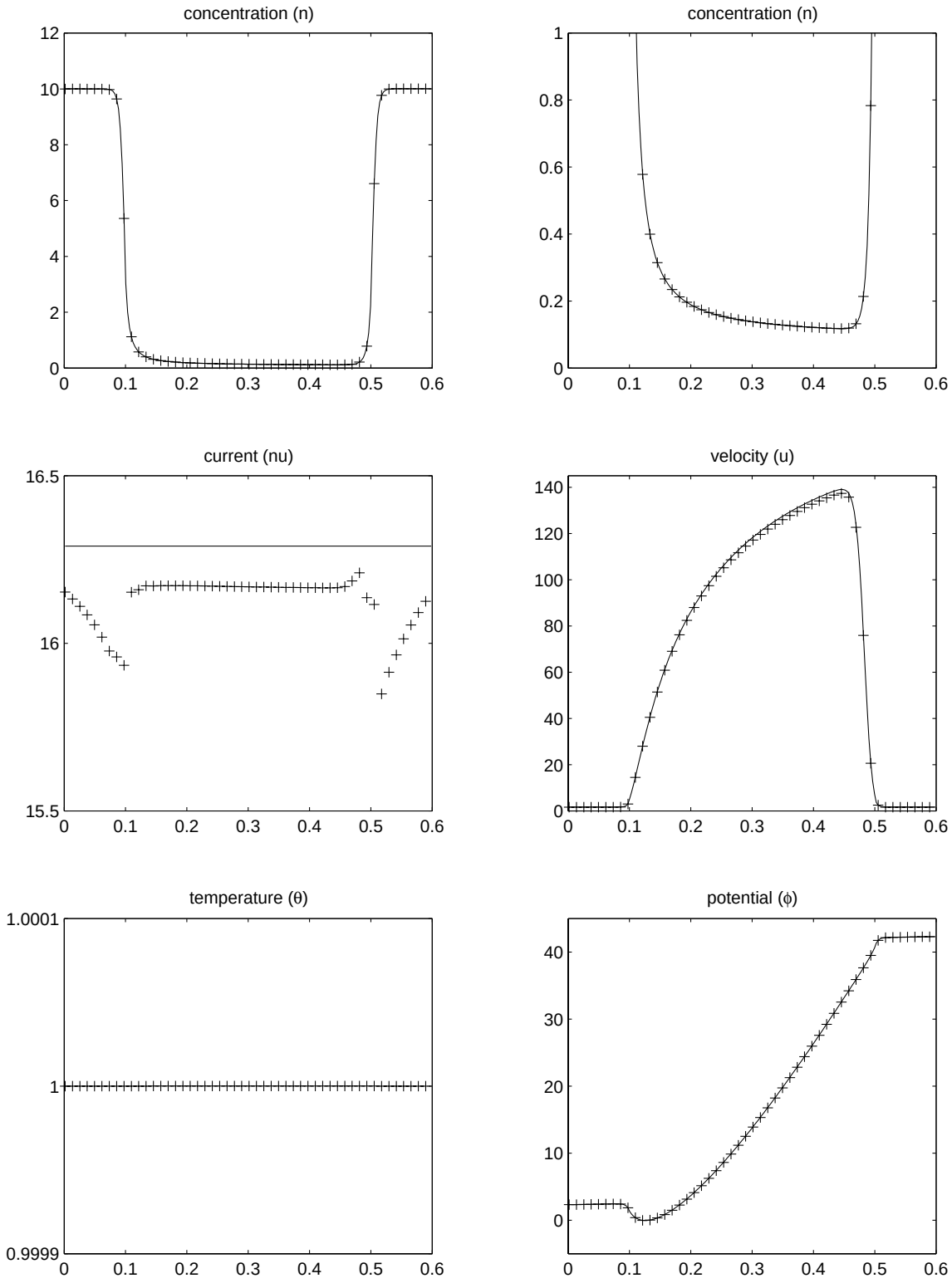

Figure 6.7: Steady state results for $S3E$-1 vs. drift-diffusion results. Channel length is $L = 0.6 \times 10^{-2}\,\mathrm{m}$. Each scheme uses 200 meshpoints. Note that the top right plot is just a magnified version of top left plot in the diode channel.

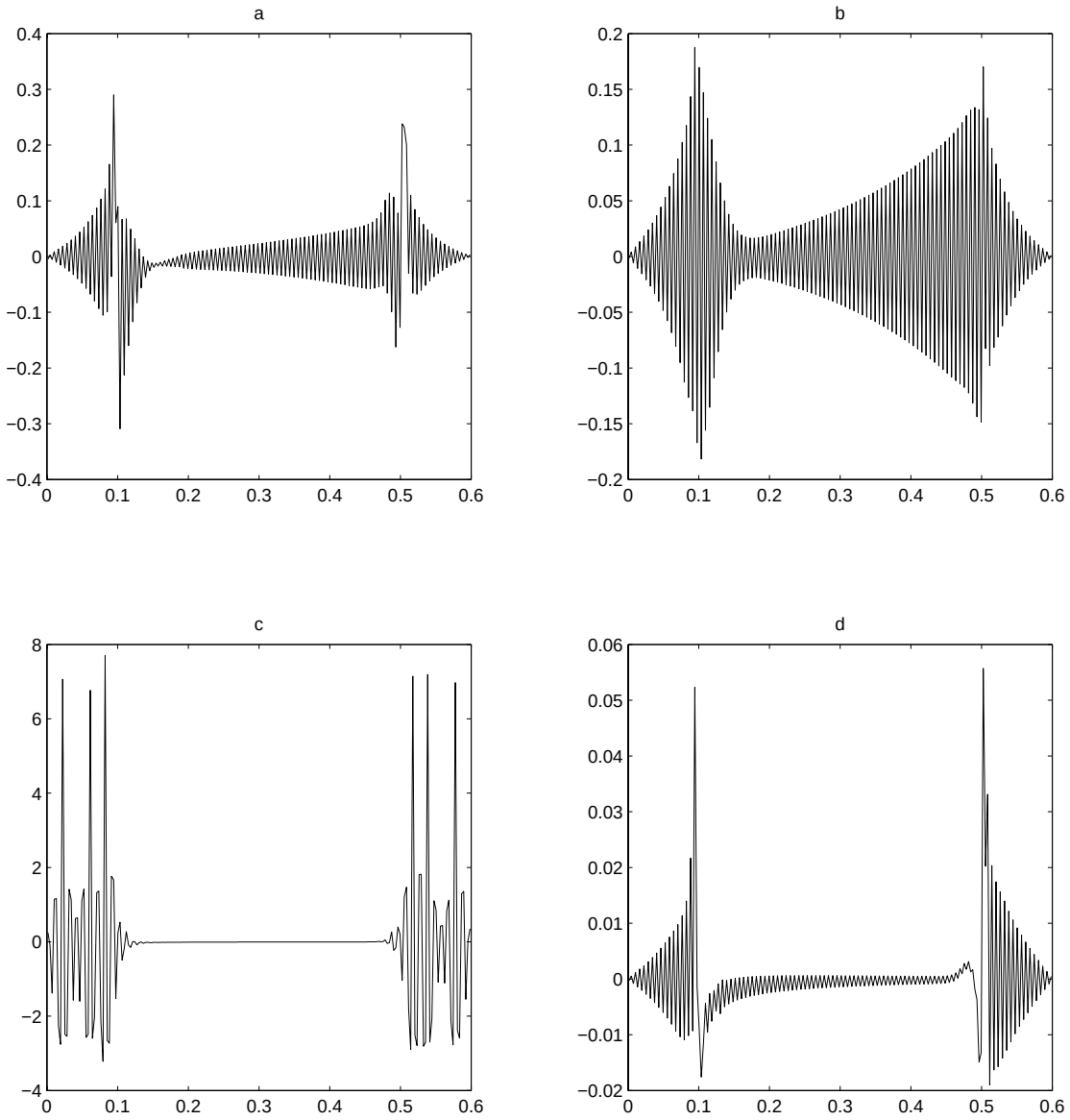

Figure 6.8: Steady state current oscillations around the spatial average. Each plot uses 200 points, with timestep given by (6.42). $L = 0.6 \times 10^{-6}$ m and $\varepsilon = 2 \times 10^{-2}$. (a) S3I-2, $c = 0.8$. (b) S3I-2, $c = 0.1$. (c) S3SI-2, $c = 0.8$. (c) S3SI-2, $c = 0.1$.

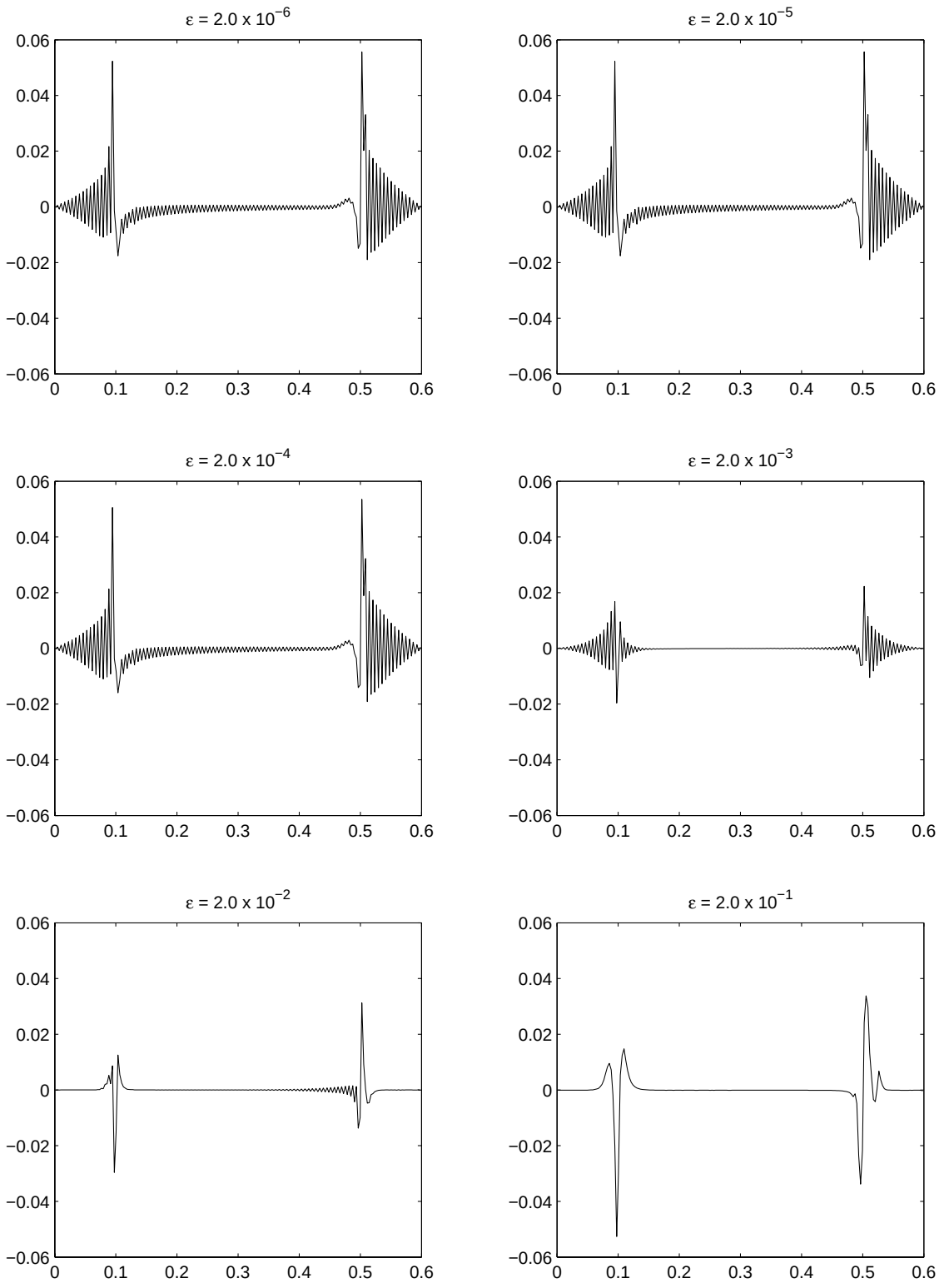

Figure 6.9: Current oscillations around average spatial value for various values of ε. Each plot uses scheme S3I-2 with 200 points and timestep given by (6.42) with $c = 0.1$.

6.6.3 Convergence Analysis

In Table 6.1, L^1 and L^∞ convergence data for a representative calculation of scheme $S3S$-I-2 is presented. Calculations are performed for a range of device lengths with meshes of 100, 200, 400, 800, and 1600 points. The error is defined as the norm of the difference between approximate solutions for two successive meshes. These approximate solutions are piecewise linear reconstruction that are generated from the mesh data. An approximate convergence rate is given by

$$rate_k = \frac{\log(error_k/error_{k+1})}{\log(\Delta x_k/\Delta x_{k+1})}\,. \tag{6.43}$$

The errors and rates are computed for the scaled versions of the variables n, m, and θ. In studying Table 6.1, the following should be noted.

1. None of the convergence rates for n, m, and θ are consistently second-order.

2. The convergence rate of n is consistently first-order, regardless of the device size or the topology in which the error is measured.

3. The convergence rate of m in the L^1 topology doesn't follow any sort of trend with respect to the mesh size. In the L^∞ topology, the convergence rate floats between one and two. (It should be noted that the L^∞ norm is essentially a measure of the non-physical current oscillations at the junctions.) Typically the rate increases as the mesh is refined near the drift-diffusion regime and decreases near the transition regime.

4. In the L^∞ topology, θ displays first-order convergence. However, in the L^1

topology, the convergence rate of θ drops off significantly in the drift-diffusion regime. This is because any error is effectively washed out by the asymptotic limit $\theta = 1 + O(\varepsilon^2)$. For larger devices, $\varepsilon \ll \Delta x$ and therefore refining the mesh has little effect. This behavior is confirmed by Figures (6.6) and (6.7), where the temperature results across the device are virtually identical and equal to one.

Finally, for purposes of comparison, we present convergence rates for schemes $S1$-2 and $S2$-2 when $L = 0.6\,\mu$m. (Analysis of larger devices is not practical given the restrictive time step for these schemes.) In most cases, the convergence rate of these schemes is comparable to that of $S3S$-I-2. There are a few cases where the convergence rate is faster than that of $S3S$-I-2, but only because the initial errors of these schemes are significantly larger. As expected, the issues that lead to poor accuracy of $S1$-2 and $S2$-2 are alleviated when the mesh is refined.

		x_0	100-200	rate	200-400	rate	400-800	rate	800-1600
L^1	n	10^{-7}	6.04E-02	9.97	3.03E-02	1.03	1.49E-02	0.96	7.65E-03
		10^{-6}	5.95E-02	1.00	2.97E-02	1.00	1.49E-02	1.00	7.43E-03
		10^{-5}	6.21E-02	1.06	2.99E-02	1.01	1.49E-02	1.00	7.43E-03
		10^{-4}	6.11E-02	1.02	3.02E-02	1.02	1.49E-02	1.00	7.43E-03
		10^{-3}	6.11E-02	1.02	3.02E-02	1.02	1.49E-02	1.00	7.43E-03
		10^{-2}	6.11E-02	1.02	3.02E-02	1.02	1.49E-02	1.00	7.43E-03
	m	10^{-7}	2.99E-02	1.98	7.55E-03	3.25	7.89E-04	2.45	1.44E-04
		10^{-6}	2.68E-02	-1.00	5.39E-02	1.83	1.52E-02	0.50	1.07E-02
		10^{-5}	1.61E-01	-0.21	1.86E-01	2.05	4.51E-02	0.38	3.46E-02
		10^{-4}	2.88E-01	0.38	2.21E-01	2.07	5.28E-02	0.44	3.88E-02
		10^{-3}	2.90E-01	0.38	2.22E-01	2.06	5.31E-02	0.45	3.88E-02
		10^{-2}	2.90E-01	0.38	2.22E-01	2.07	5.31E-02	0.45	3.88E-02
	θ	10^{-7}	7.08E-02	4.75	2.63E-03	-0.48	3.67E-03	0.30	2.97E-03
		10^{-6}	4.63E-02	1.84	1.29E-02	2.18	2.85E-03	0.43	2.12E-03
		10^{-5}	1.73E-03	1.36	6.75E-04	2.51	1.18E-04	0.10	1.10E-04
		10^{-4}	2.46E-05	1.64	7.90E-06	2.55	1.28E-06	0.07	1.30E-06
		10^{-3}	2.49E-07	1.65	7.92E-08	2.54	1.36E-08	0.08	1.28E-08
		10^{-2}	2.49E-09	1.65	7.91E-10	2.56	1.34E-10	0.06	1.29E-10
L^∞	n	10^{-7}	1.73	1.04	8.45E-01	0.94	4.41E-01	1.05	2.13E-01
		10^{-6}	2.59	0.84	1.44	1.17	6.42E-01	1.10	3.00E-01
		10^{-5}	3.03	0.84	1.70	1.33	6.73E-01	0.95	3.49E-01
		10^{-4}	3.05	0.83	1.72	1.34	6.79E-01	0.94	3.53E-01
		10^{-3}	3.05	0.83	1.72	1.34	6.78E-01	0.94	3.53E-01
		10^{-2}	3.05	0.83	1.72	1.34	6.78E-01	0.94	3.53E-01
	m	10^{-7}	2.05E-01	2.40	3.88E-02	1.77	1.14E-02	1.32	5.54E-03
		10^{-6}	1.19	2.04	2.89E-01	2.60	4.75E-02	0.78	2.76E-02
		10^{-5}	1.97	1.54	6.76E-01	2.02	1.67E-01	1.08	7.89E-02
		10^{-4}	3.59	1.68	1.12	1.19	4.90E-01	2.25	1.29E-01
		10^{-3}	3.64	1.67	1.14	1.15	5.13E-01	1.99	1.29E-01
		10^{-2}	3.64	1.67	1.14	1.15	5.13E-01	1.99	1.29E-01
	θ	10^{-7}	2.29E-01	4.59	9.53E-03	-0.69	1.54E-02	0.73	9.28E-03
		10^{-6}	3.94E-01	2.13	9.00E-02	0.75	5.36E-02	1.39	2.04E-02
		10^{-5}	1.30E-02	1.65	4.15E-03	0.82	2.35E-03	1.15	1.06E-03
		10^{-4}	1.84E-04	1.94	4.84E-05	1.06	2.33E-05	1.06	1.11E-05
		10^{-3}	1.86E-06	1.93	4.87E-07	1.06	2.33E-07	0.94	1.19E-09
		10^{-2}	1.87E-08	1.94	4.85E-09	1.04	2.35E-09	0.91	1.25E-09

Table 6.1: Convergence rate of scheme *S3S-I-2* for the $n^+ - n - n^+$ diode at various lengths. The time step is given by (6.42) with $c = 0.2$. The first column gives the topology in which the error is measured. The second column gives the variable of interest. The third column gives the device length. Columns 4, 6, 8, and 10 give the error computed when the mesh is refined by a factor of two. Columns 5, 7, and 9 give the convergence rate of the adjacent errors, computed according to (6.43).

		100-200	rate	200-400	rate	400-800	rate	800-1600
L^1	n	6.47E-02	1.10	3.01E-02	1.02	1.49E-02	1.00	7.43E-03
	m	4.07E-01	1.63	1.31E-01	2.35	2.57E-02	1.13	1.17E-02
	θ	2.58E-02	1.22	1.11E-02	1.77	3.25E-03	0.75	1.94E-03
L^∞	n	2.07	0.88	1.13	0.90	6.03E-01	1.00	3.01E-01
	m	1.34E+01	2.19	2.94	2.06	7.06E-01	2.05	1.71E-01
	θ	3.57E-01	1.51	1.26E-01	1.43	4.66E-02	1.01	2.31E-02

Table 6.2: Convergence rate of scheme $S1$-2 for the $n^+ - n - n^+$ diode. The device length is $L = 0.6\,\mu$m and the time step is given by (6.27) with $c = 0.2$.. The first column gives the topology in which the error is measured. The second column gives the variable of interest. Columns 3, 5, 7, and 9 give the error computed when the mesh is refined by a factor of two. Columns 4, 6, and 8 give the convergence rate of the adjacent errors, computed according to (6.43).

		100-200	rate	200-400	rate	400-800	rate	800-1600
L^1	n	6.57E-02	1.13	3.01E-02	1.02	1.49E-02	1.00	7.43E-03
	m	3.65E-01	1.52	1.27E-01	1.71	3.87E-02	0.89	2.09E-02
	θ	2.75E-02	1.32	1.10E-02	1.86	3.03E-03	0.65	1.93E-03
L^∞	n	2.12	0.92	1.12	0.84	6.26E-01	1.08	2.96E-01
	m	9.25	2.20	2.01	1.10	9.40E-01	0.95	5.02E-01
	θ	3.14E-01	1.28	1.29E-01	1.65	4.12E-02	0.85	2.28E-03

Table 6.3: Convergence rate of scheme $S2$-2 for the $n^+ - n - n^+$ diode. The device length is $L = 0.6\,\mu$m and the time step is given by (6.27) with $c = 0.2$. The first column gives the topology in which the error is measured. The second column gives the variable of interest. Columns 3, 5, 7, and 9 give the error computed when the mesh is refined by a factor of two. Columns 4, 6, and 8 give the convergence rate of the adjacent errors, computed according to (6.43).

6.7 Conclusions and Discussion

We have found that the drift-diffusion balance plays an important role in numerical schemes for the hydrodynamic model in both the transition and the drift-diffusion regimes. Using a splitting method that is based on this balance yields a scheme that is free from excessive numerical dissipation and stiff fluxes in the drift-diffusion limit *and* significantly reduces the oscillations that are typically found at the junctions

of the n^+-n-n^+ diode. One drawback to the three-step scheme is the presence of ringing oscillations that emanate from the junctions of large devices. Removing these oscillations will be the subject of future work.

Chapter 7

Simulation of a Unipolar MESFET Device

In this chapter, we compute two-dimensional solutions for perturbed entropy-based (PEB) models derived in Chapter 3. Recall that numerical experiments from Chapter 5 show a noticeable effect on the behavior of an n^+-n-n^+ diode with slab symmetry when $\Sigma \neq 0$, particularly near the drain junction. Even though this setting is not really natural for studying anisotropic effects (since the velocity u varies only is one dimension), it does raise our interest in the behavior of more complicated devices. Our suspicion is supported by kinetic simulations [19, 20] which confirm that the electron distribution is highly anisotropic in regions of high electric field. As expected, our results show that the anisotropy does affect the simulated behavior of a MESFET device. This device is assumed to possess translation symmetry, meaning that the electron distribution is constant along lines perpendicular to a given plane. This means that the dynamics of electron transport can be described by equations in two spatial dimensions. Anisotropic effects are most visible near material junctions, where both the drift-diffusion model and standard hydrodynamic models show particularly poor performance. In these regions, we determine that a perturbed Gaussian closure is clearly a better option than a perturbed Maxwellian

closure, based on the way that anisotropy is introduced in each case. It is therefore feasible that a perturbed Gaussian model contains enough detail to replace expensive kinetic simulations in some instances.

Recall from the derivations in Chapter 3 that corrections to the basic Maxwellian and Gaussian closures yield both diffusive and convective terms. We consider only the models that include diffusive terms, which dissipate the entropy in each system. There are two reasons for this choice, both of which are highlighted in Chapter 3. First, it is not clear if hyperbolicity or entropy dissipation is preserved when convective corrections are included. Second, corrections to the basic (non-perturbed) Maxwellian and Gaussian closures are likely to be important in regions where the spatial gradients of some or all state variables are large. In such cases, diffusive terms (two derivatives) will dominate convective terms (one derivative). Indeed, this behavior was observed in the one-dimensional experiments conducted in Chapter 3.

The Maxwellian model we will study is

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{7.1a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\theta I + \Sigma) - \frac{q_e}{m_e^*} n \nabla_x \Phi = C_v\,, \tag{7.1b}$$

$$\partial_t \left(\frac{n|u|^2}{2} + \frac{3n\theta}{2} \right) + \nabla_x \cdot \left(\frac{n|u|^2 u}{2} + \frac{5n\theta u}{2} + \Sigma u + q \right) - \frac{q_e}{m_e^*} nu \cdot \nabla_x \Phi = C_{\frac{|v|^2}{2}}\,, \tag{7.1c}$$

where the so-called fluid variables are the concentration n, velocity u, and temperature

θ, and Φ is the electrical potential which satisfies

$$\nabla_x \cdot (\epsilon \nabla_x \Phi) = q_e(D - n)\,. \tag{7.2}$$

The constants m_e^* and q_e are the electron effective mass [44] and electron charge magnitude, respectively. The quantity $\epsilon = \epsilon(x)$ is the electric permittivity of the semiconductor and $D = D(x)$ is the doping profile, a concentration of positive charges created when dopants ionize and release free electrons.

The collision terms on the right-hand side of (7.1) are

$$C_v = \frac{1}{\tau_p} nu \qquad \text{and} \qquad C_{\frac{1}{2}|v|^2} = \frac{1}{\tau_w}\left(\frac{1}{2}n|u|^2 + \frac{3}{2}n(\theta - \theta_\ell\right),$$

where τ_p and τ_w are momentum and energy relaxation times, respectively, and θ_ℓ is the lattice temperature.

The anisotropic stress Σ and the heat flux vector q are diffusive corrections to the basic Maxwellian closure. In terms of the fluid variables,

$$\Sigma = -n\theta\tau_p\left(\nabla_x u + (\nabla_x u)^T - \frac{2}{3}(\nabla_x \cdot u)\, I\right), \qquad q = -\frac{5}{2} n\theta\tau_p \nabla_x \theta\,. \tag{7.3}$$

The difference between (7.1) and most other hydrodynamic models in the fact that Σ is non-zero. However, the expressions in (7.3) are valid only when the anisotropy in the underlying kinetic distribution is small.

The Gaussian model we will study is

$$\partial_t n + \nabla_x \cdot (nu) = 0\,, \tag{7.4a}$$

$$\partial_t (nu) + \nabla_x \cdot (nu \vee u + n\Theta) - n\nabla_x \Phi = C_v\,, \tag{7.4b}$$

$$\partial_t (nu \vee u + n\Theta) + \nabla_x \cdot (nu \vee u \vee u + 3n\Theta \vee u + Q) - 2nu \vee \nabla_x \Phi = C_{v\vee v}\,, \tag{7.4c}$$

where Φ is given by (7.2) and the temperature matrix Θ is related to the anisotropy by the relation $n\Theta = n\theta I + \Sigma$. The collision terms on the right-hand side of (7.4) are

$$\begin{aligned} C_v &= -\frac{1}{\tau_p} nu\,, \\ C_{v\vee v} &= -\frac{1}{\tau_\sigma}(n\Theta - n\theta I) - \frac{1}{\tau_p}\left(nu \vee u - \frac{1}{3} n|u|^2\right) I \\ &\quad - \frac{1}{\tau_w}\left(\frac{1}{3} n|u|^2 I + n(\theta - \theta_\ell)\, I\right)\,, \end{aligned}$$

and the heat flux tensor Q is

$$Q = -3\tau_\sigma n (\Theta \cdot \nabla_x) \vee \Theta\,,$$

where τ_σ is the anisotropic relaxation time. The heat flux tensor is a diffusive correction to the basic Gaussian closure that comes from the perturbative analysis.

Like the Maxwellian model, the Gaussian model differs from most other hydrodynamic models in that the anisotropic stress tensor Σ is nonzero. However unlike the Maxwellian model, which uses a closure relation to express Σ, the Gaussian model

determines Σ with the addition of state variables that evolve according to (7.4c). Note that the trace of (7.4c) recovers (7.1c) when $\Sigma = 0$.

To see the effects of the different closures for Σ, we compare numerical simulations of (7.1) and (7.4) to a more traditional Bløtekjær-type model, which is just the basic Maxwellian closure with a diffusive heat flux added. To create such a model, we start with (7.1), set $\Sigma = 0$, and denote it the *reference* model.

The remainder of the chapter is organized as follows. In Section 7.1, we write out the equations in two dimensions and introduce the benchmark device for testing our models. In Section 7.2, we describe the numerical scheme used in our computations, and in Section 7.3 we present results of our computations.

7.1 Modeling Two Dimensional Transport

7.1.1 Equations in Two Dimensions

In two dimensions, (7.1) and (7.4) have the form

$$\partial_t \boldsymbol{\rho} + \partial_x \mathbf{f}(\boldsymbol{\rho}) + \partial_y \mathbf{g}(\boldsymbol{\rho}) + \mathbf{l}(\boldsymbol{\rho})\partial_x \Phi + \mathbf{s}(\boldsymbol{\rho})\partial_y \Phi = \mathbf{r}(\boldsymbol{\rho}) + \partial_x \mathbf{c}(\boldsymbol{\rho}) + \partial_y \mathbf{d}(\boldsymbol{\rho})\,, \tag{7.5}$$

where Φ satisfies

$$\partial_x(\epsilon \partial_x \Phi) + \partial_y(\epsilon \partial_y \Phi) = -q_e(D - n)\,. \tag{7.6}$$

The spatial densities in (7.5) have been collected into the vector $\boldsymbol{\rho}$; the fluxes in the x and y directions are given by $\mathbf{f}(\boldsymbol{\rho})$ and $\mathbf{g}(\boldsymbol{\rho})$, respectively; the vectors $\mathbf{l}(\boldsymbol{\rho})$ and $\mathbf{s}(\boldsymbol{\rho})$ are field terms; the vector $\mathbf{r}(\boldsymbol{\rho})$ contain collision terms; and the vectors $\mathbf{c}(\boldsymbol{\rho})$ and $\mathbf{d}(\boldsymbol{\rho})$

contain diffusive terms.

7.1.1.1 The Maxwellian closure The Maxwellian closure is given by (7.5) with

$$\boldsymbol{\rho} = \begin{pmatrix} n \\ nu_1 \\ nu_2 \\ \frac{1}{2}(n|u|^2 + 3n\theta) \end{pmatrix},$$

$$\mathbf{f}(\boldsymbol{\rho}) = \begin{pmatrix} nu_1 \\ nu_1^2 + n\theta \\ nu_1u_2 \\ \frac{1}{2}\left(n|u|^2 + 5n\theta\right)u_1 \end{pmatrix}, \qquad \mathbf{g}(\boldsymbol{\rho}) = \begin{pmatrix} nu_2 \\ nu_1u_2 \\ nu_2^2 + n\theta \\ \frac{1}{2}\left(n|u|^2 + 5n\theta\right)u_2 \end{pmatrix},$$

$$\mathbf{l}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ n \\ 0 \\ nu_1 \end{pmatrix}, \qquad \mathbf{s}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ 0 \\ n \\ nu_2 \end{pmatrix}, \qquad \mathbf{r}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ -\frac{1}{\tau_p}nu_1 \\ -\frac{1}{\tau_p}nu_2 \\ \frac{1}{2\tau_w}\left(n|u|^2 + 3n(\theta - \theta_\ell)\right) \end{pmatrix},$$

$$\mathbf{c}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ \frac{4}{3}\partial_x u_1 - \frac{2}{3}\partial_y u_2 \\ \partial_x u_2 + \partial_y u_1 \\ \tau_p n\theta \left[\left(\frac{4}{3}\partial_x u_1 - \frac{2}{3}\partial_y u_2 \right) u_1 + \left(\partial_x u_2 + \partial_y u_1 \right) u_2 \right] + \frac{5}{2}\tau_p n\theta \partial_x \theta \end{pmatrix},$$

$$\mathbf{d}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ \partial_x u_2 + \partial_y u_1 \\ -\frac{2}{3}\partial_x u_1 + \frac{4}{3}\partial_y u_2 \\ \tau_p n\theta \left[\left(\partial_x u_2 + \partial_y u_1 \right) u_1 + \left(-\frac{2}{3}\partial_x u_1 + \frac{4}{3}\partial_y u_2 \right) u_2 \right] + \frac{5}{2}\tau_p n\theta \partial_x \theta \end{pmatrix}.$$

7.1.1.2 The Gaussian Closure The Gaussian closure is given by (7.5) with

$$\boldsymbol{\rho} = \begin{pmatrix} n \\ nu_1 \\ nu_2 \\ \frac{1}{2}n|u|^2 + \frac{3}{2}n\theta \\ nu_1^2 + n\Theta_{11} \\ nu_2^2 + n\Theta_{22} \\ nu_1 u_2 + n\Theta_{22} \end{pmatrix},$$

$$\mathbf{f}(\boldsymbol{\rho}) = \begin{pmatrix} nu_1 \\ nu_1^2 + n\Theta_{11} \\ nu_1u_2 + n\Theta_{12} \\ \left(\frac{1}{2}n|u|^2 + \frac{3}{2}n\theta + n\Theta_{11}\right)u_1 + n\Theta_{12}u_2 \\ \left(nu_1^2 + 3n\Theta_{11}\right)u_1 \\ \left(nu_1u_2 + 2n\Theta_{12}\right)u_2 + n\Theta_{22}u_1 \\ n\Theta_{11}u_2 + \left(nu_1u_2 + 2n\Theta_{12}\right)u_1 \end{pmatrix},$$

$$\mathbf{g}(\boldsymbol{\rho}) = \begin{pmatrix} nu_2 \\ nu_1u_2 + n\Theta_{12} \\ nu_2^2 + n\Theta_{22} \\ n\Theta_{12}u_1 + \left(\frac{1}{2}n|u|^2 + \frac{3}{2}n\theta + n\Theta_{22}\right)u_2 \\ \left(nu_1u_2 + 2n\Theta_{12}\right)u_1 + n\Theta_{11}u_2 \\ \left(nu_2^2 + 3n\Theta_{22}\right)u_2 \\ n\Theta_{22}u_1 + \left(nu_1u_2 + 2n\Theta_{12}\right)u_2 \end{pmatrix},$$

$$\mathbf{l}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ n \\ 0 \\ nu_1 \\ 2nu_1 \\ 0 \\ nu_2 \end{pmatrix}, \qquad \mathbf{s}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ 0 \\ n \\ nu_2 \\ 0 \\ 2nu_2 \\ nu_1 \end{pmatrix},$$

$$\mathbf{r}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ -nu_1 \\ -nu_2 \\ \frac{1}{2}n|u|^2 + \frac{3}{2}n(\theta - \theta_\ell) \\ -\frac{1}{\tau_\sigma}n\left(\Theta_{11} - \theta\right) - \frac{1}{\tau_p}\left(\frac{2}{3}u_1^2 - \frac{1}{3}u_2^2\right) - \frac{1}{\tau_w}\left(n|u|^2 + n(\theta - \theta_\ell)\right) \\ -\frac{1}{\tau_\sigma}n\left(\Theta_{22} - \theta\right) - \frac{1}{\tau_p}\left(\frac{2}{3}u_2^2 - \frac{1}{3}u_1^2\right) - \frac{1}{\tau_w}\left(n|u|^2 + n(\theta - \theta_\ell)\right) \\ -\frac{1}{\tau_\sigma}n\Theta_{12} - \frac{1}{\tau_p}nu_1u_2 \end{pmatrix},$$

$$\mathbf{c}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \Theta_{11}\left(\frac{3}{2}\partial_x\theta + \partial_x\Theta_{11}\right) + \Theta_{12}\left(\partial_x\Theta_{12} + \partial_y\Theta_{11} + \frac{3}{2}\partial_y\theta\right) + \Theta_{22}\partial_y\Theta_{12} \\ 3\Theta_{11}\partial_x\Theta_{11} + 3\Theta_{12}\partial_y\theta_{11} \\ \Theta_{11}\partial_y\Theta_{22} + \Theta_{12}(2\partial_x\Theta_{12} + \partial_y\Theta_{22}) + 2\Theta_{11}\partial_x\Theta_{12} \end{pmatrix},$$

$$\mathbf{d}(\boldsymbol{\rho}) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \Theta_{11}\partial_x\Theta_{12} + \Theta_{12}\left(\partial_y\Theta_{12} + \partial_x\Theta_{22} + \frac{3}{2}\partial_x\theta\right) + \Theta_{22}\left(\frac{3}{2}\partial_y\theta + \partial_y\Theta_{22}\right) \\ 2\Theta_{11}\partial_x\Theta_{12} + \Theta_{12}(2\partial_y\Theta_{12} + \partial_x\Theta_{11}) + \Theta_{22}\partial_y\Theta_{11} \\ 3\Theta_{12}\partial_x\theta_{22} + 3\Theta_{22}\partial_y\Theta_{22} \end{pmatrix}.$$

7.1.2 The Benchmark Device

Computation of the Maxwellian and Gaussian closures will be performed for a MESFET (Metal Semiconductor Field Effect Transistor) device [84] that is represented on the two-dimensional domain

$$\Omega = \{(x, y) \in [0, 0.6] \times [0, 0.2]\}$$

by doping profile

$$D(x,y) = \begin{cases} 3.0 \times 10^{17}\,\mathrm{cm}^{-3}, & (x,y) \in [0.15, 0.20] \times ([0, 0.1] \cup [0.5, 0.6]) \\ 1.0 \times 10^{17}\,\mathrm{cm}^{-3}, & \text{elsewhere .} \end{cases}$$

The device, which is shown in Figure 7.1, has three contacts. The source and drain sit above the heavily doped n^+ regions of the MESFET and the gate is centered above the low doped n region.

We assume that the device is made entirely of silicon in which case $\epsilon = 1.04 \times 10^{-16}\,\mathrm{C}/\,\mu\mathrm{m}$. The effective mass is $m_e^* = 0.32 m_e$, where $m_e = 9.109 \times 10^{-31}\,\mathrm{kg}$ is the free electron mass. The lattice temperature is $\theta_\ell = \frac{k_B}{m_e^*} T_\ell$, where $T_\ell = 300\,\mathrm{K}$.

The boundary conditions are

- At the source, $n = 3.0 \times 10^{17}\,\mathrm{cm}^{-3}, \Phi = 0.0\,\mathrm{V}$, $u_1 = 0\,\mathrm{cm/\,s}$, $\theta = \Theta_{11} = \Theta_{22} = \theta_\ell$, $\Theta_{12} = 0$, u_2 satisfies Neumann condition;

- At the gate, $n = 3.0 \times 10^{17}\,\mathrm{cm}^{-3}, \Phi = -0.8\,\mathrm{V}$, $u_1 = 0\,\mathrm{cm/\,s}$, $\theta = \Theta_{11} = \Theta_{22} = \theta_\ell$, $\Theta_{12} = 0$, u_2 satisfies Neumann condition;

- At the drain, $n = 3.9 \times 10^5\,\mathrm{cm}^{-3}, \Phi = 1.0\,\mathrm{V}$, $u_1 = 0\,\mathrm{cm/\,s}$, $\theta = \Theta_{11} = \Theta_{22} = \theta_\ell$, $\Theta_{12} = 0$, u_2 satisfies Neumann condition;

- At all other boundaries, Neumannn conditions are imposed for all variables.

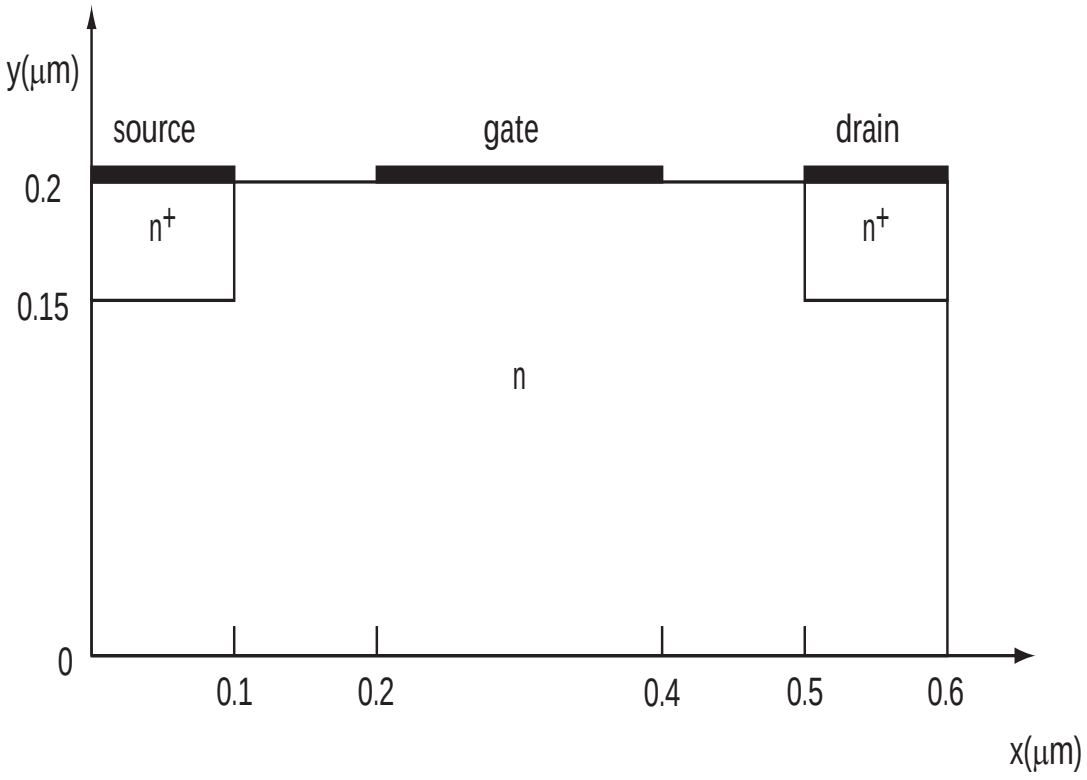

Figure 7.1: Schematic representation of the MESFET device.

The initial conditions at time $t = 0$ are

$$n = D\,, \qquad u_1 = u_2 = 0\,, \qquad \Phi = 0\,,$$

$$\theta = \Theta_{11} = \Theta_{22} = \theta_\ell\,, \qquad \Theta_{12} = 0\,.$$

Finally, the relaxation times τ_w, τ_p, and τ_s for this device are modeled as functions of energy, using a Monte-Carlo fitting [63]. One should note that these values are for a one-dimensional n^+-n-n^+diode. Therefore, our results could be improved by re-calibrating the relaxation times specifically for the MESFET device. In particular, one should expect different values for τ_p and τ_s in different directions.

7.2 Numerical Scheme

In this Section, we present details of the numerical scheme used to compute solutions of the Maxwellian and Gaussian closures. Their are four main components to the scheme: discretization of convective terms, discretization of diffusive terms, a Poisson solver for (7.6), and discretization of the collision and field terms. The scheme proceeds by computing Φ with a Poisson solver and then using Φ to update the components of $\boldsymbol{\rho}$ via a discretization of (7.5). The new value of n is then used to find Φ and the process continues until the steady state is achieved.

We introduce a uniform rectangular grid $\{(x_i, y_j)\}_{i=0,j=0}^{N_i\ N_j}$ with spacing $\Delta x = x_{i+1} - x_i$ and $\Delta y = y_{i+1} - y_i$. The domain Ω is divided into cells of the form

$$C_{ij} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}]$$

where $x_{i\pm1/2} = x_i \pm 0.5\Delta x$ and $y_{j\pm1/2} = y_j \pm 0.5\Delta y$. We will construct a scheme that is second-order in space and first-order, explicit in time. The scheme is easy to implement with higher-order Runge-Kutta methods to examine transient behavior, but first-order time steps will be sufficient for steady-state solutions.

7.2.1 Discretization of Convective Terms

For the convective terms, a shock-capturing scheme is employed with fluxes that are evaluated using a central-upwind approach [45]. For a standard conservation law in two dimensions:

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) + \partial_y \mathbf{g}(\mathbf{u}) = 0\,,$$

a second-order central-upwind scheme has the semi-discrete form

$$\frac{d}{dt}\mathbf{u}_{ij}+\frac{\mathbf{F}_{i+1/2,j}-\mathbf{F}_{i-1/2,j}}{\Delta x}+\frac{\mathbf{G}_{i,j+1/2}-\mathbf{G}_{i,j-1/2}}{\Delta y}=0\,,$$

where

$$\mathbf{u}_{ij}=\frac{1}{\Delta x\Delta y}\int_{y_{i-1/2}}^{y_{i+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}}\mathbf{u}(x,y)\,dxdy\,,$$

$$\mathbf{F}_{j+1/2,k}=\frac{a^+_{i+1/2,j}\mathbf{f}(\mathbf{u}^W_{i+1,j})-a^-_{i+1/2,j}\mathbf{f}(\mathbf{u}^E_{ij})}{a^+_{i+1/2,j}-a^-_{i+1/2,j}}+\frac{a^+_{i+1/2,j}a^-_{i+1/2,j}}{a^+_{i+1/2,j}-a^-_{i+1/2,j}}(\mathbf{u}^W_{i+1,j}-\mathbf{u}^E_{ij})\,,$$
$$\mathbf{G}_{j+1/2,k}=\frac{b^+_{i,j+1/2}\mathbf{f}(\mathbf{u}^S_{i,j+1})-b^-_{i,j+1/2}\mathbf{f}(\mathbf{u}^N_{ij})}{b^+_{i,j+1/2}-b^-_{i,j+1/2}}+\frac{b^+_{i,j+1/2}b^-_{i,j+1/2}}{b^+_{i,j+1/2}-b^-_{i,j+1/2}}(\mathbf{u}^S_{i,j+1}-\mathbf{u}^N_{ij})\,,$$

$$a^+_{i+1/2,j}=\max\left\{\lambda^+\left(\frac{\partial\mathbf{f}}{\partial\mathbf{u}}(\mathbf{u}^W_{i+1,j})\right),\lambda^+\left(\frac{\partial\mathbf{f}}{\partial\mathbf{u}}(\mathbf{u}^E_{ij})\right),0\right\}\,,$$
$$a^-_{i+1/2,j}=-\min\left\{\lambda^-\left(\frac{\partial\mathbf{f}}{\partial\mathbf{u}}(\mathbf{u}^W_{i+1,j})\right),\lambda^-\left(\frac{\partial\mathbf{f}}{\partial\mathbf{u}}(\mathbf{u}^E_{ij})\right),0\right\}\,,$$
$$b^+_{i,j+1/2}=\max\left\{\lambda^+\left(\frac{\partial\mathbf{g}}{\partial\mathbf{u}}(\mathbf{u}^S_{i,j+1})\right),\lambda^+\left(\frac{\partial\mathbf{g}}{\partial\mathbf{u}}(\mathbf{u}^N_{ij})\right),0\right\}\,,$$
$$b^-_{i,j+1/2}=-\min\left\{\lambda^-\left(\frac{\partial\mathbf{g}}{\partial\mathbf{u}}(\mathbf{u}^S_{i,j+1})\right),\lambda^-\left(\frac{\partial\mathbf{g}}{\partial\mathbf{u}}(\mathbf{u}^E_{ij})\right),0\right\}\,,$$

$$\mathbf{u}^E_{ij}=\mathbf{u}(x_{ij})+\frac{\Delta x}{2}\,(\mathbf{u}_x)_{ij}\,,\qquad \mathbf{u}^W_{ij}=\mathbf{u}(x_{ij})-\frac{\Delta x}{2}\,(\mathbf{u}_x)_{ij}\,,$$
$$\mathbf{u}^N_{ij}=\mathbf{u}(x_{ij})+\frac{\Delta y}{2}\,(\mathbf{u}_y)_{ij}\,,\qquad \mathbf{u}^S_{ij}=\mathbf{u}(x_{ij})-\frac{\Delta y}{2}\,(\mathbf{u}_y)_{ij}\,.$$

The values $(\mathbf{u}_x)_{ij}$ and $(\mathbf{u}_y)_{ij}$ are second-order approximations of the derivatives $\partial_x \mathbf{u}$ and $\partial_y \mathbf{u}$. We use a minmod-type approximation:

$$\begin{aligned}(\mathbf{u}_x)_{ij} &= \text{minmod}\left(\frac{\mathbf{u}_{i+1,j}-\mathbf{u}_{i,j}}{\Delta x}, \frac{\mathbf{u}_{i+1,j}-\mathbf{u}_{i-1,j}}{2\Delta x}, \frac{\mathbf{u}_{i,j}-\mathbf{u}_{i-1,j}}{\Delta x}\right), \\ (\mathbf{u}_y)_{ij} &= \text{minmod}\left(\frac{\mathbf{u}_{i,j+1}-\mathbf{u}_{i,j}}{\Delta y}, \frac{\mathbf{u}_{i,j+1}-\mathbf{u}_{i,j-1}}{2\Delta y}, \frac{\mathbf{u}_{i,j}-\mathbf{u}_{i,j-1}}{\Delta y}\right).\end{aligned}$$

7.2.2 Discretization of Diffusive Terms

Our approach to discretizing the diffusive terms in (7.5) is based on their entropy dissipative properties. These terms can be written in the form:

$$\partial_x \mathbf{c}(\boldsymbol{\rho}) + \partial_y \mathbf{d}(\boldsymbol{\rho}) = \nabla_x \cdot (T(\boldsymbol{\rho}) \cdot \nabla_x \boldsymbol{\beta}(\boldsymbol{\rho})) \tag{7.7}$$

(see Chapter 3, Section 5), where $\boldsymbol{\beta}$ is a tensor of Lagrange multipliers associated with $\boldsymbol{\rho}$ and T is a tensor that induces a positive, symmetric bilinear form $\mathcal{F}_T$. Given tensors $\mathbf{v}$ and $\mathbf{w}$ (of appropriate size),

$$\mathcal{F}_{T(\mathbf{u})}(\mathbf{v}, \mathbf{w}) \equiv \int_\Omega \mathbf{v} \cdot T(\mathbf{u}) \cdot \mathbf{w}\, dxdy = \int_\Omega \mathbf{w} \cdot T(\mathbf{u}) \cdot \mathbf{v}\, dxdy\,.$$

and

$$\mathcal{F}_{T(\mathbf{u})}(\mathbf{v}, \mathbf{v}) \geq 0\,.$$

For the Maxwellian closure, $\boldsymbol{\beta} = \left(\frac{u}{\theta}, \frac{1}{\theta}\right)$ and T can be written in the form of a block tensor:

$$T = \tau n \theta^2 \begin{pmatrix} 3I \vee I - \frac{5}{3} I \otimes I & 3I \vee u - \frac{5}{3} I \otimes u \\ 3I \vee u - \frac{5}{3} u \otimes I & \frac{5}{2} n\theta I + \frac{1}{3} u \vee u + |u|^2 I \end{pmatrix}.$$

For the Gaussian closure, $\boldsymbol{\beta} = \frac{1}{2}\Theta^{-1}$ and T can be written in the form of a six tensor:

$$\begin{aligned} T^{ijklmn} &= \tau n (\Theta_{il}\Theta_{jm}\Theta_{kn} + \Theta_{il}\Theta_{jn}\Theta_{km} + \Theta_{im}\Theta_{jl}\Theta_{kn} \\ &\quad + \Theta_{im}\Theta_{jn}\Theta_{kl} + \Theta_{in}\Theta_{jl}\Theta_{km} + \Theta_{in}\Theta_{jm}\Theta_{kl}) \,. \end{aligned}$$

We proceed with a weak formulation for the right-hand side of (7.7). Let ϕ be any smooth function on Ω. Then

$$\int_\Omega \phi \, \nabla_x \cdot (T(\boldsymbol{\rho}) \cdot \nabla_x \boldsymbol{\beta}) \, dxdy = -\mathcal{F}_{T(\boldsymbol{\rho})}(\nabla_x \phi, \nabla_x \boldsymbol{\beta}) + \int_{\partial\Omega} (\phi \, T(\boldsymbol{\rho}) \cdot \nabla_x \boldsymbol{\beta}) \cdot \nu \, dxdy \,, \quad (7.8)$$

where $\nu(x, y)$ is the outward normal to $\partial\Omega$ at $(x, y) \in \partial\Omega$. In particle, if we assume formally that $\boldsymbol{\beta}$ is smooth, then setting $\phi = \boldsymbol{\beta}$ gives the entropy dissipation associated with $\mathbf{c}$ and $\mathbf{d}$. Since we are only interested in discretizing in the interior of Ω, we use ghost points to implement boundary conditions and then choose ϕ with compact support on Ω. This means the boundary term on the right-hand side of (7.8) vanishes. We use the following quadrature to approximate $\mathcal{F}_{T(\boldsymbol{\rho})}(\nabla_x \phi, \nabla_x \boldsymbol{\beta})$:

$$\int_\Omega \nabla_x \phi \cdot T(\boldsymbol{\rho}) \cdot \nabla_x \boldsymbol{\beta} \, dxdy \simeq \sum_{ij} (D\phi)_{i+1/2,j+1/2} \cdot (T(\boldsymbol{\rho}))_{i+1/2,j+1/2} \cdot (D\boldsymbol{\beta})_{i+1/2,j+1/2} \,,$$

where $(D\phi)_{i+1/2,j+1/2}$ is an approximation of the gradient of ϕ at $(x_{i+1/2}, y_{j+1/2})$:

$$(D\phi)_{i+1/2,j+1/2} = \begin{pmatrix} \dfrac{(\phi_{i+1,j} - \phi_{i,j})}{2\Delta x} + \dfrac{(\phi_{i+1,j+1} - \phi_{i,j+1})}{2\Delta x} \\ \dfrac{(\phi_{i,j+1} - \phi_{i,j})}{2\Delta y} + \dfrac{(\phi_{i+1,j+1} - \phi_{i+1,j})}{2\Delta y} \end{pmatrix},$$

and $(D\boldsymbol{\beta})_{i+1/2,j+1/2}$ is defined similarly. Finally, $(T(\boldsymbol{\rho}))_{i+1/2,j+1/2}$ is the average of the four surrounding values of T:

$$T(\boldsymbol{\rho}_{i+1/2,j+1/2}) = \frac{1}{4}\sum T(\boldsymbol{\rho}_{i,j}) + T(\boldsymbol{\rho}_{i+1,j}) + T(\boldsymbol{\rho}_{i,j+1}) + T(\boldsymbol{\rho}_{i+1,j+1})\,.$$

A discretization for (7.7) at (x_i, y_j) is computed by setting $\phi_{kl} = \delta_{ik}\delta_{jl}$. The actual formulas that result are extremely long and tedious and are therefore omitted.

7.2.3 Multigrid Poisson Solver

A linear discretization for (7.6) takes the form Φ

$$(L\Phi)_{ij} = \frac{q_e}{\epsilon}(D_{ij} - n_{ij}) \tag{7.9}$$

where D_{ij} and n_{ij} are cell average values of D and n on the cell C_{ij} and $(L\Phi)_{ij}$ is a finite volume approximation of $-\nabla^2\Phi$. (Actually, pointwise values at the cell centers are equivalent since the scheme is second-order in space). Solving the linear matrix equation that arises from (7.9) is not easy because, unlike the standard one-dimensional case, the matrix representation of L is not tri-diagonal. Rather, standard discretization of the Laplacian operator in two dimensions produces a matrix with

three center diagonal bands representing differentiation in one direction plus two additional bands, one above and one below, representing differentiation in the other direction. These additional bands are spaced either N_i or N_j places from the main diagonal, depending on how the matrix variables in (7.9) are organized into vector form. In either case, unless the grid is very coarse, the matrix is sparse and inversion of (7.9) requires non-standard methods.

Popular iterative methods for solving (7.9) include the *alternate-direction implicit* method (ADI), *successive over-relaxation* (SOR), and *multigrid methods*. See [85] for a brief synopsis of each method along with algorithms and additional references. We choose a multigrid method for solving (7.9). As the name suggests, multigrid methods use a hierarchy of grids to substantially improve the classical relaxation techniques for solving (7.9). Our current presentation follows that of [86].

We endow Ω with M different grids,

$$G^m = \{(x_i^m, y_j^m)\}_{i=0,j=0}^{N_i^m \; N_j^m}, \qquad m = 1, \ldots, M\,.$$

Here G^M is the original grid for solving (7.5)-(7.6) so that

$$N_i = N_i^m \qquad \text{and} \qquad N_j = N_j^m\,,$$

and G^1 is a grid which is coarse enough for linear equations such as (7.9) to be solved

explicitly with a small number of operations. For $m < M$, G^m is defined by setting

$$(x_i^{m-1}, y_j^{m-1}) = (x_{2i}^m, y_{2j}^m)\,, \qquad 1 \le m < M\,, \quad 0 \le i \le N_i^m\,, \quad 0 \le j \le N_j^m\,.$$

Note that this definition implies that N_i and N_j are constant multiples of some power of two.

Let $\Phi_{i,j}^m \equiv \Phi(x_i^m, y_j^m)$ and let D_{ij}^m and n_{ij}^m be cell average values of D and n on the cell

$$C_{ij}^m \equiv [x_{i-1/2}^m, x_{i+1/2}^m] \times [y_{j-1/2}^m, y_{j+1/2}^m]\,,$$

where

$$\begin{aligned} x_{i\pm1/2}^m &= x_i^m \pm \frac{\Delta x^m}{2}\,, \qquad \Delta x^m = \frac{\Delta x}{2^{M-m}} \\ y_{i\pm1/2}^m &= y_i^m \pm \frac{\Delta y^m}{2}\,, \qquad \Delta y^m = \frac{\Delta y}{2^{M-m}}\,. \end{aligned}$$

We approximate (7.6) on G^m by the discretization

$$(L\Phi)_{ij}^m = \frac{q_e}{\epsilon}(D_{ij}^m - n_{ij}^m)\,, \tag{7.10}$$

where

$$(L\Phi)_{ij}^m \equiv -\frac{(\Phi_{i+1,j}^m - 2\Phi_{i,j}^m + \Phi_{i-1,j}^m)}{(\Delta x^m)^2} - \frac{(\Phi_{i,j+1}^m - 2\Phi_{i,j}^m + \Phi_{i,j+1}^m)}{(\Delta y^m)^2}\,, \tag{7.11a}$$

$$F_{ij}^m \equiv \frac{q_e}{\epsilon}(D_{ij}^m - n_{ij}^m)\,. \tag{7.11b}$$

Substituting (7.11) into (7.10) and rearranging terms gives $\Phi^m = X(\Phi^m)$, where the relaxation operator X is

$$(X(\Phi^m))_{ij} \equiv \frac{1}{2}\frac{1}{(\Delta x^m)^2 + (\Delta y^m)^2}\Big[(\Delta x^m \Delta y^m)^2 F_{ij}^m + (\Delta x^m)^2(\Phi_{i,j+1}^m + \Phi_{i,j+1}^m) + (\Delta y^m)^2(\Phi_{i+1,j}^m + \Phi_{i-1,j}^m)\Big] .$$

One approach to solving (7.10) is by relaxation—that is, by updating $\Phi^m \leftarrow X(\Phi^m)$ iteratively until it converges. In this instance, relaxation can be implemented with so-called red-black iteration, where all Φ_{ij} are computed first with $(i+j)$ even and then with $(i+j)$ odd, thereby updating all of the red (even) and then black (odd) values of Φ_{ij} in the pattern of a checkerboard.

The problem with relaxation methods is that they tend to act as a filter that effectively removes high frequency modes from the error. However, the remaining low frequency modes lead to overall slow convergence. The basic idea of multigrid methods is to improve convergence at low frequencies by solving the residual equation for (7.10) on the coarsened grid G^{m-1}, where the frequency of modes relative to the grid is doubled. The solution to the residual equation on G^{m-1} is used to correct the current approximation of Φ^m on G^{m-1}.

Let us describe a two-grid algorithm in more detail. Suppose that $\hat{\Phi}^m$ is an approximation of Φ^m after applying X a prescribed number of times. Then the error

$$\Upsilon^m \equiv \Phi^m - \hat{\Phi}^m \tag{7.12}$$

is dominated by modes that are low frequency with respect to G^m. Since L is linear, Υ^m satisfies

$$(L\Upsilon)^m_{ij} = (L\Phi)^m_{ij} - (L\hat{\Phi})^m_{ij} = \frac{q_e}{\epsilon}(D^m_{ij} - n^m_{ij}) - (L\hat{\Phi})^m_{ij}\,, \tag{7.13}$$
$$1 \le m < M\,,\ 0 \le i \le N^m_i\ ,\quad 0 \le j \le N^m_j\,,$$

where Υ and $\hat{\Phi}$ are smooth functions such that $\Upsilon^m_{i,j} = \Upsilon(x^m_i, y^m_j)$ and $\hat{\Phi}^m_{i,j} = \hat{\Phi}(x^m_i, y^m_j)$. If (7.13) can be solved for Υ^m, then Υ^m can be substituted into (7.12) to find Φ^m. However, like (7.10), (7.13) is only solved approximately for Υ^m. Moreover, relaxation techniques for solving (7.12) iteratively will converge slowly since Υ^m is composed predominantly of low frequency modes. The solution to this dilemma is to map $(L\Upsilon)^m$ onto G^{m-1} with a restriction operator R:

$$\begin{aligned}(L\Upsilon)^{m-1}_{ij} &= R((L\Upsilon)^m)_{ij} \qquad (7.14)\\ &\equiv \frac{1}{16}\left[(L\Upsilon)^m_{2i-1,2j-1} + (L\Upsilon)^m_{2i-1,2j+1} + (L\Upsilon)^m_{2i+1,2j-1} + (L\Upsilon)^m_{2i+1,2j+1}\right]\\ &\quad + \frac{1}{8}\left[(L\Upsilon)^m_{2i,2j-1} + (L\Upsilon)^m_{2i,2j+1} + (L\Upsilon)^m_{2i-1,2j} + (L\Upsilon)^m_{2i+1,2j}\right]\\ &\quad + \frac{1}{4}(L\Upsilon)^m_{2i,2j}\,.\end{aligned}$$

The frequency of error modes relative to the grid is therefore doubled, making relaxation of the residual equation more effective.

Suppose now that $\hat{\Upsilon}^{m-1}$ is an approximation of Υ^{m-1} after application of X to (7.13) on the grid G^{m-1} a prescribed number of times. The approximate error $\hat{\Upsilon}^{m-1}$ is then mapped back to G^m with an interpolation operator I. If $\Upsilon^m = I(\Upsilon^{m-1})$,

then

$$\begin{aligned}
\Upsilon^m_{2i,2j} &= \Upsilon^{m-1}_{ij} \\
\Upsilon^m_{2i+1,2j} &= \frac{1}{2}\left(\Upsilon^{m-1}_{ij} + \Upsilon^{m-1}_{i+1,j}\right) \\
\Upsilon^m_{2i,2j+1} &= \frac{1}{2}\left(\Upsilon^{m-1}_{ij} + \Upsilon^{m-1}_{i,j+1}\right) \\
\Upsilon^m_{2i+1,2j+1} &= \frac{1}{4}\left(\Upsilon^{m-1}_{ij} + \Upsilon^{m-1}_{i,j+1} + \Upsilon^{m-1}_{i+1,j} + \Upsilon^{m-1}_{i+1,j+1}\right) .
\end{aligned}$$

and $\hat{\Phi}^m$ is updated by

$$\hat{\Phi}^m \leftarrow \hat{\Phi}^m + \hat{\Upsilon}^m$$

Afterward X is applied again for a prescribed number of iterations in order to remove any high frequency errors that may have been introduced by the interpolation.

In practice there are more than two grids; and rather than accept the value of $\hat{\Upsilon}^{m-1}$ as the approximate error, the entire process described above is repeated again—this time between grids $G^{m-1}$and G^{m-2}—to find

$$\Upsilon^{m-1} \equiv \Upsilon^m - \hat{\Upsilon}^m = \Phi^m - \hat{\Phi}^m - \hat{\Upsilon}^m .$$

In this way, a nested algorithm proceeds to remove errors at lower frequencies by solving residual equations on coarser grids. The nesting terminates when $m = 1$, in which case an exact solution is easy to compute. The process of moving from the finest to coarsest grid and back again is called a cycle.

We summarize the recursive algorithm for the so-called μ-cycle $M\mu^m$.

$\underline{\Upsilon^m \leftarrow M\mu^m(\Upsilon^m, F^m, s_1, s_2)\text{:}}$

// μ, s_1 and s_2 are positive integers.

- If $m = 1$, solve $L(\Upsilon)^m = F^m$ exactly.
- Else,

  1. Relax $\Phi^m = X(\Phi^m)$ s_1 times with initial guess Υ^m.
  2. Update the residual and restrict to G^{m-1}:

$$F^{m-1} \leftarrow R(F^m - (L\Upsilon)^m) \, .$$

  3. Update $\Upsilon^{m-1} \leftarrow M\mu^m(\Upsilon^{m-1}, F^{m-1}, s_1, s_2)$ μ times with initial condition $\Upsilon^{m-1} = 0$.
  4. Interpolate Υ^{m-1} onto G^m and correct Υ^m: $\Upsilon^m \leftarrow \Upsilon^m + I(\Upsilon^{m-1})$
  5. Relax $\Phi^m = X(\Phi^m)$ s_2 times with initial guess Υ^m.

The procedure $\Phi^M \leftarrow M\mu^m(\Phi^M, F^M, s_1, s_2)$, where

$$F_{ij}^M = \frac{q_e}{\epsilon}(D_{ij} - n_{ij}) \, , \tag{7.15}$$

produces highly accurate solution for (7.9). It works particularly well for solving the Poisson equation because the restriction operator R is an approximation of the inverse Laplacian operator.

Convergence of the μ-cycle is greatly improved by a good initial guess. This is done by restricting F_{ij}^M to the grid G^{M-1} and computing

$$\Phi^{M-1} \leftarrow M\mu^m(\Phi^{M-1}, F^{M-1}, s_1, s_2)$$

which is then interpolated onto G^M as an initial guess for Φ^M. In turn, an initial guess will be needed for Φ^{M-1}. This recursive process continues until an exact solution is computed on G^1. The full multigrid algorithm is

$\underline{\Upsilon^m \leftarrow FMG(F^m, s_0, s_1, s_2)}$:

// s_0,s_1, and s_2 are positive integers.

- If $m = 1$, solve $L(\Upsilon)^m = F^m$ exactly.
- Else

  1. Restrict to G^{m-1}: $F^{m-1} \leftarrow R(F^m)$.
  2. Update $\Upsilon^{m-1} \leftarrow FMG(\Upsilon^{m-1}, F^{m-1}, s_0, s_1, s_2)$ s_0 times.
  3. Interpolate to G^m: $\Upsilon^m \leftarrow I(\Upsilon^{m-1})$.

The electric potential Φ is updated at each time step by

$$\Phi(t + \Delta t) = FMG(\Phi(t), F, s_0, s_1, s_2)),$$

where $F = F^M$ is given by 7.15 and s_0, s_1, and s_2 are chosen to optimize convergence.

7.2.4 Discretization of Field and Collision Terms

We now present the remaining details of the scheme. The collision terms are approximated by using

$$\frac{1}{\Delta x \Delta y} \int_{C_{ij}} \mathbf{r}(\boldsymbol{\rho}) = \mathbf{r}(\boldsymbol{\rho}_{ij}) + O((\Delta y)^2 + (\Delta x)^2)$$

with similar expressions for the field terms $\mathbf{l}$ and $\mathbf{s}$. The electric field is approximated with center differences:

$$\begin{aligned} \frac{1}{\Delta x \Delta y} \int_{C_{ij}} \partial_x \Phi &= \frac{\Phi_{i+1,j} - \Phi_{i+1,j}}{\Delta x} + O((\Delta y)^2 + (\Delta x)^2) \\ \frac{1}{\Delta x \Delta y} \int_{C_{ij}} \partial_y \Phi &= \frac{\Phi_{i,j+1} - \Phi_{i,j-1}}{\Delta y} + O((\Delta y)^2 + (\Delta x)^2) \; . \end{aligned}$$

7.3 Numerical Results

Numerical computations are performed on a 96×32 grid. The scheme is allowed to run until the following steady-state criteria is achieved:

$$\frac{||\boldsymbol{\rho}(t + \Delta t) - \boldsymbol{\rho}(t)||_{L^1(\Omega)}}{||\boldsymbol{\rho}(t)||_{L^1(\Omega)}} \leq 10^{-4} \Delta t \, .$$

Results are presented below in Figures 7.2-7.13. Most of these figures show results from the Maxwellian and Gaussian models and the differences of each with respect to the reference model.

Figure 7.2 is a plot of electron concentration. Both models show the same general behavior with a large drop-off in the region just below the gate. Differences are only

noticeable to the eye when comparing to the reference model. The two models differ from the reference model in a similar fashion near the drain contact of the MESFET. However, the Maxwellian model also displays major differences between the source and the gate contacts and near the artificial lateral boundaries of the MESFET at $x = 0$ and $x = 0.6$.

Figures 7.3 and 7.4 are plots of electron momentum density. The region just under the gate show very little current flow because of the charge depletion there. Instead, current flows out from the source, below the depletion region, then across the MESFET and up to the drain contact. The vertical current flow in and out of the contacts is larger for Gaussian model than it is for the Maxwellian model. The lack of current between contacts for the Maxwellian model is compensated by a horizontal flow at the lower lateral boundaries. As in Figure 7.2, the Maxwellian model displays larger variations from the reference than the Gaussian model between the source and the gate contacts and at the lateral boundaries.

Figures 7.5 and 7.6 are plots of electron bulk velocity. The Maxwellian models tends to produce velocity spikes near contacts. We note the rather large deviation of the Gaussian model from the reference in the depletion region of the MESFET, between the gate and drain.

Figure 7.7 shows the steady-state potential, and Figures 7.8 and 7.9 are electric field plots. The Gaussian model is not much different from the reference model. However, the Maxwellian model displays a noticeable drop in the x-component of the electric field, which is consistent with the horizontal current flow observed at the lower lateral boundaries.

Figures 7.10 and 7.11 are plots of electron energy and temperature, respectively. The main point to take from these plots is that the spike (in both plots) that occurs at the edge of the gate contact in both the Maxwellian and reference models is not present in the Gaussian model. A spike does appear in the temperature matrix component Θ_{11} (see Figure 7.13), which replaces θ in the pressure term of the momentum equation when passing from the Maxwellian to the Gaussian model. However, this new spike is substantially smaller.

The results displayed in Figures 7.2-7.9 follow two basic trends. First, the Gaussian model agrees with the reference model over most of the device; for the most part, major corrections appear only near the gate-drain end of the MESFET, where the electron temperature is greatest and where one might expect the greatest deviation from a drift-diffusion model. Second, the Maxwellian model appears to produce corrections similar to the Gaussian model in the gate-drain area, but it also deviates from the reference model at the lower lateral boundaries of the MESFET and especially near the source-gate region. It may be then that the Maxwellian model is not an appropriate for simulating transport in these regions of the MESFET.

We recall that the key assumption for the perturbed Maxwellian closure is that the underlying kinetic distribution is close to being anisotropic, which implies Σ is small. Therefore, in Figures 7.12, we check the validity of the Maxwellian model by computing the size of each component of Σ relative to the isotropic pressure $n\theta$. It turns out that the results are inconsistent with the smallness assumption on Σ near both sides of the gate contact. On the drain side, their is a very localized spike in

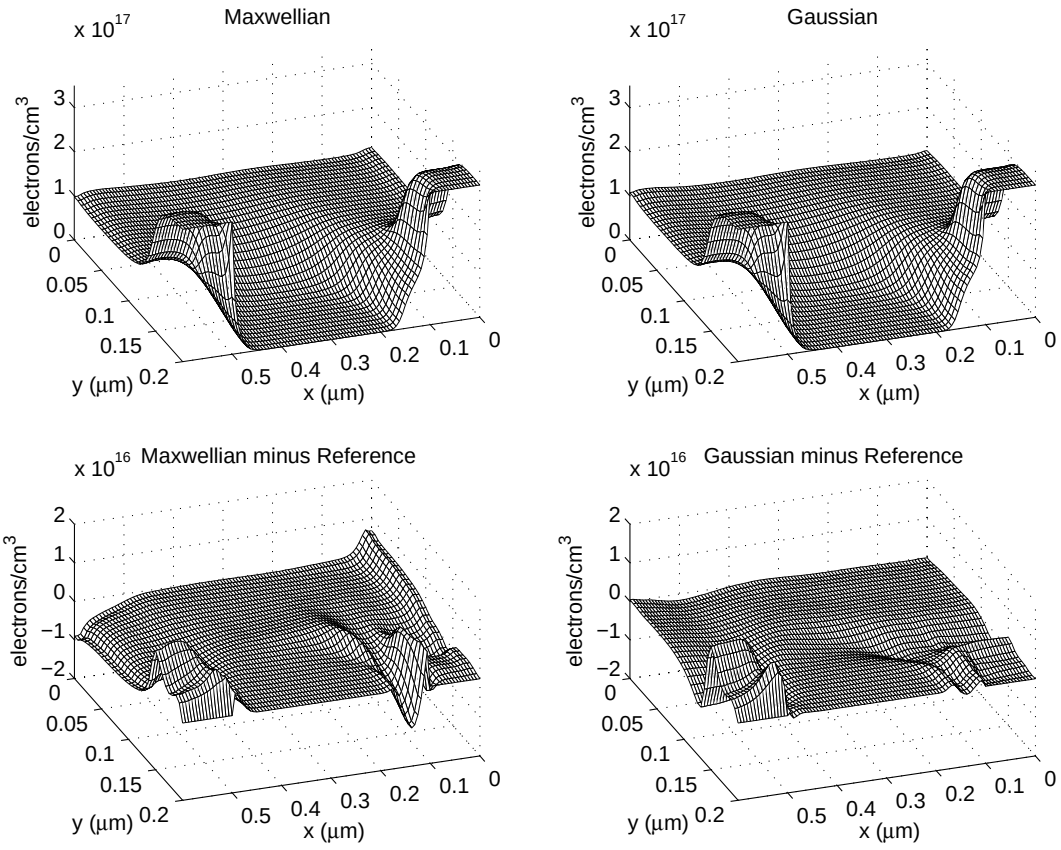

Figure 7.2: Steady-state electron concentration.

Σ. However, on the source side, Σ is quite large over a much greater area.

Analogous results for the Gaussian model are given in 7.13. Clearly Σ is much smaller than for the Maxwellian model, but still rather large: $\Sigma \sim 10^{-1}$, with the largest variations appearing close to the MESFET contacts. We therefore conjecture that the anisotropy of the underlying kinetic distribution is small, but certainly not small enough to treat it as a perturbation. Our conjecture—and the accuracy of the Gaussian model in general—must still be tested against kinetic or Monte Carlo simulations. This will be the subject of future work.

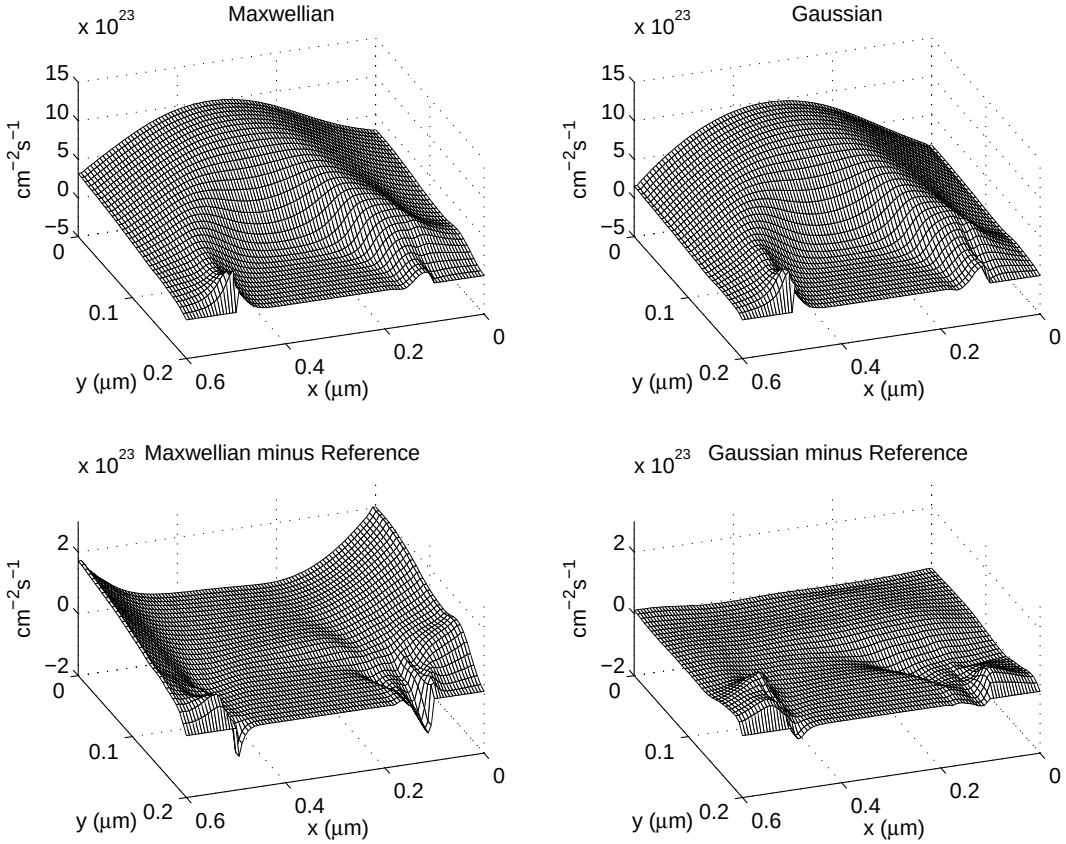

Figure 7.3: Steady-state momentum, x-component.

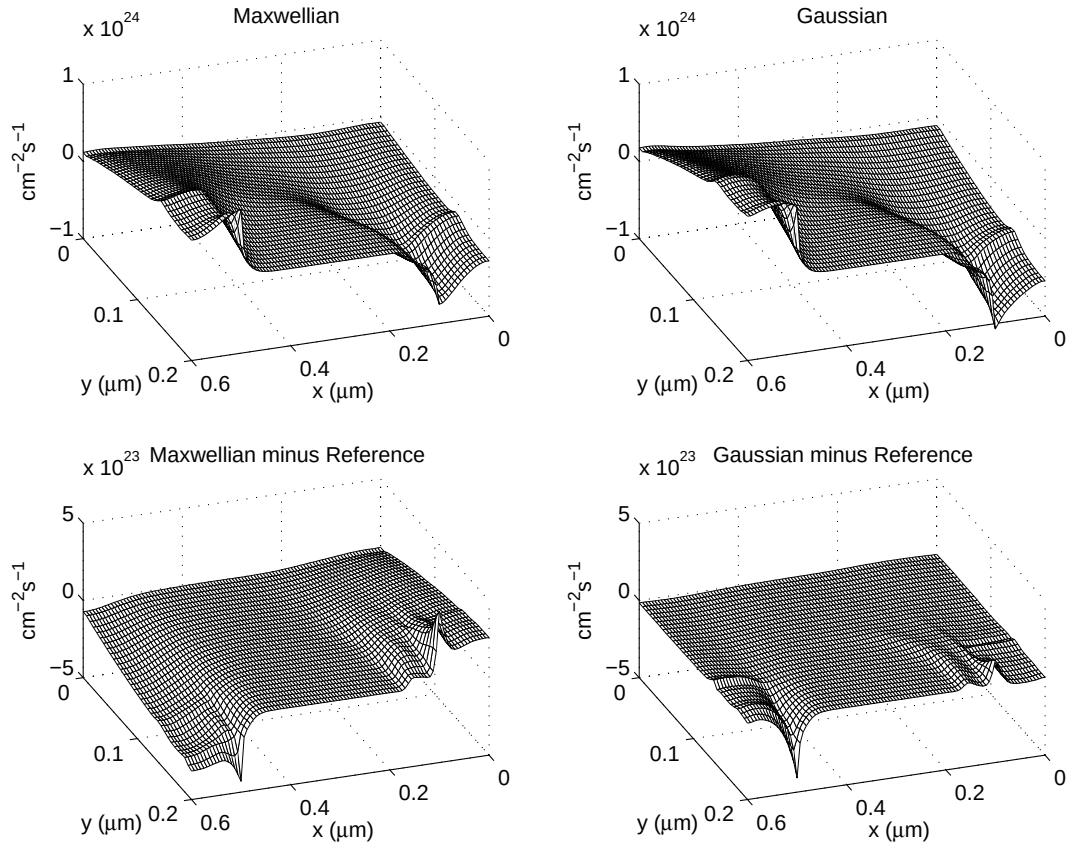

Figure 7.4: Steady-state momentum, y-component.

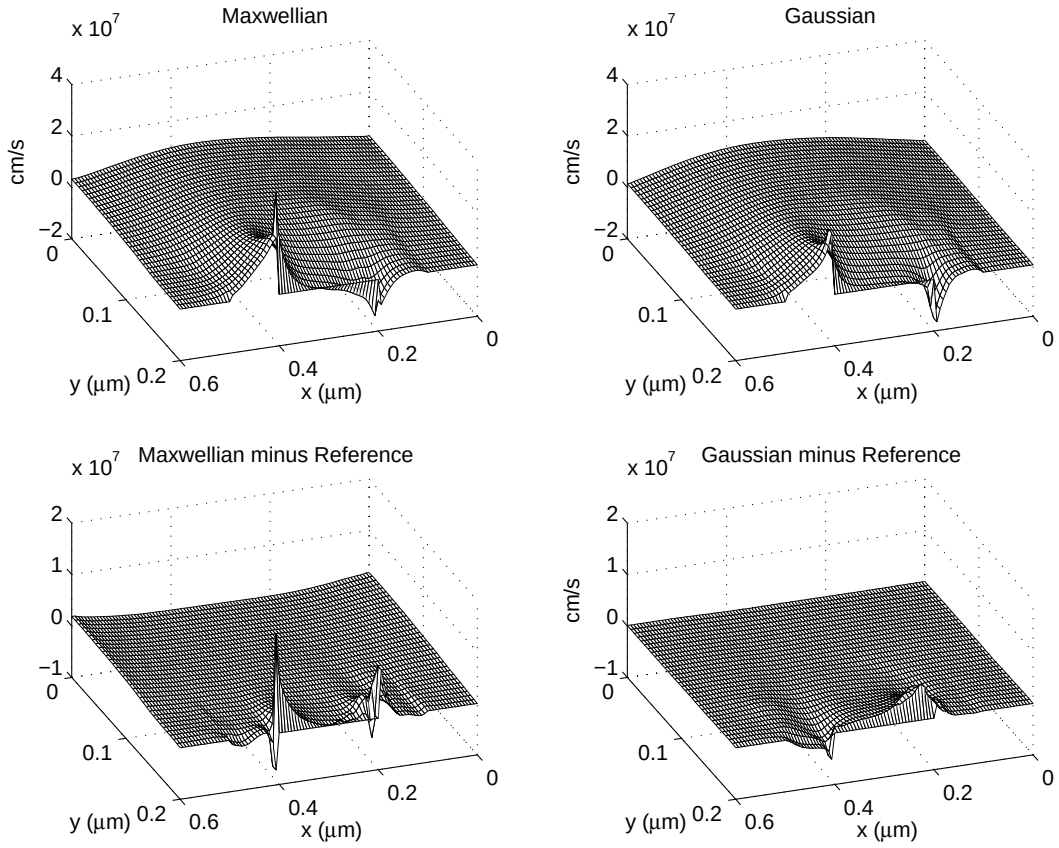

Figure 7.5: Steady-state velocity, x-component.

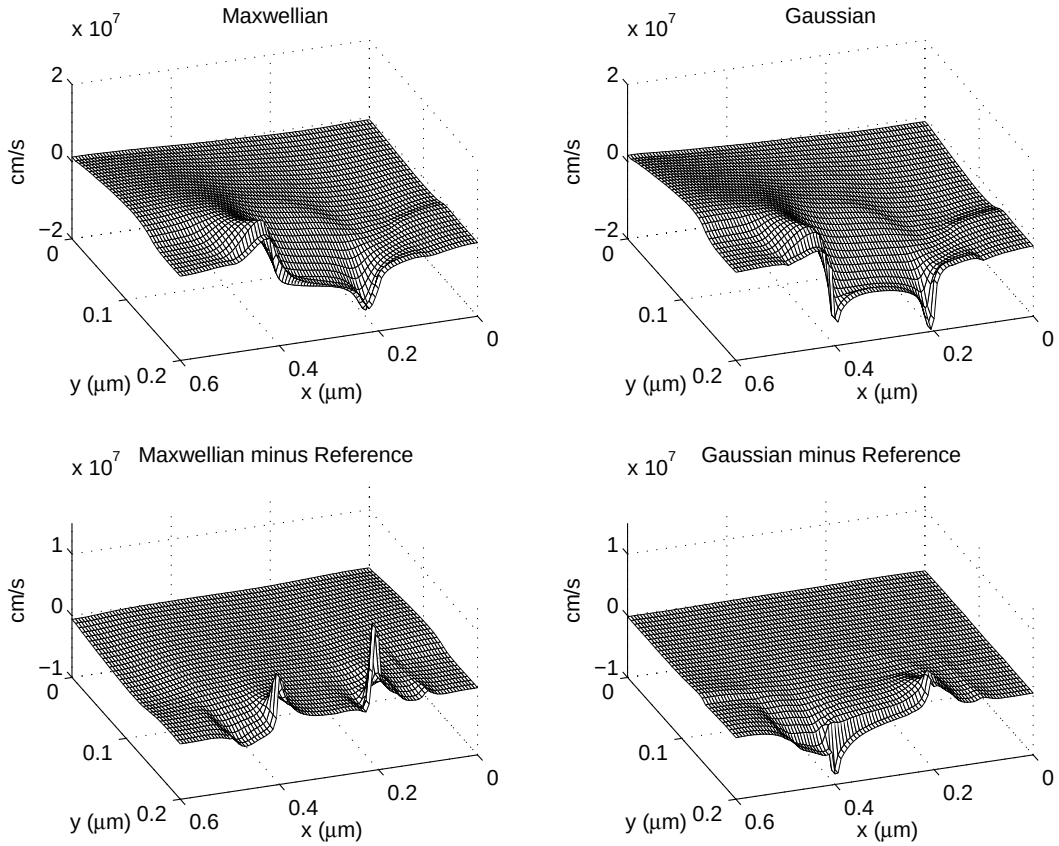

Figure 7.6: Steady-state velocity, y-component.

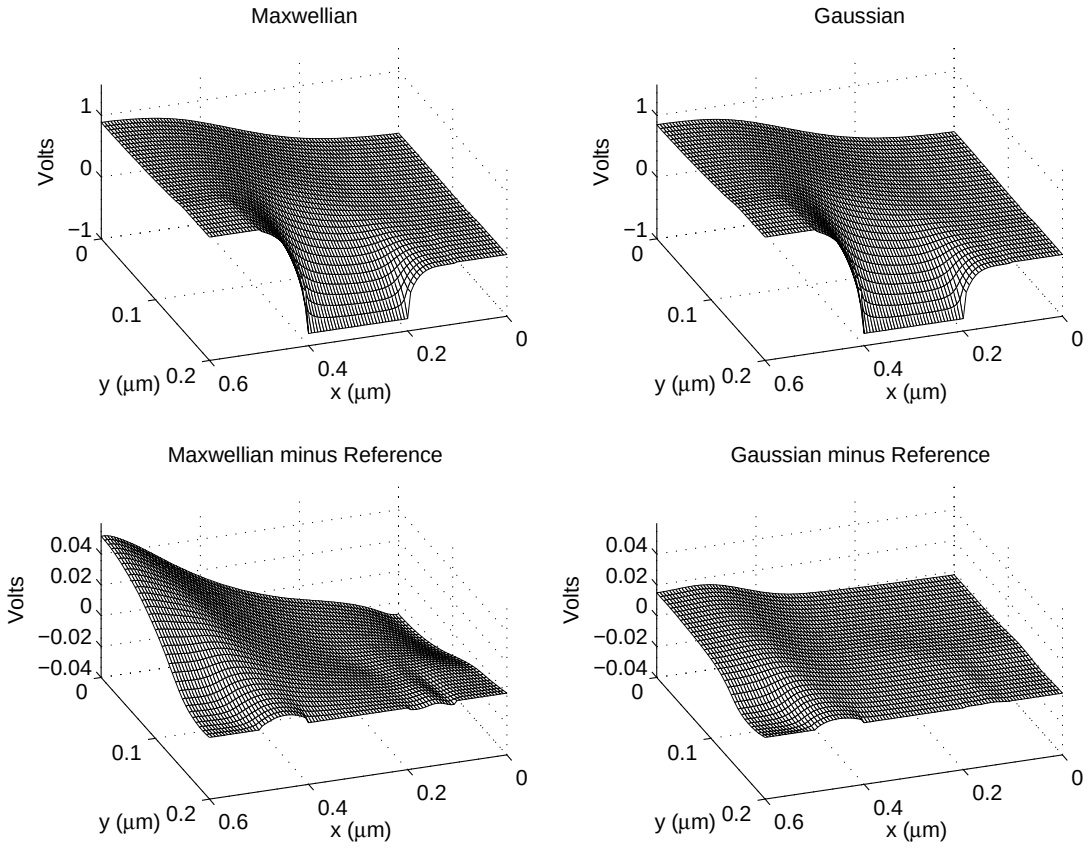

Figure 7.7: Steady-state potential.

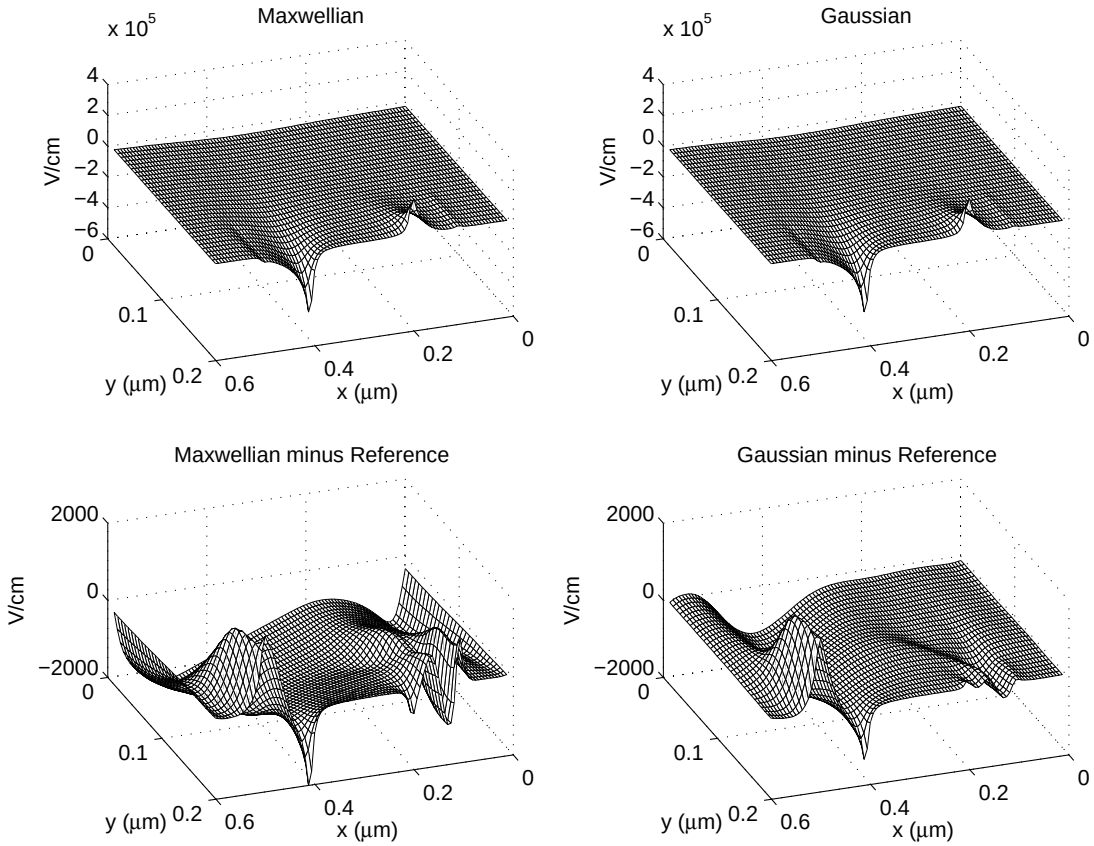

Figure 7.8: Steady-state electric field, x-component.

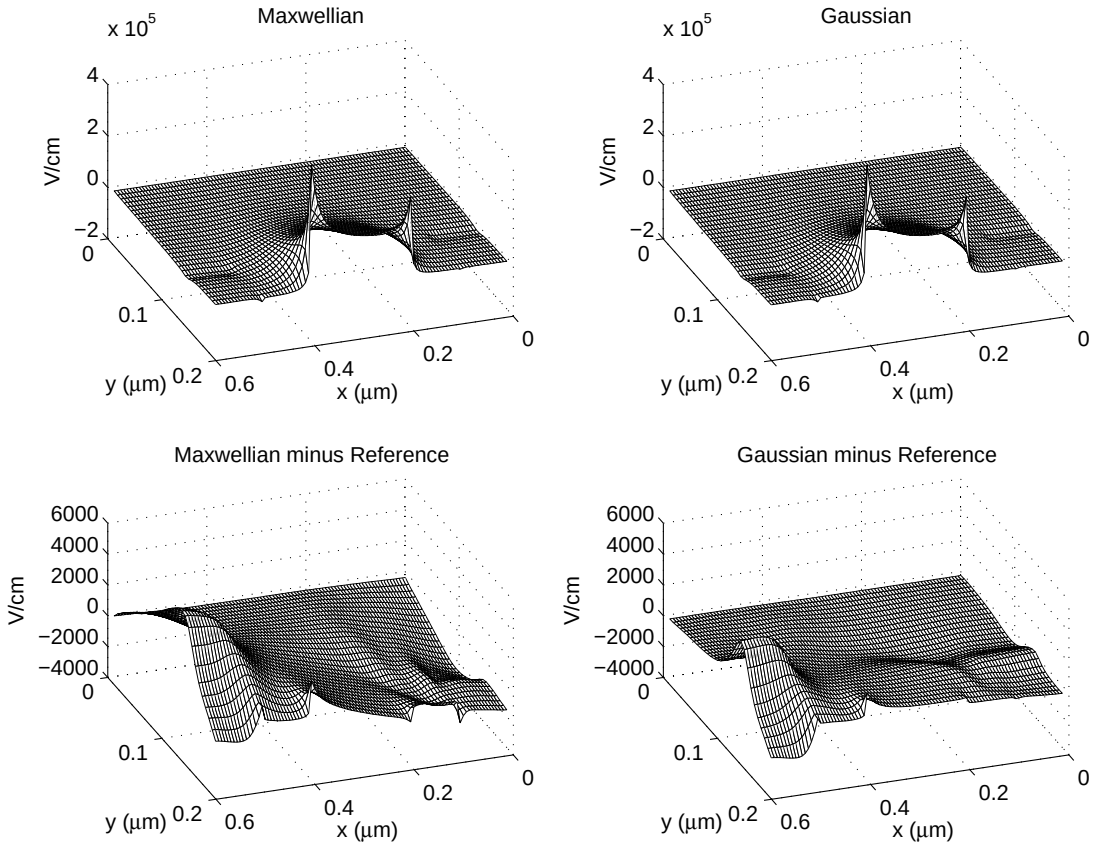

Figure 7.9: Steady-state electric field, y-component.

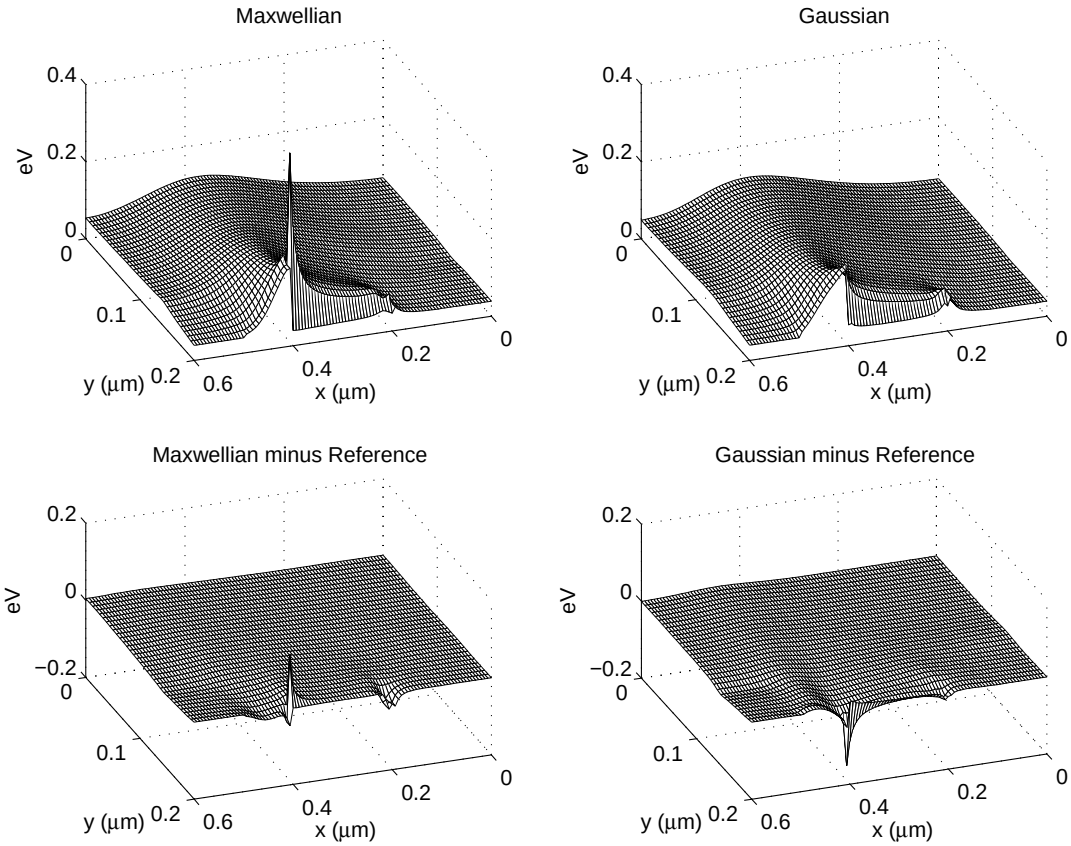

Figure 7.10: Steady-state energy profile, $\frac{1}{2}m_e^*|u|^2 + \frac{3}{2}m_e^*\theta$.

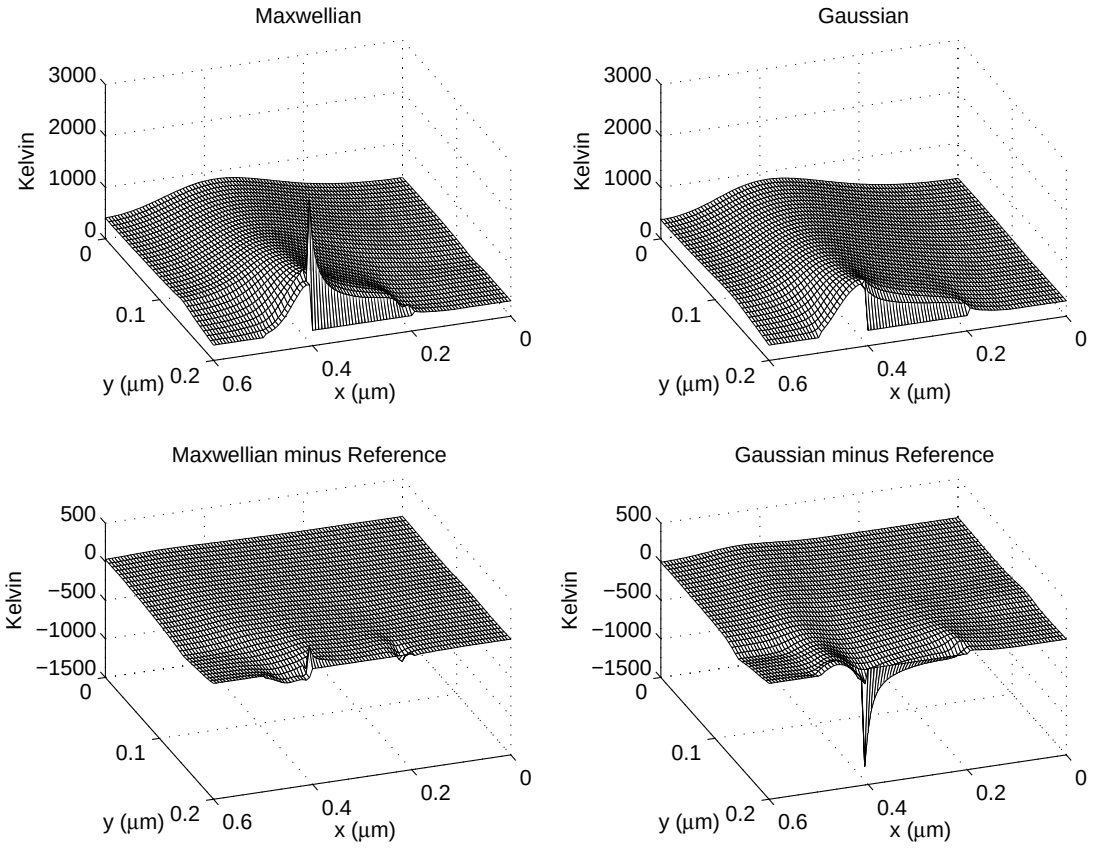

Figure 7.11: Steady-state temperature profile, $T = m_e^* \theta / k_B$.

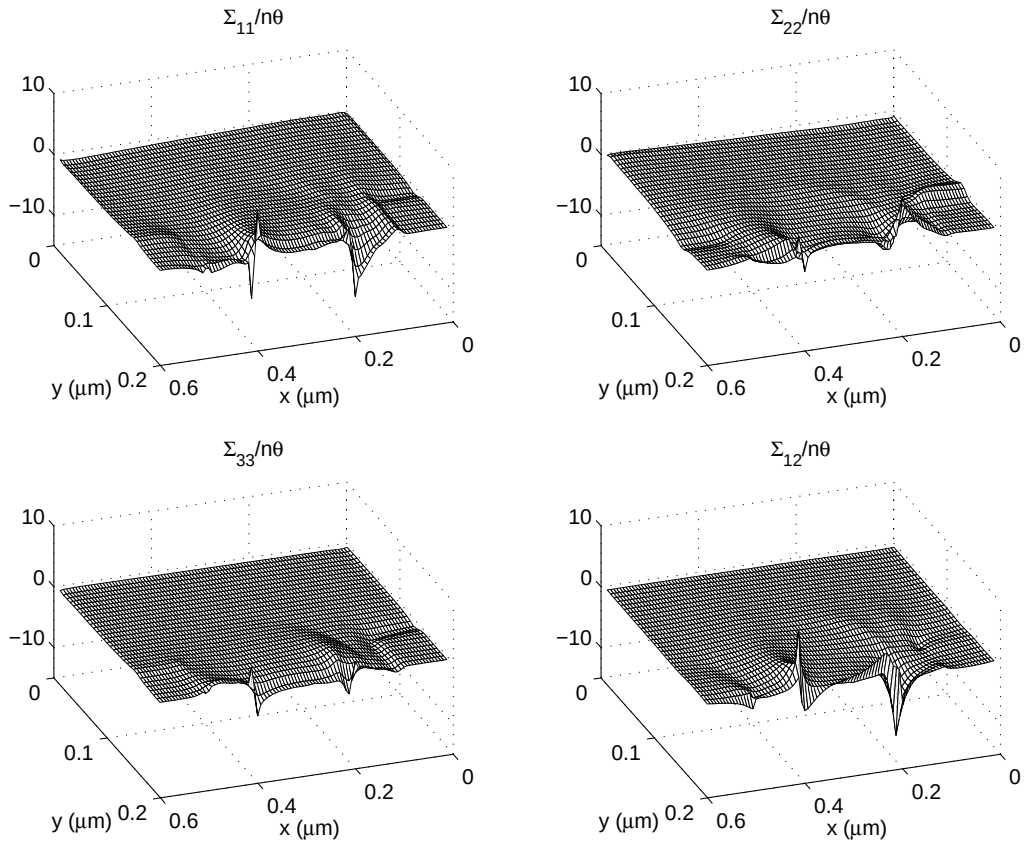

Figure 7.12: Anisotropy in the Maxwellian closure

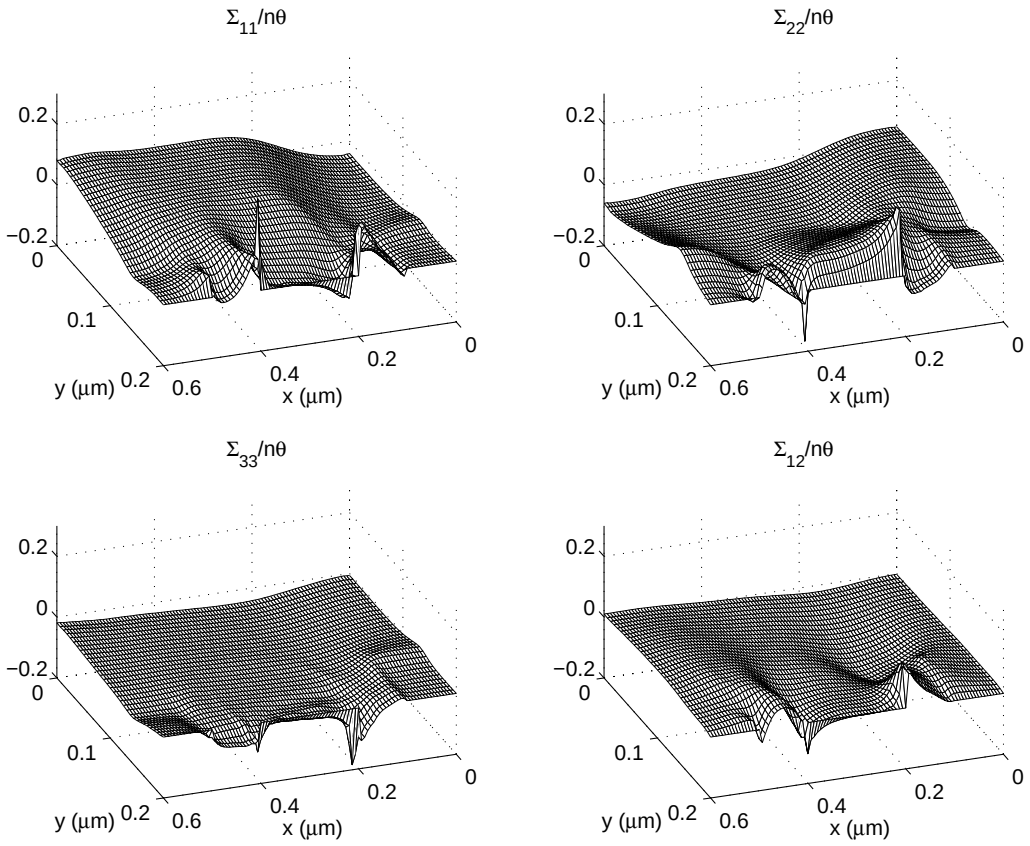

Figure 7.13: Anisotropy in the Gaussian closure.

Bibliography

[1] N. Ben Abdallah and P. Degond, *The Child-Langmuir law for the Boltzmann equation of semiconductors*, SIAM J. Math. Anal. **26** (1995), 364–398.

[2] N. Ben Abdallah, P. Degond, and A. Yamnahakki, *The Child-Langmuir law as a model for electron transport in semiconductors*, Solid-State Electronics **39** (1996), 737–744.

[3] A. M. Anile, G. Mascali, and V. Romano, *Recent developments in hydrodynamical modeling of semiconductors*, Lecture Notes in Mathematics, vol. 1823, pp. 1–56, Springer-Verlag, Berlin, 2003, Lectures given at the C.I.M.E. Summer School held in Cetraro, Italy on July 15-22, 1998.

[4] A. M. Anile and S. Pennisi, *Thermodynamic derivation of the hydrodynamical model for charge transport in semiconductors*, Phys. Rev. B **46** (1992), no. 20, 187–193.

[5] Kendall E. Atkinson, *An Introduction to Numerical Analysis*, second ed., John Wiley and Sons, New York, 1989.

[6] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water*

flows., SIAM J. Sci. Comp. **25** (2004), no. 6, 2050–2065.

[7] A. Sundstrom B. Gustafsson, *Incompletely parabolic systems in fluid dynamics*, SIAM J. Appl. Math. **35** (1978), 343–357.

[8] G. Baccarani and M. R. Wordemann, *An investigation of steady-state velocity overshoot in silicon*, Solid-State Electron. **28** (1985), 407–416.

[9] L. V. Ballestra and R. Sacco, *Numerical problems in semiconductor simulation using the hydrodynamic model: A second-order finite difference scheme*, J. Comput. Phys. **195** (2004), 320–340.

[10] R. Benedetti and Jean-Jacques Risler, *Real Algebraic and Semi-Algebraic Sets*, Actualités Mathématiques, Hermann Éditeurs des Sciences et des Arts, Paris, 1990.

[11] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, Massachusetts, 2003.

[12] P. L . Bhatnagar, E. P. Gross, and M. Krook, *A model for collision processes in gases i: Small amplitude processes in charged and neutral one-component systems*, Phys. Rev. **94** (1954), 511–524.

[13] K. Blotekjaer, *High-frequency conductivity, carrier waves, and acoustic amplification in drifted semiconductor plasmas*, Ericsson Technics **2** (1966), 125–183.

[14] ———, *Transport equations for electrons in two-valley semiconductors*, IEEE T. Electron Dev. **17** (1970), no. 1, 38–47.

[15] F. Bouchut and T. Morales, *Preprint*, (2005).

[16] J. Jerone C.-W. Shu C. Cercgnani, I.M. Gamba, *Device benchmark comparisons via kinetic, hydrodynamic, and high-field models*, Computer Methods in Applied Mechanics and Engineering **181** (2000), 381–392.

[17] H. B. Callen, *Thermodynamics and an Introduction to Thermostatstics*, second ed., John Wiley and Sons, Inc., New York, 1985.

[18] J. A. Carrillo, I. Gamba, and C.-W. Shu, *Computational macroscopic approximations to the one-dimensional relaxation-time kinetic system for semiconductors*, Physica D **146** (2000), 1–18.

[19] J. A. Carrillo, I. M. Gamba, A. Majorana, and C.-W. Shu, *A direct solver for 2d non-stationary Boltzmann-Poisson Systems for semiconductor devices: A MESFET simulation by WENO-Boltzmann schemes*, J. Comput. Electron. **2** (2003), 375–380.

[20] J.A. Carrillo, I.M. Gamba, A. Majorana, and C.W. Shu, *2d semiconductor device simulations by WENO-Boltzmann schemes: Efficiency, boundary conditions and comparison to Monte Carlo methods*, J. Comp. Phys., *to appear* (2006).

[21] C. Cercignani, I. M. Gamba, and C. D. Levermore, *A drift-collision balance for a Boltzmann-Poisson system in bounded domains*, SIAM J. Appl. Math. **61** (2001), no. 6, 1932–1958.

[22] G-.Q. Chen, J.W. Jerome, and Bo Zhang, *Particle hydrodynamic moment models in biology and microelectronics: Singular relaxation limits*, Nonlinear Anal. **30** (1997), 233–244.

[23] G.-Q. Chen, J.W. Jerome, and Bo Zhang, *Modelling and computation for applications in mathematics, science, and engineering*, ch. Existence and the Singular Relaxation Limit for the Inviscid Hydrodynamic Energy Model, pp. 189–215, Oxford University Press, 1998.

[24] C.-W. Shu F. Filbet, *Approximation of hyperbolic models for chemosensitive movement*, SIAM J. Sci. Comp. **27** (2005), 850–872.

[25] E. Fatemi, J. Jerome, and S. Osher, *Solution of the hydrodynamic device model using high-order nonoscillatory shock capturing algorithms*, IEEE T. Comput. Aid. D. **10** (1991), no. 2, 232–244.

[26] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, New Jersey, 1976.

[27] K. O. Friedrichs and P. D. Lax, *Systems of conservation equations with a convex extension*, Proc. Nat. Acad. Sci. USA **68** (1971), 1686–1688.

[28] C. L. Gardner, *Numerical simulation of a steady-state electron shock wave in a sub-micrometer semiconductor device*, IEEE T. Electron Dev. **38** (1991), 392–398.

[29] C. L. Gardner, J. W. Jerome, and D. J. Rose, *Numerical methods for the hydrodynamic device model: Subsonic flow*, IEEE T. Comput. Aid. D. **8** (1989), no. 5, 501–507.

[30] I. Gasser and R. Natalini, *The energy transport and the drift diffusion equations as relaxation limits of the hydrodynamic model for semiconductors*, Quart. Appl. Math **57** (1999), 269–282.

[31] C.G. Gibson, *Construction of Canonical Stratifications*, Lecture Notes in Mathematics, vol. 552, ch. 1, pp. 9–34, Springer-Verlag, 1976.

[32] A. Gnudi, F. Odeh, and M. Rudan, *Investigation of non-local transport phenomena in small semicondcutor devices*, Eur. J. Telecom. **1** (1990), 307–312.

[33] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, *A review of hydrodynamic and energy-transport models for semiconductor device simulation*, P. IEEE **91** (2003), no. 2, 251–274.

[34] D. J. Griffiths, *Introduction to Quantum Mechanics*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[35] W. Hänsch, *Drift-Diffusion Equation and its Application in MOSFET Modeling*, Springer-Verlag, New York, 1991.

[36] M. W. Hirsch, *Differential Topology*, Springer-Verlag, New York, 1988.

[37] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106** (1957), no. 4, 620–630.

[38] G.-S. Jiang, D. Levy, C.-T. Lin, and E. Tadmor, *High-resolution nonoscillatory central schemes with nonstagggered grids for hyperbolic conservation laws*, SIAM J. Numer. Anal. **35** (1998), no. 6, 2147–2168.

[39] S. Jin, *Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation terms*, J. Comput. Phys. **122** (1995), 51–67.

[40] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal. **35** (1998), no. 6, 2405–2439.

[41] A. Junger and S. Tang, *A relaxation scheme for scheme for the hydrodynamic equations for semiconductors*, Appl. Num. Math. **43** (2002), 229–252.

[42] M. Junk, *Domain of definition of Levermore's five moment system*, J. Stat. Phys. **93** (1998), no. 5-6, 1143–1167.

[43] ———, *Maximum entropy for reduced moment problems*, Math. Mod. Meth. Appl. S. **10** (2000), no. 7, 1001–1025.

[44] C. Kittel, *Introduction to Solid State Physics*, 7th ed., John Wiley and Sons, Inc., New York, 1996.

[45] A. Kurganov, S. Noelle, and G. Petrova, *Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations*, SIAM J. Sci. Comput. **23** (2001), no. 3, 707–740.

[46] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys. **160** (2000), 241–282.

[47] C. B. Laney, *Computational Gasdynamics*, Cambridge University Press, New York, 1998.

[48] S. E. Laux and M. V. Fischetti, *Monte Carlo simulation of submicron Si n-MOSFETs at 77 and 300*, IEEE Electron Device Letters **9** (1991), 467–469.

[49] S.-C. Lee and T.-W. Tang, *Transport coefficients for a silicon hydrodynamic model extracted from inhomogeneous Monte-Carlo data*, Solid-State Electron **35** (1992), no. 4, 561–569.

[50] R. J. Leveque, *Numerical Methods for Conservation Laws*, second ed., Lectures in Mathematics, Birkhäuser, Boston, 1992.

[51] C. D. Levermore, *Moment closure hierarchies for kinetic theory*, J. Stat. Phys. **83** (1996), 1021–1065.

[52] ———, *Entropy-Based Moment Closures for Kinetic Equations*, Trans. Th. and Stat. Phys. **26** (1997), 591–606.

[53] ———, *Moment closure hierarchies for the Boltzmann-Poisson equation*, VLSI Design **6** (1998), 97–101.

[54] S. F. Liotta, V. Romano, and G. Russo, *Central schemes for balance laws of relaxation type*, SIAM J. Numer. Anal. **38** (2000), no. 4, 1337–1356.

[55] R. B. Lowrie and J. E. Morel, *Methods for hyperbolic systems with stiff relaxation*, Int. J. Numer. Meth. Fluids **40** (2002), 413–423.

[56] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, Inc., New York, 1969.

[57] A. Majorana, *Space homogeneous solutions of the Boltzmann equation describing electron-phonon interactions in semiconductors*, Trans. Th. and Stat. Phys. **20** (1991), 261–279.

[58] ———, *Equilibrium solutions of the nonlinear Boltzmann for an electron gas in a semiconductor*, Il Nuovo Cimento **108B** (1993), 871–877.

[59] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, New York, 1990.

[60] J. Milnor, *Singular Points of Complex Hypersurfaces*, Princeton University Press and the University Press of Tokyo, Princeton, New Jersey, 1968.

[61] I. Müller and T. Ruggeri, *Rational Extended Thermodynamics*, second ed., Springer Tracts in Natural Philosophy, vol. 37, Springer-Verlag, New York, 1993.

[62] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, John Wiley and Sons, Inc., New York, 1986.

[63] O. Muscato, R. M. Pidatella, and M. V. Fischetti, *Monte Carlo and hydrodynamic simulation of a one dimensional $n^+ - n - n^+$ silicon diode*, VLSI Design **6** (1998), 247–250.

[64] G. Naldi and L. Pareschi, *Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation*, SIAM J. Numer. Anal. **37** (2000), no. 4, 1246–1270.

[65] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys. **87** (1990), no. 2, 408–463.

[66] F. Poupaud, *On a system of nonlinear Boltzmann equations of semiconductor physics*, SIAM J. Appl. Math. **50** (1990), no. 6, 1593–1606.

[67] ———, *Derivation of a hydrodynamic systems hierarchy from the Boltzmann equation*, Appl. Math. Lett. **4** (1991), 75–79.

[68] ———, *Diffusion approximation of the linear semiconductor Boltzmann equation: Analysis of boundary layers*, Asymptotic Anal. **4** (1991), 293–317.

[69] ———, *A half-space problem for a non-linear Boltzmann equation arising in semiconductor statistics*, Math. Method. Appl. Sci. **14** (1991), 121–137.

[70] ———, *Runaway phenomena and fluid approximation under high fields in semiconductor kinetic theory*, Z. Angew. Math. Mech., **72** (1992), 359–372.

[71] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[72] V. Romano and G. Russo, *Numerical solution for hydrodynamical models of semiconductors*, Math. Meth. Appl. S. **10** (2000), no. 7, 1099–1120.

[73] J. Schneider, *Entropic approximation in kinetic theory*, Math. Model. Numer. Anal. **38** (2004), 541–561.

[74] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, New York, 1984.

[75] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423 and 623–656.

[76] C.-W. Shu, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, Lecture Notes in Mathematics, vol. 1697, ch. Essentially Non-oscillatory and Weighted Essentially Non-oscillatory Schemes for Hyperbolic Conservation Laws, pp. 285–372, Springer Verlag, New York, 1998, Lectures given at the second session of the C.I.M.E. held in Cetraro, Italy in June, 1997.

[77] M.S. Shur and L.F. Eastman, *Near ballistic transport in GaAs devices at 77K*, Solid-State Electronics **24** (1991), 11–18.

[78] S.Jin and L.Pareschi, *Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes*, J. Comp. Phys. **161** (200), 312–330.

[79] N. J. A. Sloan and A . D. Wyner, *Claude Elwood Shannon: Collected Papers*, IEEE Press, New York, 1993.

[80] M. A. Stettler, M. A. Alam, and M. S. Lundstrom, *A critical examination of the assumptions underlying macroscopic transport equations for silicon devices*, IEEE T. Electron Dev. **40** (1993), no. 4, 733–740.

[81] R. Stratton, *Diffusion of hot and cold electrons in semiconductor barriers*, Phys. Rev. **126** (1962), no. 6, 2002–2014.

[82] J. C. Strikwerda, *Initial boundary value problems for incompletely parabolic systems*, Comm. Pure Appl. Math. **30** (1977), 797–822.

[83] E. Tadmor, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, Lecture Notes in Mathematics, vol. 1697, ch. Approximate solutions of nonlinear conservation laws, pp. 1–149., Spinger Verlag, New York, 1998, Lectures given at the second session of the C.I.M.E. held in Cetraro, Italy in June, 1997.

[84] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.

[85] W. T. Vetterling B. P. Flannery W. H. Press, S. A. Teukolsky, *Numerical Recipes in C*, second ed., Cambridge University Press, New York, 1992.

[86] S. F. McCormick W. L. Briggs, V. E. Henson, *A Multigrid Tutorial*, second ed., SIAM, Philadelphia, 2000.

[87] W. Wagner, *Stochastic models and Monte Carlo algorithms for Boltzmann type equations.*, Springer-Verlag, New York, 2004.

[88] C.-W. Shu Y. Xing, *High order well-balanced finite difference WENO schemes for a class of hyperbolic systems with source terms*, J. of Sci. Comput, to appear.