Simplifying Semiclassical Electron Transport in Highly Inhomogeneous Fields

By
Thomas Christiaan Minnée

A thesis
submitted to Victoria University of Wellington
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Victoria University of Wellington
2018

Abstract

The Boltzmann transport equation describes the dynamics of electrons via the time evolution of a $6-$D scalar field. This semiclassical description is valid in any device where the external field is relatively constant over the decoherence length. Near equilibrium, the electron distribution can be characterized by a local chemical potential and the lattice temperature, leading to a simplified electron state described by a $3-$D scalar field. When external fields are large, this approximation breaks down as the electrons are accelerated far away from a lattice temperature thermal equilibrium. If the external field is quasi-homogeneous, the local field or average kinetic energy can be used to characterize the shape of the distribution function, leading to an electron state described by one or two $3-$D scalar fields. However, if the external field is not quasi-homogeneous, there is currently no significant simplification of the Boltzmann transport equation that is widely accepted as being theoretically sound.

In this thesis, we derive significant simplifications of the semiclassical electron state in devices where the external fields are highly inhomogeneous. We argue that in many materials where the crystal momentum transfer in one scattering time is large, the scattering operator will drive the electron distribution toward a local *elastically-constrained quasi-equilibrium*, which is characterised by a $4-$D scalar field. Furthermore, we argue that in materials where the energy spectrum of scattering is *not narrow*, electrons in the device are efficiently driven toward one of three quasi-equilibria at different temperatures, one of which is also constrained by an upper limit on the chemical potential. The result is an electron state characterized by four to six $3-$D scalar fields, depending on whether electron-electron scattering between two high energy electrons can or cannot be neglected. We present equations of motion for these simplified electron states that can be solved efficiently if the transport parameters— associated with well-defined weighted integrals of the scattering operator and bandstructure over these quasi-equilibria— are precomputed and stored in a look-up table.

Acknowledgements

A thesis is a formal document, but in order to properly acknowledge the all-too-human relationships that provided me with the stability required to complete this long research project, it's necessary for this acknowledgements section to be informal. The academic reader should consider themselves warned: unlike the rest of this thesis, this section has not been written for you.

First and foremost, I'd like to thank my supervisor Ben Ruck. There are few supervisors in the world with the patience and trust required to tolerate a stubborn student taking the kind of wide-ranging— and often deluded— intellectual journey that led to this thesis. I am forever grateful that I found one. I wandered far from your area of expertise, and yet you had the intellectual curiosity to remain deeply interested in this project, and the physical insight to always ask the important questions. On top of this, you were also just an awesome, interesting guy to get to frequently interact with.

Secondly, I'd like to thank Petrik Galvosas for his work as Associate Dean. The wide-ranging intellectual journey I took in this research did not endear me to many in the modern academy, but you understood me and put yourself on the line to ensure that there was continued room for me in the PhD program at Victoria University of Wellington, despite the fact I must have frustrated you many times. I've heard it said that good people management is the combination of care and directness: you embody that combination as much as I've seen in anyone. This thesis would never have been completed at Victoria without your support.

Thirdly, I'd like to thank Patricia Stein for her work as Postgraduate Student Advisor.

Many former thesis students agree with me that you have an almost supernatural ability to raise people's morale. On many occasions I went to your office anxious and sure that I was in for some uncomfortable administrative drama, and many times I left a short time later relaxed and ready to focus on my work. You're an asset to the university.

And now I'd like to thank an absurd number of people in my personal life.

Firstly, I'm enormously grateful to my awesome family. To my siblings Sam Minnee, Diana Minnee II, Tash Steenhof, Casey Pickett, Jeremy Pickett and Rhiana Clarke: you have all been incredible and kept me laughing (often at myself!) while the going was a little rough. To my nieces and nephews Charlotte, Eva and Abigail (when they're in Welly!) and especially Madeleine, Ollie and Ruby (who are always in Welly): thank you for bringing many smiles to my face. I'd like to thank all my very loving grandparents, so to Oma and Opa Minnee: thank you for giving me your ambition (and your cheek!); and to Oma and Opa Weeda: thank you for never failing to show me the beauty in simple living (especially since keeping my costs low was essential to finishing this). Finally to my Mum and Dad: I love you both so much, and I couldn't have finished this thesis without parents as supportive of this wild goose chase of mine as you were. Thank you for your belief in me, for loving me despite our philosophical differences, and for raising me to have the core values that I still hold dear.

Secondly, I'd like to thank some incredible friends. To start, I'll give a huge shout-out to the wider Yule/Liardet St whanau: Jerry "Jazz" Van Lier, Scotty "Bon/Bom/Boom-body" Richardson, Kelly-Ann Barrett, James "Jimmy" Webb, Claudia Ducrot, Michaela "Emcee" Llyod, Jamie Milne, Hayden "Cuz" Richardson, Qariah Omar, Kent Barrett, Harry A'Court, Kristie "Bella" Richardson, Clint "Buzzy" Williams, Johnny Elsmore and many others. Walking up to the house on a Friday night and hearing Scotty rocking the bongos, with Cuz soulfully jamming on his guitar, with Clint shredding on his, or with Harry singing an incredible tune is a truly beautiful thing. Walking into that living room to receive a scramble of hugs from you all has always felt like the sweetest welcome home. You guys are all amazing and have all brought so much joy to my life over these last years, and I'm so grateful for that. Jazz, making that decision that I'd catch up with you most Fridays all those years ago was one of my favourite decisions: thank you for being my bro all these years, and for introducing me to and sharing all

these wonderful people with me.

Closer to the science sphere, I'd like to thank Rufus Boyack for sharing with me a constantly evolving conversation on the mysteries of physics that is so intense you could probably submit for peer-review. You're a formidable physicist and a source of frequent inspiration. I'd like to thank Thomas Bevan for many entertaining late night explorations of the similarities and differences in the chemist's and physicist's viewpoints in our shared office. And I'd like to thank Ben Wylie Van Eerd, Ruth Corkill and Jacqui Barber for their rich fellowship, and ever interesting perspectives on everything science related and not. You're all such interesting and awesome people, and I'm grateful to have met all of you through our shared love (and occasional hate!) of science.

For my final personal notes I'd like to thank the awesome guys I spent all the spare hours of my childhood and adolescence with: Finn O'Brien, Edward Oosterbaan, Dan Byrne and once again Jerry Van Lier. There is nothing quite like hanging out with you guys, the guys who were all there while we inelegantly struggled to figure out who we each were. And I'd like to thank Ellie McKenzie for being a powerful source of joy and strength for a few of the hardest years of this PhD. Thank you for being you lass.

I will end this epic by acknowledging the financial support for this work. I'd like to thank the New Economy Research Fund and the Victoria Submission Scholarship Program for their partial support of this work, as well as the New Zealand Student Loan system and the Bank of D&G for both supporting this work with near-zero interest loans.

Contents

Chapter 1

Introduction

1.1 Economic Value of Semiconductor Device Models

Electronics have created a significant fraction of total wealth of humankind. The functionality of almost all modern electronics is enabled by the integrated circuits produced by the $\sim \$350\text{B}$ USD semiconductor industry [1].

We expect that innovation in integrated circuit design will continue to have an outsized effect on the rest of the economy. The efficiency of this innovation is limited by the speed, accuracy and flexibility of the modelling and simulation tools the semiconductor industry possesses. These modelling and simulation tools are known as ELECTRONIC DESIGN AUTOMATION (EDA) tools, and the large subset of these tools that are directly based on the physical simulation of matter in space are known as TECHNOLOGY COMPUTER AIDED DESIGN (TCAD) tools.

One might assume given the enormous economic value of improving the physics underlying TCAD tools that private-sector EDA companies would be strongly incentivised to invest in basic research. However on closer examination of the incentives, it is far from obviously that this is actually the case. Any improvements the private sector make in their understanding of the underlying physics is not patentable, and therefore—

roughly speaking— will only be of commercial value if kept secret. This is problematic for two reasons. Firstly, this secrecy greatly diminishes the reliability of the physics knowledge private companies can gain, as it is not subject to outside criticism and improvement. Secondly, this secrecy greatly diminishes the value of the physics knowledge, since in order to keep this knowledge from competitors it must be hidden *even from the users of TCAD products*; that is, any new physics discovered by the EDA industry must be hidden from the semiconductor industry itself if it is to remain a commercially valuable asset.

As such, the EDA industry should not be thought of as a reliable contributor to the basic physics that underlies TCAD tools. Instead, the $\sim$ \$8B USD EDA industry [1] is more accurately thought of as a *powerful and efficient distribution network*, transforming academic advances in fundamental physics that are relevant to TCAD into modelling tools that are convenient and practical for engineers in the semiconductor industry to use [2]. The existence of this distribution network means academic advances in the physics underlying TCAD tools will be leveraged to create wealth at a large scale, in a short timeframe, with high probability. These impact characteristics are not common in academic physics research, and make this area of research attractive for academic physicists who are interested in having a unusually tangible economic impact.

We can divide TCAD tools into two[1] basic levels:

The process level, concerned with modelling the processes that are used to fabricate the integrated circuit.

The device level, concerned with modelling electronic, thermal, and mechanical characteristics of a small physical subregion of a fabricated integrated circuit, typically one associated with an individual device such as a diode or transistor.

In this thesis we are concerned with the electronic part of the device level. That is, we are concerned with modelling electron transport in semiconductor devices.

[1]The EDA tools expand on the two TCAD levels by adding a *circuit level*, concerned with modelling the abstraction of devices and interconnects known as a circuit.

1.2 Transport in the Quasi-Homogeneous Regime

Historically the DRIFT-DIFFUSION model[2] has been the back-bone of TCAD electron transport models, and it remains in heavy use even today [3]. In its domain of validity the drift-diffusion model is accurate, flexible, and fast. The *accuracy* of the drift-diffusion model is primarily due to the fact it can be derived from first-principles using well-justified assumptions. The *flexibility* of the drift-diffusion is primarily due to the fact that these assumptions have a broad range of validity. The *speed* of the drift-diffusion model is primarily due to the fact that the electron state[3] is defined uniquely by only a single parameter at each point in space— the electron density. In addition the accuracy, flexibility, and speed of the drift-diffusion model is secondarily enhanced by the fact that an unconditionally stable discretization of it can be defined on arbitrary $1-$D, $2-$D or $3-$D unstructured meshes [4].

The derivation of the drift-diffusion model from first principles relies on being able to define the MOBILITY— a transport parameter that relates the local driving forces to the local electron flux— in terms of the device state[4] and an electron state characterized only by position-dependent electron density. This can be done in a theoretically sound[5] manner only if the electron distribution is in LOCAL DYNAMIC EQUILIBRIUM with the external field; that is, so long as the electron distribution function is approximately[6] the same

[2]The STATE of an entity can generally be understood as the information concerning that entity at some time t that is necessary to predict the entity's state at $t + \Delta t$. This subtle, seemingly-circular definition is a surprisingly powerful guide for creating useful physics.

[3]Unless explicitly stated otherwise, the reader should assume that the transport models mentioned are only closed when coupled to Poisson's equation, which defines the state of the external electromagnetic field via an electric potential at each point in space.

[4]By DEVICE STATE, we mean the state of the physical environment that the electron state is subject to. Typically this is defined by a position-dependent lattice temperature, bandstructure, electric potential and doping density.

[5]We use the term THEORETICALLY SOUND heavily in this thesis to refer to a consistency between a macroscopic model and the underlying microscopic reality. A model is theoretical sound if it has a derivation from first principles using well-justified assumptions. Since most of the macroscopic models of electron transport in the literature we discuss make assumptions that *obviously contradict* the underlying microscopic physics, "well-justified assumptions" should be understood as relative to the state-of-the-art. As such, we will freely refer to a macroscopic model as theoretically sound if its assumptions *do not obviously contradict* the underlying microscopic physics.

[6]This qualifier is used because there is a perturbation to the homogeneous electron distribution function due to the diffusion current.

as it would be in a homogeneous external field[7] which is the same as the local external field [5].

The assumption that the distribution function is determined uniquely by the local field requires that the field changes sufficiently slowly in space that the effect of non-local fields on the distribution function is negligible. The problem with this assumption is that as devices change scale by a factor s, the average external field in a device changes by a factor $\frac{1}{s}$, and the *average spatial rate of change of the external field* in a device changes by a factor $\frac{1}{s^2}$, assuming the voltage change is kept relatively constant. Therefore, every time Moore's Law has halved the length scale of a device, it has roughly *doubled* the average external field, and has roughly *quadrupled* the spatial rate of change in the external field. When the characteristic length scale of devices crosses below ~ 300nm, the local dynamic equilibrium approximation has usually begun to break down [6, Fig. 23] [7, Fig. 4].

One obvious symptom that the local dynamic equilibrium approximation is breaking down comes from examining the energy density of the electron distribution. In a steady-state homogeneous field, the rate the distribution gains energy density w from the external field, $\left(\frac{\partial w}{\partial t}\right)_{\text{field}}$, must be equal to the rate at which the distribution loses energy density to the lattice via scattering, $\left(\frac{\partial w}{\partial t}\right)_{\text{scat}}$:

$$\overbrace{\frac{\partial w}{\partial t}}^{\text{Zero in steady state}} = \left(\frac{\partial w}{\partial t}\right)_{\text{field}} + \left(\frac{\partial w}{\partial t}\right)_{\text{scat}}. \tag{1.1}$$

The rate the distribution gains energy density from the field is easy to define in terms of the particle flux density $\mathbf{j}$ and the external force $\mathbf{F}$:

$$\left(\frac{\partial w}{\partial t}\right)_{\text{field}} = \mathbf{j} \cdot \mathbf{F}. \tag{1.2}$$

The rate at which the distribution loses energy density to the lattice via scattering is more complex to define, but in a thermal distribution it increases if the local energy density of the distribution increases. This behaviour can also be expected for distri-

[7] The external field/force is simply the part of the electric— or more generally, electromagnetic— field/force in a device that remains *once one has removed the part of the electric field/force associated with the bandstructure and scattering operator*.

butions which are very roughly thermal, such as those in local dynamic equilibrium. Therefore, in a homogeneous external field, local dynamic equilibrium is reached when the average energy of the distribution gets large enough that the rate of energy loss due to scattering is equal to the rate of energy gain from the field.

The assumption underlying this analysis of a homogeneous field— and underlying the local dynamic equilibrium assumption— is that in steady-state the rate of change of energy density in a small volume is zero. However, imagine a small volume between two cross-sections of a $1-$D device where the field increases rapidly in the direction of current flow. The external field on the *upstream* face of this volume will be smaller than the external field on the *downstream* face by construction. According to the local dynamic equilibrium approximation, we can expect that the average energy of electrons will be smaller on the upstream side of the small volume than on the downstream side. In steady state, the *particle flux* through both these faces will be the same.[8] But due to the difference in average energy, the *energy flux* entering the cross-sectional volume from the *upstream* face will be *smaller* than the energy flux exiting that small volume to the *downstream* face.[9] This fact is significant, because according to the general continuity equation, this implies that the net rate of energy change in the volume due to fields and scattering *must be positive in steady-state*. Therefore local dynamic equilibrium is only valid when the *average energy of the distribution changes slowly enough that we can neglect the divergence of the energy flux*.

Given the above analysis, a natural approach to move beyond the drift-diffusion equation is to find a transport equation where the energy density is also part of the electron state. The resulting model is known as an ENERGY TRANSPORT MODEL. This approach was pioneered by Stratton in 1962 [8], though most of the assumptions he used to derive the energy transport model have been superseded [9].

To form an energy transport model, we have to add the energy density continuity equation described earlier to our set of equations. As Bløtekjær showed in 1970 [10], one can derive the terms in the energy density continuity equation rigorously by simply multi-

[8] Ignoring net particle creation or destruction inside the cross-sectional volume.

[9] It is technically possible for a distribution to have a larger or identical energy flux than another distribution with the same particle flux and a larger average energy. This technical possibility is highly unlikely to be realized in realistic situations, and can be regarded as a pathological case.

plying both sides of the Boltzmann transport equation by energy, and integrating over all bands and crystal momenta. The terms in the energy density continuity equation are therefore weighted integrals of the distribution function, bandstructure and scattering operator. Apart from the weighted integral that defines the energy density, these weighted integrals are excess unknown variables. In order to close this equation, we therefore need some technique for estimating these excess unknown variables. This is known as the CLOSURE PROBLEM.

It is difficult to solve the closure problem in a theoretically sound manner [6, 9]. The fundamental problem is that a theoretically sound closure is only possible if there is a well-defined mapping from the total state variables— a position-dependent electric potential, electron particle density, and electron energy density— to an approximation of the distribution function at each point. Currently, there are only two special cases where such a mapping is known to exist.

The first special case is if the local dynamic equilibrium is just beginning to break down. When the local dynamic equilibrium approximation is just beginning to break down, the distribution can be approximated as being in local dynamic equilibrium with an external field *slightly upstream* of the local external field. A simple, unbiased estimate of this slightly upstream external field is the homogeneous field that creates a electron distribution which has the same average energy as the electron state. When this approximation is valid, the field of the homogeneous simulation associated with the local average energy will always be similar to the actual local field. If there is a large discrepancy between the local field and the homogeneous field that gives the same average energy as the local electron state, then this is a clear sign that transport is far from the local dynamic equilibrium regime, in which case the approximation described has no physical justification.

The second special case is if electron-electron collisions are sufficiently dominant to ensure the distribution *efficiently relaxes to an internal equilibrium*. In this case, the distribution can be assumed to be approximated by small perturbations to a heated Maxwell-Boltzmann distribution (or Fermi-Dirac distribution, if degeneracy effects matter). And this will be true *whether or not the that heated distribution is created by a homogeneous or inhomogeneous field*. Thus, if the electron-electron collisions are strong throughout the

device, then the energy transport equations can be closed in a theoretically sound way by using homogeneous field data with the same electron density.

This first special case is of limited theoretical significance. The energy transport equation was proposed in order to overcome the weaknesses of the local dynamic equilibrium approximation, so proposing a model based on the *near* local dynamic equilibrium approximation does not greatly expand the range of device geometries we can model [6, Fig. 23]. This exact same criticism can be levelled at the so-called "six moment model" [11], which has also gained popularity in TCAD applications [12]. The six moment model expands on the energy transport model by allowing for a small perturbation to the *square energy* predicted by the local dynamic equilibrium. Accordingly, we may expect that a six moment model correctly closed using homogeneous field data will give better results than the energy transport model or drift diffusion model when the average energy and square average energy diverge *slightly* from their local dynamic equilibrium values. But once again we have no theoretically sound reason to expect better results from this model when the perturbations from local dynamic equilibrium are large and electron-electron collisions are insufficiently strong to drive the distribution toward internal equilibrium.

On the other hand, the second special case— where frequent electron-electron collisions bring electrons into an internal equilibrium— allows us to model a genuinely new regime of transport. The transport parameters can be gleaned from a homogeneous simulation which possesses the same local energy and density, *even when this homogeneous field is completely different from the local field* [13, Fig. 4]. In the presence of dominant electron-electron collisions, the energy-transport model significantly expands the length scale of modelling capability beyond the drift-diffusion model in a theoretically sound manner.

Unfortunately it is seldom the case that this internal equilibrium approximation holds everywhere in a device. Most practical devices contain enormous spatial variations in the density of electrons, often including regions that are effectively *depleted* of electrons. The internal equilibrium approximation of the second special case is only a valid a few mean free paths into the interior of regions with sufficiently high electron density. Unfortunately, the field is its *most inhomogeneous* outside the interiors described, since the

relative absence of mobile charge is precisely what allows large inhomogeneities in the external field to exist. This means the energy transport model is only valid for a device as a whole if the near local dynamic equilibrium approximation is valid in these highly inhomogeneous, relatively depleted regions.

The theoretical viewpoint we have described in this section can be used to interpret, for example, the results of Vasicek et al. [14, Fig. 6], shown in Fig. 1.1. These results show the drain current prediction error of macroscopic models for Double-Gate MOSFETs of various channel lengths. The macroscopic models are essentially those proposed by Grasser et al. [15], which are distinctive because unlike many other proposed closures [9] these closures are constructed from near-optimal use of homogeneous field data, without any ad-hoc alterations or unnecessary bandstructure or scattering operator simplifications. As such, if we are in a regime where a sound closure using homogeneous field data is theoretically possible, we expect these particular macroscopic models to actually achieve that theoretically possible accuracy.

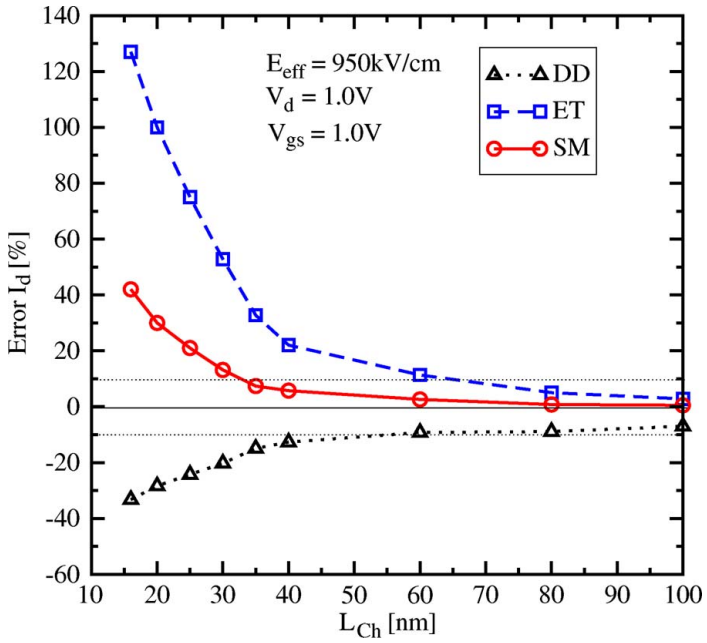


Figure 1.1: The results of Vasicek et al. [14, Fig. 6]. The error of a Drift-Diffusion (DD), Energy Transport (ET), and Six Moment (SM) in predicting the drain current of a Double Gate MOSFET with various channel lengths relative to a detailed Monte Carlo simulation. Regardless of channel length, the Double Gate MOSFET is subject to a source-drain bias of 1V, a gate-source bias of 1V, and an average transverse external field of 950kV/cm.

An important caveat is that the closure of these macroscopic models is more complicated than comparing to data generated by homogeneous field simulations in a bulk semiconductor. This is because the majority of conduction electrons in the channel of a MOSFET exist in a INVERSION LAYER parallel to a semiconductor–oxide heterojunction[10]. In effect, the conduction electrons in the channel do not exist in a bulk silicon crystal, but in a small portion of silicon crystal *parallel to the heterojunction*.

The extent to which conduction electron states are "squashed" against the face of the silicon crystal depends on the size of the relatively homogeneous field perpendicular to the heterojunction. At zero field, no such squashing occurs and the material is bulk silicon. When the field is large, the conduction electrons in the channel experience more of a $2-$D bandstructure, and experience new surface scattering mechanisms in addition to the ordinary bulk ones. Put loosely, the effective "material" that the conduction electrons in the channel experience depends on the field strength perpendicular to the heterojunction. As a result, the homogeneous transport simulations required to close the macroscopic models need to involve the entire spectrum of different effective "materials" that channel conduction electrons experience. To achieve this, the transport parameters are derived from an infinite planar Si–SiO_2 junction— rather than bulk silicon— subject to various homogeneous fields. The homogeneous[11] field perpendicular to the junction defines the effective "material", and the homogeneous field parallel to the junction defines the driving force on the electron distribution.

[10] Which in this case is a silicon–silicon-dioxide heterojunction.

[11] A minor caveat is that these fields are only *nearly* homogeneous, since the heterojunction creates small field inhomogeneities in silicon. The size of the fields perpendicular to the heterojunction are therefore characterized by the *average perpendicular field* strength the channel conduction electron density is subject to.

The results of Vasicek et al. are in the regime where local dynamic equilibrium is well-broken. This is ensured by the fact that the results for a device with a relatively large source-drain bias ($\sim$ 1V) and a characteristic length scale which is at most 100nm—considerably shorter than the $\sim$ 300nm required for local dynamic equilibrium. Accordingly, the drift-diffusion model does not accurately predict the drain current for any of the devices shown, since the largest device has a channel length of 100nm, and the gradient of error with respect to channel length ($\sim 1\%$ per 20nm) suggests that the drift-diffusion model will not be accurate until the channel length is ~ 300nm.

The results of Vasicek et al. are also in a regime where electrons crossing the device are subject to fairly strong electron-electron scattering throughout their entire transit. This is ensured by the high gate-source voltage (1V), which implies that the conductive part of the channel is a high-density inversion region. We expect then that the domain of validity of the energy transport and six moment model is extended beyond the near local dynamic equilibrium approximation, and that this fact is responsible for the accuracy of the energy transport model down to $\sim$ 100nm, and the accuracy of the six transport model all the way down to ~ 80nm. We expect then that in MOSFET's where the gate-source voltage is *small*, and the source-drain voltage is similarly large, the assumptions underlying both the energy transport and six moment model would break down at larger length scales then this. Unfortunately, we cannot illustrate this using actual results from the paper, because the paper omits all results with $V_g < 0.2$V. However, the paper does acknowledge that "the macroscopic models overestimate the drain current at low V_g" which supports our expectations.

Finally, the results of Vasicek et al. also show a regime where the assumptions underlying all models are clearly broken down: specifically, below ~ 50nm. It is valuable to show this regime because it makes clear the point that once models are not theoretically sound there is no reason to expect a more sophisticated model to be more accurate than a less sophisticated model. That is, when models are fundamentally theoretically unsound, it is just as plausible that additional terms lead to larger errors as it is that they lead to smaller errors. This is illustrated by the fact that the energy transport model is *far less* accurate than the drift diffusion model in the regime where all models are broken. In this particular example, the six moment transport model remains more accurate than the drift-diffusion model down to ~ 20nm, but this provides no guarantee that the

six moment model will not generate even larger errors than the energy transport model in another device when the assumptions underlying both models are broken.

To conclude this section, the most popular models in TCAD applications are the drift-diffusion model, the energy transport model and— more recently— the six moment model. We refer to these models as QUASI-HOMOGENEOUS MACROSCOPIC TRANSPORT MODELS: *macroscopic* in the sense that the electron state is defined by simple macroscopic densities, *quasi-homogeneous* in the sense that the closure of these models relies on an similarity between the local distribution function in a device and the distribution function in a homogeneous field. In the case of the drift-diffusion model, this quasi-homogeneous closure is only theoretically sound when the external field is sufficiently homogeneous for the local dynamic equilibrium approximation hold. In the case of the energy transport and six moment models, this local closure is only theoretically sound when *either* the external field is sufficiently homogeneous for the near local dynamic equilibrium approximation to hold, *or* when the electron-electron collisions are sufficiently strong to ensure the electron distribution is near internal thermal equilibrium. We refer collectively to these assumptions as QUASI-HOMOGENOUS CLOSURE ASSUMPTIONS. We refer to the regime of transport in which none of the quasi-homogeneous closure assumptions hold as the INNATELY INHOMOGENEOUS regime of transport.[12]

1.3 Transport in the Innately Inhomogeneous Regime

When we begin to investigate transport in the innately inhomogeneous regime it quickly becomes apparent that quasi-homogeneous macroscopic transport models not only provide a solution to the primary problem of finding a sound mapping between the state variables and the distribution function, they also provide a very elegant solution to a *secondary* difficulty in overcoming the closure problem. The secondary problem is that

[12]In the existing literature, the quasi-homogeneous and innately inhomogeneous regimes of transport will often be referred to respectively as the LOCAL and NON-LOCAL regimes. We avoid this terminology, because local and non-local are important mathematical terms that we may want to use in other contexts without confusion. For instance, the full Boltzmann transport equation— which is a valid model of transport for *both* the quasi-homogeneous and innately inhomogeneous regimes of transport in the semiclassical regime— is *local* in position-space, and *non-local* in momentum-space. This statement would becomes confused if we use the existing terminology.

in any device much smaller than a micron which is subject to modest supply voltages of a few volts, the average energy of the distribution function inside the device will be large enough that low energy approximations of the bandstructure and scattering operator will no longer be valid for many semiconductors— including silicon. Accordingly, for these semiconductors the precise nature of the full band-structure and full scattering operator will have important effects on electron transport [16]. The elegant solution quasi-homogeneous macroscopic models provide is that they *neatly factor this problem out of the device modelling step*. When quasi-homogeneous closure approximations are valid, quasi-homogeneous macroscopic models incorporate *almost all* the effects of a complex scattering operator and bandstructure on electron transport in a device simply by extracting a small set of transport parameters from highly detailed simulations of homogeneous field transport [15].

Unfortunately, in the innately inhomogeneous regime there is currently no known theoretically sound approach that similarly separates the problem of modelling the effect of complex scattering operators and bandstructures from the problem of modelling the effect of the highly inhomogeneous external fields inside modern semiconductor devices [6, 17, 12]. This is problematic, because encoding the net effect of a complex bandstructure and scattering operator on electron transport into a set of device-agnostic transport parameters that can be *precomputed and tabulated* greatly increases the speed with which a device model can be solved. Accordingly, all currently proposed theoretically sound models of electron transport in the innately inhomogeneous regime are much, much slower to solve than the quasi-homogeneous macroscopic models.

At present, the only method that can claim to be truly capable of simultaneously incorporating detailed scattering operators, arbitrary bandstructure, and an arbitrary position dependent external potential is the Monte Carlo method [16, 18, 17]. Therefore, the Monte Carlo method is currently the only theoretically sound model of innately inhomogeneous transport available, as all other models make assumptions that contradict the microscopic reality of the situation. The problem with using the Monte Carlo method to evaluate the transport characteristics of a device is that, compared to macroscopic modelling, it is CPU intensive. For example, in the paper of Vasicek et al. from which Fig. 1.1 is taken, finding the *entire I-V curve* on an Intel Core i7 CPU 3.4 GHz machine with the drift diffusion, energy transport and six-moment models took ~ 3,

~ 5 and ~ 15 *minutes* respectively. In contrast finding the current *at a single voltage point* with the Monte Carlo model they used as the control took to 6-7 *hours* [14]. While much faster Monte Carlo simulations are possible— Vasicek et al. admit in the paper the Monte Carlo simulation they use is not optimized for speed— this illustrates common problem that Monte Carlo models which do not ignore crucial physics have CPU-times that are several orders-of-magnitude greater than macroscopic models.

The extent to which a model is used in TCAD applications is determined primarily by the speed with which it generates the transport characteristics of a given device [12, 2]. Accordingly theoretically unsound device models that are empirically calibrated to give accurate results in a narrow set of device geometries will always be necessary in TCAD applications as long as theoretically sound models of the innately inhomogeneous regime take significantly longer to solve than quasi-homogeneous macroscopic models. And as long as theoretically sound models are much slower than macroscopic models, their role in TCAD applications will always be limited to determining tuning parameters for these empirical models, and providing insight and intuition into the physics of transport [13, 19].

Despite the pragmatic workaround afforded by empirical models, speed improvements to theoretically sound models of the innately inhomogeneous regime are still incredibly valuable to the semiconductor industry. This is especially true in an environment where device designers are exploring complex $3-$D designs with novel geometries that are constantly exposing novel physical effects [19]. This trend means more and more theoretically sound simulations are required, both to understand this novel physics and to recalibrate theoretically unsound empirical models that are only narrowly reliable by their very nature. This trend means that the time taken to perform theoretically sound simulations in the innately inhomogeneous regime is becoming more and more of a bottleneck to innovation [2].

The need for theoretically sound models of innately inhomogeneous electron transport that are much faster to solve than Monte-Carlo is precisely what we aim to address in this thesis. In this thesis we propose two novel innately inhomogeneous electron transport models that we argue are theoretically sound: the ELASTICALLY-CONSTRAINED EQUILIBRIUM MODEL, and the THREE-EQUILIBRIA MODEL. Both models rely on an ansatz for the

distribution function, which allows us to *precompute and tabulate* transport parameters that describe the effect of the complicated scattering operators and bandstructures on these ansatz. The two models differ primarily by the fact that the ansatz used in the elastically-constrained equilibrium model makes fewer assumptions than in the three-equilibrium model. As a result, the elastically-constrained equilibrium model has a broader range of validity than the three-equilibria model, but it is considerably slower to solve. However both models are expected to be much much faster to solve than a Monte Carlo simulation with the same level of detail, it is simply that the THREE-EQUILIBRIUM MODEL is expected to have a speed comparable to quasi-homogeneous macroscopic models. Both models are valid well into the innately inhomogeneous regime, and can incorporate the effects of arbitrary bandstructure and a wide range of scattering operators. As such, both approaches seem to offer significant value to the semiconductor industry and therefore are worthy of further research and development.

1.4 The Structure of this Thesis

The two models of transport we present in this thesis are quite general. The focus of this thesis, however, is to elucidate the kernel of both transport models in a clear and concrete manner. To achieve this, we limit ourselves to modelling transport in simple, concrete conditions. Specifically, we will model devices in which electrons travel via pure semiclassical transport through a single crystal of silicon.

We expect with further research and development, these models can be adapted beyond these specific conditions using similar techniques that are used to adapt local transport models and Monte Carlo models to beyond these conditions. That is, we expect that the two models presented in this thesis can be adapted for:

- devices in which holes contribute significantly to transport,
- devices based on semiconductors other than silicon,
- devices involving heterojunctions,

- devices in which the bandstructure is altered by mechanical stress,
- devices in which the bandstructure is altered by quantum confinement effects, and
- devices in which the external electric potential is replaced by an external effective potential that includes first-order quantum corrections.

It is well beyond the scope of this thesis to explicitly attempt to demonstrate any such adaptions, however. In this thesis, we focus entirely on monopolar, pure semiclassical transport in silicon.

In the background chapter, we derive a model of transport in this regime that is well-accepted to be theoretically sound [17]. Specifically, we derive the Boltzmann–Poisson transport equation system subject to the scattering operator and bandstructure used in the IBM DAMOCLES program [16, 20, 6]. We take pains to describe clearly the many assumptions that underlie this model, relying more heavily on first principles than on established explanations. This effort is not as gratuitous as it may seem. The rapid pace of device innovation and scale-reduction in the semiconductor industry means that in the long term it is likely all assumptions underlying any regime of electron transport will eventually be stretched or broken by new technologies. If our understanding of these assumptions is not as clear as possible, we will underestimate or overestimate the regimes of validity for these assumptions. Underestimation leads to premature model complexity, and overestimation leads to unexpected inaccuracy. Both are undesirable. Therefore new perspectives on old assumptions are useful as they help to generate additional clarity within the field.

The background section ends with a brief review of innately inhomogeneous electron transport models proposed in the literature. Each model is described in terms of the simplifying assumptions it makes to the state, bandstructure and scattering operator of the full Boltzmann transport equation. The models are then judged in terms of the theoretical soundness of these assumptions, and in terms of the computational resources required to solve the model.

In the theoretical framework chapter, we describe the approach this thesis endorses

for simplifying the state of the Boltzmann transport equation. Since this theoretical framework is essentially a novel non-equilibrium statistical mechanical argument that is potentially of broader interest, we will give a fairly detailed outline of it here.

The basic argument begins with the observation that there are two types of terms in the Boltzmann transport equation. The parts of the device environment that are precisely controlled are associated with REPRODUCIBLE HAMILTONIAN terms, which in the semiclassical regime are associated with an external field and bandstructure, whereas the parts of the device environment that are imprecisely controlled are associated with NON-REPRODUCIBLE HAMILTONIAN terms, which in the semiclassical regime are associated with a spatially localized scattering operator.

The *reproducible Hamiltonian* terms result in the classical Hamiltonian evolution of the distribution function. This is a reversible, one-to-one mapping of the entire distribution function to another distribution function at any particular later time. Accordingly any distribution function which has been subject to *only these classic dynamics* will be the result of *only one* distribution function at a given earlier time. Thus if a distribution function is only subject to these classic dynamics, *all the information encoded into the distribution function is strictly required* to determine the future distribution function of the electrons. The electron state is intrinsically unable to be simplified without discarding essential information. As such, if electrons are subject to only these classical dynamics, there would be no theoretically sound manner to simplify the Boltzmann transport equation.

The *non-reproducible Hamiltonian* terms are associated with an irreversible, many-to-one mapping of the local distribution function. Accordingly any distribution function which has been subject to scattering dynamics can be the result of *multiple* distribution functions at earlier times. Thus if a distribution is subject to scattering, *only some subset of the information encoded in the distribution function is strictly required to determine the future state* of the electrons. Put another way, the imprecisely defined environment the electrons are subjected to results in a continuous *erasure* of the physical information contained in the distribution function. The result of this information destruction means that the space of physically possible local distribution functions is no longer accurately described by the function space of arbitrary $3-$D scalar functions. In intuitive terms, the "information

storage capacity" of this function space is much larger than physically required, and it is possible to use a much smaller function space to encode the same physical information. As such, we have a theoretically sound reason for arguing that a careful evaluation of the scattering operator will lead to a simplification of the electron state function. The remainder of the theoretical framework section goes toward understanding the precise physical information that scattering erases.

The first layer of information erasure we describe is intuitively well-known: the erasure of crystal momentum information. We formalize this process by arguing scattering drives the local distribution efficiently toward an ELASTICALLY CONSTRAINED EQUILIBRIUM. An elastically constrained equilibrium is a state in which all microstates consistent with a given *energy-dependent density function* are equally probable. In the first results chapter, we use essentially only this assumption in order to derive a corresponding model of innately inhomogeneous electron transport in which the electron state is defined by a position and energy dependent distribution function, which is subject to diffusion at constant total energy, and an arbitrary, purely inelastic scattering operator.

The second layer of information erasure we describe is the erasure of energy distribution information. The erasure of energy information by the purely inelastic component of scattering is far less complete than the erasure of crystal momentum information by the purely elastic component of scattering. The reason is that a very broad range of crystal momenta are typically exchanged between an electron and the environment during scattering, while the energy exchanged between an electron during scattering is generally much more narrowly defined. As a result, the energy distribution function information is typically inefficiently erased by the inelastic scattering operator. There are two well-known exceptions. The first is electron-electron scattering events which, unlike other scattering mechanisms, are associated with a very broad range of energy exchanges for the electrons involved. These scattering events efficiently erase all energy distribution function information apart from the total energy and particle density. The second is electron-phonon scattering events for subpopulations of electrons that are close to being thermally distributed at the lattice temperature. In this specific situation, the inelastic scattering does in fact efficiently erase all perturbations from thermal equilibrium in the energy distribution function.

The two efficient energy distribution information erasure mechanisms associated with the inelastic scattering operator are insufficient to meaningfully simplify the electron energy distribution function in the innately inhomogeneous regime. There is however, a *third* mechanism of energy distribution erasure that is associated with the elastic component of scattering rather than the inelastic: diffusion at constant total energy. Diffusion at constant total energy is an irreversible, many-to-one mapping of the distribution function. This diffusion process acts to erase information concerning *short-range distribution function fluctuations at constant total energy*. As a result, the occupation rate at constant total energy will only tend to reflect the *long-range* spatial trends in sources and sinks for electrons entering and exiting the total energy level, rather than the short range trends in these quantities. If the external field is very small, the *precise kinetic-energy dependent details* of the inelastic scattering operator can generate the long-range spatial trends in the sources and sinks of electrons entering or exiting a total energy level. However, when the field is large, the precise kinetic-energy dependent details of the inelastic scattering operator can only generate short range trends in the sources and sinks entering or exiting a total energy level, which will not be reflected in the distribution function. As a result, in devices where external fields are large, we have a theoretically sound reason for expecting that information regarding the precise kinetic-energy dependent details of inelastic scattering will not be reflected in the shape of the distribution function.

We propose a simple three-equilibria ansatz for the energy distribution function that we argue roughly captures the small amount of energy-distribution function information that is not erased by the three aforementioned efficient energy-distribution information erasure processes. This ansatz relies on some simple qualitative assumptions about the inelastic scattering operator and the shape of the external potential, the former of which is believed to be valid in silicon, the latter of which is believed to be valid in many devices of interest. In the second results chapter, we derive a complete transport model for silicon devices based on this ansatz.

In the discussion and conclusion sections we discuss the caveats of the two models, and the urgent need to test these models. We then describe broader implications of this work, both for the semiconductor industry and for theoretical physics as a discipline.

On the practical side, the potential impact of these models on the semiconductor industry is very large. But this impact will not materialize until these models are made into flexible, easy to use tools that can be used to model a wide range of devices. This will require significant additional effort both from more research and development, both within the TCAD industry and in academia.

On the theoretical side, the framework used to derive the models proposed in this thesis is quite broadly applicable. The approach presented to simplifying the scattering operator is quite general, and may find application in other transport problems.

Statement of Contribution

The major original contributions in this thesis are as follows:

1. In the Theoretical Framework chapter: the definition of— and description of scattering mechanisms that lead to— the WEDGE CONSTRAINED, ELASTICALLY CONSTRAINED, and CHEMICALLY CONSTRAINED QUASI-EQUILIBRIA. These novel forms of quasi-equilibrium far from thermal equilibrium are likely to be of wide interest.

2. In the Results I chapter: the derivation of the ELASTICALLY CONSTRAINED TRANSPORT model from first principles. This demonstrates that the model of Dmitruk et al. [21] is a much more generally applicable class of model than the field has yet appreciated.

3. In the Results II chapter: the derivation of a completely novel THREE EQUILIBRIUM TRANSPORT model from first principles. No macroscopic model has ever been derived from an ansatz that is so realistic in the innately inhomogeneous regime.

4. In the Results II chapter: the derivation of ansatz-parameter dependent diffusion and inelastic scattering parameters directly from a scattering operator and bandstructure with a DAMOCLES level-of-detail [16]. No macroscopic model has ever directly incorporated such a realistic scattering operator and bandstructure.

Chapter 2

Background

2.1 Introduction

Entropy is an increasing function of the expected number of precise microscopic states that can be associated with a system characterized by a non-precise specification of state [22, 23]. If we have two systems, and the total entropy of both systems can be increased by moving particles from one system to the other then, on average, particles will be transferred between these systems in this direction of increasing entropy. However, the actual rate at which particles are transferred from one system to the other is determined by the specific mechanism for particle transfer. Entropy considerations alone can only reveal the *average direction* of particle transfer.

In semiconductor transport theory, we study the entropy driven transfer between systems of particles called TERMINALS, via the specific mechanism of the movement of CARRIER QUASIPARTICLES through a SEMICONDUCTOR DEVICE. In this thesis, we study devices in which there is sufficient disorder that the carrier state can be described by a SEMICLASSICAL DISTRIBUTION FUNCTION, and in which the external field is so large and inhomogeneous that this distribution function at any given point *cannot be assumed* to be small perturbation from the distribution function in any homogeneous field simulation. This is the INNATELY INHOMOGENEOUS regime of SEMICLASSICAL TRANSPORT. In this chapter,

we seek to provide a solid introduction to a model of transport in this regime that is well-accepted to be theoretically sound, the BOLTZMANN TRANSPORT EQUATION subject to a DETAILED SCATTERING OPERATOR and FULL BANDSTRUCTURE. We refer to this model as the FULL BOLTZMANN TRANSPORT EQUATION.

2.2 Roots of the Semiclassical Regime

In this section we briefly review how disorder in a device leads to carriers which are governed by semiclassical— instead of fully quantum— dynamics.

We begin by examining the basic concept of a *carrier*. In the ground state of an ideal, pure semiconductor, each primitive cell is filled with a suitable number of pairs of electrons such the next available state for an electron is precipitously higher in energy than the last available state for an electron, by a BAND GAP. This band gap inhibits the net transfer of these ground state VALENCE ELECTRONS from one primitive cell to another, even in the presence of quite large electric fields. In order for there to be a transfer of charge that is not inhibited by this precipitous energy cost, we either need to add an electron which has already paid this precipitously higher energy cost to the CONDUCTION BAND of the system, or we need to remove an electron from system and leave a HOLE in the VALENCE BAND which can transport charge from primitive cell to primitive cell without paying the precipitous energy cost. This mobile subpopulation of conduction electrons and holes are known collectively known as carriers.

Next, we observe that a semiconductor device is a *non-precise specification* of a physical state. Fundamentally, a device is a collection of particles that has *some reproducible organization* due to process control during fabrication or experimental control during use, juxtaposed with degrees of freedom that are beyond process or experimental control.[1] The *reproducible organization* in a semiconductor device is:

- the position-dependent ideal semiconductor crystal the device locally approxi-

[1]The author learned the general strategy of first identifying the reproducible and non-reproducible parts of a physical system from Jaynes [24].

mates,

- the lattice temperature,
- the position-dependent dopant/defect density, and
- a reproducible position-dependent "external" force.

The degrees of freedom that can be assumed to be beyond process or experimental control are:

- the effect of a local dopant/defect density on each precise primitive cell, and
- the effect of a known lattice temperature on each precise primitive cell at a each precise time.

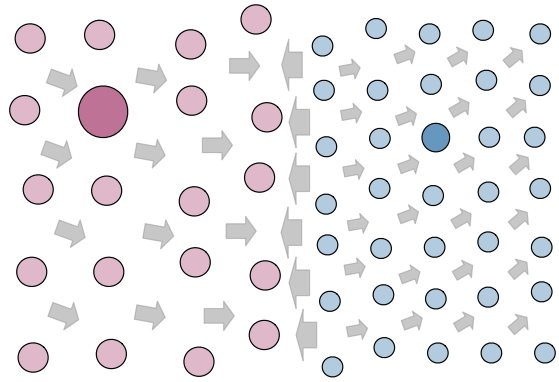
It is true that the carrier state associated with the Hamiltonian of a *single manifestation* of a device at a particular increment of time may have a complex pattern of constructive and destructive quantum interference that is intractably complex to compute. However, our model will only ever be required to predict the statistics associated with reproducible experiments, which by their very nature involve a *population* of different device manifestations tested at different times. Accordingly, it is appropriate that our DEVICE MODEL is a statistical mixture of the pure Hamiltonians that would be associated with a particular device manifestation at a particular time. This statistical mixture represents the population of possible precise device manifestations a carrier is statistically expected to encounter in a set of reproducible experiments. If there are many degrees of freedom that are beyond process and experimental control, then many of the intricate patterns of constructive and destructive quantum interference that exist for the carrier state of a single device manifestation will be averaged out when we consider the carrier state of the *device model*.

The only details in the quantum interference patterns that will remain are those that are reproducible across all the pure Hamiltonians in our statistical mixture. These reproducible interference patterns can only be due to the sub-Hamiltonians common to all

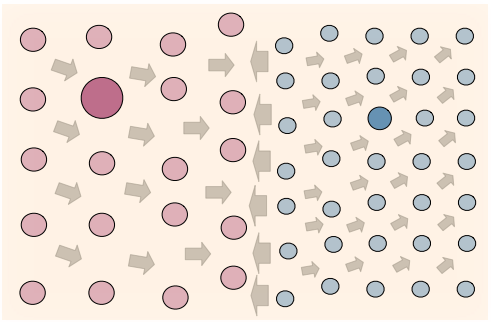
Hamiltonians in our statistical mixture. Since averaging over the uncontrolled degrees of freedom *will not* average out the interference patterns associated with the common sub-Hamiltonian, it is important that the time evolution of the carrier that is associated with this common sub-Hamiltonian is treated as *fully* quantum mechanical. The physical effect of the remaining non-reproducible Hamiltonians in our statistical mixture can be described as simply scattering carrier probability density *incoherently*; that is, without phase information, since phase information will be averaged out over the device population. The first-order effect of incoherent perturbations is a standard result of quantum mechanics known as FERMI'S GOLDEN RULE. We will discuss these incoherent scattering perturbations later, and will first discuss the reproducible order in a population of devices that leads to a common sub-Hamiltonian across all Hamiltonians in our statistical mixture.

We separate the device–carrier interaction common to all Hamiltonians in our statistical mixture into two parts: a quickly varying potential due to the reproducible crystal, and a slowly varying "external" force. The external force can be thought as the slowly-varying macroscopic vector force field that remains after averaging over all uncontrolled degrees of freedom in the device, and removing the quickly varying, approximately periodic, crystal potential. The use of the word "external" is conventional but confusing terminology: the external force will typically be attributed to unbalanced charge *inside* the semiconductor device.

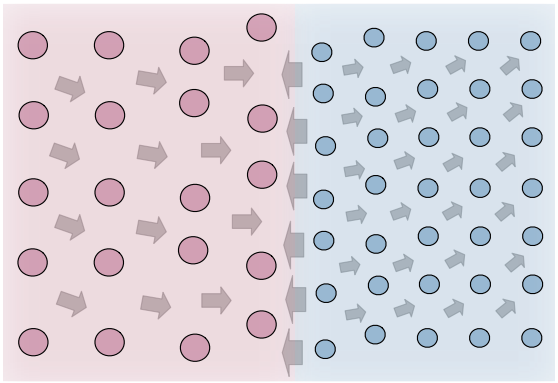
If it helps the reader, in Fig. 2.1 we show a crude schematic illustrating the step-by-step paring back of the pure Hamiltonian associated with a particular device manifestation, to the sub-Hamiltonian associated with the reproducible order common to all device manifestations in the device population. All the unreproducible effects that are removed will be treated later as perturbations that cause incoherent scattering. Note the role of the blue crystal in this figure could just as easily be played by a vacuum or other insulator, and so the crystal interface can also represent the edge of a finite semiconductor device. The purpose of this process is to illustrate the sub-Hamiltonian which creates patterns of constructive and destructive interference that are *not* averaged out across the population of particular device manifestations; that is, the sub-Hamiltonian that will effect carriers in a purely quantum mechanical manner.



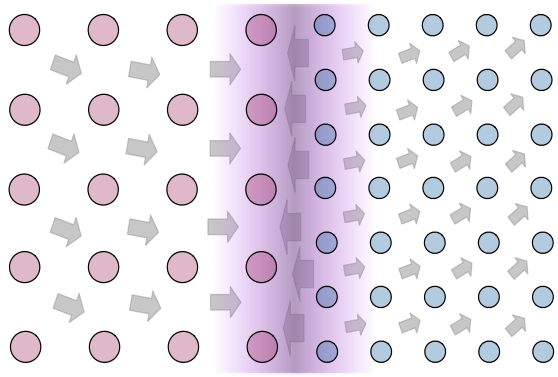
(a) A particular device manifestation. The schematic representation of the pure Hamiltonian a carrier experiences due to a particular instance of a device near a crystal interface at a particular time. The reproducible, non-periodic "external" force is represented by grey arrows. (Technical note: the discontinuity in external force at the interface typical, and is caused by the band energy of a carrier in the blue crystal being higher than that of a carrier in the red. The scale of this discontinuity will in general be **k** dependant, which is atypical as outside this kind of discontinuities the external force is roughly independent of **k**.)



(b) Removing phonons. The precise momentary change to the carrier potential due to the excitation of atoms from their equilibrium positions is not coordinated between other possible pure Hamiltonians in the statistical mixture, so its reproducible effect on carriers can treated by an incoherent scattering perturbation (represented by light orange) to a phononless potential. Note, the anharmonic effects of these excitations will lead to a uniform thermal expansion of the average ion positions, which *is* reproducible across the statistical mixture for a given lattice temperature.



(c) Removing dopants. The position dependent forces due to the precise location of dopants are not reproducible across other possible pure Hamiltonians, as the precise atomic position of dopants are not experimentally controlled. So the reproducible effect on carriers of dopants can be found by a incoherent scattering perturbation (represented by light blue and red) to a dopantless potential. Note, the contribution of dopant atoms to the occupation of carrier states is a reproducible effect, as well as their contribution to the external force when their contrasting nuclear charge is not balanced by similarly contrasting local electron charge.



(d) Removing defects. The position dependent forces due to defects are in general not a reproducible effect, and therefore as before we treat their effect as an incoherent scattering perturbation (represented by light purple) a defectless potential. Depicted in the diagram is the diagram is a borderline case of this reasoning— that of surface reconstructions— which in very controlled device growth conditions could cause reproducible patterns of carrier interference as the forces would be reproducible across the pure Hamiltonians in our statistical mixture.

Figure 2.1: Removing non-universal force fields in a statistical mixture of Hamiltonians as incoherent scattering perturbations.

Finding the solution for carrier states in this electromagnetic field common to all devices in the population is still a fairly intractable problem if tackled directly. To make progress,

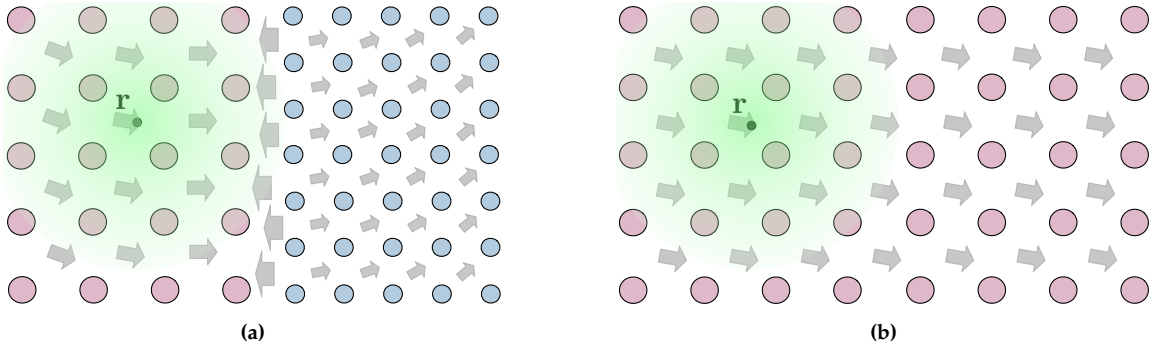


Figure 2.2: A schematic illustration of the assumption that a coherent carrier wavepacket centred at $\mathbf{r}$ is sufficiently localized that it interacts with the reproducible part of the Hamiltonian as if it were the superposition of a perfectly periodic crystal and a uniform force.

we presume that the *non-reproducible* parts of the device Hamiltonian localize the carrier to a small enough area of the crystal such that the carriers experiences a quickly varying crystal potential that is roughly equivalent to perfectly periodic potential, and experiences the slowly varying external force as a uniform force. This presumption is illustrated schematically in Fig. 2.2, *and is the essential assumption that leads to semiclassical dynamics*.

It is well-known that there are no stable eigenstates to this simple Hamiltonian involving a perfectly periodic lattice and a uniform external field, if the uniform external field is non-zero [25]. Accordingly, we begin in a standard manner by first finding the eigenstates of the zero external force problem, and only then will we investigate their time-dependant evolution when the uniform field is turned on.

The single carrier eigenstates in a perfectly periodic potential are known as BLOCH WAVES, and are labelled by a CRYSTAL MOMENTUM $\mathbf{k}$— which is restricted by convention to a BRILLOUIN ZONE— and a BAND INDEX ν.[2] The energy eigenvalues— $\varepsilon_{\mathbf{k}\nu}$— corresponding to each of these single carrier eigenstates— $|\mathbf{k}\nu\rangle$— are given by a function known as the BANDSTRUCTURE. The background of these terms is described in Ap-

[2]We will present the variables crystal momentum $\mathbf{k}$ and band index ν as a combined pair of variables $\mathbf{k}\nu$. This notational convention highlights the fact that the *pair* of variables is associated with a single $3-$D degree of freedom, since it is possible to create an extended zone scheme where $\mathbf{k} \in \mathbb{R}^3$ and there is only one band, so $\nu = 1$. We will not use this extended zone scheme unless explicitly specified, as in most cases a reduced zone scheme in which $\mathbf{k} \in$ BZ (where "BZ" is the first Brillouin Zone) and $\nu \in \mathbb{Z} \setminus \{0\}$ is much more physically intuitive. Nevertheless, it is useful to have a notational reminder that fundamentally it is the *pair* of variables that is associated with a *single* $3-$D degree of freedom.

pendix A. In near-equilibrium carrier transport, the carriers tend to be concentrated in the bottom of the conduction band and top of the valence band. By Taylor's theorem, this allows one to approximate the energy eigenvalues as a function of displacement from the crystal momentum at the extrema as a $3-$D quadratic. In far-from-lattice-temperature-equilibrium transport no such approximation can be relied upon— carriers are widely dispersed throughout conduction and valence band energies— and so the full bandstructure is required.

The problem of calculating a full bandstructure from first principles is an area of active research that can be considered to be completely disconnected from this thesis. As explained in Section 2.4, this is because there is a phenomenological approach— known as the EMPIRICAL PSEUDOPOTENTIAL METHOD— that enables one to approximate the full bandstructure from experimental measurements of the optical band gaps. However, we would like to note here a possible point of confusion with regard to bandstructure calculation. From the perspective of bandstructure physicists, the bandstructure is the solution to a many-body electron problem that is *concerned with* every electron but the carriers. Whereas from the perspective of a carrier transport physicist, it is simply a statement of the kinetic energy to crystal momentum relationship for carrier quasiparticles that allows us to *ignore* every electron but the carriers. It is worth briefly outlining why these two views are not contradictory.

The field due to a highly localized carrier will be a locally significant perturbation to the ideal crystal. As such, we expect a correlation between the position of a carrier and the local state of the ideal crystal. This correlation typically takes the form of a polarization of the ideal lattice, since the local electrons in the crystal can only rearrange themselves in a manner that conserves the charge in each primitive cell. This is because— as discussed earlier— charge transport between primitive cells requires the precipitous energy cost defined by the band gap. Thus carrier–ideal crystal correlations can typically be described by an altered dielectric constant. We note that, if a carrier has a kinetic energy greater than the band gap, then valence band electrons can respond dynamically. Thus to complete the description of the carrier–ideal crystal correlation we must also include an IMPACT IONIZATION term in the scattering operator, which is associated with carriers that have an energy greater than the band gap.

We define a perfect bandstructure calculation as providing the set of energy eigenstates of a single valence electron state at a given crystal wavevector, taking into account the perfectly periodic interaction of the electron with ground state of the crystal ions and every other valence electron. To create a single carrier quasiparticle, we flip the ground-state occupation number at a single wavevector and band index. The eigenvalues of a carrier will only be the same as the eigenvalues of the bandstructure if the correlation term is negligible.[3] In a monatomic semiconductor such as silicon, we expect the polarizability of a primitive cell to be small, and therefore expect the bandstructure to yield the energy eigenvalues of isolated carriers.

The Bloch waves have a well-defined average velocity. It can be shown (for instance, see Ashcroft and Mermin [26]) that this average velocity is proportional to the $\mathbf{k}$-space gradient of the bandstructure:

$$\mathbf{v}(\mathbf{k}\nu) = \frac{1}{\hbar}\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}. \tag{2.1}$$

There are two different ways in which this equation can be seen as physical intuitive. On the one hand, if we think of the bandstructure (divided by $\hbar$) as a dispersion relation, then eq. (2.1) is the equation for the group velocity of a wavepacket made of frequencies near $\frac{\varepsilon_{\mathbf{k}\nu}}{\hbar}$. On the other hand, if one thinks of $\mathbf{k}$ (multiplied by $\hbar$) as a quasi-momentum[4] $\mathbf{p} = \hbar\mathbf{k}$, and the energy eigenvalues as the kinetic energy of the carrier $\varepsilon_{\text{kinetic}}(\mathbf{k}\nu) = \varepsilon_{\mathbf{k}\nu}$, it can be considered to be analogous to the formula in classical mechanics $\mathbf{v} = \nabla_{\mathbf{p}}\varepsilon_{\text{kinetic}}$.

Having found the carrier Bloch eigenstates of the zero external force problem, and written down an expression for their velocity, it is time to examine the dynamic evolution of these carriers when we turn on the uniform external force. When we turn on the uniform external force we are immediately confronted with the INTERBAND TRANSITION issue: the uniform external force will stimulate electrons to make transitions between bands. For instance, if the valence band and conduction band at $\mathbf{k}$ are separated by

[3] Generally, we only refer to a carrier as a carrier if the carrier–ideal crystal correlation is negligible. If the carrier–ideal crystal correlation is significant, we generally refer to the carrier and induced local polarization as a POLARON to emphasize the fact that its eigenvalues are not given by the bandstructure.

[4] Often called the CRYSTAL MOMENTUM, but $\mathbf{k}$ *by itself* is also called this at least as often, including in the vast majority of this thesis. This convention probably originates from unit systems where $\hbar = 1$. We also note that this is the only time in this thesis we use $\mathbf{p}$ to refer to this quasi-momentum. In the rest of this thesis, it will be used to refer to the Bloch wavevector associated with a secondary carrier.

an energy $\varepsilon_{\text{gap}}^{\text{direct}}(\mathbf{k})$, an electron previously stuck in the valence band can always ZENER TUNNEL to a state in the conduction band a distance $d = \frac{\varepsilon_{\text{gap}}^{\text{direct}}}{|\mathbf{F}|}$ downfield that will have the same total energy. This process will create two carriers: a conduction electron and a hole. However, unless the field is very large and the current very small, the rate at which interband transitions mediate the transport of carriers is usually negligible compared to the rate at which carriers flow due to other transport mechanisms. As such, we assume in this thesis that the rate of external force induced interband transitions is zero, and simply note that this approximation can be corrected for by adding a Zener term to the scattering operator if required.

It is shown in Appendix B that when the interband transition rate is set to zero, the uniform force has an astoundingly simple dynamic effect on a Bloch carrier state: the carrier wavevector will change in accordance with eq. (2.2). This equation is known as NEWTONS LAW FOR BLOCH STATES because the external force affects $\hbar\mathbf{k}$— the quasi-momentum associated with Bloch states— in exactly the same way that a real force affects real momentum:

$$\mathbf{F} = \frac{\partial \hbar \mathbf{k}}{\partial t}. \tag{2.2}$$

We assumed that in the region that a carrier is localized to, the reproducible device potential can be approximated as the superposition of an ideal crystal and a uniform external field. This localization implies that we assume a single carrier is a *coherent superposition* of Bloch waves, as opposed to a pure Bloch state which would be perfectly delocalized. In Appendix B, we demonstrate an important fact about the effect of an external force on coherent superpositions of Bloch waves. Namely, the effect of a uniform field on a coherent superposition of Bloch waves is *exactly the same as the effect of a uniform field on a classic statistical mixture of Bloch waves*. Since scattering is incoherent, we similarly have that the effect of scattering on a coherent superposition of Bloch waves is *exactly the same as the effect of scattering on a classic statistical mixture of Bloch waves*. These two facts enable us to model the evolution of any statistical collection of coherent Bloch carrier wavepackets by stripping that collection of its phase information and only tracking the movement of a *classical* distribution of fictitious "Bloch particles"— each of which have *both* an exact position *and* an exact Bloch wavevector— which has the same probability density as the actual carrier wavefunction in a coarse grained phase space.

The probability density associated with the statistical mixture of fictitious Bloch particles is expressed by the carrier's DISTRIBUTION FUNCTION, $f(\mathbf{k}\nu, \mathbf{r}, t)$. Since the carrier state is defined by a classical phase space distribution function over states with fairly classical dynamics except for idiosyncratic kinetic details defined by pure quantum mechanics via the bandstructure, this is referred to as a SEMICLASSICAL MODEL of a carrier.

Our derivation so far has made it appear that the semiclassical model requires an external force which is uniform over the localization length scale. However, we note that we have not actually ruled out the validity of semiclassical model in the more general condition where the carrier is localized to a non-uniform external force. The most intuitive way to explore this question is to reformulate Schrödinger's equation in terms of Wigner's quasiprobability distribution [27, 28, 29, 30].

The Wigner quasiprobability distribution is a representation of the density matrix on phase space. It has the attractive feature that the expectation value of any physical quantity is the same weighted integral that would determine that same physical quantity for a classical probability distribution. The only difference from an ordinary probability distribution is that the quasiprobability distribution is permitted to have values greater than one and less than zero. This does not cause impossible results with respect to experimentally verifiable questions because— as has been known since the dawn of quantum mechanics [31]— there is a limit to which we can experimentally localize particle density in phase space.

We can transform Schrödinger's equation into an equation for the evolution of the Wigner quasiprobability distribution. From this transformation, we find that the evolution of the quasiprobability distribution in an external potential $V(\mathbf{r})$ can *always* be modelled as the classical evolution of a simple *probability distribution* of fictitious particles *unless* there exists a $\lambda \in \{3, 5, \dots\}$ such that $\frac{\partial^\lambda V}{\partial \mathbf{r}^\lambda} \neq 0$ [32].[5]

It may be thought, then, that we can rigorously strengthen our claim of the validity of the semiclassical regime when the external potential takes a quadratic form over the region the carrier is localized to. However, we note that the derivation of Newton's law for Bloch states *does not involve perturbation theory*, and therefore the result holds for arbitrar-

[5]We note that if these derivatives are not defined, then they are not equal to zero.

ily strong fields, so long as one also includes a scattering term which models interband transitions. In the case of a non-uniform external field, there is no known analogous *non-perturbative separation of the periodic part of the potential from the non-periodic part*. As such, the result of the Wigner formulation can only be taken to suggest that, if a carrier is localized to an external potential that can be approximated by a quadratic, this external potential will be consistent with the semiclassical model *to a first-order approximation*.

So far, we have only really dealt with the dynamics of a *single* carrier, so we now consider the problem of many carriers. These carriers will interact with one another, and thus their positions will be correlated and not independent as implicitly assumed so far. As is becoming a habit, we begin by separating this interaction Hamiltonian into reproducible and non-reproducible parts. We consider the reproducible part to be associated with a coarse-grained, position-dependent density of carriers, and the non-reproducible part to be associated with the precise instantaneous position of the carriers. The reproducible bulk density of carriers is associated with long-range, reproducible carrier–carrier correlations that can be captured in the definition of the external field, while the precise instantaneous position of any given carriers results in short-range, non-reproducible carrier–carrier correlations that can be captured by the definition of an incoherent, carrier–carrier scattering rate.

In order for the reproducible carrier–carrier correlations to be captured by the external field, it is clear that the external field must be updated at a frequency significantly greater than the frequency of plasma oscillations [16, 6]. This only leaves open the question of an appropriate size for the coarse grain. It is shown by the rigorous analysis of jellium by Bohm and Pines [33] that the Debye screening length is a natural length scale for the coarse-grain. This is physically intuitive for two complimentary reasons, that are really just two ways of looking at the same physics. Firstly, the Debye screening length is the length-scale associated with the canonical example of a *reproducible* carrier–carrier density correlation: screening. Secondly, incoherent carrier–carrier scattering— associated with the *non-reproducible*, precise position of carriers— quickly vanishes at length scales larger than the Debye screening length *because of that screening*, and therefore correlations that occur above this length scale must not be due to the precise position of carriers.

The purpose of this section was to give a conceptual introduction to the semiclassical model of carriers. It is clear that the validity of the semiclassical model relies on the carriers being localized to regions over which the external field is reasonably uniform. To safely use the semiclassical model of carriers, we must justify this assumption in the devices we investigate in this thesis.

In this thesis we investigate non-equilibrium transport that is PHONON COLLISION DOMINATED. By phonon collision dominated, we mean that most scattering events are phonon scattering events, and that the mean free path for a carrier between phonon scattering events is much smaller than the characteristic length scale of the device. The crucial point about transport in this regime is that, while there *is* a phonon state associated with every crystal momentum, there *is not* a phonon state associated with every energy. In particular, there is no phonon that has a higher energy that the highest energy optical phonon state. The result is that in a device with a large field, scattering with phonons tends to relax the crystal momentum gained between collisions effectively, while not effectively relaxing the energy gained between collisions. The result is a distribution of carriers that has a high average energy, is highly non-equilibrium, but is fairly isotropic.

The primary localization mechanism for carriers in this regime is decoherence [34, 35] via phonon interaction. However, it is the authors understanding that the most important insight as to how this localization actually works in this regime has not been directly discussed before. For the completeness of this section, we will roughly describe it here. A phonon of wavevector $\mathbf{q}$, in band η, which has a frequency of $\omega_{\mathbf{q}\eta}$ can scatter with a carrier in eigenstate $|\mathbf{k}\nu\rangle$ if and only if there a carrier band ν' such that the phonon annihilation/creation conserves energy and crystal momentum; that is, such that $\varepsilon_{(\mathbf{k}\pm\mathbf{q})\nu'} - \varepsilon_{\mathbf{k}\nu} = \pm\hbar\omega_{\mathbf{q}\eta}$. As such, the set of phonons that *can* scatter with carriers is therefore a peculiar function of the crystal momentum/kinetic energy of the carrier, and so the phonon annihilated or created in a scattering event reveals a lot of information about the crystal momentum/kinetic energy of the carrier state. In a strong field, *the crystal momentum/kinetic energy of a carrier is strongly coupled to the position of the carrier*. Accordingly, the carrier can be expected to leak fairly precise information about its position to the phonon system once every mean free scattering time. This essentially means that the position of the carrier is expected to be *measured*, and therefore the carrier localized, with a precision roughly proportional to the field strength, every mean scattering

time.

As the field tends to infinity, localization will tend toward a delta function in position. However, such a carrier will— according the uncertainty principle— be a superposition of an infinite number of crystal momenta eigenstates. As such, in a mean free time, the wavepacket will delocalize into a wavepacket with a width roughly twice that of the mean free path. The *average* localization of the wavepacket in the infinite field limit is therefore a wavepacket with a width roughly halfway between perfect localization and this two mean free path wavepacket. As such, as the field tends to infinity, the expected localization of carriers will be approximately equal to the mean free path between phonon scattering events. We will call this the LARGE FIELD APPROXIMATION of localization in the phonon collision dominated regime.

In the large field approximation, the semiclassical model will hold so long as the mean free path is much larger than a primitive lattice cell, and the field is approximately uniform over the mean free path between phonons. Additionally, the semiclassical— as opposed to full quantum— treatment of *plasma oscillations* via frequent updates to the external potential will hold if the wavelength of plasma oscillations is *much greater* than the mean free path between phonon collisions, as this implies the plasma oscillations are largely incoherent. Thus it is crucial to know that in room temperature silicon, the mean free path between phonon collisions is on the order of $\sim 2 - 3$nm [12].

2.3 The Boltzmann–Poisson System

In the previous section, we argued that non-equilibrium, phonon collision dominated carrier transport can be modelled semiclassically. That is, the mechanics of carriers in a device can be modelled using a classical time-dependent distribution function $f(\mathbf{k}\nu, \mathbf{r}, t)$ on a semiclassical phase space characterized by crystal momentum $\mathbf{k}$, band index ν and position $\mathbf{r}$. The distribution function itself can be understood as the probability density associated with the existence of a given "Bloch particle": a fictitious, highly classical carrier particle that has a specific position, crystal momentum and band index.

The semiclassical model is an extremely powerful and intuitive model of carrier transport when it is valid. Its elements are as follows. The reproducible, periodic crystal potential leads to the bandstructure, which defines the kinematics of Bloch particles.[6] The reproducible, aperiodic external force leads to Newton's law for Bloch states, which states that the external force affects crystal momentum in exactly the same way that the net force affects momentum in classical mechanics. Finally the unreproducible carrier–carrier and carrier–device interactions lead to the definition of the rate at which Bloch particles abruptly scatter into states with different crystal momentum and band index, which is not unlike classical particle scattering.

The abstract, high-level origin of this beautiful simplification of quantum mechanics into semiclassical mechanics can be understood as follows. The *classical uncertainty* with respect to the carrier–carrier and carrier–device interaction Hamiltonians creates enough "wiggle-room" for us to be able to create this *fictional, nearly classical microscopic mechanics* that leads to the same predictions as the *experimentally verified, fully quantum microscopic mechanics*.

In this section, we will describe how to find a closed equation of motion for the semiclassical carrier distribution function. As is often the case in carrier transport theory, we begin by forming an *open* equation of motion for the carrier state using little more than a simple continuity argument.

Suppose that, at time t, we have a Bloch particle with wavevector/crystal momentum $\mathbf{k}$, band index ν, at a position $\mathbf{r}$. If we wait a time δt, *in the absence of scattering*, this Bloch particle:

- will have a Bloch wavevector $\mathbf{k} + \frac{\mathbf{F}}{\hbar}\delta t$ according to Newtons law for Bloch states,
- will be in the same band ν since interband transitions are always incorporated into scattering, and

[6] These kinematics are admittedly complex: the effective mass of the Bloch particle is neither scalar, nor constant, nor positive definite. So one must vigilantly fight the classical habit of slipping into thinking of velocity and (crystal) momentum as interchangeable quantities, or thinking there is a simple quadratic relation between these quantities and the kinetic energy.

- will be at a position $\mathbf{r} + \mathbf{v}\delta t$ by the definition of velocity.

Thus, *in the absence of scattering*, the probability of a Bloch particle at $(\mathbf{k}\nu, \mathbf{r}, t)$ would be the same as the probability of a Bloch particle at $\big((\mathbf{k} + \frac{\mathbf{F}}{\hbar}\delta t)\nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t\big)$. If the latter probability is smaller, there must be a net positive probability of diverting carriers *out of* the trajectory that leads toward $\big((\mathbf{k} + \frac{\mathbf{F}}{\hbar}\delta t)\nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t\big)$; whereas if the latter probability is larger, there must be a net positive probability of diverting carriers *into* the trajectory that leads toward $\big((\mathbf{k} + \frac{\mathbf{F}}{\hbar}\delta t)\nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t\big)$. To reiterate, if we follow a single Bloch particle trajectory, *in the absence of scattering*, we can only attribute a *single* probability to the *entire locus of time-dependent phase space points* that defines a Bloch particle trajectory.

This does not seem to imply that, *in the absence of scattering*, we can only attribute a single probability *density*, $f(\mathbf{k}\nu, \mathbf{r}, t)$, to the entire locus of phase space points that defines a trajectory. That is, it seems naïvely possible that Bloch particle trajectories might be such that they are more dense in some parts of phase space and less dense in others. However, it is well-known result known as Liouville's theorem that this compression and expansion of trajectories in phase space *does not* occur [36]. As such, we can apply the reasoning of the previous paragraph to probability *densities* also. Thus the reasoning of the previous paragraph leads to the following equation for the phase space probability density $f(\mathbf{k}\nu, \mathbf{r}, t)$, where $\left(\frac{\partial f(\mathbf{k}\nu, \mathbf{r}, t)}{\partial t}\right)_{\text{scat}}$ is the net density of Bloch particles per unit time that scatter into a small volume of points near $(\mathbf{k}\nu, \mathbf{r}, t)$:

$$f\Big(\big(\mathbf{k} + \tfrac{\mathbf{F}}{\hbar}\delta t\big)\,\nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t\Big) - f(\mathbf{k}\nu, \mathbf{r}, t) = \left(\frac{\partial f(\mathbf{k}\nu, \mathbf{r}, t)}{\partial t}\right)_{\text{scat}} \delta t. \tag{2.3}$$

We now perform explicitly a step-by-step Taylor expansion of the leftmost term of eq. (2.3), which can be rigorously truncated to first order as δt becomes infinitesimal:

$$
\begin{aligned}
& f\Big(\left(\mathbf{k} + \tfrac{\mathbf{F}}{\hbar}\delta t\right) \nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t \Big) \\
&= f(\mathbf{k}\nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t) + \frac{\mathbf{F}}{\hbar}\delta t \cdot \nabla_{\mathbf{k}} f(\mathbf{k}\nu, \mathbf{r} + \mathbf{v}\delta t, t + \delta t) + \mathcal{O}(\delta t^2) \\
&= f\left(\mathbf{k}\nu, \mathbf{r}, t + \delta t\right) + \frac{\mathbf{F}}{\hbar}\delta t \cdot \nabla_{\mathbf{k}} f(\mathbf{k}\nu, \mathbf{r}, t + \delta t) + \mathbf{v}\delta t \cdot \nabla_{\mathbf{r}} f\left(\mathbf{k}\nu, \mathbf{r}, t + \delta t\right) + \mathcal{O}(\delta t^2) \\
&= f(\mathbf{k}\nu, \mathbf{r}, t) + \mathbf{v}\delta t \cdot \nabla_{\mathbf{r}} f(\mathbf{k}\nu, \mathbf{r}, t) + \frac{\mathbf{F}}{\hbar}\delta t \cdot \nabla_{\mathbf{k}} f(\mathbf{k}\nu, \mathbf{r}, t) + \delta t \frac{\partial}{\partial t} f(\mathbf{k}\nu, \mathbf{r}, t) + \mathcal{O}(\delta t^2).
\end{aligned}
\tag{2.4}
$$

We substitute eq. (2.4) truncated to first order in δt into eq. (2.3) and divide by δt. This leads to the following expression:

$$
\frac{\partial}{\partial t} f(\mathbf{k}\nu, \mathbf{r}, t) + \mathbf{v} \cdot \nabla_{\mathbf{r}} f(\mathbf{k}\nu, \mathbf{r}, t) + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}\nu, \mathbf{r}, t) = \left(\frac{\partial f(\mathbf{k}\nu, \mathbf{r}, t)}{\partial t} \right)_{\text{scat}}. \tag{2.5}
$$

Eq. (2.5) is known as the BOLTZMANN TRANSPORT EQUATION. It is the general equation of motion for carriers in the semiclassical regime. When the Boltzmann Transport Equation is presented from here on out, we will generally suppress the domain of the distribution function, and therefore write it simply as follows:

$$
\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{r}} f + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f = \left(\frac{\partial f}{\partial t} \right)_{\text{scat}}. \tag{2.6}
$$

In order to close this equation, we need to find independent expressions for the external force $\mathbf{F}$, the velocity $\mathbf{v}$, and the scattering term $\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}$. We begin with the external force $\mathbf{F}$.

The external force is a Lorentz force, which is a function of the local electromagnetic field at a given time, the charge of the Bloch particle, and the velocity of the Bloch particle. Accordingly, the external force is generally a function of $(\mathbf{k}\nu, \mathbf{r}, t)$. However, this thesis is concerned exclusively with carrier transport in non-magenetic homogeneous semiconductor devices. Since the device is non-magnetic, we assume the external magnetic field is negligible. Additionally, since semiconductor transport typically involves relatively small current densities, we assume that the magnetic field induced by current flow is negligible. Since both the induced and external magnetic field are negligible,

the Lorentz force associated with the magnetic field is also negligible, and as such the Lorentz force *does not depend on the velocity* of the Bloch particle.

Furthermore, according to Maxwell's equations, the electric field has a curl-less component due to the instantaneous charge distribution, and a divergence-less component due to the rate of change of the magnetic field. Since the magnetic field is negligible, we assume the electric field $\mathbf{E}$ and— by extension— the external force $\mathbf{F}$ are curl-less, indicating the external force can be related to the gradient of the electric potential $\phi(\mathbf{r}, t)$:

$$\begin{aligned} \mathbf{F}\big(\text{sign}(\nu), \mathbf{r}, t\big) &= -\text{sign}(\nu) e \mathbf{E}(\mathbf{r}, t) && \text{by definition of Lorentz force,} \\ &= \text{sign}(\nu) e \nabla_{\mathbf{r}} \phi(\mathbf{r}, t) && \text{by definition of electric potential.} \end{aligned} \tag{2.7}$$

Here e is the fundamental charge, and we have used the convenient convention that the band index ν is a positive integer for conduction electrons and a negative integer for holes. The electric potential can be determined by POISSON'S EQUATION, where $\rho(\mathbf{r}, t)$ is the coarse-grained charge density:

$$\nabla_{\mathbf{r}}^2 \phi(\mathbf{r}, t) = -\frac{\rho(\mathbf{r}, t)}{\epsilon}. \tag{2.8}$$

Suppose we have a number of different doping fields, each indexed by an integer i, and each described by a position-dependent dopant density $N_{\text{dop}}^i(\mathbf{r})$. We suppose Z_{dop}^i is the number of electrons per dopant added to the conduction band, which is equivalent to the number of additional protons the dopant contributes to the local charge compared to the ideal crystal. We note that by this definition Z_{dop}^i is positive if the dopant is a donor, and negative if the dopant is an acceptor.

The full valence band balances the charge of the ideal crystal. Accordingly, the local charge density is proportional to the local net difference of added dopant proton density *and* hole density *from* subtracted dopant proton density *and* conduction electron density:

$$\frac{\rho(\mathbf{r}, t)}{e} = p(\mathbf{r}, t) - n(\mathbf{r}, t) + \sum_i N_{\text{dop}}^i(\mathbf{r}) Z_{\text{dop}}^i. \tag{2.9}$$

Here the density of holes $p(\mathbf{r}, t)$ and conduction electrons $n(\mathbf{r}, t)$ is defined by integrating the distribution function over the Brillouin zone for all negative bands and all positive

bands respectively, and multiplying by Γ— the density of carrier states per unit volume of $\mathbf{k}$-space, per band[7]:

$$p(\mathbf{r},t) = \Gamma \sum_{\nu<0} \int_{\mathrm{BZ}} f(\mathbf{k}\nu,\mathbf{r},t)\mathrm{d}\mathbf{k}, \tag{2.10a}$$

$$n(\mathbf{r},t) = \Gamma \sum_{\nu>0} \int_{\mathrm{BZ}} f(\mathbf{k}\nu,\mathbf{r},t)\mathrm{d}\mathbf{k}. \tag{2.10b}$$

Through the set of prior equations, we have related the external force to the distribution function f. Next, we wish to find an expression for the velocity. We simply note that in this thesis we are concerned with devices made from a homogeneous material, which implies the local bandstructure is independent of position $\mathbf{r}$. As such, the velocity of the distribution function at crystal wavevector $\mathbf{k}$ is simply given by eq. (2.1), and is position independent:

$$\mathbf{v}(\mathbf{k}\nu) = \frac{1}{\hbar}\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}. \tag{2.11}$$

Finally, we note that at this point we will not simplify the problem of relating the scattering term to the distribution function any further. Instead, we now combine eq. (2.6) – eq. (2.11), in order to define the tightly coupled pair of equations known as the BOLTZMANN–POISSON SYSTEM:

$$\boxed{\begin{aligned} \frac{\partial f}{\partial t} &= \left(\frac{\partial f}{\partial t}\right)_{\text{scat}} - \frac{1}{\hbar}\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}\cdot\nabla_{\mathbf{r}}f - \frac{\mathrm{sign}(\nu)e}{\hbar}\nabla_{\mathbf{r}}\phi\cdot\nabla_{\mathbf{k}}f \qquad (2.12a)\\ \nabla^2_{\mathbf{r}}\phi &= \frac{e}{\epsilon}\left(\overbrace{\Gamma\sum_{\nu}\mathrm{sign}(\nu)\int_{\mathrm{BZ}} f\mathrm{d}\mathbf{k}}^{n-p} - \sum_i N^i_{\text{dop}}Z^i_{\text{dop}}\right) \qquad (2.12b) \end{aligned}}$$

In the Boltzmann–Poisson system, the Boltzmann transport equation (2.12a) is the equation of motion for the distribution function f in terms of the external electric potential ϕ, and the Poisson equation (2.12b) is the equation of motion for the external electric potential ϕ in terms of the distribution function f. We note that the Boltzmann–Poisson system is closed if and only if we can find expressions for the bandstructure $\varepsilon_{\mathbf{k}\nu}$ and the net scattering rate term $\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}$ in terms of distribution function f and/or the external

[7]For spin-degenerate bands of silicon, $\Gamma = \frac{1}{4\pi^3}$

electric potential ϕ.

In conventional, near-lattice-temperature-equilibrium carrier transport, it is trivial to close the Boltzmann–Poisson system with analytic expressions for the bandstructure and scattering rate term [26]. We outline one robust approach here. The scattering rate term is given by the LATTICE TEMPERATURE EQUILIBRIUM RELAXATION TIME APPROXIMATION. In this relaxation time approximation, we presume there exists a local equilibrium distribution for each band characterized by a band-dependent quasifermi level[8] $\varepsilon_\nu^F(\mathbf{r},t)$ and the lattice temperature T_L:

$$f_{\text{equilibrium}}(\mathbf{k}\nu,\mathbf{r},t) = \frac{1}{e^{\frac{\varepsilon_{\mathbf{k}\nu}-\varepsilon_\nu^F(\mathbf{r},t)}{kT_L}}+1}. \tag{2.13}$$

Here k is Boltzmann's constant, and the size of the quasifermi level is determined by the conservation of local carrier density. We then assume that perturbations from the local equilibrium distribution exponentially decay in time, with the decay constant conventionally represented by a equilibrium relaxation time $\tau_{\text{relax}}^{\text{equilibrium}}$:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}} = -\frac{f(\mathbf{k}\nu,\mathbf{r},t) - f_{\text{equilibrium}}(\mathbf{k}\nu,\mathbf{r},t)}{\tau_{\text{relax}}^{\text{equilibrium}}}. \tag{2.14}$$

For the near-lattice-temperature-equilibrium bandstructure, we proceed as follows. First, we note the number of valleys within $3kT_L$ of the conduction band minima or the valence band maxima. Second, we change the band index convention so that only the valleys noted are associated with a unique band index ν. Finally, we apply a band-index-dependent crystal momentum translation to both sides of the Boltzmann transport equation, such that all the valleys noted now occur at the gamma point of the transformed domain. It is clear from inspection that such a translation does not effect the LHS or RHS of the Boltzmann transport equation in the relaxation time approximation except via the transformed bandstructure. Thus in the relaxation time approximation, we can use a fictitious "centred" bandstructure.

Finally, we can approximate each of valleys in the centred bandstructure as being parabolic in $\mathbf{k}$ over the range of small thermal range of energy where carriers are concentrated:

[8] Typically, the quasifermi level is assumed to be single-valued for all hole bands, and single-valued for all conduction electron bands.

$$\varepsilon_{\mathbf{k}\nu} = \varepsilon^0_\nu + \frac{1}{2}\hbar^2(m_\nu^{-1}\mathbf{k})\cdot\mathbf{k}. \tag{2.15}$$

Here m_ν^{-1} is the inverse effective mass tensor at the kinetic energy minima of the ν band. We note that, if desired, we can define a small non-parabolic correction by mulitplying the RHS by $\left(1+\alpha\varepsilon_{\mathbf{k}\nu}\right)$, where α is a non-parabolicity factor [9].

While closing the Boltzmann-Poisson system in the near-lattice-temperature-equilibrium regime is straightforward, in the far-from-thermal-equilibrium regime relevant to this thesis, there is no similarly simple analytic closure that is theoretically sound. Instead, considerable effort is required to define the full bandstructure and the full scattering rate term [16]. Thus, we now turn our attention to defining the full bandstructure and a full scattering rate term.

2.4 Generating the Full Bandstructure

In far-from-lattice-temperature-equilibrium transport, it is typical for the average kinetic energy of carriers to be on the order of an electron volt, rather than on the order of the lattice thermal energy $kT_L \ll 1\text{eV}$. As a result, carriers are spread throughout a wide range of energies in the conduction and/or valence bands— and correspondingly spread widely throughout the entire Brillouin zone— instead of being concentrated around a few kinetic energy minima. As such for many materials— including silicon— we cannot approximate the bandstructure as parabolic or near parabolic, and must instead incorporate the full bandstructure into our carrier transport model if the model is to be theoretically sound [6]. In this section, we will therefore give an overview of the theory required to generate a reliable estimate of the full $3-$D bandstructure.

A full bandstructure calculation generates the expected energy eigenvalues associated with carrier eigenstates of fixed crystal wavevector. This calculation is difficult to do from first principles, as there are an enormous number of non-negligible effects one must take into account. Since bandstructure calculation is not the focus of this work,

we would ideally like to be able to simply assume the bandstructure is a measured experimental quantity. However, since the bandstructure defined by M $3-$D scalar fields, where M is the number of bands of interest, this is not the most pragmatic approach. Instead, we will use a simple phenomenological approach known as the EMPIRICAL PSEUDOPOTENTIAL METHOD. This method enables us to generate a full bandstructure from a small set of phenomenological constants, which can be determined implicitly from empirical optical absorption data.

We begin by expanding a valence electron Bloch eigenstate $\psi_{\mathbf{k}\nu}$ into two parts: a valence pseudoeigenstate $\varphi^{\text{pseudo}}_{\mathbf{k}\nu}$ and a core orthogonalization term. The core orthogonalization term is equal to the projection of the pseudoeigenstate onto the core Bloch eigenstates $\psi_{\mathbf{k}\nu^{\text{core}}}$ multiplied by -1:

$$\psi_{\mathbf{k}\nu} = \varphi^{\text{pseudo}}_{\mathbf{k}\nu} - \sum_{\nu^{\text{core}}} \left\langle \psi_{\mathbf{k}\nu^{\text{core}}} | \varphi^{\text{pseudo}}_{\mathbf{k}\nu} \right\rangle \psi_{\mathbf{k}\nu^{\text{core}}}. \tag{2.16}$$

The point of the orthogonalization term is to ensure that the real valence electron Bloch eigenstate $\psi_{\mathbf{k}\nu}$ is automatically orthogonal to the core states *no matter how non-orthogonal the pseudoeigenstate $\varphi^{pseudo}_{\mathbf{k}\nu}$ might be to the core states*:

$$\begin{aligned} \langle \psi_{\mathbf{k}\nu} | \psi_{\mathbf{k}\mu^{\text{core}}} \rangle &= \left\langle \varphi^{\text{pseudo}}_{\mathbf{k}\nu} | \psi_{\mathbf{k}\mu^{\text{core}}} \right\rangle - \sum_{\nu^{\text{core}}} \delta_{\nu^{\text{core}}\mu^{\text{core}}} \left\langle \varphi^{\text{pseudo}}_{\mathbf{k}\nu} | \psi_{\mathbf{k}\nu^{\text{core}}} \right\rangle \\ &= 0. \end{aligned} \tag{2.17}$$

The valence electron pseudoeigenstate is therefore a valence electron eigenstate with *an arbitrary amount of the core electron eigenstates mixed in*. The point of this is as follows. There is a *minimum* rate of spatial variation for the real valence electron Bloch eigenstate, enforced by the fact that the valence electron eigenstate must vary fast enough to be orthogonal with all the core electron Bloch eigenstate at the same crystal momentum. There is, however, no such minimum rate of spatial variation for the *pseudoeigenstate* of a valence electron. This raises the possibility that the pseudoeigenstate for a valence electron might be able to be described in terms of a much smaller basis of plane waves than the real Bloch eigenstate of a valence electron. A smaller set of basis functions implies that such eigenstates will be less computationally demanding to find.

We would like then to derive a pseudo-Hamiltonian $\hat{H}_{\text{pseudo}}$ that has the pseudoeigen-

states as its eigenstates, and that has the *same energy eigenvalues* as the real valence electron Bloch eigenstates. We begin by expanding the expression for the real Hamiltonian operating on the real valence electron Bloch eigenstate in terms of the pseudoeigenstate:

$$\hat{H}\psi_{\mathbf{k}\nu} = \hat{H}\varphi^{\text{pseudo}}_{\mathbf{k}\nu} - \sum_{\nu^{\text{core}}} \left\langle \psi_{\mathbf{k}\nu^{\text{core}}} | \varphi^{\text{pseudo}}_{\mathbf{k}\nu} \right\rangle \hat{H}\psi_{\mathbf{k}\nu^{\text{core}}}. \tag{2.18}$$

We can now substitute in the energy eigenvalues for eigenstates of the real Hamiltonian, and eliminate the dependence of eq. (2.18) on the valence electron eigenstates $\psi_{\mathbf{k}\nu}$. The core electron energy eigenvalues are written as $\varepsilon_{\mathbf{k}\nu^{\text{core}}}$ while the valence electron eigenvalues are written as $\varepsilon_{\mathbf{k}\nu}$:

$$\varepsilon_{\mathbf{k}\nu}\left(\varphi^{\text{pseudo}}_{\mathbf{k}\nu} - \sum_{\nu^{\text{core}}} \left\langle \psi_{\mathbf{k}\nu^{\text{core}}} | \varphi^{\text{pseudo}}_{\mathbf{k}\nu} \right\rangle \psi_{\mathbf{k}\nu^{\text{core}}}\right) = \hat{H}\varphi^{\text{pseudo}}_{\mathbf{k}\nu} - \sum_{\nu^{\text{core}}} \varepsilon_{\mathbf{k}\nu^{\text{core}}} \left\langle \psi_{\mathbf{k}\nu^{\text{core}}} | \varphi^{\text{pseudo}}_{\mathbf{k}\nu} \right\rangle \psi_{\mathbf{k}\nu^{\text{core}}}. \tag{2.19}$$

If we simply rearrange so that the energy eigenvalue of the state $\psi_{\mathbf{k}\nu}$ and the pseudoeigenstate are on the LHS, then the RHS must automatically define the desired pseudo-Hamiltonian $\hat{H}_{\text{pseudo}}$:

$$\begin{aligned}\varepsilon_{\mathbf{k}\nu}\varphi^{\text{pseudo}}_{\mathbf{k}\nu} &= \hat{H}\varphi^{\text{pseudo}}_{\mathbf{k}\nu} + \sum_{\nu^{\text{core}}} (\varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathbf{k}\nu^{\text{core}}}) \left\langle \psi_{\mathbf{k}\nu^{\text{core}}} | \varphi^{\text{pseudo}}_{\mathbf{k}\nu} \right\rangle \psi_{\mathbf{k}\nu^{\text{core}}} \\ &= \hat{H}_{\text{pseudo}}\varphi^{\text{pseudo}}_{\mathbf{k}\nu}.\end{aligned} \tag{2.20}$$

This means that the pseudo-Hamiltonian is the real Hamiltonian plus a repulsive artificial potential term that *adds the difference in energy* between the valence electron eigenstate and the core electron eigenstate, weighted by the overlap of these functions:

$$\hat{H}_{\text{pseudo}} = \hat{H} + \hat{V}_{\text{artificial}}, \tag{2.21a}$$

$$\text{where} \qquad \hat{V}_{\text{artificial}} = \sum_{\nu^{\text{core}}} \frac{(\varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathbf{k}\nu^{\text{core}}}) \left\langle \psi_{\mathbf{k}\nu^{\text{core}}} | \varphi^{\text{pseudo}}_{\mathbf{k}\nu} \right\rangle |\psi_{\mathbf{k}\nu^{\text{core}}}\rangle}{\left|\varphi^{\text{pseudo}}_{\mathbf{k}\nu}\right\rangle}$$

$$= \sum_{\nu^{\text{core}}} (\varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathbf{k}\nu^{\text{core}}}) \left|\psi_{\mathbf{k}\nu^{\text{core}}}\right\rangle\!\left\langle\psi_{\mathbf{k}\nu^{\text{core}}}\right|. \tag{2.21b}$$

This can be understood as follows. The pseudoeigenstate is nothing more than a partic-

ular superposition of the core electron eigenstates and the valence electron eigenstate. The energy expectation of the pseudoeigenstate according to the actual Hamiltonian will be the weighted average of the valence and core electron eigenstate energies. Since we want the pseudo-Hamiltonian to assign the pseudowavefunction the *same* expected energy as the valence eigenstate, we need to "make up" the deficit of the energy by adding the repulsive artificial potential $\hat{V}_{\text{artificial}}$ the real Hamiltonian, which grows the greater the difference between core and valence energies, and grows the greater the amount of core state that is incorporated into the pseudowavefunction.

This notion of an artificial potential was originally proposed by Phillips and Kleinman [37]. There are three critical points to note about the artificial potential, and the pseudo-Hamiltonian it is associated with.

Firstly, the artificial potential— and by extension the pseudo-Hamiltonian— is only diagonal in a basis containing the core eigenstates if we associate the artificial potential with a single distinct energy eigenvalue at $\mathbf{k}$. If we associate more than one energy eigenvalue with $\mathbf{k}$, then the artificial potential is no longer diagonal in a basis containing the core eigenstates.

Secondly, if we do associate the pseudo-Hamiltonian with a single distinct eigenvalue at $\mathbf{k}$, then the pseudo-Hamiltonian associates an infinite number of degenerate pseudoeigenstates with that single eigenvalue $\varepsilon_{\mathbf{k}\nu}$. More precisely, if there is N core electrons per primitive cell, then there is N degrees of freedom associated with the pseudoeigenstate, as the pseudoeigenstates can possess an arbitrary amount of each of the N core electron eigenstates.

Thirdly, if we do associate the pseudo-Hamiltonian with a single distinct eigenvalue at $\mathbf{k}$, then it is inherently non-local. Thus, we cannot associate the artificial potential with a *universal* position-dependant artificial potential. We can only express the artificial potential as a position dependent function for a given pseudoeigenstate $\varphi_{\mathbf{k}\nu}^{\text{pseudo}}$ that has a well-defined amount of each of the N core electron eigenstates:

$$V_{\text{artificial}}^{\varphi_{\mathbf{k}\nu}^{\text{pseudo}}}(\mathbf{r}) = \frac{\left\langle \mathbf{r} \left| \hat{V}_{\text{artificial}} \right| \varphi_{\mathbf{k}\nu}^{\text{pseudo}} \right\rangle}{\left\langle \mathbf{r} | \varphi_{\mathbf{k}\nu}^{\text{pseudo}} \right\rangle}. \tag{2.22}$$

The critical conclusion is this. There is a N dimension space of pseudoeigenstates that can be associated with *any single* energy eigenvalue in the bandstructure, and each pseudoeigenstate in this space will be associated with a *different* position-dependent artificial potential $V^{\varphi^{\text{pseudo}}_{\mathbf{k}\nu}}_{\text{artificial}}(\mathbf{r})$. The position-dependent artificial potentials *will always be near zero in the space where there is very little core electron density*, and will be a pseudoeigenstate-dependent positive function in the space where there is substantial core electron density.

We can write the real Hamiltonian in terms of the kinetic energy $\hat{T}$ and the "natural" potential $\hat{V}_{\text{natural}}$. This natural potential that the valence Bloch electrons experience is dominated by the ionic cores. The *net* potential the pseudoeigenstate experiences is thus the sum of the natural potential and the artificial potential. We call the net potential the PSEUDOPOTENTIAL:

$$\hat{H}_{\text{pseudo}} = \hat{T} + \hat{V}_{\text{natural}} + \hat{V}_{\text{artificial}}, \tag{2.23a}$$

$$= \hat{T} + \hat{V}_{\text{pseudo}}. \tag{2.23b}$$

We note that the natural potential is most quickly varying near the nuclei, exactly where there is also significant core electron density. Since the space of possible artificial potentials is so large when there is significant core electron density, we expect that there exists some pseudoeigenstates for which this quickly varying natural potential is *essentially cancelled* by the quickly varying artificial potential. We expect this to be the case *regardless of the energy eigenvalue* the pseudoeigenstate is associated with. We just need more core state wavefunction in the pseudoeigenstates associated with smaller energy eigenvalues, and less core state wavefunction in the pseudoeigenstate associated with larger energy eigenvalues.

Inside the enormous space of all possible pseudoeigenstates, we expect that there exists a tiny subset of pseudoeigenstates— one pseudoeigenstate associated with each eigenvalue— such that all pseudoeigenstates across the entire bandstructure are subject to roughly *the same slowly-varying, position-dependent pseudopotential*: a pseudopotential which only matches the natural potential in the space outside the core states. In fact, we expect there to be *many* such subsets of pseudoeigenstates, where each subset is associated with a particular, idiosyncratically shaped pseudopotential inside the core region. This allows us to define a MODEL HAMILTONIAN where we replace the exact pseudopo-

tential operator $\hat{V}_{\text{pseudo}}$ in the pseudo-Hamiltonian, with a *universal*, slowly-varying, localized pseudopotential $V_{\text{pseudo}}(\mathbf{r})$. The solution of the time-independent Schrödinger equation for this model Hamiltonian will generate this very special collection of pseudoeigenstates, and by extension, generate the eigenvalues for the entire bandstructure.

What we have outlined is an *approximate existence theorem* that was first noted by Cohen and Heine [38]. Stated simply, we can expect there exists a universal local pseudopotential $V_{\text{pseudo}}(\mathbf{r})$ that is much more slowly-varying than the natural potential is inside the core region, and that approximately reproduces the eigenvalues of the entire valence electron bandstructure. While many existence theorems are futile, this particular existence theorem leads directly to two very useful conclusions.

The first useful conclusion is that the pseudopotential $V_{\text{pseudo}}(\mathbf{r})$ should be easy to characterize using empirical data. We begin by firstly noting that the universal local pseudopotential must respect all crystal symmetry. Thus a Fourier series expansion of the pseudopotential will only require reciprocal lattice vector components, and furthermore many of these components will be forced by the point symmetries of the crystal to be equal to other components or to be zero. Secondly, a *slowly-varying* universal local pseudopotential can be fully described by an reciprocal lattice vectors that have a relatively small magnitude. The net result is that universal local pseudopotential can be characterized by an extremely small number of Fourier coefficients. Accordingly, the universal local pseudopotential should be able to be determined from a small number of pieces of experimental data about the bandstructure. The result is known as a LOCAL EMPIRICAL PSEUDOPOTENTIAL. The local empirical pseudopotential we use for silicon is that of Chelikowsky and Cohen [39], and is characterized by only *three* distinct Fourier components.

The second useful conclusion is that the resulting model Hamiltonian is easy to solve. We are generally only interested in the lowest energy bands for valence electrons, since these are the bands that carriers with several eV of kinetic energy will occupy. Since the magnitude of kinetic energy increases with increasing spatial variation in the wavefunction, the lowest energy bands will be associated with pseudoeigenstates that do not vary much more quickly than the pseudopotential. As such, the time-independent Schrödinger equation associated with these pseudoeigenstates can be solved using a

basis of plane waves that is only slightly larger than the basis used to describe the pseudopotential.

The argument of Cohen and Heine effectively shows that we can expect a universal local pseudopotential to exist, that is easy to characterize empirically, and that generates a robust estimate of the bandstructure without requiring significant computational power. The empirical pseudopotential method is thus very appropriate for our purposes, and the details of the specific calculation done to generate the bandstructure is discussed in Appendix C. The basic conclusion is that the non-core bandstructure $\varepsilon_{\mathbf{k}\nu}$ can be generated by solving the following time-independent Schrödinger equation:

$$\varepsilon_{\mathbf{k}\nu}\varphi^{\text{pseudo}}_{\mathbf{k}\nu}(\mathbf{r}) = \left(-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}} + \sum_{\mathbf{G}} \cos(\mathbf{G}\cdot\boldsymbol{\tau})\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G})e^{i\mathbf{G}\cdot\mathbf{r}}\right)\varphi^{\text{pseudo}}_{\mathbf{k}\nu}(\mathbf{r}), \quad \text{where}$$

$$\boldsymbol{\tau} = \frac{a}{8}(1,1,1), \quad \text{and}$$

$$\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}) = \begin{cases} -0.2241\,\text{Rydbergs} & \text{for } |\mathbf{G}|^2 = 3 \\ 0.0551\,\text{Rydbergs} & \text{for } |\mathbf{G}|^2 = 8 \\ 0.0724\,\text{Rydbergs} & \text{for } |\mathbf{G}|^2 = 11 \\ 0 & \text{otherwise.} \end{cases} \tag{2.24}$$

2.5 Generating the Full Scattering Operator

2.5.1 The Breakdown of the Relaxation Time Approximation

We have discussed the coherent quantum mechanical evolution of a carrier due to the *reproducible* sub-Hamiltonian in the device, which is associated with Newton's law for Bloch states and the bandstructure. We now turn to discuss the incoherent quantum mechanical evolution due to the *non-reproducible* sub-Hamiltonian in the device, which is associated with the scattering term $\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}$.

Near thermal equilibrium,[9] we can use the relaxation time approximation:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}} = -\frac{f - f_{\text{equilibrium}}}{\tau_{\text{relax}}^{\text{equilibrium}}} = -\frac{f}{\tau_{\text{relax}}^{\text{equilibrium}}} + \frac{f_{\text{equilibrium}}}{\tau_{\text{relax}}^{\text{equilibrium}}}. \tag{2.25}$$

The relaxation time approximation can be seen as the generic linear response of a stable system to a perturbation from a stable point. It is therefore almost always valid for small perturbations from thermal equilibrium [40, 41]. However, in eq. (2.25) we have presented the relaxation time approximation in a form that makes the crudeness of this assumption for large perturbations explicit. The relaxation time approximation is equivalent to a scattering operator that simply annihilates carriers in the actual distribution f in a mean free time $\tau_{\text{relax}}^{\text{equilibrium}}$, and generates carriers in an equilibrium distribution $f_{\text{equilibrium}}$ in the same mean free time. As such, the relaxation time approximation is incapable of describing any process where scattering transforms the actual distribution function into an *intermediate distribution* that is distinct from the equilibrium distribution. Accordingly, in order for the relaxation time approximation to be theoretically sound, scattering must be such that it drives carriers from the actual distribution f into the equilibrium distribution $f_{\text{equilibrium}}$ in a single scattering event.

If we examine the distribution function at a position inside a non-equilibrium transport device, we will typically find that the carriers require approximately *ten* scattering events in order to become thermalized with the lattice [42]. It is also typical in these devices that there are positions where carrier–carrier scattering is not sufficiently strong to ensure the carrier distribution is always only a small perturbation from an internal thermal equilibrium [6]. Carriers are not, therefore, a single scattering event from being in equilibrium with the lattice or being in equilibrium with one another. In fact,

[9] Note that being near *thermal* equilibrium is a more general condition than being near *lattice temperature* equilibrium. The thermal equilibrium relaxation time approximation is thus more generally applicable in this regard than the parabolic or near-parabolic bandstructure approximation. For the thermal equilibrium relaxation time approximation the average energy of the distribution can be large— i.e. on the order of $\sim$ 1eV— so long as the shape of the distribution is approximately thermal; in contrast, for many materials— including silicon— the parabolic or near-parabolic approximation of the bandstructure actually requires the vast majority of distribution function density is associated with kinetic energies that are much smaller than 1eV. On the other hand, the parabolic or near-parabolic approximation of the bandstructure is more general in another regard: within this low kinetic energy constraint, the distribution function can be any shape.

from our perspective this characteristic *defines* the far-from-thermal-equilibrium regime that is the focus of this thesis. Using the relaxation time approximation to describe the manner in which scattering drives the carriers toward thermal equilibrium is therefore unphysical *by definition* in the far-from-thermal-equilibrium regime, and we must use a more detailed expression for scattering term.

This section is devoted to deriving an expression for the scattering operator in far-from-thermal-equilibrium transport in silicon that is well-accepted as being reliable and fairly accurate. The scattering operator we will use as an example in this thesis is roughly that which was implicitly proposed in 1988 by Fischetti and Laux [16], which itself was used as the original basis for the IBM DAMOCLES Monte Carlo simulation program [20]— the gold-standard semiclassical transport model [17]. The derivation given here is essentially this authors own attempt to clarify and make explicit Fischetti and Laux's scattering model, so little of this derivation and the actual scattering operator derived can be found in Fischetti and Laux's paper. Instead, this derivation more closely follows the work Jacoboni and Reggiani [43] and Ziman [44].

We note that in 1995 Fischetti, Laux and Crabbé wrote an update to Fischetti and Laux's 1988 paper [6]. In this update, several adjustments were made to the original 1988 scattering operator in order to increase its accuracy. We will largely ignore these adjustments as they add considerable complexity to what is already fairly complex discussion, and they are peripheral to the central point of this thesis. The central point of this thesis is that the full Boltzmann transport equation can be quite accurately mapped to a macroscopic model that is much quicker to solve at run time. Demonstrating the existence of this mapping does not require that all features of carrier scattering are perfectly described. It simply requires that the scattering operator we use is *sufficiently detailed to act as a reasonable representative of a class of fairly general scattering operator that can appear in the full Boltzmann transport equation*. We have made the judgement that the scattering operator proposed in the original 1988 DAMOCLES model is sufficiently detailed for this purpose, and thus only make some very minor adjustments to it that we believe increase accuracy without adding additional complexity.

2.5.2 General Scattering Theory

We will begin our discussion of carrier scattering in very general terms. We note first that the non-kinetic part of the carrier Hamiltonian is associated with *forces* on the carrier, and forces act between two bodies and act equally on both bodies. Thus, when we speak of the non-kinetic part of carrier Hamiltonian we are speaking of a collection of forces, where each force is between the carrier and another body. We will refer to the other body involved in an interaction with a carrier as the PARTNER body of the interaction Hamiltonian. A partner body is, in general, a useful abstraction rather than a fundamental entity. In the devices we investigate, we will express the non-reproducible part of the carrier Hamiltonian as a collection of interactions between the carrier and the following partner bodies:

- a dopant ion,
- another carrier,
- a valence band electron,[10] or
- a phonon mode.

We are only able to ignore the effect of the carrier–partner interaction on the partner state if the effective mass of the carrier is negligible compared to the effective mass of the partner body. In the list of partner bodies given, the only partner body that has a large enough effective mass that its state change can be ignored is the dopant ion. As such, the carrier–partner interactions we explore in this section will typically cause both interacting bodies to change state appreciably.

According to Fermi's golden rule [44, 43]— the canonical equation of first-order incoherent perturbation theory— the rate of transition from an initial state to a final state of a system is non-zero if and only if the zeroth-order total energy of initial state and the final state is the same.[11] Fermi's golden rule does not simply apply to the initial and

[10] The partner body associated with impact ionization is a valence band electron.

[11] By zeroth order total energy, we mean the total energy attributed to the system by the Hamiltonian

final state of the *carrier subsystem*— which would imply inelastic carrier scattering is impossible— Fermi's golden rule applies to the initial and final state of the *total system*. If we express the non-reproducible Hamiltonian as a sum of carrier–partner interaction Hamiltonians $\hat{H}_{\text{car–par}}$ for all possible partners, then we can characterize the initial and final total state by the initial and final carrier–partner state since the carrier–partner interaction will only cause state changes between the carrier and the partner. Fermi's golden rule can then be expressed as follows:

$$S_{\text{par}}(\mathbf{k}\nu, \mathfrak{s}_{\text{par}}; \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}) = \frac{2\pi}{\hbar}\left|\left\langle \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}\middle|\hat{H}_{\text{car–par}}\middle|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle\right|^2 \times\delta\Big(\varepsilon(\mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}) - \varepsilon(\mathbf{k}\nu, \mathfrak{s}_{\text{par}})\Big). \tag{2.26}$$

Here $S(\mathbf{k}\nu, \mathfrak{s}_{\text{par}}; \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}})$ is the probability per unit time that an initial carrier state $|\mathbf{k}\nu\rangle$ and initial partner state $|\mathfrak{s}_{\text{par}}\rangle$ becomes a final carrier state $|\mathbf{k}'\nu'\rangle$ and a final partner state $|\mathfrak{s}'_{\text{par}}\rangle$, and $\varepsilon(\mathbf{k}\nu, \mathfrak{s}_{\text{par}})$ is the first-order energy associated with the initial total system. It is common to refer to $\left\langle \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}\middle|\hat{H}_{\text{car–par}}\middle|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle$ simply as the MATRIX ELEMENT for the scattering transition.

We note that we are not so much interested in the carrier–partner scattering operator in and of itself, but in the scattering term it generates in the Boltzmann transport equation. These two quantities are connected by the single-carrier scattering operator $S(\mathbf{k}\nu; \mathbf{k}'\nu')$, which defines the rate at which occupation probability is transferred from $|\mathbf{k}\nu\rangle$ to empty states at $|\mathbf{k}'\nu'\rangle$. The single-carrier scattering operator is determined by taking the sum (or integral) of the carrier–partner scattering $S(\mathbf{k}\nu, \mathfrak{s}_{\text{par}}; \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}})$ over all *occupied* initial and *available* final partner states, for every partner body present in the device:

$$\left(\frac{\partial f(\mathbf{k}\nu, \mathbf{r}, t)}{\partial t}\right)_{\text{scat}} = \sum_{\nu'}\int_{\text{BZ}} S(\mathbf{k}\nu; \mathbf{k}'\nu') f(\mathbf{k}\nu, \mathbf{r}, t)\big(1 - f(\mathbf{k}'\nu', \mathbf{r}, t)\big)\mathrm{d}\mathbf{k}', \tag{2.27a}$$

$$S(\mathbf{k}\nu; \mathbf{k}'\nu') = \sum_{\mathfrak{s}_{\text{par}}}^{\text{occupied}} \sum_{\mathfrak{s}'_{\text{par}}}^{\text{available}} S_{\text{par}}(\mathbf{k}\nu, \mathfrak{s}_{\text{par}}; \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}). \tag{2.27b}$$

While we derive Fermi's golden rule in detail in Appendix D, it is so central to scattering theory that it is worth spending a few paragraphs here to describe the essential

that does not contain interaction terms.

physics behind it. We can express a given carrier–partner interaction Hamiltonian as a *sum* of single-matrix-element Hamiltonians. Since Fermi's golden rule is a one-to-one mapping of each individual matrix element of $\hat{H}_{\text{car–par}}$ to an individual transition, we can consider Fermi's golden rule to deal with a perturbation due to a particular single-matrix-element Hamiltonian. We also abstract out all the details of the initial and final total state, and consider the initial and final *total* state simply as names for two states in a generic two-state system.

We begin by turning off the interaction Hamiltonian, and we will assume that in this case the energy eigenvalues of initial and final *total* states are the same. Accordingly, any superposition of initial and final total state will define a eigenstate with the same energy eigenvalue. If we now turn on the interaction Hamiltonian, this space of superpositions is no longer a space of degenerate eigenstates. Rather, there will be only two eigenstates in the new system, where one is an equally weighted superposition of initial and final total state with a particular relative phase, and the other an equally weighted superposition of initial and final total state with the opposite phase. The eigenvalues of the new eigenstates will be greater and less than the the zeroth order eigenvalue by the modulus of the matrix element.

Notice then, that when the interaction is turned on, *the initial total state of the carrier–partner system must now be an equal superposition of the two new total eigenstates*. That is, the initial state is an equal combination of the two new total eigenstates that has a *particular relative phase*. The final state is also an equal combination of the new eigenstates, except that it has the *opposite* relative phase. Since the new total eigenstates have slightly different eigenvalues, *a superposition of total eigenstates will not have a constant relative phase*. Thus the initial state of the carrier–partner system must be a *transient* phenomena, and the probability amplitude given purely to the initial state at time zero will be transferred to the final state at some rate. This rate at which *probability amplitude* is transferred to the final state is proportional to the rate at which relative phase changes. The rate at which probability amplitude is transferred to the final state is therefore proportional to the "beat frequency" or difference in eigenvalues of the new eigenstates, and thus proportional to the modulus of the single matrix element in the single-matrix-element interaction Hamiltonian.

For convenience, the details of the discussion so far have relied on the fact that the zeroth-order energy eigenstates are the same.[12] But it is only now that the equivalence of zeroth-order energy of the total eigenstates becomes truly essential. This is because the *phase* of the probability amplitude transferred to the final state at a given time depends on the phase of the initial state at that given time. Therefore, the probability amplitudes transferred to the final state a given time will generally cancel a probability amplitude transferred at some past time *unless* the phase of the initial and final state rotates at the same frequency. That is, there will only a significant probability amplitude transferred to the final state in the characteristic time-scale of carrier transport if the zeroth-order energy eigenvalues of initial and final state are essentially the same.

We have shown then, that there is only a significant *probability* transferred from an initial state to a final state if there zeroth order energy eigenvalues of the initial and final state are the same. We have also shown that if the zeroth order energy eigenvalues are the same, then the rate probability is transferred from the initial state to the final state is proportional to the modulus of the matrix element *squared*, since the *probability amplitude* transfer is proportional to the modulus of the matrix element. Thus we have described the essential physics behind Fermi's golden rule.

The problem of calculating $S_{\text{par}}(\mathbf{k}\nu, \mathfrak{s}_{\text{par}}; \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}})$ is, according to eq. (2.26), essentially equivalent to the problem of calculating the appropriate matrix element of the interaction Hamiltonian. We now wish to separate the dependence of this matrix element on the states of the partner and of the carrier. Since carriers will primarily interact via the Coulomb interaction, we can expect that any interaction Hamiltonian associated with carrier will depend only on the position of the carrier (and its charge), and not on its crystal momentum or band index. As such, we can expand the interaction Hamiltonian in terms of a carrier-position dependent Hamiltonian $\hat{H}^{\mathbf{r}}_{\text{car-par}}$, which is a function of the position $\mathbf{r}$ of the carrier, but that technically only acts on the partner state:

$$\hat{H}_{\text{car–par}} = \int_V |\mathbf{r}\rangle\langle\mathbf{r}| \otimes \hat{H}^{\mathbf{r}}_{\text{car–par}} \mathrm{d}\mathbf{r}. \tag{2.28}$$

Here V is the entire volume occupied by the crystal sample. It may help the reader

[12] If the zeroth-order energy eigenstates were not the same, then the new eigenstates would be *uneven* superpositions of the initial and final state, and the difference in new eigenvalues would no longer be proportional to the modulus of the matrix element.

to note that when the carrier–partner Hamiltonian is expressed in this form, it can be represented by a block diagonal matrix, where the block that occurs at the set of rows and columns associated with carrier position $\mathbf{r}$ is defined by the operator $\hat{H}^{\mathbf{r}}_{\text{car–par}}$.

We make the assumption that the total state of the carrier partner system can be expressed as the tensor product of the state of the carrier system and the partner system: $\left|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle = \left|\mathbf{k}\nu\right\rangle\left|\mathfrak{s}_{\text{par}}\right\rangle$. That is, we make the assumption that the carrier and partner states are so weakly entangled/correlated that we can calculate the matrix element associated with their interaction as if they were independent. This is known as the BORN APPROXIMATION, and it is reasonable so long the carrier–partner interaction is small compared to the zeroth-order carrier and partner kinetic energies. The Born approximation allows us to rewrite the matrix element for the interaction as follows:

$$\begin{aligned}\left\langle \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}\middle|\hat{H}_{\text{car–par}}\middle|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle &= \left\langle \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}\middle| \int_V |\mathbf{r}\rangle\langle\mathbf{r}| \otimes \hat{H}^{\mathbf{r}}_{\text{car–par}} \mathrm{d}\mathbf{r}\middle|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle \\ &= \int_V \langle \mathbf{k}'\nu'|\mathbf{r}\rangle \langle \mathbf{r}|\mathbf{k}\nu\rangle \left\langle \mathfrak{s}'_{\text{par}}\middle|\hat{H}^{\mathbf{r}}_{\text{car–par}}\middle|\mathfrak{s}_{\text{par}}\right\rangle \mathrm{d}\mathbf{r}. \end{aligned} \tag{2.29}$$

In order to simplify this expression further, we will express $\hat{H}^{\mathbf{r}}_{\text{car–par}}$ in terms of a Fourier series. To properly understand this Fourier series expansion, it is helpful to note two points. First, for any chosen partner basis, a single matrix element of $\hat{H}^{\mathbf{r}}_{\text{car–par}}$ can be viewed as a simple scalar function of $\mathbf{r}$, which according to periodic boundary conditions has the same periodicity as the entire crystal sample.[13] Therefore, for any chosen partner basis, it is clear then that we can express the $\mathbf{r}$ dependence of each matrix element of $\hat{H}^{\mathbf{r}}_{\text{car–par}}$ in terms of a Fourier series over reciprocal space vectors $\mathbf{q}$. Second, for any chosen partner basis, we can combine the set of all Fourier components associated with a wavevector $\mathbf{q}$ into a matrix. Therefore, for any basis, there is a matrix representation of the operator $\hat{H}^{\mathbf{q}}_{\text{car–par}}$ defined in eq. (2.30). Thus it clear that $\hat{H}^{\mathbf{q}}_{\text{car–par}}$ is a well-defined operator:

$$\hat{H}^{\mathbf{r}}_{\text{car–par}} = \sum_{\mathbf{q}} \hat{H}^{\mathbf{q}}_{\text{car–par}} e^{i\mathbf{q}\cdot\mathbf{r}}. \tag{2.30}$$

[13] By "the same periodicity as the entire crystal sample", we do not mean periodicity with respect to *lattice* vectors, but periodicity with respect to a unit cell defined by V, the volume occupied by the entire crystal sample.

Noting that $\langle \mathbf{r}|\mathbf{k}\nu\rangle = \psi_{\mathbf{k}\nu}(\mathbf{r})$ we can rewrite eq. (2.29):

$$\left\langle \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}\middle|\hat{H}_{\text{car–par}}\middle|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle = \sum_{\mathbf{q}} \left\langle \mathfrak{s}'_{\text{par}}\middle|\hat{H}^{\mathbf{q}}_{\text{car–par}}\middle|\mathfrak{s}_{\text{par}}\right\rangle \int_V \psi^*_{\mathbf{k}'\nu'}(\mathbf{r})\psi_{\mathbf{k}\nu}(\mathbf{r})e^{i\mathbf{q}\cdot\mathbf{r}}\mathrm{d}\mathbf{r}. \tag{2.31}$$

Bloch's theorem states that $\psi_{\mathbf{k}\nu} = \frac{1}{\sqrt{N}}u_{\mathbf{k}\nu}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}$ where $u_{\mathbf{k}\nu}(\mathbf{r}) = u_{\mathbf{k}\nu}(\mathbf{r}+\mathbf{R})$ if $\mathbf{R}$ is a lattice vector and N is the number of unit cells in the crystal sample.[14] If we examine the integral over the crystal sample, we note that the integral over any unit cell can only differ from the integral over any other unit cell by a phase factor. As such, we can take the integral of the volume occupied by the unit cell at $\mathbf{R} = 0$— which we refer to as Ω— and multiply this integral by the sum of all phase factors relative to this unit cell:

$$\int_V \psi^*_{\mathbf{k}'\nu'}(\mathbf{r})\psi_{\mathbf{k}\nu}(\mathbf{r})e^{i\mathbf{q}\cdot\mathbf{r}}\mathrm{d}\mathbf{r} = \frac{1}{N}\sum_{\mathbf{R}} e^{i(\mathbf{q}+\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}} \int_\Omega u^*_{\mathbf{k}'\nu'}(\mathbf{r})u_{\mathbf{k}\nu}(\mathbf{r})e^{i(\mathbf{q}+\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}\mathrm{d}\mathbf{r}. \tag{2.32}$$

Each phase factor in the sum can be viewed as a step in a $2-$D plane with an absolute angle determined $(\mathbf{q}+\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}$ mod 2π. Roughly speaking, a "walk" of evenly spaced lattice of $\mathbf{R}$ vectors is transformed into an kind of "walk" of evenly changing angles in the $2-$D plane. Thus, the $2-$D vector sum of steps will form closed "circles" *unless there is no change in angle for each step*. Thus we can understand the well-known result that the sum will vanish unless $\mathbf{q}+\mathbf{k}-\mathbf{k}'$ is a reciprocal lattice vector:

$$\frac{1}{N}\sum_{\mathbf{R}} e^{i(\mathbf{q}+\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}} = \begin{cases} 1 & \text{if } \mathbf{q}+\mathbf{k}-\mathbf{k}' = \mathbf{G}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

We can therefore substitute eq. (2.32) and eq. (2.33) into our expression for the matrix element in eq. (2.31):

$$\left\langle \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}\middle|\hat{H}_{\text{car–par}}\middle|\mathbf{k}\nu, \mathfrak{s}_{\text{par}}\right\rangle = \sum_{\mathbf{G}} \left\langle \mathfrak{s}'_{\text{par}}\middle|\hat{H}^{\mathbf{k}'-\mathbf{k}+\mathbf{G}}_{\text{car–par}}\middle|\mathfrak{s}_{\text{par}}\right\rangle \mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu, \mathbf{k}'\nu'), \tag{2.34a}$$

$$\text{where} \qquad \mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu; \mathbf{k}'\nu') = \int_\Omega u^*_{\mathbf{k}'\nu'}(\mathbf{r})u_{\mathbf{k}\nu}(\mathbf{r})e^{i\mathbf{G}\cdot\mathbf{r}}\mathrm{d}\mathbf{r}. \tag{2.34b}$$

We refer to the quantity $\mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu; \mathbf{k}'\nu')$ as the OVERLAP INTEGRAL [43]. This new expression for the matrix element in turn creates a new expression for Fermi's golden rule from that

[14] We have normalized $u_{\mathbf{k}\nu}(\mathbf{r})$ over a unit cell of the crystal.

expressed in eq. (2.26):

$$S_{\text{par}}(\mathbf{k}\nu, \mathfrak{s}_{\text{par}}; \mathbf{k}'\nu', \mathfrak{s}'_{\text{par}}) = \frac{2\pi}{\hbar} \left| \sum_{\mathbf{G}} \left\langle \mathfrak{s}'_{\text{par}} \middle| \hat{H}^{\mathbf{k}'-\mathbf{k}+\mathbf{G}}_{\text{car–par}} \middle| \mathfrak{s}_{\text{par}} \right\rangle \mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu; \mathbf{k}'\nu') \right|^2 \times \delta\Big(\varepsilon_{\mathbf{k}'\nu'} + \varepsilon_{\mathfrak{s}'_{\text{par}}} - \varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathfrak{s}_{\text{par}}} \Big). \tag{2.35}$$

There are two important points to note about this updated form of Fermi's golden rule. The first point is that the overlap integral $\mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu; \mathbf{k}'\nu')$ is *entirely independent of the partner state*, and therefore can be calculated purely from knowledge of the carrier system alone. For reasons that will become clear latter, we will discuss overlap integral only after we discuss electron-phonon scattering. The second point is that, from a high level perspective, eq. (2.35) is simply the combination of Fermi's golden rule with the Born approximation and Bloch's theorem. It is therefore a fairly general starting point for describing many semiclassical scattering processes.

2.5.3 General Theory of Scattering with a Charged Partner

We remind the reader that in Section 2.2, we noted the general scheme for incorporating interactions between carriers and charged partners into the semiclassical model. Firstly, the reproducible, long-range interactions between carriers and charged partners are to be treated by frequently coupling the Boltzmann transport equation to a coarse grained solution of the Poisson equation. Secondly, the non-reproducible, short-range interactions between carriers and charged partners are to be treated as incoherent scattering perturbations. In this section, we now turn to discussing the short-range interaction between carriers and charged partners in detail.

The short-range interaction between a charged partner and a carrier can be assumed to take the form of a screened Coulomb potential in a dielectric. The dielectric effect— or polarization of the bound charge in response to an applied electric field— is characterized by a dielectric constant ϵ. The screening effect— or redistribution of the free charge in the response to a (static) applied electric field— is characterized by an inverse screening length β_S. The interaction Hamiltonian for a screened Coulomb potential in a dielectric is written as follows:

$$\hat{H}_{\text{car–cha}} = Z_{\text{cha}} e^2 4\pi\epsilon |\hat{\mathbf{r}} - \hat{\mathbf{r}}_{\text{cha}}| e^{-\beta_S |\hat{\mathbf{r}} - \hat{\mathbf{r}}_{\text{cha}}|}. \tag{2.36}$$

Where Z_{cha} is the ratio of the charge of partner to the charge of the carrier, and $\hat{r}_{\text{cha}}$ is the position operator for the charged partner. In eq. (2.28) we introduced the concept of an interaction operator that depends on carrier position but acts on the partner state, $\hat{H}^{\mathbf{r}}_{\text{car–cha}}$. For the interaction Hamiltonian in eq. (2.36), this operator is written as follows, where $\hat{I}$ is the identity operator in the partner state:

$$\hat{H}^{\mathbf{r}}_{\text{car–cha}} = \frac{Z_{\text{cha}} e^2}{4\pi\epsilon |\mathbf{r}\hat{I} - \hat{\mathbf{r}}_{\text{cha}}|} e^{-\beta_S |\mathbf{r}\hat{I} - \hat{\mathbf{r}}_{\text{cha}}|}. \tag{2.37}$$

We can interpret eq. (2.37) as an *operator-valued function* of carrier position. The Fourier components of this operator-valued function can be calculated in the same way Fourier components are calculated for any other function of carrier position. If we assume periodic boundary conditions about the crystal sample volume V, the Fourier components are given as follows:

$$\hat{H}^{\mathbf{q}}_{\text{car–cha}} = \frac{1}{V} \int_V \hat{H}^{\mathbf{r}}_{\text{car–cha}} e^{-i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}. \tag{2.38}$$

We note that in eq. (2.38), $\mathbf{r}$ now plays the role of a dummy variable that is replaced by every point in the unit cell V. If we take a unit cell of any periodic lattice, and shift every point by an arbitrary constant vector $\mathbf{r}'$, we *have defined another unit cell*. As such, the integral over the unit cell $V + \mathbf{r}'$ can equally be used to define the Fourier component of eq. (2.38). By the simple manipulation shown below, this leads to the conclusion that the Fourier component of $\hat{H}^{\mathbf{r}+\mathbf{r}'}$ is only a phase shift from the Fourier component of $\hat{H}^{\mathbf{r}}$:

$$\begin{aligned}
\hat{H}^{\mathbf{q}}_{\text{car–cha}} &= \frac{1}{V} \int_{V+\mathbf{r}'} \hat{H}^{\mathbf{r}}_{\text{car–cha}} e^{-i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} \\
&= \frac{1}{V} \int_V \hat{H}^{\mathbf{r}+\mathbf{r}'}_{\text{car–cha}} e^{-i\mathbf{q}\cdot(\mathbf{r}+\mathbf{r}')} \mathrm{d}\mathbf{r} \\
&= e^{-i\mathbf{q}\cdot\mathbf{r}'} \frac{1}{V} \int_V \hat{H}^{\mathbf{r}+\mathbf{r}'}_{\text{car–cha}} e^{-i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}.
\end{aligned} \tag{2.39}$$

In retrospect, eq. (2.39) should be fairly obvious. The *magnitude* of a Fourier component *cannot* depend the choice for the origin of the position coordinate, but the *phase* of a Fourier component *can*.

The purpose of the preceding discussion is that it enables us to write down calculate the matrix representation of $\hat{H}^{\mathbf{q}}_{\text{car–cha}}$ in the position basis of the charged partner state. This matrix representation is diagonal, and each of the diagonal matrix elements can be shown to be a phase shift to the same analytic integral by setting $\mathbf{r}' = \mathbf{r}_{\text{cha}}$ in each case. The result is the following:

$$\hat{H}^{\mathbf{q}}_{\text{car–cha}} = \frac{Z_{\text{cha}} e^2}{4\pi\epsilon V} e^{-i\mathbf{q}\cdot\hat{\mathbf{r}}_{\text{cha}}} \int_V \frac{e^{-\beta_S|\mathbf{r}|}}{|\mathbf{r}|} e^{-i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}. \tag{2.40}$$

The integral in eq. (2.40) is a well-known standard integral, sometimes referred to as the YUKAWA POTENTIAL integral. Evaluating this integral we are led to the following algebraic expression for $\hat{H}^{\mathbf{q}}_{\text{car–cha}}$:

$$\hat{H}^{\mathbf{q}}_{\text{car–cha}} = \frac{Z_{\text{cha}} e^2}{\epsilon V} \frac{1}{\beta_S^2 + \mathbf{q}^2} e^{-i\mathbf{q}\cdot\hat{\mathbf{r}}_{\text{cha}}}. \tag{2.41}$$

It is well-known that using the Born approximation ignores the Coloumb correlations between the carrier and the charged partner, and can lead to miscalculating the scattering rate by a factor of ~ 5 for scattering processes which involve only low energy carriers [45]. We note in passing that Fischetti, Laux and Crabbé improved upon the Born approximation by introducing a phase-shift correction in their 1995 update to the DAMOCLES model [6], and this is a relatively important improvement that could be made to the scattering model we describe here.

In addition to the problems with the Born approximation, we also note that accurately modelling the inverse screening length β_S in the far-from-thermal-equilibrium regime is challenging. The inverse screening length is fundamentally a function of the entire distribution function, as well as the states of both the carrier and charged partner involved in scattering [26]. One of the most important effects that was discussed in the 1995 update paper is that the inverse screening length is greatly reduced for pairs of charged particles that have a large relative velocity— owing to the fact that the background carriers cannot respond dynamically to the interactions between such partners— resulting

in significantly stronger scattering for these pairs than was predicted in the original 1988 paper.

Our approach for modelling the inverse screening length in this thesis differs slightly from any approach Fischetti et al. have suggested in any of the major papers relating to the DAMOCLES model [16, 46, 6]. It is a very crude model in which we simply assume that, given we are most interested in the scattering with carriers that have a high energy, we assume that *charge density never responds dynamically at all*. Instead, the only screening effect of the free charge is a *statistical screening* effect due to the random distribution of charge.

This statistical screening effect was first noted by Ridley [47], in relation to ionized dopants. Note that in this section, we will only deal with scattering between a single charged partner and the carrier. The tacit assumption we will later use is that the total carrier scattering with many charged can be viewed as the sum of carrier scattering with all charged partners. However, if two charged partners are close together, the electric fields due to each ion will often cancel out one another due to their vectorial nature. In the devices of interest, the density of charged partners is typically sufficiently high that this kind of cancellation is common. Ridley was the first to attempt to account for this kind of cancellation.

Ridley accounted for this cancellation by making the assumption that a carrier could only interact with the ionized dopant nearest to it [47], which we will modify into the assumption that a carrier can only interact with the *charged partner* nearest to it. To understand this assumption, consider the following. The electric force on the carrier due the nearest charged partner is larger than the charged partners at larger distances. We can view the charged partners at larger distances as adding a perturbation, of random size and orientation, to this nearest charged partner field. We can make the argument then that there is no *reproducible* effect on the nearest charged partner field, due to the other charged partners. This, in turn implies that carriers can be modeled as being scattered only by the nearest charged partner field.

When the density of charged partners is small, and carrier screening ensures there is negligible overlap between adjacent charged partner fields, Ridleys statistical screening

effect is negligible. However, when the charged partner density is heavy, and there is significant overlap between adjacent charged partner fields, Ridleys statistical screening assumption creates an effective radial exponential cutoff on the perturbation field due to a dopant, in a similar manner that the dynamic response of carriers creates a radial exponential cutoff in ordinary dynamic screening theory.

Ridley himself was interested in the direct effect of this statistical screening on mobilities rather than on the inverse screening length, and so we do not use his expression directly. Instead, we follow and expand the simple method used by Fischetti, Frank and Laux in their 1990 DAMOCLES paper [46]. In this paper, Fischetti et al. model the statistical screening effect of the dopant fields on the inverse screening length β_S. We simply expand this model to include not only the statistical screening effect of dopant fields, but also the statistical screening effect of carrier fields, and we then *ignore the dynamic screening effect* of the carrier fields. The result is the following:

$$\boxed{\begin{aligned}\beta_S(\mathbf{r},t) &= \left(\frac{e\rho_{\text{unsigned}}(\mathbf{r},t)}{\epsilon k T_L}\right)^{\frac{1}{2}}, \quad \text{where} \\ \rho_{\text{unsigned}}(\mathbf{r},t) &= e\left(\sum_i N^i_{\text{dop}}(\mathbf{r})|Z^i_{\text{dop}}| + \overbrace{\Gamma\sum_\nu \int_{\text{BZ}} f(\mathbf{k}\nu,\mathbf{r},t)\mathrm{d}\mathbf{k}}^{n(\mathbf{r},t)+p(\mathbf{r},t)}\right).\end{aligned}} \tag{2.42}$$

2.5.4 Carrier–Dopant Scattering

We are interested here in the scattering rate with a single ionized dopant atom at position $\mathbf{R}_{\text{dop}}$ that has Z_{dop} additional protons relative to the ideal lattice, where Z_{dop} might be negative. We assume the dopant is sufficiently large to be safely treated as inert during the interaction with the carrier. As such, the initial and final state of a partner dopant can characterized by a single lattice position $\left|\mathbf{R}_{\text{dop}}\right\rangle$, and the initial and final energy eigenvalue of the dopant can be assumed to be identical. By substituting eq. (2.41) into eq. (2.35), we are led to the following expression for the scattering rate:

$$S_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}'\nu',\mathbf{R}_{\text{dop}}) = \frac{2\pi Z^2_{\text{dop}}e^4}{\hbar\epsilon^2V^2}\left|\sum_{\mathbf{G}}\frac{e^{-i(\mathbf{k}-\mathbf{k}'+\mathbf{G})\cdot\mathbf{R}_{\text{dop}}}\mathscr{J}^{\mathbf{G}}(\mathbf{k}\nu;\mathbf{k}'\nu')}{\beta_S^2+(\mathbf{k}-\mathbf{k}'+\mathbf{G})^2}\right|^2\delta(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon_{\mathbf{k}\nu}). \tag{2.43}$$

Fischetti and Laux make the approximation that the sum is dominated by the $\mathbf{G}^*$ term, where $\mathbf{G}^*$ is defined such that $\mathbf{k} - \mathbf{k}' + \mathbf{G}^*$ is in the first Brillouin zone. This leads to the following expression for the scattering operator for a carrier and an ionized dopant partner:

$$\boxed{S_{\text{dop}}(\mathbf{k}\nu, \mathbf{R}_{\text{dop}}; \mathbf{k}'\nu', \mathbf{R}_{\text{dop}}) = \frac{2\pi Z_{\text{dop}}^2 e^4}{\hbar\epsilon^2 V^2} \frac{\left|\mathscr{I}^{\mathbf{G}^*}(\mathbf{k}\nu; \mathbf{k}'\nu')\right|^2}{\left(\beta_S^2 + (\mathbf{k} - \mathbf{k}' + \mathbf{G}^*)^2\right)^2} \delta(\varepsilon_{\mathbf{k}'\nu'} - \varepsilon_{\mathbf{k}\nu}).} \tag{2.44}$$

2.5.5 Carrier–Carrier Scattering

It is well-known that the Born approximation ignores the exchange correlation in calculating the scattering rate between similar carriers, which can lead to miscalculating the scattering rate by a factor of ~ 2 [48]. However, the effect of this exchange correlation is much less important for carriers which have a large relative wavevector. While Fischetti, Laux and Crabbé included these correlation effects in the 1995 update to the DAMOCLES model [6], this is not a particularly important addition to the scattering model we describe here. The reason is that the carrier–carrier scattering processes that actually affect transport tend to involve carriers which have a large relative wavevector. This will be discussed in more detail in the results chapters of this thesis.

With this in mind, we are interested here in the scattering rate of a carrier with a single carrier partner in initial state $|\mathbf{p}\mu\rangle$ which transforms to a final state $|\mathbf{p}'\mu'\rangle$ after scattering.[15] Inserting eq. (2.41) into eq. (2.35) for this situation leads to the following expression:

$$\begin{aligned} S_{\text{car}}(\mathbf{k}\nu, \mathbf{p}\mu; \mathbf{k}'\nu', \mathbf{p}'\mu') = \frac{2\pi e^4}{\hbar\epsilon^2 V^2} \left| \sum_{\mathbf{G}} \frac{\langle \mathbf{p}'\mu' | \, e^{-i(\mathbf{k}'-\mathbf{k}+\mathbf{G})\cdot\hat{\mathbf{r}}_{\text{par}}} \, |\mathbf{p}\mu\rangle \, \mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu; \mathbf{k}'\nu')}{\beta_S^2 + (\mathbf{k}' - \mathbf{k} + \mathbf{G})^2} \right|^2 \\ \times \delta(\varepsilon_{\mathbf{k}'\nu'} + \varepsilon_{\mathbf{p}'\mu'} - \varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathbf{p}\mu}). \end{aligned} \tag{2.45}$$

We can use eq. (2.32) and eq. (2.33) to rewrite the matrix element over the partner carrier:

$$\langle \mathbf{p}'\mu' | \, e^{-i(\mathbf{k}'-\mathbf{k}+\mathbf{G})\cdot\hat{\mathbf{r}}_{\text{par}}} \, |\mathbf{p}\mu\rangle = \sum_{\mathbf{G}'} \delta_{\mathbf{K}-\mathbf{G}'} \mathscr{I}^{\mathbf{K}-\mathbf{G}}(\mathbf{p}\mu; \mathbf{p}'\mu'). \tag{2.46}$$

[15] In this thesis, the symbol μ is slightly overloaded. It is used to reference both the band index of a partner state, and later, the chemical potential of various quasi-equilibria.

Here $\mathbf{K}$ is defined as total initial crystal momentum minus total final crystal momentum $\mathbf{K} = \mathbf{k} + \mathbf{p} - \mathbf{k}' - \mathbf{p}'$ and the sum over $\mathbf{G}'$ is a sum over reciprocal lattice vectors. Substituting eq. (2.46) into eq. (2.45) yields the following:

$$S_{\text{car}}(\mathbf{k}\nu, \mathbf{p}\mu; \mathbf{k}'\nu', \mathbf{p}'\mu') = \frac{2\pi e^4}{\hbar\epsilon^2 V^2} \left| \sum_{\mathbf{G},\mathbf{G}'} \frac{\delta_{\mathbf{K}-\mathbf{G}'} \mathscr{I}^{\mathbf{K}-\mathbf{G}}(\mathbf{p}\mu; \mathbf{p}'\mu') \mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu; \mathbf{k}'\nu')}{\beta_S^2 + (\mathbf{k}' - \mathbf{k} + \mathbf{G})^2} \right|^2 \times \delta(\varepsilon_{\mathbf{k}'\nu'} + \varepsilon_{\mathbf{p}'\mu'} - \varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathbf{p}\mu}). \tag{2.47}$$

Fischetti and Laux make the assumption that the sum over $\mathbf{G}$ is dominated by the term with the smallest denominator, at $\mathbf{G} = \mathbf{G}^*$. We note that $\mathbf{G}^*$ is reciprocal lattice vector such that $\mathbf{k}' - \mathbf{k} + \mathbf{G}$ is in the first Brillouin zone. We separate out the sum over $\mathbf{G}'$, so it is clear that the only effect of this series is to ensure that $\mathbf{K}$ is equal to some reciprocal lattice vector. Accordingly we have the following expression for the scattering operator of a carrier with another carrier:

$$\boxed{S_{\text{car}}(\mathbf{k}\nu, \mathbf{p}\mu; \mathbf{k}'\nu', \mathbf{p}'\mu') = \frac{2\pi e^4}{\hbar\epsilon^2 V^2} \frac{\left|\mathscr{I}^{\mathbf{K}-\mathbf{G}^*}(\mathbf{p}\mu; \mathbf{p}'\mu')\right|^2 \left|\mathscr{I}^{\mathbf{G}^*}(\mathbf{k}\nu; \mathbf{k}'\nu')\right|^2}{\left(\beta_S^2 + (\mathbf{k}' - \mathbf{k} + \mathbf{G}^*)^2\right)^2} \times \delta(\varepsilon_{\mathbf{k}'\nu'} + \varepsilon_{\mathbf{p}'\mu'} - \varepsilon_{\mathbf{k}\nu} - \varepsilon_{\mathbf{p}\mu}) \sum_{\mathbf{G}'} \delta_{\mathbf{K}-\mathbf{G}'}.} \tag{2.48}$$

2.5.6 Impact Ionization

Impact ionization occurs due to the Coulomb interaction between a carrier and an electron in the valence band. The scattering between a carrier and a *vacancy* in the valence band or an electron in the conduction band is accounted for by carrier–hole or carrier–conduction electron scattering respectively. The scattering between a carrier and an electron in the valence band can only take place if the carrier has a kinetic energy larger than the band gap ε_{gap}, since the only available final states for the valence band electron are in the conduction band. When an electron is moved from the valence band to the conduction band two additional carriers are generated: a conduction electron and a hole.

The 1995 update paper describes the updated impact ionization model as "major progress"

on the original 1988 model [6]. Since the new model is trivial to implement, we use the updated impact ionization model of Cartier et al. [49]. In both the 1988 and the 1995 models, rather than directly modelling the interaction Hamiltonian relevant to impact ionization, the impact ionization rate itself is modelled as an empirical function of initial carrier kinetic energy, and all energy-conserving final states as presumed to be equally probable.

We note that this is equivalent to modelling the matrix elements as a function of the carriers initial kinetic energy only, and is equivalent to ignoring entirely crystal momentum pseudo-conservation: an approximation which was first justified by Kane [50]. The idea behind this approximation is that the total impact ionization rate of a given initial state $|\mathbf{k}\nu\rangle$ is dominated by the effect of the simultaneous density of energy-conserving final states, which itself is only a function of initial kinetic energy. The conservation of crystal momentum on the other hand does not significantly adjust this, on that basis that a fixed initial and final carrier crystal momentum state imposes no restriction on the crystal momentum of the created electron, since— in the absence of energy conservation restrictions— the created hole can always ensure crystal momentum conservation. The impact ionization rate is also not strongly impacted by variations in matrix element, simply because the impact ionization processes associated with a given initial state $|\mathbf{k}\nu\rangle$ have so many possible final states, and accordingly the matrix elements are averaged over an enormous space of final states. For this reason, we will not "reverse-engineer" the implicitly assumed formula for the matrix elements as it is not of direct interest to us and is not likely to be accurate. Instead, we will simply state the implicitly assumed scattering operator for impact ionization as a function of $\frac{1}{\tau_{ii}(\varepsilon_{\mathbf{k}\nu})}$, the empirically modelled impact ionization time:

$$S_{ii}(\mathbf{k}\nu;\mathbf{k}'\nu',\mathbf{k}'_e\nu'_e,\mathbf{k}'_h\nu'_h) = \left(V^3\int_0^{\varepsilon_{\mathbf{k}\nu}-\varepsilon_{\text{gap}}}\int_0^{\varepsilon_{\mathbf{k}\nu}-\varepsilon_{\text{gap}}-\varepsilon_e} D_{\text{con}}(\varepsilon_e)D_{\text{val}}(\varepsilon_h)D_{\text{car}}(\varepsilon_{\mathbf{k}\nu}-\varepsilon_{\text{gap}}-\varepsilon_e-\varepsilon_h)\mathrm{d}\varepsilon_h\mathrm{d}\varepsilon_e\right)^{-1} \times \frac{1}{\tau_{ii}(\varepsilon_{\mathbf{k}\nu})}\delta(\varepsilon_{\mathbf{k}'\nu'}+\varepsilon_{\mathbf{k}'_e\nu'_e}+\varepsilon_{\mathbf{k}'_h\nu'_h}+\varepsilon_{\text{gap}}-\varepsilon_{\mathbf{k}\nu}). \tag{2.49}$$

Here $D_{\text{car}}(\varepsilon)$ is the density of states per kinetic energy, per volume for a given carrier type at kinetic energy ε, similarly $D_{\text{con}}(\varepsilon_e)$ is the conduction band density of electrons

at kinetic energy ε_e, and $D_{\text{val}}(\varepsilon_h)$ is the valence band density of holes at kinetic energy ε_h. The formula above can be derived as follows. Suppose we have an initial state $|\mathbf{k}\nu\rangle$. The rate of impact ionization of $|\mathbf{k}\nu\rangle$ is $\frac{1}{\tau_{ii}(\varepsilon_{\mathbf{k}\nu})}$, by definition. The number of energy conserving final states after impact ionization of $|\mathbf{k}\nu\rangle$ is given by the term in the large parenthesis. The rate of transition to a *particular* energy-conserving final state must be proportional to $\frac{1}{\tau_{ii}(\varepsilon_{\mathbf{k}\nu})}$ and inversely proportional to the number of final states since we assume all energy conserving final states are equally likely. Finally the delta function ensures the rate of transition to states that do not conserve energy is zero.

The empirical form for the net impact ionization rate is that determined by Cartier et al. [49], and is a Keldysh-type formula [51] with multiple threshold energies:

$$\boxed{\frac{1}{\tau_{ii}(\varepsilon)} = \sum_i \theta\left(\varepsilon - \varepsilon_i^{\text{thr}}\right) R_i \left(\frac{\varepsilon - \varepsilon_i^{\text{thr}}}{\varepsilon_i^{\text{thr}}}\right)^2.} \tag{2.50}$$

Here $\theta(x)$ is the Heaviside step function, i is the number of threshold energies, and R_i and $\varepsilon_i^{\text{thr}}$ are the empirically determined threshold rate parameters and threshold energies respectively.

2.5.7 General Phonon Theory

In order to properly describe carrier–phonon scattering, we first need to briefly describe the theoretical basis of phonons themselves.

A general deformation of the ions in an ideal crystal lattice can be described by associating each of the ion j ions with a displacement vector $\mathbf{u}(\mathbf{R}_j)$, where $\mathbf{R}_j$ is the ideal position of the j^{th} ion. The result is a vector field of deformations $\mathbf{u}(\mathbf{R})$ defined on a $3-$D lattice of discrete points. Hypothetically, suppose we propose that the net restoring force on the j^{th} ion is linearly proportional to $\mathbf{u}(\mathbf{R}_j)$. Notice, that in the case that every ion is displaced by the same vector, this proposal implies there will be a non-zero restoring force. Since physically this deformation is associated with a translation of the entire crystal sample in the room, such a restoring force is clearly unphysical. As such, a more physical first-order approximation of the effect of ion displacements is to assume

that the net restoring force on the j^{th} ion is linearly proportional to the *changes in nearest neighbour distances* as a result of the discrete vector field of deformations $\mathbf{u}(\mathbf{R})$. As discussed in Appendix E, in a continuum, this change in nearest neighbour distances is measured by the STRAIN TENSOR FIELD $\mathbf{e}(\mathbf{R})$.

While we have rejected the assumption that the j^{th} ion experiences a net force linearly proportional $\mathbf{u}(\mathbf{R}_j)$, it is useful to notice that *under this incorrect assumption* the motion of *each individual ion* would define an independent harmonic oscillator. In the *accepted assumption* that each ion experiences a restoring force that is linearly proportional to the local *strain*, the independent harmonic oscillators cannot be localized to individual ions, and are instead *inherently collective* motions. These *collective* independent harmonic oscillators are known as PHONON MODES. The phonon modes can, by Bloch's theorem, be associated with a precise Brillouin zone crystal momentum $\mathbf{q}$ and a band η. The evenly-spaced energy eigenstates in each of these *collective* independent harmonic oscillators/phonon modes are interpreted as referring to the "number of phonons" which occupy the phonon mode.[16] Accordingly an eigenstate of the phonon mode at $\mathbf{q}\eta$ is written $|n_{\mathbf{q}\eta}\rangle$.

Since ions obey the laws of quantum mechanics, an ion cannot simultaneously have an arbitrarily precise position and a bounded momentum. Since the energy eigenstates of a phonon mode have a bounded momentum, they cannot simultaneously be eigenstates of ion position. Thus we not cannot associate an eigenstate of a phonon mode with a precise discrete deformation vector field. We can however associate a phonon mode with a deformation vector field *operator* $\hat{\mathbf{u}}_{\mathbf{q}\eta}(\mathbf{R})$. It is a standard textbook result [52] that the deformation vector operator associated with a phonon mode can be expressed in terms of the *energy eigenstate ladder operators* of that phonon mode $\hat{a}^\dagger_{\mathbf{q}\eta}$ and $\hat{a}_{\mathbf{q}\eta}$, which are equivalent to the *phonon creation and annihilation operators* respectively:

$$\hat{\mathbf{u}}_{\mathbf{q}\eta}(\mathbf{R}) = \left(\frac{\hbar}{2\rho V \omega_{\mathbf{q}\eta}}\right)^{\frac{1}{2}} \left(\hat{a}_{\mathbf{q}\eta} e^{\,i\mathbf{q}\cdot\mathbf{R}} + \hat{a}^\dagger_{\mathbf{q}\eta} e^{\,-i\mathbf{q}\cdot\mathbf{R}}\right) \boldsymbol{\xi}_\eta. \tag{2.51}$$

[16] It is worth noting that the historical decision to reify the gaps in the energy eigenstates of collective phenomena into BOSON PARTICLES such as phonons or photons was, from a conservative perspective, somewhat eccentric. It is not that difficult to imagine a physics which evolved without referring to these gaps in energy eigenstates of collective phenomena as particles. Nevertheless, when properly understood the idea has proven to be useful, efficient, intuitive and fruitful. This author merely believes that one part of this proper understanding is noting the extent to which these bosonic particles are particles *simply by agreed convention*.

Here $\boldsymbol{\xi}_\eta$ is the polarization vector for a phonon band, ρ is the (mass) density of the crystal sample, and as previously mentioned V is the volume of the crystal sample and $\omega_{\mathbf{q}\eta}$ is the frequency of the phonons associated with the $\mathbf{q}\eta$ mode. The various phonon bands indexed by η which are associated with a single crystal momentum $\mathbf{q}$ arise as follows. Firstly, in a $3-$D crystal, there three are independent phonon modes associated with different polarizations $\boldsymbol{\xi}_\eta$. There is one LONGITUDINAL band associated with displacement vector fields that are parallel to $\mathbf{q}$, and two TRANSVERSE bands that are associated with displacement vector fields that are perpendicular to $\mathbf{q}$. Secondly, in a crystal in which the unit cell contains two basis atoms, there is an ACOUSTIC band associated with oscillations where these basis atoms move in phase and an OPTICAL MODE associated with oscillations where these phonons move 180° out of phase. As such, each crystal momentum point in the conventional Brillouin zone of a material such as silicon— which has a two atom basis if associated with face centred cubic unit cell, and therefore a base centred cubic Brillouin zone— is associated with six phonon bands.

For the phonon bandstructure, Fischetti and Laux fit very simple analytic expressions to known phonon bandstructure data [16]. For acoustic bands, they assume the bands are of the following form.

$$\boxed{\omega^{\text{ac}}_{\mathbf{q}\eta} = \begin{cases} \omega^{\max}_\eta \left(1 - \cos \frac{|\mathbf{q}|a}{4}\right)^{\frac{1}{2}} & \text{for } |\mathbf{q}| < \frac{2\pi}{a}, \\ \omega^{\max}_\eta & \text{for } |\mathbf{q}| \geq \frac{2\pi}{a}. \end{cases}} \tag{2.52a}$$

The simplicity of this form is more apparent when we note that the obviously crude $\frac{|\mathbf{q}a|}{2\pi}$ is never more than 9% from the function $\left(1 - \cos \frac{|\mathbf{q}|a}{4}\right)^{1/2}$ on its specified domain. For the optical bands, Fischetti and Laux make the strong assumption that all optical phonons have the same energy:

$$\boxed{\omega^{\text{op}}_{\mathbf{q}\eta} = \omega^{\max}_\eta.} \tag{2.52b}$$

2.5.8 Carrier–Phonon Scattering

We now turn to the discuss the interaction between a phonon mode and a carrier. By direct analogy to the argument used in the discussion of phonons in and of themselves,

we expect that the carrier energy eigenstates will not be changed by a uniform displacement of every ion. As such, we begin by exploring the effect of lattice strain on the carrier eigenstates.

A crystal lattice subject to a *uniform* strain tensor field is simply a crystal lattice with a different unit cell. As such, a uniformly strained crystal lattice will still have energy eigenvalues associated with each crystal momenta and band, as Bloch's theorem will still apply. That there exists an energy eigenstate which can be labelled by $|\mathbf{k}\nu\rangle$ both before and after the uniform strain is applied is most easily visualized in the extended zone scheme rather than the reduced zone scheme, since the Brillouin zone will change size but the extended zone scheme will continue to occupy the entire $3-$D space and associate a single band with each position in that space.

For small uniform strains, Taylor's theorem tells us we can expect an approximately linear relationship between a uniform strain tensor $\mathbf{e}$ and an energy eigenvalue shift $\Delta\varepsilon_{\mathbf{k}\nu}$ to the carrier state $|\mathbf{k}\nu\rangle$. An arbitrary linear relationship between a scalar and a symmetric tensor can itself always be characterized by a similarly sized symmetric tensor, which we will refer to as the DEFORMATION POTENTIAL $\mathbf{\Xi}_{\mathbf{k}\nu}$, where the subscript reflects the fact that it will generally depend on the carrier eigenstate:

$$\begin{aligned}\Delta\varepsilon_{\mathbf{k}\nu} &= \sum_{i,j}^{3,3} \Xi^{ij}_{\mathbf{k}\nu} e_{ij} \\ &= \mathbf{\Xi}_{\mathbf{k}\nu} : \mathbf{e}\end{aligned} \tag{2.53}$$

Here the colon operator ":" represents the double dot, or tensor contracted product. We make the ADIABATIC or BORN-OPPENHEIMER approximation, that the carrier finds the new eigenstates on a time-scale that is negligible to the time-scale of the variation of the strain. We also make the BAND-DEPENDENT DEFORMATION POTENTIAL APPROXIMATION, that the deformation potential is only a function of the band index of the carrier. This latter approximation can be understood as follows. We first note that a carrier localized to a given unit cell must be an equal superposition of eigenstates with every different crystal momentum in the Brillouin zone. As such, if a carrier is located to a unit cell, the only other information it can possess is a band index. The relationship of energy shift to strain for a carrier in a given band, located to a unit cell at $\mathbf{R}$ is given as follows:

$$\Delta\varepsilon_\nu(\mathbf{R},t) = \boldsymbol{\Xi}_\nu : \mathbf{e}(\mathbf{R},t), \tag{2.54}$$
$$\text{where} \qquad \boldsymbol{\Xi}_\nu = \frac{\int_{\mathrm{BZ}} \boldsymbol{\Xi}_{\mathbf{k}\nu}\mathrm{d}\mathbf{k}}{\int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}}.$$

In the band-dependent deformation potential approximation, we make the assumption that a single band-dependent deformation potential determines the position-dependent energy shift for *any* carrier state, not only for carrier states that are localized a unit cell. That is, we make the approximation:

$$\boldsymbol{\Xi}_{\mathbf{k}\nu} \approx \boldsymbol{\Xi}_\nu. \tag{2.55}$$

We now wish to explore the process by which strain leads to scattering in the band-dependent deformation potential approximation. To elucidate the situation, suppose we have a sinusoidally varying strain of wavevector $\mathbf{q}$. As a function of lattice position vector and time, there will be an energy shift $\Delta\varepsilon_\nu(\mathbf{R},t) = \boldsymbol{\Xi}_\nu : \mathbf{e}(\mathbf{R},t)$. Thus the expected net energy shift of a carrier eigenstate $|\mathbf{k}\nu\rangle$ is zero since the energy increases will be balanced by equal and opposite energy decreases elsewhere in the crystal.

In fact, for states which are delocalized across the entire crystal sample, a non-zero energy shift can only be associated with a *superposition* of carrier eigenstates that differ in crystal momentum by $\mathbf{q}$ mod $\mathbf{G}$— only in such a superposition can we, for instance, systematically constructively interfere to increase probability density in the unit cells where strain reduces eigenstate energy, and systematically destructively interfere to decrease probability density in the unit cells where strain increases eigenstate energy. The amount of constructive and destructive interference in a weighted superposition of carrier eigenstates separated by $\mathbf{q}$ is determined by the overlap integral. For equally weighted superpositions of carrier eigenstates in different bands, the net energy shift is determined simply by averaging the deformation potentials $\boldsymbol{\Xi}_\nu$ over the two bands. Accordingly, there is an interaction Hamiltonian between pairs of phonon states that generate sinusoidal strains of wavevector $\mathbf{q}$, and pairs of carrier states seperated by crystal momentum $\mathbf{q}$ mod $\mathbf{G}$. The analagous expression to eq. (2.34) for the interaction Hamiltonian matrix element relevant to scattering between a carrier and a partner body defined by $\mathbf{q}\eta$ phonon mode, is therefore the following:

$$\left\langle \mathbf{k}'\nu', n'_{\mathbf{q}\eta} \right| \hat{H}_{\text{car–pho}} \left| \mathbf{k}\nu, n_{\mathbf{q}\eta} \right\rangle = \sum_{\mathbf{G}} \frac{\boldsymbol{\Xi}_{\nu} + \boldsymbol{\Xi}_{\nu'}}{2} : \left\langle n'_{\mathbf{q}\eta} \right| \hat{\mathbf{e}}^{\mathbf{k}'-\mathbf{k}+\mathbf{G}} \left| n_{\mathbf{q}\eta} \right\rangle \times \mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu, \mathbf{k}'\nu'). \tag{2.56}$$

Here $\hat{\mathbf{e}}^{\mathbf{k}'-\mathbf{k}+\mathbf{G}}$ is an operator that acts on the phonon state and determines the Fourier component of the strain tensor field at wavevector $\mathbf{k}'-\mathbf{k}+\mathbf{G}$. The next step in determining this carrier–phonon matrix element is to write down an explicit expression for this operator in terms of the state of the $\mathbf{q}\eta$ phonon mode. We first note that the strain tensor field operator associated with acoustic modes and optical modes is qualitatively different, because in optical modes the strain is dominated by the relative displacements of basis atoms, whereas in acoustic modes it is dominated by the relative displacements of lattice points. We will discuss acoustic modes first. Hypothetically, if the displacement operator in eq. (2.51) was associated with a *continuous* vector field of displacements, rather than a discrete vector field of displacements, the strain operator would be given by eq. (2.57):

$$\hat{\mathbf{e}}(\mathbf{r}, t) = \frac{1}{2}\left(\nabla \otimes \hat{\mathbf{u}}(\mathbf{r}, t) + (\nabla \otimes \hat{\mathbf{u}}(\mathbf{r}, t))^{T}\right). \tag{2.57}$$

Thus the strain operator is simply a symmetrized Jacobian of the displacement operator. The simplicity of the strain operator in eq. (2.57) motivates us to make a SEMI-CONTINUUM APPROXIMATION for acoustic modes. This approximation consists of two transformations. The first transformation is to extend our discrete displacement operator $\hat{\mathbf{u}}(\mathbf{R})$ into an operator defined for all positions $\hat{\mathbf{u}}(\mathbf{r})$, so that the expression for strain given in eq. (2.57) can be used. The second transformation is effectively to remove the meaningless smooth variation of strain within a unit cell that the first transformation creates, and replace it with a constant strain within a unit cell. This meaningless variation is associated with the factor $e^{i\mathbf{q}\cdot\mathbf{r}}$ in the overlap integral, and therefore is removed by multiplying the overlap integrand in eq. (2.35) by $e^{-i\mathbf{q}\cdot\mathbf{r}}$. We refer to this transformed overlap integral as the MODIFIED OVERLAP INTEGRAL $\mathscr{I}^{\text{mod}}(\mathbf{k}\nu, \mathbf{k}'\nu')$.[17]

Due to the first transformation in the semi-continuum approximation, we can combine

[17] The modified overlap integral does not depend on $\mathbf{G}$, since according to eq. (2.33) $e^{i(\mathbf{G}-\mathbf{q})\cdot\mathbf{r}} = e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}$. The same result can be seen by multiplying eq. (2.32) by $e^{-i\mathbf{q}\cdot\mathbf{r}}$.

eq. (2.51) and eq. (2.57) in order to generate the equation for the continuous strain field operator associated with an acoustic mode:

$$\hat{\mathbf{e}}_{\mathbf{q}\eta}(\mathbf{r}) = i\left(\frac{\hbar}{2\rho V\omega_{\mathbf{q}\eta}}\right)^{\frac{1}{2}}\left(\hat{a}_{\mathbf{q}\eta}e^{i\mathbf{q}\cdot\mathbf{r}} - \hat{a}^{\dagger}_{\mathbf{q}\eta}e^{-i\mathbf{q}\cdot\mathbf{r}}\right)\frac{\mathbf{q}\otimes\boldsymbol{\xi}+\boldsymbol{\xi}\otimes\mathbf{q}}{2}. \tag{2.58}$$

We note that the fact that the strain operator is always linear in crystal momentum is an artifact of the pseudocontinuum approximation. In a discrete lattice, the strain operator would only be linear in crystal momentum for long wavelength phonons.

Since the initial and final state of the phonon mode are zeroth order eigenstates, we wish to evaluate the strain operator in the eigenstate basis of a phonon mode. We note that in this basis, all matrix elements of the creation and annihilation operators are zero except those associated with a pair of eigenstates that differ by a single phonon. More specifically, the creation and annihilation operators are defined as follows:

$$\hat{a}^{\dagger}\left|n_{\mathbf{q}\eta}\right\rangle = \sqrt{n_{\mathbf{q}\eta}+1}\left|n_{\mathbf{q}\eta}+1\right\rangle, \tag{2.59a}$$

$$\hat{a}\left|n_{\mathbf{q}\eta}\right\rangle = \sqrt{n_{\mathbf{q}\eta}}\left|n_{\mathbf{q}\eta}-1\right\rangle. \tag{2.59b}$$

Accordingly, there is only a non-zero expected strain tensor field associated with a superposition of phonon mode eigenstates that differ by one phonon, since all phonon mode eigenstates— associated with different values of $n_{\mathbf{q}\eta}$— are orthogonal. Note this implies that *the expected strain tensor field for any pure phonon eigenstate is zero*. Thus the carrier–phonon matrix element must be zero unless the initial and final phonon mode eigenstates differ by one phonon. For the non-zero matrix elements, the expected strain for the matrix element is given by eq. (2.60):

$$\left\langle n_{\mathbf{q}\eta}\pm 1\right|\hat{\mathbf{e}}\left|n_{\mathbf{q}\eta}\right\rangle = i\left(\frac{\hbar}{2\rho V\omega_{\mathbf{q}\eta}}\left(n_{\mathbf{q}\eta}+\tfrac{1}{2}\pm\tfrac{1}{2}\right)\right)^{\frac{1}{2}}\frac{\mathbf{q}\otimes\boldsymbol{\xi}_{\eta}+\boldsymbol{\xi}_{\eta}\otimes\mathbf{q}}{2}e^{\mp i\mathbf{q}\cdot\mathbf{r}}. \tag{2.60}$$

We note that all matrix elements are thus associated with a single Fourier component of the strain tensor field. In the case $\mathbf{q}\eta$–phonon creation, the Fourier component is associated with a $-\mathbf{q}$ wavevector. In the case of $\mathbf{q}\eta$–phonon annihilation, the Fourier component is associated with a $+\mathbf{q}$ wavevector. By combining eq. (2.60) with a version of eq. (2.56) that contains the modified overlap integral $\mathscr{I}^{\mathrm{mod}}(\mathbf{k}\nu,\mathbf{k}'\nu')$ we obtain the following:

$$\left\langle \mathbf{k}'\nu', n'_{\mathbf{q}\eta} \right| \hat{H}_{\text{car–pho}} \left| \mathbf{k}\nu, n_{\mathbf{q}\eta} \right\rangle = i \left(\frac{\hbar}{2\rho V \omega_{\mathbf{q}\eta}} \left(n_{\mathbf{q}\eta} + \tfrac{1}{2} \pm \tfrac{1}{2} \right) \right)^{\frac{1}{2}} \mathscr{I}^{\text{mod}}(\mathbf{k}\nu, \mathbf{k}'\nu') \delta_{\mathbf{q}\pm\mathbf{k}'-\mathbf{k}+\mathbf{G}} \times \left(\frac{\boldsymbol{\Xi}_\nu + \boldsymbol{\Xi}_{\nu'}}{2} : \frac{\mathbf{q} \otimes \boldsymbol{\xi}_\eta + \boldsymbol{\xi}_\eta \otimes \mathbf{q}}{2} \right). \tag{2.61}$$

Fischetti and Laux make the ISOTROPIC COUPLING CONSTANT, or ISOTROPIC BAND-DEPENDENT DEFORMATION POTENTIAL APPROXIMATION, which builds upon the band-dependent deformation approximation to also assume that the energy shift does not depend on the direction of $\mathbf{q}$:

$$\boldsymbol{\Xi}_\nu : \frac{\mathbf{q} \otimes \boldsymbol{\xi}_\eta + \boldsymbol{\xi}_\eta \otimes \mathbf{q}}{2} = \Delta_{\eta,\nu} |\mathbf{q}|. \tag{2.62}$$

This leads to the following expression for carrier scattering by acoustic phonons:

$$S_{\text{pho}}(\mathbf{k}\nu, n^{\text{ac}}_{\mathbf{q}\eta}; \mathbf{k}'\nu', n^{\text{ac}}_{\mathbf{q}\eta} \pm 1) = \frac{\pi}{\rho V \omega_{\mathbf{q}\eta}} \left(n_{\mathbf{q}\eta} + \tfrac{1}{2} \pm \tfrac{1}{2} \right) \left(\frac{\Delta_{\eta,\nu} + \Delta_{\eta,\nu'}}{2} \right)^2 \mathbf{q}^2 \left| \mathscr{I}^{\text{mod}}(\mathbf{k}\nu, \mathbf{k}'\nu') \right|^2 \times \delta(\varepsilon_{\mathbf{k}'\nu'} \pm \hbar\omega_{\mathbf{q}\eta} - \varepsilon_{\mathbf{k}\nu}) \delta_{\mathbf{q}\pm\mathbf{k}'-\mathbf{k}+\mathbf{G}^*}. \tag{2.63}$$

For the coupling of optical phonons to carriers via the strain operator, the argument differs only slightly. In optical phonons, the strain is typically dominated by the internal strain inside the unit cell, rather than the strain due to the changing size of the unit cell. Such a strain is not associated with a new unit cell, but with a new *basis*. By parallel reasoning to the case for a changing unit cell, we can again assume a change in carrier eigenstate energy that is linearly proportional to the strain. We will assume however, that in the case of optical phonons, this internal strain in a unit cell is directly proportional to the displacement operator, and thus we assume that the magnitude of internal strain is independent of the phonon wavevector. We note that the wavevector of the phonon still plays precisely the same role in determining the superposition of carriers that have a non-zero expected eigenstate energy shift. That is, the wavevector of the phonon still determines which superposition of carriers are capable of systematically increasing carrier probability density in low eigenstate energy positions and systematically decreasing density in high energy eigenstate positions. Thus, following a parallel analysis for optical phonons similar to the analysis for acoustic phonons, we are led to the following expression:

$$S_{\rm pho}(\mathbf{k}\nu, n^{\rm op}_{\mathbf{q}\eta}; \mathbf{k}'\nu', n^{\rm op}_{\mathbf{q}\eta} \pm 1) = \frac{\pi}{\rho V \omega_{\mathbf{q}\eta}} \left(n_{\mathbf{q}\eta} + \tfrac{1}{2} \pm \tfrac{1}{2}\right) \left(\frac{(\Delta q)_{\eta,\nu} + (\Delta q)_{\eta,\nu'}}{2}\right)^2 \left|\mathscr{I}^{\rm mod}(\mathbf{k}\nu, \mathbf{k}'\nu')\right|^2 \times \delta(\varepsilon_{\mathbf{k}'\nu'} \pm \hbar\omega_{\mathbf{q}\eta} - \varepsilon_{\mathbf{k}\nu}) \delta_{\mathbf{q}\pm\mathbf{k}'-\mathbf{k}+\mathbf{G}^*}. \quad (2.64)$$

In the case of optical phonons, the isotropic coupling constant is denoted as $(\Delta q)_{\eta,\nu}$, for no other reason than to emphasize the dimensional difference between this isotropic coupling constant and the acoustic isotropic coupling constant $\Delta_{\eta,\nu}$. If, for bands η corresponding to acoustic phonons, we define $(\Delta q)_{\eta,\nu} = \Delta_{\eta,\nu}|\mathbf{q}|$, then we have the following universal expression for acoustic *or* optical phonon scattering:

$$\boxed{S_{\rm pho}(\mathbf{k}\nu, n_{\mathbf{q}\eta}; \mathbf{k}'\nu', n_{\mathbf{q}\eta} \pm 1) = \frac{\pi}{\rho V \omega_{\mathbf{q}\eta}} \left(n_{\mathbf{q}\eta} + \tfrac{1}{2} \pm \tfrac{1}{2}\right) \left(\frac{(\Delta q)_{\eta,\nu} + (\Delta q)_{\eta,\nu'}}{2}\right)^2 \left|\mathscr{I}^{\rm mod}(\mathbf{k}\nu, \mathbf{k}'\nu')\right|^2 \times \delta(\varepsilon_{\mathbf{k}'\nu'} \pm \hbar\omega_{\mathbf{q}\eta} - \varepsilon_{\mathbf{k}\nu}) \delta_{\mathbf{q}\pm\mathbf{k}'-\mathbf{k}+\mathbf{G}^*}.} \quad (2.65)$$

2.5.9 The Overlap Integral

It is important to distinguish the ordinary and modified overlap integrals. The modified overlap integral, for which the phase factor $e^{i\mathbf{q}\cdot\mathbf{r}}$ is removed from the overlap integrand, is associated only with carrier–phonon scattering. The reason is that, in the case of phonons, the $e^{i\mathbf{q}\cdot\mathbf{r}}$ term is associated with a smooth variation of the Hamiltonian across a unit cell. However, this concept makes no physical sense when we use deformation potentials, which express the relationship between a change in unit cell (or basis) and the change in eigenstate energy. The fictitious variation associated with the $e^{i\mathbf{q}\cdot\mathbf{r}}$ term arises as a consequence of the continuum approximation, in which we associate the phonons with a continuous field of displacements, rather than a discrete field. To undo this effect of the continuum approximation, Thus instead of a pure continuum approximation, we removing this fictitious variation of the Hamiltonian within a unit cell resulting in a modified overlap integral. We have called this approximation, the *quasicontinuum* approximation.

For the modified overlap integral, we use the very simple approach of Ziman [44] refer-

enced of the original 1988 paper [16]. We begin with the fundamental definition of the modified overlap integral:

$$\mathscr{I}^{\text{mod}}(\mathbf{k}\nu, \mathbf{k}'\nu') = \int_{\Omega} u^*_{\mathbf{k}'\nu'} u_{\mathbf{k}\nu} e^{\,i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} \mathrm{d}\mathbf{r}. \tag{2.66}$$

Following Ziman, we make the strong assumption that carriers are free electrons, and as such $u_{\mathbf{k}\nu}$ is of uniform magnitude for any $\mathbf{k}\nu$, and is normalized over a unit cell. This leads to the following:

$$\mathscr{I}^{\text{mod}}_{\text{rigid ion}}(\mathbf{k}\nu, \mathbf{k}'\nu') = \frac{1}{\Omega} \int_{\Omega} e^{\,i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} \mathrm{d}\mathbf{r}. \tag{2.67}$$

We also assume that the integral over a Wigner-Seitz unit cell can be approximated by an integral over a sphere of radius a_{sph} which has the same volume as the unit cell. We orient the coordinate system such that the z axis along $\mathbf{k} - \mathbf{k}'$, and then transform the system to spherical coordinates:

$$\mathscr{I}^{\text{mod}}_{\text{rigid ion}}(\mathbf{k}\nu, \mathbf{k}'\nu') = \frac{3}{4\pi a^3_{\text{sph}}} \int_0^{a_{\text{sph}}} \int_0^{\pi} \int_0^{2\pi} e^{\,i|\mathbf{k}-\mathbf{k}'|r\cos\theta} r^2 \sin\theta \mathrm{d}\phi \mathrm{d}\theta \mathrm{d}r \tag{2.68}$$

Evaluating the integral integral over ϕ, and substituting $u = \cos\theta$, and $\Theta = |\mathbf{k} - \mathbf{k}'|r$ leads to the following:

$$\mathscr{I}^{\text{mod}}_{\text{rigid ion}}(\Theta_{\text{sph}}) = \frac{3}{2\Theta^3_{\text{sph}}} \int_0^{\Theta_{\text{sph}}} \int_1^{-1} -\Theta^2 e^{\,i\Theta u} \mathrm{d}u \mathrm{d}\Theta. \tag{2.69}$$

Where $\Theta_{\text{sph}} = a_{\text{sph}}|\mathbf{k} - \mathbf{k}'|$. Both integrals can now be computed trivially. The result is the following:

$$\mathscr{I}^{\text{mod}}_{\text{rigid ion}}(\Theta_{\text{sph}}) = \frac{3}{\Theta^3_{\text{sph}}} \left(\sin\Theta_{\text{sph}} - \Theta_{\text{sph}}\cos\Theta_{\text{sph}}\right). \tag{2.70}$$

It is worth explicitly noting that the limit of this equation as $\Theta_{\text{sph}} \to 0$ does not diverge as it might naively appear on first glance:

$$\begin{aligned}\lim_{\Theta_{\text{sph}}\to 0}\left[\mathscr{I}^{\text{mod}}_{\text{rigid ion}}(\Theta_{\text{sph}})\right] &= \lim_{\Theta_{\text{sph}}\to 0}\left[\frac{3}{\Theta^3_{\text{sph}}}\left(\left(\Theta_{\text{sph}} - \frac{1}{6}\Theta^3_{\text{sph}}\right) - \left(\Theta_{\text{sph}} - \frac{1}{2}\Theta^3_{\text{sph}}\right) + \mathcal{O}(\Theta^5_{\text{sph}})\right)\right] \\ &= 1.\end{aligned}$$

For the ordinary overlap integral, we simply use the same expression, except we replace $|\mathbf{k}-\mathbf{k}'|$ with $|\mathbf{G}|$. This leads to the following expressions for the overlap integrals:

$$\mathscr{I}^{\mathbf{G}}(\mathbf{k}\nu,\mathbf{k}'\nu') = \frac{3}{(a_{\text{sph}}|\mathbf{G}|)^3}\Big(\sin(a_{\text{sph}}|\mathbf{G}|) - (a_{\text{sph}}|\mathbf{G}|)\cos(a_{\text{sph}}|\mathbf{G}|)\Big),$$
$$\mathscr{I}^{\text{mod}}(\mathbf{k}\nu,\mathbf{k}'\nu') = \frac{3}{(a_{\text{sph}}|\mathbf{k}-\mathbf{k}'|)^3}\Big(\sin(a_{\text{sph}}|\mathbf{k}-\mathbf{k}'|) - (a_{\text{sph}}|\mathbf{k}-\mathbf{k}'|)\cos(a_{\text{sph}}|\mathbf{k}-\mathbf{k}'|)\Big). \tag{2.71}$$

In the 1995 update, Fischetti, Laux and Crabbé improve upon these expressions near band minima by using $\mathbf{k}\cdot\mathbf{p}$ theory. Once again, we will largely ignore such improvements in this thesis.

2.5.10 Scattering Parameters of Conduction Electrons in Silicon

In this section, we have discussed the scattering of non-equilibrium carriers in a non-polar semiconductor. However, this thesis concerned with the more specific case of the transport of non-equilibrium conduction electrons in silicon. In Table 2.1 we list all the give the values of parameters that are pertinent to the non-equilibrium conduction electron scattering in silicon, extracted largely from Fischetti and Laux's 1988 paper [16].

We note that having defined these parameters, if we now collect all the boxed equations given in this chapter, we form a closed semiclassical model of non-equilibrium transport in homogeneous silicon would be well-accepted by the field as being close to theoretically sound. For convenience, we refer to this model as a FULL BOLTZMANN TRANSPORT EQUATION, noting that the coupling to Poisson's equation is taken for granted. We note that there are also many other models that would qualify as a "full" Boltzmann transport equation, all that is required is that the model makes a reasonable effort to include a realistic scattering operator and bandstructure into the Boltzmann transport equation.

Quantity	Value
a	0.543 nm
a_{sph}	$a\left(\frac{3}{16\pi}\right)^{1/3}$
Γ	$\frac{1}{4\pi^3}$
ϵ	11.9 ϵ_0
ρ	2.328 g/cm^3
$\varepsilon_1^{\text{thr}}$	1.2 eV
$\varepsilon_2^{\text{thr}}$	1.8 eV
$\varepsilon_3^{\text{thr}}$	3.45 eV
R_1	6.25 x 10^{10}
R_2	3.0 x 10^{12}
R_3	6.8 x 10^{14}
$\hbar\omega_{\text{TA}}^{\text{max}}$	22.1 meV
$\hbar\omega_{\text{LA}}^{\text{max}}$	44.3 meV
$\hbar\omega_{\text{TO}}^{\text{max}}$	62.0 meV
$\hbar\omega_{\text{LO}}^{\text{max}}$	62.0 meV
$(\Delta q)_{\text{TA},\nu=1}$	1.2 eV x $\lvert\mathbf{q}\rvert$
$(\Delta q)_{\text{TA},\nu\neq1}$	1.7 eV x $\lvert\mathbf{q}\rvert$
$(\Delta q)_{\text{LA},\nu=1}$	1.2 eV x $\lvert\mathbf{q}\rvert$
$(\Delta q)_{\text{LA},\nu\neq1}$	1.7 eV x $\lvert\mathbf{q}\rvert$
$(\Delta q)_{\text{TO},\nu=1}$	1.75 x 10^8 eV/cm
$(\Delta q)_{\text{TO},\nu\neq1}$	2.10 x 10^8 eV/cm
$(\Delta q)_{\text{LO},\nu=1}$	1.75 x 10^8 eV/cm
$(\Delta q)_{\text{LO},\nu\neq1}$	2.10 x 10^8 eV/cm

Table 2.1: Table of empirical parameters pertinent to conduction electron scattering. In Fischetti and Laux's 1988 paper, $\hbar\omega_{\text{LA}} = 22.1$ meV and a $\hbar\omega_{\text{TA}} = 44.3$ meV, but this is likely to be a typographical error.

Chapter 3

State Of The Art

3.1 Introduction

The aim of this thesis is to model innately inhomogeneous semiclassical electron transport in a theoretically sound manner that is considerably less computationally intensive than the current state of the art. In the background chapter we have argued that the Boltzmann transport equation— subject to a full bandstructure and complex scattering operator— coupled to Poisson's equation is universally accepted as a theoretically sound model of electron transport in the semiclassical regime,[1] and that the chief assumption of the semiclassical regime in high field transport is that the external field must be approximately constant on the length scale of electron coherence. The problem with the Boltzmann–Poisson system is that it is infeasible to solve the full Boltzmann transport equation numerically without making further assumptions. In this chapter we review the state of the art with respect to simplifying this equation into a form that is feasible to solve numerically.

The reason that the full Boltzmann transport equation is manifestly infeasible to solve numerically is because the corresponding ELECTRON STATE— that is, the *minimal* set of

[1] Assuming magnetic fields are negligible. In the general case, the full Boltzmann transport equation needs to be coupled to Maxwell's equations and the Lorentz force equation.

time-dependent "unknown variables" (or degrees of freedom) associated with electrons— simply requires too much memory to describe. In the full Boltzmann transport equation, the electron state is defined by an arbitrary $6-$D scalar field. If we suppose each scalar in the field is a 64-bit (8-byte) number, and each dimension of the scalar field is discretized into 100 points, then the instantaneous state of an arbitrary $6-$D scalar field requires 100^6 8-byte numbers, or 8Tb to store.[2] Since solving a moderately complex equation requires frequent reading and rewriting of this memory— and manipulating the hard drive memory is very slow— these 8Tb would need to primarily be RANDOM ACCESS MEMORY (RAM) in order for the computation to be able to be completed in a reasonable time frame. This is not feasible with current computer architectures.

The electron state can obviously be simplified if we make the assumption that the device can be defined in terms of $1-$D or $2-$D doping field and boundary conditions. We will ignore this method of state simplification in this review. A sufficient reason for doing this is that more and more modern devices are inherently $3-$D in nature, meaning reducing the dimensionality is inherently unphysical. It is also instructive to mention a second, slightly more subtle point. The problem with solving the Boltzmann equation numerically is not *just* that the electron state is infeasibly large, it is *also* that the Boltzmann transport equation is also much more computationally expensive to solve than a pure partial differential equation with a similar size state. This is a consequence of the fact that while the Boltzmann transport equation is local in real space like a pure partial

[2]When comparing the computational resources for the various models considered in this chapter, our main consideration is the number of scalars required to define the discretized electron state. Essentially— with the exception of the ensemble Monte Carlo model which is unusually computationally expensive even with a relatively low-information electron state— we make the crude assumption that computational expense is roughly linearly related to the information/number of scalars required to define a discretized electron state. This is likely to generally underestimate the differences between models since the computational expense is likely to be *superlinear* in the size of the state— especially given that scattering is non-local in crystal momentum/energy space. Regardless of the exact details, the irrefutable point is that as we reduce the number of dimensions of the electron state from $6-$D (full semiclassical electron state) to $5-$D (high-order spherical harmonic electron state) to $4-$D (energy-dependent electron state) to $3-$D (macroscopic electron state), the computational expense of the model *will drop dramatically*, given the information required to define the electron state will change by between one or two orders of magnitude (i.e. require $10-100$ times fewer scalars) each time the electron state changes by $1-$D, and there will be a similar decrease in the number of simultaneous scalar equations to solve. The difference in computational expense between the detailed ensemble Monte Carlo models and the deterministic models described here is harder to quantify from abstract analysis, so estimates given are not reliable and sensitive to the exact details of the ensemble Monte Carlo algorithm. In the end, the author leaves the problem of measuring the exact difference in computational expense between all styles of model mentioned here for future work.

differential equation, it is *non-local* in crystal momentum space. That is, in a small time step, while the value of the occupation rate at a particular point in phase space is only affected by the occupation rate at *nearby points* in *real* space, it is potentially affected by the occupation rate at *all other points* in *crystal momentum* space because of scattering. This means that reducing the dimensionality of *crystal momentum space* decreases the number of computations that are necessary to solve the equation by orders of magnitude more than a similar reduction to the dimensionality of *real space*.

We will review the Ensemble Monte Carlo Approach[3], the Spherical Harmonic Approximation approach, the Elastic Relaxation Time approach, the Fokker Planck approach, and the *innately inhomogeneous* Macroscopic approaches. We describe each approach in terms of the simplifications made to the full Boltzmann transport equation *state function*. In addition, we discuss the simplifications made to the full bandstructure and full scattering operator. The models are then judged on the basis of the theoretical soundness of these simplifications and the solution speed of the resulting model.

3.2 The Ensemble Monte Carlo Electron State

The state simplification in the MONTE CARLO METHOD is usually described as following the phase-space trajectories of a random sample of particles [53]. This an excellent mental picture for understanding the method so long as one remains cognizant of the fact that the "particles" involved are the fictitious Bloch particles described in the background chapter. That is, it is important to remain aware that in the semiclassical regime, the behaviour of a *single* electron is dictated by a *distribution* of the "particles" whose trajectories we track using the Monte Carlo method. The qualifier ENSEMBLE refers to the fact that all "particles" need to be simultaneously tracked because the Boltzmann–Poisson system is non-linear, meaning that the evolution of a "particle" depends on the positions of other "particles" [48, 54]. In the simpler models, this non-linearity may not be modelled at all [55], of occur only indirectly through the long-range electron–electron

[3]Which, in fact, simplifies the spatial dependence in a manner that does not require assuming the dopant field and boundary conditions are $1-\mathrm{D}$ or $2-\mathrm{D}$.

interactions that are modelled by via Poisson's equation [43].[4] In more accurate models this non-linearity will also occur directly in the Boltzmann Transport Equation through short-range electron–electron scattering and degeneracy effects [48, 16, 6].

While the mental picture of following "particles" is extremely useful as a starting point, it is also useful to understand the Monte Carlo method in a slightly more abstract fashion. We first note that the $6-$D distribution function associated with the full Boltzmann transport equation can be viewed as a $6-$D list of weight scalars associated with a basis delta function at each point in real space and crystal momentum space. In the Monte Carlo method, we reduce the electron state function significantly by following the time evolution of only a random sample of the basis delta functions, where the scalar weight associated with each delta function is typically forced to be uniform.[5] The uniform weight of delta functions in the Monte Carlo sample means that the probability of a particular delta belonging to the sample is proportional to the distribution function weight at that point, and vice versa that the number of delta functions in a given region of phase space is proportional to an estimate of the average distribution function in that region.

The full Boltzmann transport equation dictates that the full $6-$D electron state is subject to two two types of time-evolution terms: the pure Hamiltonian evolution terms, and the scattering terms. The pure Hamiltonian terms, according to Liouville's theorem [36], describe a complicated incompressible "flow" on the set of delta functions, where the weight associated with each delta function is carried with it. Thus following the effect of the pure Hamiltonian terms on a delta function is simple: we just do exactly what we would do for a classical particle subject to the same Hamiltonian. The effect of scatter-

[4] This indirect coupling will model plasma oscillations if and only if the Poisson equation is updated significantly more frequently than the plasma oscillation frequency [16].

[5] In more sophisticated Monte Carlo models [43, 16, 53], the weight functions are typically *multi-step functions*, that are only locally uniform in phase space rather than globally uniform. The problem with a globally uniform weight function is that the relative statistical error in estimating the distribution function in a given region of phase space is inversely proportional to the size of the distribution function. This is problematic because in many devices unusually high energy electrons have outsized effects because they can overcome an energy barrier that otherwise stop a physical process from occurring. In such devices, a uniform weight function is inappropriate. A better approach is to have a much smaller weight function in these important but rarely occupied high-energy regions of phase-space, which results in so called STATISTICALLY ENHANCED REGIONS. Upon entering a statistically enhanced region in which the weight function is reduced by a factor M, a delta function is cloned into M independent delta functions, and upon exiting such a region, a delta function is deleted with a probability $\frac{1}{M}$.

ing on a delta function is a little more subtle. The scattering operator is a one-to-many function that spreads the weight associated with an delta function to many different delta functions. These delta functions all have the same real space coordinates, but not the same crystal momentum space coordinates. However, since the weight associated with all delta functions in the Monte Carlo sample is uniform by decree, scattering must preserve the number of delta functions if it is to preserve the total distribution function weight associated with the initial delta function. Therefore we treat the scattering operator as determining the statistics for an *indeterministic random process*, and thus model the evolution of the distribution due to scattering as an indeterministic one-to-one mapping of delta functions.

The Ensemble Monte Carlo process occurs as follows. The initial distribution function is randomly sampled to produce an initial sample of delta functions. These delta functions evolve according to both deterministic Hamiltonian evolution and indeterministic scattering. At any point in time, the number of delta functions in a given region of phase space can be assumed to be roughly proportional to the distribution function in that region. Various conditions can be used to define the interaction of delta functions with the boundaries of the device, the most common being reflecting boundary conditions and ohmic boundary conditions. In reflecting boundary conditions, the crystal momentum space vector of the delta function is reflected along an axis. In ohmic boundary conditions, an expected density of delta functions is defined for a given region of space, and delta functions are randomly deleted when there are too many, and randomly created from a lattice temperature distribution when there are too few.

The first problem the Ensemble Monte Carlo method addresses is that it massively reduces the memory requirements needed to solve the full Boltzmann transport equation. The reason is that the number of simultaneous delta functions needed for a physically sound simulation is relatively small, and so one can achieve the accuracy of an arbitrarily large sample of delta functions by compiling the results of an arbitrarily large number of smaller simulations. In the case of transient simulations, this compilation is achieved by literally running multiple independent simulations. In the the case of steady-state simulations, this compilation can be achieved simply by taking a long time-average of a sample of positions of the delta functions of a single simulation once the transients have disappeared. Either of these techniques is fine so long as the delta func-

tions in each of the simulations being compiled are subject to physically realistic evolution.

The number of delta functions required in order for the evolution of each to be physically realistic is not very large. Indeed, if the Boltzmann–Poisson system was linear, the evolution of a solitary delta function could be modelled in a physically sound manner. While it is true that the Boltzmann–Poisson system is non-linear, it is also true the non-linearity can be largely predicted from only the *real space distribution* of electrons; that is, simulating the precise *crystal momentum distribution* of electrons is unnecessary for modelling non-linear effects. This is crucial, because we can get a fairly accurate estimate of the spatial distribution of electrons— and therefore a physically accurate simulation— using fairly small sample ($\sim 10^4 - 10^5$) of delta functions [17]. Tracking 6 scalars that define the phase space position for each of $\sim 100,000$ delta functions only takes $600,000$ 64-bit numbers, or 4.8Mb, so the memory requirements for a valid Ensemble Monte Carlo simulation are very modest compared to those required for a full Boltzmann transport equation solution.

As stated previously however, the Boltzmann transport equation is not simply infeasible to solve due to memory issues, it is also very CPU intensive. The massive reduction in memory required does also offer a practical improvement to processing speed, since the smaller the memory requirements the more feasible it is to store the relevant data in locations closer to the CPU with faster read/write properties. However, the Monte Carlo approach also has disadvantages. Adding new additional sample points at random phase space is generally a suboptimal way to improve a mesh, and therefore the ensemble Monte Carlo method is theoretically expected to require *more* total CPU time in order to solve than a direct Boltzmann solution of similar accuracy.[6]

This leads to the second problem the ensemble Monte Carlo approach solves. It alleviates the need to estimate the full $6-$D distribution function *at all*. Typically the physical quantities of interest are not the full $6-$D distribution function, but certain weighted integrals over the crystal momentum states of the distribution function— most obviously densities and net currents. Therefore, we do not need our Monte Carlo simulations to reach the level of statistical accuracy required to define an accurate $6-$D distribution

[6] Assuming the speed of memory access is identical.

function, we only need the accuracy required to define the $3-$D weighted integrals of the distribution function. This reduces the accuracy of the simulation required by several orders of magnitude, and therefore the CPU requirements of the ensemble Monte Carlo approach by several orders of magnitude.

For a review of the early developments of the Monte Carlo approach to electron transport see Jacoboni and Reggiani [43]. This describes the initial implementation of the Monte Carlo approach to electron transport before the inclusion of very detailed scattering operators and bandstructures, which are described well in the book edited by Hess [53], and in the original DAMOCLES papers [16, 46, 6]. For a review of the early impact of the DAMOCLES model, see the 1996 review of the field by Fischetti and Laux [56]. This review describes the "standard model" that emerged for modelling electron transport in bulk silicon, including a detailed scattering operator and full band structure. For a slightly more recent developments aimed toward increasing the statistical accuracy of the basic algorithm, see the 2000 review by Kosina et al. [18]. Finally for a broad historical overview of the role of ensemble Monte Carlo approaches in TCAD, see the 2004 review by Fischetti et al. [13].

The strength of the ensemble Monte Carlo approach is its minimal memory requirements and the ease with which one can include arbitrary scattering operators and bandstructures. It is a truly theoretically sound approach to solving the Boltzmann transport equation. The weakness of the approach is that— even only when weighted integrals of the distribution function are computed and not the entire distribution function— it is vastly more computationally expensive to solve than macroscopic models, and therefore plays a limited role in TCAD applications.

3.2.1 The Spherical Harmonic Electron State

Unlike the Monte Carlo approach, the spherical harmonic approach involves a relatively orthodox simplification of the $6-$D scalar function state in the full Boltzmann equation. The spatial coordinates are left unchanged, and the crystal momentum dependence is expressed as a sum of spherical harmonic terms.

It is simplest to visualize the transformation of the crystal momentum dependent distribution function in terms of spherical harmonics in three steps. Let us examine the distribution function for a single band at a single point in space, which is a $3-$D scalar function of crystal momentum and band index (k_x, k_y, k_z, ν).

- First, let us apply a band-dependent affine transformation[7] to the crystal momentum coordinates designed to transform the equipotential surfaces into surfaces that are as close as possible to concentric spheres centred at the origin. In general, the only equipotential surfaces such a transformation can distort into perfect concentric spheres is a set of off-centre concentric ellipses [57]. The distribution function for a single band at a single point in space is now a $3-$D scalar function of Shifted And Stretched (SAS) crystal momentum (k'_x, k'_y, k'_z, ν).

- Second, let us transform this function into pseudo-spherical polar coordinates. If one uses the magnitude of SAS crystal momentum as the radial distance, then the resulting coordinate system is an ordinary spherical polar coordinate system and the transformation is always possible. However, this is much less physically sound than using the energy as the radial coordinate, since it is the energy that relaxes at a different rate to the crystal momentum. The problem with using the energy as the radial coordinate is that the distribution function is not necessarily a single-valued function of polar and azimuthal angle. Therefore, using the energy as the radial coordinate is only possible if the ray exiting the SAS crystal momentum origin at a each polar and azimuthal angle only crosses each constant energy surface once. That is, the coordinate $(\varepsilon, \theta_{\mathbf{k}'}, \phi_{\mathbf{k}'}, \nu)$ must refer to one, and only one point. If this is not the case, then the psuedo-spherical polar coordinates cannot be well-defined.

- Third, we express the distribution functions on each constant energy surface, defined as a scalar function on the $2-$D space $(\theta_{\mathbf{k}'}, \phi_{\mathbf{k}'}, \nu)$, using a truncated set of spherical harmonic coordinates. This is the step that actually reduces the size of the distribution function.

The strength of the spherical harmonic approach is manifested when the contours of

[7] That is, a linear transformation with an origin shift

the bandstructure are approximately concentric ellipsoids. In such a case, the affine transformation transforms these into concentric spheres centred on the origin, and the distribution function on each sphere is likely to be able to be modelled using only a few low order spherical harmonics. This means that the $2-$D scalar function associated distribution function on the constant energy surfaces at a single point in real space is reduced to a few scalars, and the $3-$D scalar function associated with the distribution function at a point is reduced to a few $1-$D scalar fields. Local scattering thus becomes a non-local mapping of only a few $1-$D scalar fields rather than of a $3-$D scalar field, and the entire electron state is reduced to a few $4-$D scalar fields. Thus, if the energy contours of the bandstructure can be approximated by concentric ellipsoids, the spherical harmonic approximation radically reduces the memory and computational effort required to solve the full Boltzmann transport equation by several orders of magnitude, while still being technically able to incorporate detailed scattering operators.

The weakness of this spherical harmonic approach is that its theoretical validity is predicated on the idea that constant energy surfaces of the bandstructure can be approximated by concentric ellipsoids. In non-equilibrium transport there is generally *no reason to expect this approximation to be remotely true*. While at low energies almost all bandstructures can be approximated as a set of concentric ellipsoids, this is seldom similarly true at higher energy. And one of the major difficulties nearly all modern device models must reckon with is that it is *normal* for electrons to have energies well above the energies that simple analytic approximations to the bandstructure are remotely accurate [6]. That is, in modern devices it is *normal* for electrons to have energies in which the constant energy surfaces are not remotely elliptical. The only general truth about the shape of these high constant energy surfaces is that they reflect the point symmetry of the underlying crystal; for instance, the high energy constant energy surfaces of silicon reflect the $48-$ fold dioctahedral point-symmetry associated with the silicon lattice. No matter what affine transformation one applies to the bandstructure, the resulting shape of the higher constant energy surfaces *are nothing like spheres*, and the distribution functions on these constant energy surfaces *has nothing to do* with spherical harmonics. Indeed, often the constant energy surfaces cannot even be parameterized using pseudo-spherical polar coordinates because the tuple $(\varepsilon, \theta_{\mathbf{k}'}, \phi_{\mathbf{k}'})$ defines zero or multiple points [58].

The constant energy contours of valence bands are typically easier to parameterize us-

ing pseudo-spherical coordinates. Therefore, the spherical harmonic expansions approach to *hole* transport in silicon has been successfully implemented with a full valence bandstructure [59]. Unfortunately this does not change the fact that at high hole energy the contour surfaces reflect the dioctahedral symmetry of silicon. Therefore, it is found that *hundreds* of spherical harmonic terms are needed for an accurate description of the distribution function [59].

For the conduction band, where typically the actual constant energy surfaces at high energies cannot be parameterized using spherical polar coordinates, a different approach must be taken. The best way forward is to redefine the bandstructure using concentric ellipsoids which are separated by energies *that are defined such that the density of states of the actual bandstructure is reproduced*. If the bandstructure is simultaneously redefined such that the average speed of states at a given total energy is reproduced, the results can be quite accurate for a small number of spherical harmonics [60]. It is far from clear that this approach is internally consistent however, and it is unclear what theoretical foundation these approximations are founded on as the resulting distribution functions in crystal momentum space do not remotely resemble the actual distribution functions in crystal momentum space. We believe this is a contrived, superfluous attempt to insert spherical harmonics into a theory based that can be based purely on the *elastic relaxation time* approximation, which we describe in the next section.

3.3 The Energy-Dependent Electron State

3.3.1 Elastic Relaxation Time Approach

Given the theoretical weakness of the spherical harmonic approach in complex bandstructures, it is surprising how popular the approach is. This is especially true given the fact that a reduction of the electron state to an energy-dependent distribution function can be achieved in a theoretically sound way.

The proper technique was introduced in a 1992 paper by Dmitruk et al. [21], which was

also used in the work of Vecchi and Rynan [61]. Rather than arbitrarily assuming a spherical harmonic expansion, the approach begins instead with splitting the distribution function into an odd and even part. The even distribution is further split into an energy-dependent part, and an even anisotropic part. The effect of scattering on the *odd part* and the *even anisotropic part* of the distribution function is to make it decay in a relaxation time— which we refer to in this thesis as the ELASTIC RELAXATION TIME. The energy dependent part of the scattering operator is then left to be described by a more detailed scattering model.

It is notable that a $4-$D electron state model of non-equilibrium electron transport can be derived from the full Boltzmann transport equation essentially using *only the elastic relaxation time approximation*. The fact that this technique has *nothing to do with spherical harmonics* seems not to be properly appreciated in the literature [62, 60, 63]. The ELASTICALLY-CONSTRAINED EQUILIBRIUM model we present in the Results I chapter of this thesis is an attempt to isolate, clarify and expand upon this powerful and underappreciated elastic relaxation time approximation of Dmutrik et al. In contrast to the simple scattering operator of Dmutrik et al., the scattering operator we use to define the transport parameters is the DAMOCLES style [16, 46, 6] scattering operator introduced in the Background chapter.

The strength of this approach is that, with a single well-justified approximation, the electron state of the full Boltzmann transport equation is reduced by two dimensions. That is, rather than having to model the evolution of $f(\mathbf{k}\nu, \mathbf{r}, t)$, we only have to model the evolution of $f(\varepsilon_{\mathbf{k}\nu}, \mathbf{r}, t)$. One weakness of this approach is that, while faster than ensemble Monte Carlo, the approach is still more than an order magnitude more computationally expensive to solve than most macroscopic models owing to the simple fact that the electron state function is still a dimension larger than the $3-$D macroscopic electron state function, and that scattering is still non-local in energy. However in the view of this author, the real weakness is simply a lack of proper recognition by the field: it is not recognized that the approach is fully independent of the spherical harmonic approximation, it is not recognized that the approach is capable of incorporating arbitrary band structure and scattering operators, and it is not recognized that the approach has a theoretically sound, intuitive basis.

3.3.2 Fokker–Planck Approach

Other models have been proposed in which the electron state is characterized by an energy-dependent scalar function. Apart from the first-order spherical harmonic expansions already criticized, there is the Fokker–Planck approach pioneered by Bringuier [64, 65] and later developed into a complete device simulator by Kolobov [66]. The basic model underlying this is simple and physically intuitive: electrons undergo $4-$D drift–diffusion in real space *and* energy space. Accordingly, the electron state is similarly a simple $4-$D scalar field and– unlike the elastic relaxation time approach— inelastic scattering is always local in energy.

The strength of the Fokker–Planck approach is that it is physically intuitive, is capable of including the effects of an arbitrary bandstructure, and is fast— being only an order of magnitude more computationally expensive to solve than a macroscopic transport model, owing to the fact that the scattering operator is local in energy [66]. The fundamental theoretical weakness of the approach is that it is innately incapable of modeling highly inelastic scattering processes. Thus the model cannot take into account important highly inelastic processes such as impact ionization and electron–electron scattering. Secondly, even in cases where the energy exchanged is small, we argue that the present form of the Fokker–Planck model is not theoretically ideal. The Fokker–Planck approach of Bringuier was not derived from the full Boltzmann transport equation using well articulated assumptions, but instead is presented as a distinct formalism [64, 65]. If instead, as will be shown later in this thesis, one attempts to derive a model similar to the Fokker–Planck model by formally separating the elastic and inelastic components of scattering, the resulting model has *pure diffusion* at constant *total* energy which is driven by the gradient in *occupation rate* rather than the gradient in particle density. If inelastic scattering is limited to events that involve small energy exchanges with a lattice, the result is again *pure diffusion* in energy space that is driven by the gradient in *lattice-temperature chemical potential* rather than by the gradient in particle density.[8] However, these can be seen as minor technical flaws of the *current iteration* of the Fokker–Planck approach, rather than fundamental weaknesses of the main idea. The only true fundamental weakness of the Fokker–Planck approach is the fundamental to inability accurately model highly inelastic processes such as impact ionization and

electron–electron scattering.

3.4 The Macroscopic Electron State

In the introduction, we discussed in relative detail *quasi-homogeneous* macroscopic models which are theoretically sound in a well-defined regime. In this literature review, we focus on macroscopic models which have been proposed which claim to be accurate in the *innately inhomogeneous* regime.

All INNATELY INHOMOGENEOUS MACROSCOPIC MODELS share the simplification of state that is characteristic of macroscopic models. In macroscopic models the electron state is defined by a small number of $3-$D scalar fields, that are defined by simple weighted integrals of the distribution function over crystal momentum. Examples of the resulting macroscopic quantities are particle density, average square velocity, average energy, average square energy etc. All macroscopic models share a similar strength that— due to this very simple electron state— they are very fast to solve. And all innately inhomogeneous macroscopic models share a similar weakness— that they are not theoretically sound. In order to understand why this is the case, we need to look at the derivation of macroscopic models in more detail.

The first step of deriving a macroscopic model of non-equilibrium transport is to use Bløtekjær 's technique of taking various weighted integrals of the Boltzmann transport equation over all crystal momentum states and all conduction bands [10]. This solves the problem of defining rigorous macroscopic transport equations. However, as stated in the introduction, this process always creates a larger set of unknowns than the set of equations. As a result, the equation set is open, and needs to be closed by specifying additional relationships between the set of macroscopic unknowns. Defining these ad-

[8]The importance of this is relatively easy to understand if one thinks of the end-state equilibrium these diffusion processes are driving toward. In the case of diffusion at constant total energy, the irreversible diffusion process is attempting to equalize the occupation rate at constant total energy. In the case of the inelastic scattering, the diffusion process is attempting to equalize the lattice temperature chemical potential. When the density of states is a complex function of energy, these entropy-increasing pure diffusion processes have no simple link to gradients in particle density, and no simple link to the drifting of particle density.

ditional relationships for macroscopic transport coefficients, is known as the CLOSURE PROBLEM of macroscopic models. The *innately inhomogeneous* closure techniques can be roughly categorized into two types.

Inhomogenous Simulation Based Closure : In this type of closure, the transport coefficients are derived from a set of inhomogenous simulations.

Ansatz Based Closure : In this type of closure, an ansatz is proposed for the distribution function and on the basis that ansatz the transport model is closed without reference to inhomogeneous simulation data.

We will briefly review both closure approaches.

3.4.1 Inhomogeneous Simulation Based Closure

Inhomogeneous simulation based closure is a closure technique similar to the quasi-homogeneous closure technique, in that excess variables are removed by extracting parameters from a set of Monte Carlo simulations. There are, however, significant differences in using *inhomogeneous* simulation data to close an equation set. The set of possible *homogeneous* simulations is naturally restricted, and each simulation is mapped to a unique single set of macroscopic densities; for example, in a given material there is only one homogeneous simulation that will produce a particular local particle and energy density. In contrast, the set of possible *inhomogeneous* simulations is effectively infinite, and furthermore a *range* of particle and energy densities are associated with *each* simulation. This means that— unlike in case of quasi-homogeneous simulation closures— there is no known set of macroscopic densities that has a one-to-one mapping with the approximate distribution function specified in a particular inhomogeneous simulation. Instead, the most obvious sets of macroscopic densities such as a *particular specification* of particle density, energy density and a few higher order moments can generally be associated with a *large range* of very different inhomogeneous simulations, implying a *large range* of very different distribution functions underlying these macroscopic densities. This implies that the remaining transport parameters used to close the model— associated with integrals of the distribution function that are weighted by terms that de-

pend on the complex bandstructure and/or scattering operator— never have any solid theoretical reason for being single-valued functions of the macroscopic properties used to close the model. In every innately inhomogeneous closure proposed, this fundamental theoretical problem is hidden somewhere. This basic theoretical issue was noted more than twenty years ago by Fischetti and Laux on their discussion of these models, and has still not been properly addressed [6].

We can demonstrate this theoretical problem for a few examples of popular innately inhomogeneous closures. The most popular innately inhomogeneous closures of macroscopic models generally share the following two properties.

1. Hot carrier effects— defined as effects which involve a high threshold energy, such as impact ionization— are determined by coupling to an independent model designed only to determine the density of electrons above the threshold energy.

2. They are constructed to be consistent with *homogeneous* Monte Carlo simulations [9].

The first property regarding the independent modelling of hot carrier effects has resulted in the trend that it is typical to publish models of impact ionization independently from the more general transport model, due to the very loose coupling between these models.[9] Generally, the most that is required from the macroscopic model is a position-dependent electron density, which must be generated by *every* macroscopic model in order to close Poisson's equation. Examples of broadly applicable innately inhomogeneous impact ionization models are the models of Ridley [42], Meinerzhagen [67], Schrobohaci and Tang [68], and Ahn et al. [69]. The exception to this rule is the class of innately inhomogeneous impact ionization models that *can only be coupled* to a six-moment model of transport. The models tend to define the impact ionization via an ansatz for the distribution function which requires all the local moments defined by the six moment model [70, 71]. Beyond this requirement however, the impact ionization models are still *independent* of the macroscopic transport model since the ansatz used to define the impact ionization rate typically has *no relation* to the assumptions used to close the six moment transport model [72, 11, 14].

[9] By "loose coupling" in this context, we simply mean that we can match a wide range of impact ionization models with a wide range of macroscopic models of electron transport.

Regarding the second property, consistency with homogeneous Monte Carlo data is achieved using the same abstract method [9, 7, 15]. One takes one or more quasi-homogeneous closure relations, adds a term or two which is always zero in a homogeneous simulation, and weights each of these terms by an undefined scalar the value of which can be determined by comparing to a set of inhomogeneous simulations. The details and justification for the precise alterations made to the quasi-homogeneous closure relations differ greatly between various models closed using inhomogeneous simulation data, but beneath these details the basic trick used is always the same. Similarly, the theoretical criticism of such approaches is always the same: there are *no theoretically compelling reasons* why the tuning parameters added to the model are expected to be single-valued scalars across a range of inhomogeneous devices. Accordingly, the models produced by this closure method are unreliable when used on devices that are not similar to those used in the set of inhomogeneous Monte Carlo simulations used to tune them.

An example of a very simple, transparent application of the core principle used to build an innately inhomogeneous macroscopic model is given by Bork et al [73]. These authors built upon the quasi-homogeneous macroscopic model of Thoma et al. [74]. As is common to many quasi-homogeneous macroscopic models, in the model of Thoma et al., there is a HEAT FLUX term in the definition of energy flux which is proportional to the gradient in average electron temperature.[10] In a homogeneous simulation this heat flux term is zero due to the average temperature being a constant function of position. As such Bork et al. transformed the quasi-homogeneous model of Thoma et al. into an innately inhomogeneous macroscopic model simply by multiplying the heat flux term by a tuning scalar. The value of the tuning scalar could then be determined by comparing the net current predictions of the macroscopic model with various values for the tuning scalar, to the predictions made by Monte Carlo simulation, for nMOSFETs with channel lengths of $100 - 500$nm.

While the approach of Bork et al. is manifestly empirical, there have been much more thoughtful efforts to incorporate inhomogeneous simulation data into a macroscopic closure. An example is the model of Tang et al. [75]. We note that in general, the greater

[10] This term is consistent with quasi-homogeneous assumption, since in a uniform field the distribution function is typically close to a thermal distribution [6]. We know from basic transport theory that in a distribution that is thermal, gradients in the temperature drive both particle and energy currents.

the number of data points fitted, and the fewer the number of empirical tuning parameters, the stronger the circumstantial evidence that the model is robust. According to this metric, Tang et al. produce an excellent model of both the (first order) stress-energy tensor, and the second order stress-energy tensor (which is more commonly known as the fourth order moment).[11] Their model of the stress-energy tensor as a function of average energy requires *zero* inhomogeneous tuning parameters and is shown to be consistent with the stress energy tensor found at hundreds of average energies found at various points in three n^+nn^+ diodes of different lengths. Their model of the second order stress-energy tensor requires *one* inhomogeneous tuning parameter to fit the same data.[12] Since Lee and Tang earlier used similar inhomogeneous data as evidence that the energy relaxation time is a single-valued function of average energy that can be determined without inhomogeneous tuning parameters [76], this reduces the problem of closing the macroscopic model in an innately inhomogeneous manner to modelling the (particle flux) mobility and energy flux mobility.

In order to model particle and energy flux mobilities, Tang et al. relied on the usual empirical technique of adding terms to the quasi-homogeneous closure that are zero in a homogeneous field. The scattering term associated with particle flux mobility— the momentum scattering term— was assumed to be equal to the homogeneous scattering term at the same average energy, plus a term proportional to the divergence in stress-energy tensor. The scattering term associated with energy-flux mobility— the momentum-energy scattering term— was assumed to be equal to the homogeneous scattering term at the same average energy, plus a term proportional to the divergence in the second order stress-energy tensor. The constants of proportionality in each of these relations are assumed to be piecewise functions, each of which is determined by *two* scalars. The net result of these assumptions in the resulting model is simply that

[11] The STRESS-ENERGY TENSOR is defined by $\hat{U} = \langle \mathbf{v} \otimes \mathbf{p} \rangle$, where $\mathbf{v}$ is velocity, $\mathbf{p}$ is crystal momentum and $\langle\rangle$ is the distribution function weighted average over all bands and crystal momenta. The SECOND-ORDER STRESS-ENERGY TENSOR or "fourth order moment" is defined by $\hat{R} = \langle \varepsilon \mathbf{v} \otimes \mathbf{p} \rangle$ where ε is (kinetic) energy. These represent the net flux of *crystal momentum* carried by the current, and the flux of *energy-momentum* carried by the flowing electrons. These are tensors since the crystal momentum can be in a different direction to the current.

[12] One slightly confusing issue in the paper of Tang et al. is that the stress energy tensor and second order stress energy tensor are plotted as if they are scalars [75, Fig. 5,7]. Furthermore, the divergence in these quantities is talked about as being "greater or less than zero", even though it is technically a vector. This only seems to make sense if these first and second order stress-energy tensors can always be written down as the product of a dynamic scalar and a fixed invariant matrix, such as the identity matrix. And yet such a formulation does not seem to be consistent with the assumptions that Tang et al. make.

both the heat flux and the thermodynamic particle flux are weighted by two empirical tuning parameters each. The evidence that these four empirical tuning parameters are indeed single valued across the devices investigated is equivocal [75, Fig. 13,14]. Accordingly, the robustness of the model is compromised by the fact that only a moderate number of data points have been reproduced from a comparatively large number of empirical tuning parameters.

Finally, there is scant evidence at all that *any* of the assumptions made in the model are true in devices *other* than nn^+n diodes. Nevertheless, the techniques used in this paper are particularly admirable for a single reason: compared to Bork et al., the authors showed how one may extract an *enormous amount* of tuning data from just *three* inhomogeneous Monte Carlo simulations. Given that each high quality Monte Carlo simulation is extremely computationally expensive, this is an important advance in reducing the computational cost of tuning empirical models.

3.4.2 Ansatz Based Closure

The terms required to close a macroscopic model are integrals of the distribution function weighted by bandstructure and/or scattering operator dependent terms. In cases where the bandstructure and scattering operator are complicated, it is essentially impossible to directly define a theoretically sound relationship between the macroscopic densities and these weighted integrals unless one can also propose a theoretically sound relationship between the macroscopic densities and the distribution function. This requires an ansatz: an assumption about the functional form the distribution function takes.

The classic starting point for the ansatz used to close the distribution function is the expression of the distribution function as a sum of a zeroth order and first order spherical harmonic [8]. However, the fundamental theoretical unsoundness of assuming the distribution function can be expressed in terms of low order spherical harmonics in complex bandstructures has already been discussed. A much better starting point is the elastic relaxation time approximation. This is the starting point used by Chen et al. [77], and used by Vecchi and Reyna [61].

As is the case outside the macroscopic closure literature, the elastic relaxation time approximation is generally misunderstood inside the macroscopic closure literature. This can occur by being conflated with the *far stronger* relaxation time approximation that the distribution relaxes to a thermal distribution in a relaxation time [9], or by being otherwise believed to only be theoretically valid in unnecessarily restricted circumstances [75]. A failure to properly understand the elastic relaxation time approximation is even evident in the original paper of Chen et al., which contains the following non-sequiter:

> "For materials like Si, where the mean kinetic energy[13] $\frac{1}{2}m\left\langle v\right\rangle^2$ is much smaller than the mean thermal energy $\frac{1}{2}m\left\langle v^2\right\rangle$, the correction from the field-induced anisotropy is generally not important. Therefore, [the even part of the distribution function] f_0 can be approximated by an isotropic distribution $f_0(\varepsilon_{\mathbf{k}\nu})$ and [the macroscopic transport parameters] *become scalar functions of the mean carrier energy* $\langle\varepsilon_{\mathbf{k}\nu}\rangle$." [emphasis added on non-sequiter]

This quote suggests that the authors conflate the assumption that the even distribution function is an arbitrary energy-dependent function $f_0(\varepsilon)$ with the far, far stronger assumption that the even distribution function is an *average* energy dependent function $f_0(\langle\varepsilon\rangle)$.

The elastic relaxation time approximation alone is obviously not enough to close a macroscopic model. One must also make exactly the kind of mapping Chen et al. describe, and define the even distribution function in terms of the even macroscopic moments. Unlike the elastic relaxation time approximation which is under-appreciated, the assumptions made about this part of the ansatz are correctly acknowledged as being very poor [9].

Vecchi and Reyna use the distribution function in a homogeneous field which has the same average energy as the actual distribution function [61], making their closure an example of the theoretically sound quasi-homogeneous closures described in the Introduction chapter. Stratton used a heated Maxwellian [8]. Chen et al. used the product of a heated Maxwellian and an arbitrary linear function of energy [77]. No convincing case has ever been made for why any of these are theoretically sound descriptions of the even

[13] Note that in this thesis, "mean kinetic energy" refers to $\langle\varepsilon_{\mathbf{k}\nu}\rangle$, and not the quantity given here.

part of the distribution function in highly inhomogeneous devices, and empirical results suggest this is because they are simply inaccurate [6]. Despite this empirical evidence to the contrary, crude ansatz such as a heated Maxwell-Boltzmann distribution have often been theoretically justified on the basis that they maximize the entropy subject to the constraints provided by the macroscopic densities [78, 79], making them the "least bias" way to close a macroscopic transport model [80]. Unfortunately, there is no theoretically sound reason why ignoring the microscopic physics underlying non-equilibrium electron transport in the "least bias" manner possible will lead to an accurate model: nature will act in a manner that is most certainly biased by the microscopic physics [81]. It is the perspective of this thesis that arguments on the basis of entropy maximization are only theoretically sound when one can point to a *physical mechanism* that actually acts to maximize the entropy that is assumed to be maximized. But even ignoring this viewpoint, Jaynes himself— the father of the Maximum Entropy Principle— always maintained that any maximum entropy ansatz which contradicts empirical evidence must immediately be abandoned [80].

One of the few attempts to capture innately inhomogeneous effects in a physically sound manner is the model of Bordeleon et al. [82]. In their model, the even distribution function is assumed to be the sum of a lattice temperature Maxwellian— associated with the low energy population of electrons approximately in chemical equilibrium with the drain terminal— and a quasi-homogeneous distribution characterized by an average energy— associated with the electrons from the source terminal. Unfortunately, there is no theoretically sound reason for why the part of the distribution function associated with electrons originating from the source terminal should have a quasi-homogeneous distribution. In addition, the model ignores electron-electron scattering events which mix the source and drain populations. The THREE EQUILIBRIUM MODEL of non-equilibrium transport we propose in the second results chapter can be considered as an attempt to address both of these concerns with the simple two population closure of Bordeleon et al.

Chapter 4

Theoretical Framework

4.1 Introduction

In the Background chapter we derived a valid model of innately inhomogeneous non-equilibrium transport in bulk silicon: a Boltzmann transport equation, subject to a full band structure and scattering operator. In the State Of The Art chapter we then argued that the rough consensus of the TCAD field is that one must choose one of the following paths.

High Computational Cost: Accept the computational price for solving the full Boltzmann transport equation in a stochastic manner via full-band detailed ensemble Monte Carlo.

Intermediate Computation Cost: Accept the memory price for a model that is a little faster to solve[1] based on a truncated spherical harmonic expansion, which is physically dubious in complex bandstructures.

Low Computational Cost: Accept the serious unphysical assumptions associated with macroscopic models that are much faster to solve, and mitigate their impact on accuracy via empirical tuning.

We agreed with the consensus regarding detailed ensemble Monte Carlo simulation. Detailed programs such as DAMOCLES [16], are capable of producing accurate simulations for a high computational cost.

We disagreed with the consensus regarding spherical harmonics. We argued that basing a transport model on a spherical harmonics expansion is inappropriate given the complex bandstructure of silicon, and that a much better solution is to base a theory purely on what we called the *elastic relaxation time* approximation which makes a spherical harmonic expansion superfluous. We argued that this viewpoint is consistent with a short paper written in 1992 by Dmitruk et al. [21], but is inconsistent with the vast body of more modern literature that endorses the spherical harmonic expansion approach [83, 84, 62, 85, 86, 87, 63, 59, 58].

We agreed with the consensus regarding currently proposed macroscopic models, but we argued that this is not fundamental, and suggested that a theoretically-sound closure of a macroscopic model in the innately inhomogeneous regime is possible if it combines the elastic relaxation time approximation with a theoretically-sound ansatz for the energy distribution function. We argued that Bordelon et al. were on the right track with their two population ansatz [82], but that this ansatz fell short of theoretical soundness.

The aim of the remainder of this thesis is to define our proposed theoretically-sound replacements for the intermediate computational cost model and the low computational cost model. This is the ELASTICALLY-CONSTRAINED EQUILIBRIA MODEL, and the THREE QUASI-EQUILIBRIA MODEL respectively. Both ansatz rely on the existence of presently unrecognized, atypical forms of quasi-equilibrium. The aim of this theoretical framework chapter is to define these equilibria and to explain why the Boltzmann transport equation drives the electron distributions toward them.

[1]The memory and computational costs obviously depend on the order of the spherical harmonics expansion used. The computation costs are much smaller than an equivalent ensemble Monte Carlo for very low order expansion, such as that proposed by Vecchi and Rudan [60], but will often be higher than an equivalent ensemble Monto Carlo for high order expansions such as that proposed by Jungemann et al. [59].

4.2 On Simplifying the Electron State

We define a VALID ELECTRON STATE FUNCTION as a set of degrees of freedom that if specified at time t, provide sufficient information about electrons to predict the electron state function at a later time $t + \Delta t$ in the class of devices the transport model is used for. This definition of state might naïvely seem tautological and empty, but on the contrary *almost all choices of electron state function are invalid* since predicting the future of an arbitrary hypothetical state function almost always requires information about electrons that is not contained in the state function. This is precisely the problem that manifests when previous authors have tried to use a position-dependent particle and energy density to predict the future values of position-dependent particle and energy density in innately inhomogeneous non-equilibrium transport.

Recognizing this, simplifying the state function in the semiclassical regime then appears to be difficult because the reproducible Hamiltonian terms in the Boltzmann transport equation move all degrees of freedom in the $6-$D distribution function $f(\mathbf{k}\nu, \mathbf{r}, t)$ along different phase space vectors in a short time step. According to Hamilton's equation, these phase space velocity vectors are defined by the $6-$D gradient partial derivative of the Hamiltonian *rotated* 90° in each $2-$D subspace defined by each canonical conjugate pair of coordinates $\hat{\mathbf{x}}_i \times \hbar\hat{\mathbf{k}}_i$.[2] Notice that this means that the product of the velocity along a phase space coordinate and the directional gradient along that component will always be equal in magnitude and opposite in sign to the product of these two factors along the canonical conjugate coordinate. Since this product is equal to the partial time derivative of the Hamiltonian, any rate of change in the Hamiltonian associated with the movement along a phase space coordinate will *always* be compensated for by an equal and opposite change in the Hamiltonian associated with the movement along the canonical conjugate phase space coordinate. These dynamics ensure that the only changes in the Hamiltonian at any point along a phase space trajectory are those associated with the explicit time-dependence of the Hamiltonian.

[2]This is a very useful— but not immediately obvious— geometric way to understand Hamilton's equations of motion. The partial derivative defined by $\frac{\partial H}{\partial \hbar k_i}$ is rotated 90° in the $(x_i, \hbar k_i)$ plane to define the velocity along x_i, and partial derivative defined by $\frac{\partial H}{\partial x_i}$ is rotated 90° in the $(x_i, \hbar k_i)$ plane to define the velocity along $-\hbar k_i$.

To put it another way, each possible reproducible Hamiltonian is associated with a unique, intricate pattern of "stirring" it performs on the full phase space distribution function as time increases. This stirring occurs by moving the distribution function values at each point in phase space along a $5-$D collection of $1-$D trajectories defined by placing the phase space velocity vectors end-to-end. The degrees of freedom— the distribution function scalars attached to each position in phase space— are never averaged in any way by this reproducible Hamiltonian-based "stirring", just *moved*.[3] In a finite time, the effect of this stirring is to permute the set of scalars associated with each point in phase space, and as this finite time tends to zero this permutation smoothly becomes arbitrarily close to the identity operation. The complexity[4] of this phase-space stirring is defined by the complexity of the phase space gradient of the reproducible component of the Hamiltonian.

This complexity is limited by the fact that the spatial component of the phase-space gradient is not a function of $\mathbf{k}$, and the crystal momentum component of the phase space gradient is not a function of $\mathbf{x}$. However, beyond this restriction, there is little we can say. The spatial velocity as a function of $\mathbf{k}$ is innately complicated due to the innate complexity of the bandstructure. The crystal momentum velocity as a function of $\mathbf{x}$ is also relatively complicated due to fact the electric potential is a fairly arbitrary $3-$D function of position. Without enormous simplicity of either the spatial or crystal momentum gradients, the effect of each becomes tightly intertwined, and two particles starting arbitrarily close in phase space can become separated by large distances in phase space over finite times, due to a positive feedback loop encoded into Hamiltonian dynamics. For instance, two particles with the same crystal momentum, at slightly different points in space will gain crystal momentum at slightly different rates, leading to a difference in spatial velocity, leading to larger differences in space, leading to larger differences in spatial velocity.

Since the complicated bandstructure and position-dependent electric potential dictate that the $5-$D space of trajectories all evolve differently from one another, a simplification of the semiclassical electron state function is only possible if we can approximately recreate the *entire distribution function* from a smaller set of degrees of freedom. This recreation needs to be universal in the sense that the mapping needs to work across a

[3] This is Liouville's theorem.

[4] Synonym for "complicatedness". Nowhere in this thesis do we refer to complex numbers.

large range of devices. That is, we require that the distribution function is *universally* constrained to a sub-domain of the full six-dimensional scalar function space that can be relatively simply described. But even if we limit the boundary or initial distribution function to a *simple universal* subdomain of the full distribution function, the "permutation" of the scalars over finite times associated with the reproducible Hamiltonian terms will change the initial simple universal function space into one which is incredibly complicated and *tightly coupled to the precise reproducible Hamiltonian associated with a given device*. The reproducible Hamiltonian terms cannot therefore be the basis for universal simplification of the electron state.

We have described then, why it is the *very nature* of the reproducible device Hamiltonians in a class of devices to drive the distribution function *away from* universal subdomains of $6-$D scalar-function space that might serve as simpler state functions. Thankfully then, it is also in the *very nature* of the non-reproducible device Hamiltonians to drive the distribution function *toward* universal subdomains of the $6-$D scalar-function space.

As described in the background chapter, the non-reproducible device Hamiltonians are associated with local scattering. This local scattering is an irreversible process, meaning that it maps a large set of input distribution functions to a smaller set of output distribution functions. Thus it formally erases degrees of freedom from the electron state. One simple way to understand this is that local scattering causes a complicated "averaging" process of a subset of the local distribution function values, and the weighted sum of several degrees of freedom results in one degree of freedom. The caveat of this is that since there is a complicated average— or unique weighted sum— associated with every crystal momentum state (and band index), it is formally possible for the effect of the scattering operator to be reversible if the complicated averaging process associated with each crystal momentum state is linearly independent of the complicated averaging process associated with every other crystal momentum state.[5] This however, is easily shown not to be the case for the scattering operator.

[5]For simplicity we are describing cases where the scattering operator is a linear transformation of the distribution function. However the fundamental argument we are making is also true for electron–electron scattering. In fact, for high average energy distributions electron–electron scattering is the *most* irreversible scattering mechanism in its average scattering time, since it relaxes the distribution to a drifting internal thermal equilibrium on the timescale of one scattering time.

The product, or sequential application, of a set of reversible transformations is another reversible transformation. And yet, if we re-apply the scattering operator for long enough[6] to *any* distribution function, in the absence of an external force the end result will *always* be a lattice temperature equilibrium distribution. That is, the end distribution will be inside a function space characterized by a *single scalar*— the local chemical potential— rather than a $6-$D *field* of scalars. Our model of how this incredibly irreversible mapping occurs is that in each average scattering time, the distribution function maps the set of input distribution functions to a *significantly smaller* function space of output distributions, until lattice temperature equilibrium is reached. The aim of this chapter is primarily to understand and accurately approximate the function space the distribution function is reduced to on the timescale of a *single* average scattering time in semiclassical silicon devices.

We refer to the reduced function space the full distribution function is driven toward on the order of a single scattering time as ATYPICALLY CONSTRAINED QUASI-EQUILIBRIA. The basic mental picture is that the scattering operator maximizes the entropy of the degrees of freedom associated with the atypical constraints *significantly slower* than a scattering time, whereas it maximizes the entropy associated with the remaining degrees of freedom on the order of a scattering time. This is a simple generalization of the same process that underlies the established idea that when the electron–electron component of scattering is strong it drives the local distribution to a warmed internal thermal equilibrium. In this case, the entropy associated with local average energy is maximized on a timescale significantly slower than the average scattering time, whereas the entropy associated with all other degrees of freedom is maximized on a timescale on the order of the average scattering time. In our extension from *classically constrained* quasi-equilibria to *atypically constrained* quasi-equilibria, all we are doing is *abandoning* the notion that the degrees of freedom for which *entropy is maximized the slowest* are *only allowed* to be the conserved quantities which are the focus of classical thermodynamics, such as total particle number and total energy.

The successful definition of relevant atypically constrained quasi-equilibria leads directly to a simplification of the Boltzmann transport equation. This is essentially because we can view transport as a sequence of perturbations to the quasi-equilibria. To a first

[6]The scattering operator acts continuously, which means that it is reapplied to the distribution function every instant.

order approximation, the distribution function can be described by the balance between the reproducible Hamiltonian terms associated with the local field and gradient in the degrees of freedom of the *quasi-equilibrium distribution*, and the non-reproducible Hamiltonian evolution associated with the relaxation toward quasi-equilibrium in a scattering time. We expect this first approximation to be accurate to the extent that the local field and gradients in quasi-equilibrium degrees of freedom are constant over the length scale associated with scattering, and the extent to which the boundary conditions are constant in the time-scale associated with scattering. The local field is already guaranteed by the semiclassical assumptions to be relatively constant over the length scale associated with scattering, since this is equal or shorter than the length scale associated with decoherence. Similarly, the timescale at which the boundary conditions of a device change are slow compared to the scattering time.[7] Accordingly, a valid simplified electron state can be defined in terms of the degrees of freedom of the quasi-equilibria that the scattering operator drives the distribution function toward. This motivates our extensive investigation of novel forms of quasi-equilibria in the rest of this chapter. The explicit demonstration that degrees of freedom of quasi-equilibria are sufficient to define valid electron state functions is given in the two results chapters that follow this chapter.

Finally it is worth briefly noting the relationship of this theoretical framework to the work of Jaynes [80], since both argue that entropy maximization subject to constraints occurs. Simply put, this framework should *not* be viewed as an application of Jaynes' infamous MAXIMUM ENTROPY PRINCIPLE. We are very much concerned with understanding the *physical mechanism* of entropy maximization, whereas Jaynes was not. The reason we do not appeal to Jaynes Maximum Entropy Principle is because of its unrestricted flexibility. The Maximum Entropy Principle can be used to justify *literally any distribution function*, because the space of possible constraints is larger than the $6-$D scalar function space of the distribution function.[8] Thus Jaynes Maximum Entropy Principle is primarily a technique for generating empirical models and is orthogonal to the first-principles approach of this thesis.

[7] Note, this is not true if we follow Fischetti and Laux [16] and treat plasmons semiclassically by updating Poisson's equation at a faster rate than the plasma frequency. As such, the models we derive in this thesis can only incorporate plasmons via a contribution to the scattering operator.

[8] For instance, if we define the distribution function to be the maximum entropy distribution subject to the constraint that it has a given occupation rate at every point in the six dimensional phase space.

4.3 Wedge Constrained Quasi-Equilibria

The first quasi-equilibria we will investigate is the WEDGE-CONSTRAINED QUASI-EQUILIBRIA and it is by far the *weakest* we will investigate in the sense that it reduces the degrees of freedom of the distribution function by the smallest amount. The wedge-constrained quasi-equilibria is a result of what could be called the UNBIASED WALK APPROXIMATION, which is a simplification of the linear component of scattering.[9] In addition to the linear component, the scattering operator will have a non-linear contribution due to carrier–carrier scattering and degeneracy effects:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}} = S_{\text{linear}}[f] + S_{\text{non-linear}}[f]. \tag{4.1}$$

We can express the linear component of scattering as the sum of in-scattering— associated with the transfer of occupation rate from all other states to a state $\mathbf{k}\nu$— and out-scattering— associated with the transfer of occupation rate at $\mathbf{k}\nu$ to all other states[10]:

$$S_{\text{linear}}[f] = S^{\text{in}}_{\text{linear}}[f] + S^{\text{out}}_{\text{linear}}[f]. \tag{4.2}$$

For a linear scattering operator, the out-scattering operator is independent of the occupation rate at all other states. Accordingly, we can express it in terms of a state dependent scattering rate $\frac{1}{\tau_{\text{scat}}(\mathbf{k}\nu)}$:

$$S^{\text{out}}_{\text{linear}}[f] = -\frac{f}{\tau_{\text{scat}}(\mathbf{k}\nu)}. \tag{4.3}$$

The linear in-scattering operator is far more complicated, since it is a functional of all states that conserve total energy and total crystal momentum with a partner state transition. For most scattering operators, there is a set of partner state transitions associated with every possible change in electron crystal momentum. However, it is not the case that all these transitions conserve total energy. Since typically if a $n-$D scalar field intersects another $n-$D scalar field, the intersection will be a possibly disconnected,

[9] If we scale the entire distribution function by some factor, the "linear component of scattering" is the component of the scattering rate that scales by the same factor.

[10] Since scattering is local, in this chapter, we generally take for granted that all the dynamics described are occurring at some fixed position in the device $\mathbf{r}$, and that scattering parameters can depend on this position.

$(n-1)-$D hypersurface, the in-scattering operator for the state $\mathbf{k}\nu$ is dependent on the distribution function on a set of complicated, $2-$D surfaces embedded in the Brillouin zone.

Let us examine where this set of $2-$D surfaces intersects a particular change in partner energy of $\Delta\varepsilon$, and electron energy of $-\Delta\varepsilon$. This will be a complicated, disconnected set of closed $1-$D loops embedded on the *constant energy surface* of the electron bandstructure at $\varepsilon_{\mathbf{k}\nu} + \Delta\varepsilon$. The rate of in-scattering from the electron states at $\varepsilon_{\mathbf{k}\nu} + \Delta\varepsilon$ is defined by the line integral of the distribution over these $1-$D loops, weighted by a coupling term defined in the scattering operator.

If the standard deviation of crystal momentum exchange with the partner system in a scattering time is typically small, the subset of these closed loops which is most heavily coupled to the state $\mathbf{k}\nu$, will typically be confined to a subregion of the constant energy surface near $\mathbf{k} + \langle\Delta\mathbf{k}\rangle$, where $\langle\Delta\mathbf{k}\rangle$ is the average crystal momentum change. However, if the standard deviation of the crystal momentum exchange in a scattering time is a sizeable fraction of the radius of the Brillouin zone, the most heavily coupled of these closed loops will typically traverse a wide sample of the constant energy surface.

Suppose we now express the local distribution function as the sum of the following terms.

- An energy dependent distribution function f_ε, which has the same energy-dependent particle density as the actual distribution function:

$$f_\varepsilon(\mathbf{k}'\nu') = f_\varepsilon(\mathbf{k}''\nu'') \quad \text{if} \quad \varepsilon_{\mathbf{k}'\nu'} = \varepsilon_{\mathbf{k}''\nu''}, \quad \text{and}$$
$$\Gamma \sum_{\nu'} \int_{\mathrm{BZ}} f_\varepsilon(\mathbf{k}'\nu')\delta(\varepsilon - \varepsilon_{\mathbf{k}'\nu'})\mathrm{d}\mathbf{k}' = \Gamma \sum_{\nu'} \int_{\mathrm{BZ}} f(\mathbf{k}'\nu')\delta(\varepsilon - \varepsilon_{\mathbf{k}'\nu'})\mathrm{d}\mathbf{k}'.$$

- An antisymmetric/odd distribution function f_A:

$$f_A(\mathbf{k}'\nu') = -f_A(-\mathbf{k}'\nu').$$

- A symmetric/even perturbation to the energy dependent distribution $f_{S/\varepsilon}$:

$$f_{S/\varepsilon}(\mathbf{k}'\nu') = f_{S/\varepsilon}(-\mathbf{k}'\nu').$$

In an inversion-symmetric bandstructure, the density at constant energy associated with both f_A and $f_{S/\varepsilon}$ is equal zero. In a non inversion-symmetric bandstructure, the density associated with f_A and $f_{S/\varepsilon}$ is equal and opposite, so it is better to think of them as a single perturbation to the energy dependent bandstructure that has zero density at each constant energy. We express them separately since inversion symmetry is extremely common, and is present in silicon. Either way, if we examine the entire constant energy surface of the bandstructure at $\varepsilon_{\mathbf{k}\nu} + \Delta es$, the distribution function is now expressed as a constant function of all the states on the entire energy surface, plus a perturbation to that function which is positive just as often as it is negative.

In the case where the standard deviation of crystal momentum exchange in a scattering time is large, we can make the Unbiased Walk Approximation. The Unbiased Walk Approximation is that as one "walks" along these widely dispersed $1-$D loops of points embedded on the constant energy surface, weighting the local distribution function by a coupling term that is not sharply peaked as this is inconsistent with a large standard deviation of crystal momentum, and adding this scalar to the sum that determines the rate of in-scattering to $\mathbf{k}\nu$ from states at $\mathbf{k}\nu + \Delta\varepsilon$, the effect of the perturbations to the energy dependent distribution function approximately cancel, since one is just as likely for the "walk" to cross a state where they are negative as it is to cross a state where they are positive.

That is, if we express the linear in-scattering rate as the sum of the in-scattering rate due to the energy-dependent distribution and the in-scattering rate due to the perturbations to the energy-dependent distribution, the Unbiased Walk Approximation is that the in-scattering rate due to the perturbations to the energy-dependent distribution can be neglected:

$$\begin{aligned} S^{\text{in}}_{\text{linear}}[f] &= S_{\text{in-scat}}[f_\varepsilon] + S_{\text{in-scat}}[f_A + f_{S/\varepsilon}] \\ &\approx S_{\text{in-scat}}[f_\varepsilon]. \end{aligned} \tag{4.4}$$

This is a simple but extremely powerful, theoretically sound simplification which can

be applied to many scattering operators, since the assumption that the standard deviation in crystal momentum exchanged in a scattering time is a sizeable fraction of the Brillouin zone is extremely commonly satisfied.

The Unbiased Walk Approximation can even model the effect of asymmetric scattering operators and asymmetric bandstructures simply by examining their effect on the $1-$D local *energy* distribution function rather than on the full $3-$D local distribution function. This however is tangential to the aim of this thesis. More to the point is that each state is therefore subject to a linear out-scattering rate, equal to the linear scattering time, a linear in-scattering rate, equal to the linear in-scattering rate acting on the energy distribution function. The relevant picture of electron transport that emerges from the Unbiased Walk Approximation is that, *apart from the effects on the energy distribution*, the perturbations to a distribution function induced by the Hamiltonian flow are *wiped clean* in a scattering time.

This is not to say necessarily that we can describe the distribution function as an energy-dependent distribution function subject to perturbations due to Hamiltonian flow. The linear in-scattering rate of states at a given point in the distribution function associated with a change in the partner energy of $\Delta\varepsilon$, is therefore proportional simply to the "length" of the $1-$D manifold that conserves energy, weighted by the coupling strength. This does not imply that the distribution of in-scattering is the same for a state $\mathbf{k}'\nu'$ which has the same energy as $\mathbf{k}\nu$, since the "length" of the $1-$D total energy and momentum conserving manifold can easily be a function of position on a constant energy surface $\varepsilon_{\mathbf{k}\nu}$.

However, we do have the important result that if the scattering partner distribution in crystal momentum space has the same or higher point symmetry than the symmetry of the local crystal, this means than the in-scattering distribution must have at least the point symmetry of the local crystal. Thus, for instance, in the case of scattering in silicon with a distribution of phonons in lattice equilibrium, the in-scattering distribution to an energy state can be described by an energy dependent density function, and a perturbation to each density function that has the dioctrahedral point-symmetry of the silicon crystal.

As a result, in the case where the distribution of scattering partners has the same higher symmetry as the crystal, this reduces the effective electron state from a distribution function on the entire Brillouin zone to a distribution function on the irreducible wedge of the Brillouin zone, which is perturbed by the Hamiltonian flow terms to create less symmetric perturbations to the distribution function which are deleted on the scattering time-scale. We refer to this as a Wedge-Constrained Quasi-Equilibria, since it is the maximum entropy distribution subject to the set of constraints defined by the distribution function on an irreducible wedge. This nomenclature may seem a little clumsy in this case, but we use it as it is part of the general pattern of trying to understand the distribution function information that scattering destroys efficiently, and which information is not destroyed efficiently.

4.4 Elastically Constrained Quasi-Equilibria

The Unbiased Walk Approximation is applicable to a wide range of semiclassical transport problems since the assumption that the standard deviation in the crystal momentum exchanged with the scattering partners in a scattering time is a significant fraction of the size of the Brillouin zone is very commonly valid. The strength of the approximation is that it clearly shows that in scattering operators with this quality, the only effects due to the reproducible Hamiltonian terms that last on a longer time-scale than the scattering time are those terms that effect the *energy distribution* of electrons. The weakness of the approximation is that the degrees of freedom for the quasi-equilibrium distribution are still described by a $6-$D scalar field, it is just a scalar field that is smaller by a factor determined by the number of point symmetries the semiconducting crystal and scattering partner distribution have in common. The reason for this is that while the quasi-equilibrium distribution does not reflect the perturbations to the energy distribution produced by the reproducible Hamiltonian terms, it will reflect the perturbations to an energy distribution associated with the non-reproducible Hamiltonian terms. These perturbations exist because in general, any given state $\mathbf{k}\nu$ may be coupled with a particular energy level $\varepsilon_{\mathbf{k}\nu} + \Delta\varepsilon$ more or less strongly than any other given state $\mathbf{k}'\nu'$ at the same energy level $\varepsilon_{\mathbf{k}'\nu'} = \varepsilon_{\mathbf{k}\nu}$ but at a different position in the irreducible wedge. A quasi-equilibrium defined only by an energy distribution requires stronger assumptions

about the scattering operator.

In this section we will investigate a stronger form of quasi-equilibria than the Wedge-Constrained Quasi-Equilibria which we refer to as the ELASTICALLY CONSTRAINED QUASI-EQUILIBRIA. An Elastically-Constrained Quasi-Equilibrium would be defined in the classically statistical mechanical sense as the state where all microstates *consistent with a given energy dependent density* are equally probable, or equivalently can be defined as the distribution function which maximizes the entropy subject to this same constraint. In this section, we investigate two different mechanisms that can lead to such a form of quasi-equilibrium.

4.4.1 Equilibrating Mechanism of the First Type

The most obvious scattering operator that will lead to an Elastically-Constrained Quasi-Equilibrium is a scattering operator dominated by purely elastic scattering, such as that due to dopants. In such a case, at any given energy level, scattering will tend to maximize the entropy of the distribution, subject to the constraint that the number of particles is conserved, since this is the constraint that purely elastic scattering preserves. This idea is completely independent of the earlier Unbiased Walk Approximation and holds so long as all pairs of states at the same energy are coupled directly or indirectly[11] by a coupling strength that has inversion symmetry.[12] Under these conditions, the distribution of occupation rates at constant energy is stable if and only if every occupation rate is the same. If the initial distribution is a large perturbation from elastically constrained equilibrium, the relaxation to elastically constrained equilibrium may take slightly longer than the scattering time since the crystal momentum exchange associated with dopant scattering is often biased toward small crystal momentum exchanges. However, as argued in the 1995 update to the DAMOCLES model of Fischetti et al., at the high energies typical of non-equilibrium transport even dopant scattering is associated with large crystal momentum exchanges [6]. Regardless, the fact remains that scattering involving negligible energy transfer drives the distribution toward an elastically constrained equilibrium at some time-scale, and this time-scale will approach the scattering rate in the case where the typical momentum exchanges in scattering are large, or in the case where the perturbations from elastically constrained equilibrium

are small.

If dopant scattering dominates, the assumption that the distribution relaxes to an elastically constrained equilibrium on a timescale that is at most a small multiple of the scattering time, and is most likely very similar to the scattering time, has a sound theoretical basis. The more important question is whether this also applies in the far more common case where *phonon* scattering is dominant, and if so under what conditions.

Phonon scattering is not elastic, and the distribution of phonons is usually assumed to be in thermal equilibrium at the lattice temperature. This means that— unlike dopant scattering— the formal equilibrium that phonon scattering tends to drive the distribution function toward is a lattice temperature equilibrium. It is also well known, however, that phonon creation and destruction typically involves large crystal momentum changes— since phonon states are distributed widely across the entire Brillouin zone— and small energy changes— since all phonon states have relatively small energies. As a result, it is reasonable to suspect that phonon scattering may drive the distribution toward an elastically constrained quasi-equilibria efficiently on the order of a single scattering time, and then to a lattice thermal equilibrium much more slowly.

While intuitively appealing, in order for this to actually happen by the same mechanism as applied for genuinely elastic scattering, we need the same formal conditions to hold albeit over an finite energy range. We need that all states within a small energy band are directly or indirectly coupled by a dominant scattering process which has a symmetric coupling strength. This symmetry is in conflict with the fact that the phonon coupling strength between states with an energy difference $\Delta\varepsilon$ differs by an amount proportional to the Boltzmann factor $e^{\frac{-\Delta\varepsilon}{kT_L}}$.[13] Thus the phonon coupling between states is only approximately symmetric in the case where $\Delta\varepsilon$— the energy of the phonon being created or destroyed— is negligible compared to the lattice thermal energy. Unfortunately this is typically only a reasonable approximation for a minority of phonon scattering events

[11] By indirect coupling we simply mean that state A couples to state B by passing occupation rate probability through a set of intermediate states at the same energy.

[12] By an inversion symmetric coupling strength, we simply mean that the coupling strength from electron state A to electron state B is the same as the coupling strength from B to A for states at the same energy. For this not to be true requires that the probability of a scattering partner transition involving negligible energy change and a transfer of crystal momentum $\Delta\mathbf{k}$ is not symmetric to scattering partner transition involving negligible energy change and a change of crystal momentum of $-\Delta\mathbf{k}$.

[13] For details on this, see Section 4.5, particularly the Aside.

in the far-from-equilibrium regime. For instance, in room-temperature silicon when electrons are far-from-lattice-temperature-equilibrium, the majority of phonon scattering events involve large wavevector transverse acoustic phonons. Such phonons are associated with an energy that is close to the maximum energy of the phonon band (~ 25meV), which is approximately equal to the lattice thermal energy.

Thus we cannot typically rely on the phonon scattering to drive the distribution efficiently toward an elastically constrained equilibrium according to the same mechanism used for dopants.

4.4.2 Equilibrating Mechanism of the Second Type

The equilibrating mechanism of the first type is a conventional equilibration mechanism as it leads to a conventional detailed balance, where on average there is no net transfer of particles from one state to another. In this section, we wish to describe a dynamic mechanism of equilibration to an Elastically-Constrained Quasi-Equilibrium, which does not rely on detailed balance.

The starting point for this mechanism is the Wedge-Constrained Quasi-Equilibrium. The essential point that came from the Wedge-Constrained Quasi-Equilibria is that the in-scattering rate, while a functional only of the energy distribution function, is generally a function of the position in the irreducible Brillouin zone. Note, however, that the assumption that this innately drives a distribution toward a state that also depends explicitly on the position in the irreducible Brillouin zone relies on the assumption *that the out-scattering rate does not have a similar dependence on the position in the irreducible wedge as the in-scattering rate does*. In the case that the dependence *is* similar, then the additional particles scattering into a state kν which is associated with a particularly strong coupling to higher energy states, will only result in a greater "flux" of particles flowing through the state and *not* a particularly large buildup of particles in the state, as the state will also have a proportionally strong coupling to lower energy states. We refer to this as the EASY COME–EASY GO approximation, and a scattering operator which possesses Easy Come–Easy Go dynamics will be driven toward an Elastically-Constrained Quasi-Equilibrium, rather than just a Wedge-Constrained Quasi-Equilibrium by the scattering

operator $S_{\text{linear}} = S^{\text{in}}_{\text{linear}}[f_\varepsilon] + \frac{f}{\tau_{\text{scat}}(\mathbf{k}\nu)}$.

If the most frequent energy exchange $\Delta\varepsilon$ with the scattering partners in a scattering time is small enough, we expect Easy Come–Easy Go dynamics to exist. The reason is as follows. Suppose that the *shape* of a constant energy surface does not change much over an energy scale $2\Delta\varepsilon$, so that the $1-$D manifold of final electron states associated with subtracting an energy of $\Delta\varepsilon$ from the partner distribution and adding to an electron state is similar in shape to the $1-$D manifold associated with adding an energy of $\Delta\varepsilon$ to the partner distribution and removing it from the electron state. We suppose that the total line integral of the scattering operator on the $1-$D manifold on the constant energy surface at $\varepsilon+\Delta\varepsilon$ maintains the same rough shape but is scaled by some factor close to unity for the $1-$D manifold on $\varepsilon - \Delta\varepsilon$. The change in scale is determined primarily by the fractional change in the total density of states at $(\varepsilon + \Delta\varepsilon)$ and $(\varepsilon - \Delta\varepsilon)$, since the phonons involved— and by extension the coupling strength— are approximately inversion symmetric to one another. Accordingly, the out-scattering rate is approximately proportional (though not necessarily equal) to the in-scattering rate across the irreducible Brillouin zone, which is the condition required for the Easy Come–Easy Go approximation to hold. This is shown schematically in Fig. 4.1.

As $\Delta\varepsilon$ tends to zero, the Easy Come–Easy Go assumption becomes more and more accurate by virtue of the shape of the $1-$D manifolds on the constant energy surfaces becoming more and more identical, and the phonons involved becoming closer and closer to being related by perfect inversion symmetry, and therefore the out-scattering rate becomes more perfectly proportional to the in-scattering rate (and infact tending to become equal, but this is not required). We schematically illustrate this limit in Fig. 4.2.

The Easy Come–Easy Go approximation is more accurate at higher energies, since the shape of the bandstructure will change less over the energy scale associated with $2\Delta\varepsilon$, than at the very low energies where the bandstructure can change radically. Thus the Easy Come–Easy Go approximation becomes more accurate precisely for high average energy non-equilibrium distributions where the ordinary relaxation time breaks down, making the assumption that electrons are in an Elastically-Constrained Quasi-Equilibrium theoretically sound at both extremes. In distributions that are only a single scattering event from a lattice temperature equilibrium, the ordinary relaxation time en-

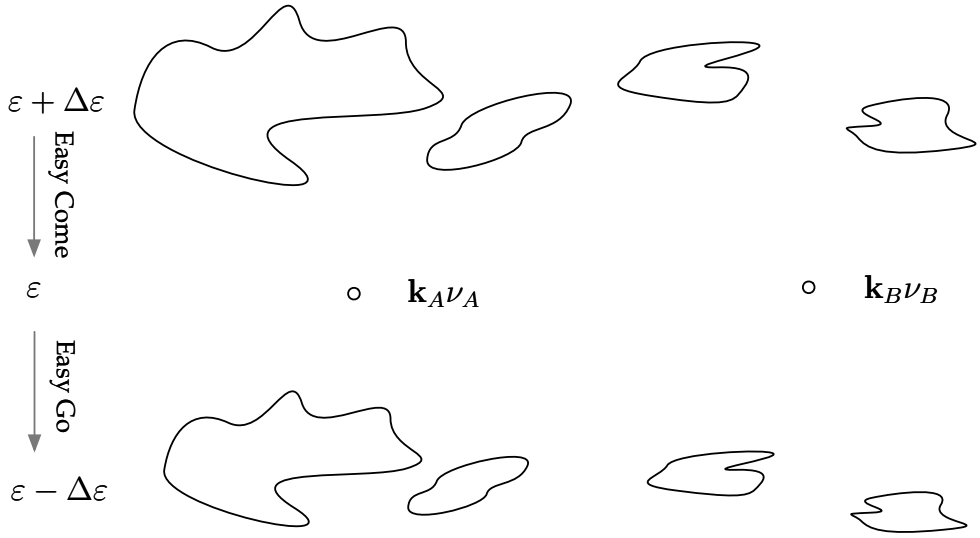


Figure 4.1: An illustration of the Easy Come–Easy Go approximation. If the shape of the constant energy surface of the bandstructure does not vary much over an energy scale $2\Delta\varepsilon$, then the shape of the $1-$D manifolds that intersect the partner states of energy $\Delta\varepsilon$ will not vary much either, so long as the partner energy bandstructure possesses inversion symmetry. As a result, the ratio of in-scattering of state $\mathbf{k}_A\nu_A$ to state $\mathbf{k}_B\nu_B$ at the same energy, will be approximately the same as the ratio of out-scattering of state $\mathbf{k}_A\nu_A$ to state $\mathbf{k}_B\nu_B$.

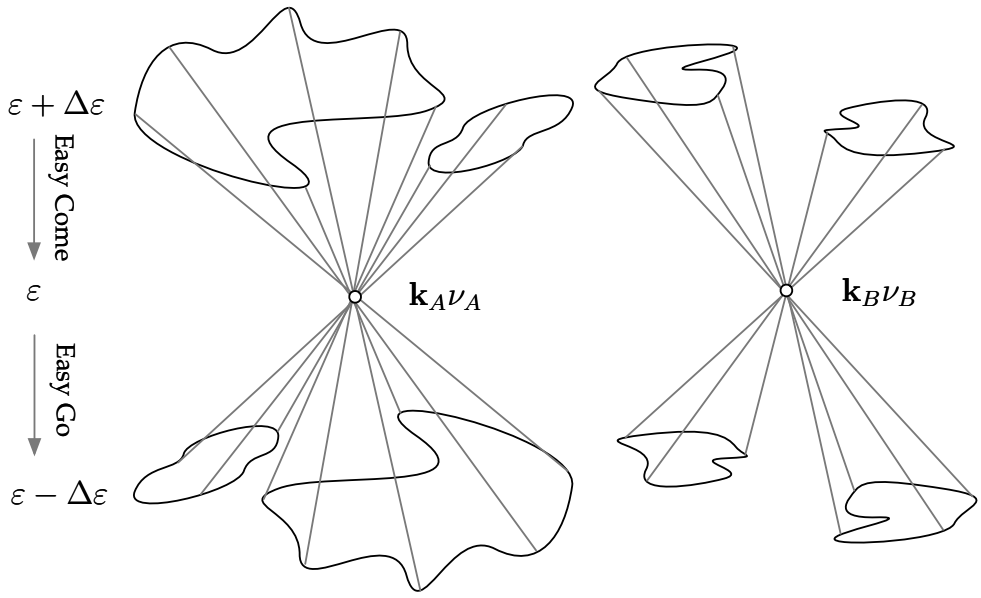


Figure 4.2: An illustration of the fact that phonons involved in the transition from $\varepsilon + \Delta\varepsilon \rightarrow \varepsilon$ and $\varepsilon \rightarrow \varepsilon - \Delta\varepsilon$ become closer and closer to being related by inversion symmetry as the constant energy surfaces at $\varepsilon + \Delta\varepsilon$ and $\varepsilon - \Delta\varepsilon$ become more and more similar in shape. Shown here is the same phonon absorbed twice, which results in a manifold at energy $\varepsilon - \Delta\varepsilon$ which is equal to the manifold at $\varepsilon + \Delta\varepsilon$ inverted about intermediate state at ε. If the second phonon was inverted about the origin, then the $1-$D manifolds at $\varepsilon - \Delta\varepsilon$ would be identical to the $1-$D manifolds at $\varepsilon + \Delta\varepsilon$.

sures it, while in distributions that are many scattering events from a lattice temperature equilibrium transport, the Easy Come–Easy Go approximation ensures it.

4.5 Chemically Constrained Quasi-Equilibria

The Elastically-Constrained Quasi-Equilibrium approximation can be used as the basis of a model of semiclassical transport in the innately inhomogeneous regime, as will be shown in the Results I chapter. In this ELASTICALLY-CONSTRAINED TRANSPORT model, the electron state is determined by a position and energy dependent distribution function $f_\varepsilon(\varepsilon, \mathbf{r}, t)$. This electron state is then subject to pure diffusion driven by the gradient of the occupation rate at constant *total*— rather than kinetic— energy, and a purely inelastic scattering operator.

The result is a theoretically sound, flexible model of phonon collision-dominated, semiclassical electron transport that is several orders of magnitude faster to solve than the full Boltzmann transport equation. However, this model still has the problem that it is $1-$D larger— and therefore one or two orders of magnitude more computationally intensive to solve— than a macroscopic transport model. In order to simplify the model, we need to further make assumptions about the shape of the energy distribution. In this section, we will accept the validity of the Elastically-Constrained Transport model, and try to find a functional form for the energy dependent distribution function. Our focus in this section is on describing an ansatz for the *subpopulation of carriers that are injected from a high potential energy terminal into a device with a highly inhomogeneous field, and which are subject to scattering with partners that are in thermal equilibrium with the lattice*. The reason for focussing on this particular subpopulation of carriers is that it is *precisely the subpopulation* that is *not* described by a conventional thermal equilibrium distribution, since these carriers are typically neither in thermal equilibrium with one another nor with the lattice. Since this subpopulation is the focus of this entire section, we will refer to the energy distribution function of this subpopulation simply as $f_\varepsilon(\varepsilon, \mathbf{r}, t)$ without any qualifying subscripts or superscripts.

4.5.1 Chemical Potential Equalization by Inelastic Scattering

The first thing to note is that when non-equilibrium carriers scatter with partners which are in thermal equilibrium with the lattice, the *direction* of net particle transfer between any two non-equilibrium carrier states is fixed. Namely, between *any two energy states* in the subpopulation the net effect of scattering will be move carriers from energy states with *high effective chemical potential* to the energy states with *low effective chemical potential*, where the EFFECTIVE CHEMICAL POTENTIAL is defined as follows:

$$\mu_{\varepsilon}^{\text{eff}}(\varepsilon, \mathbf{r}, t) = \varepsilon + kT_L \ln\left(\frac{f_{\varepsilon}(\varepsilon, \mathbf{r}, t)}{1 - f_{\varepsilon}(\varepsilon, \mathbf{r}, t)}\right). \tag{4.5}$$

Note that according to this definition, the energy-dependent distribution function is defined as follows:

$$f_{\varepsilon}(\varepsilon, \mathbf{r}, t) = \frac{1}{1 + e^{\frac{\varepsilon - \mu_{\varepsilon}^{\text{eff}}(\varepsilon, \mathbf{r}, t)}{kT_L}}}. \tag{4.6}$$

Put a different way, if we have different carrier states separated by an energy of magnitude $|\Delta\varepsilon|$, and the occupation rates of the carrier states are equal, the energy decreasing transitions will be a factor $e^{\frac{|\Delta\varepsilon|}{kT_L}}$ more frequent than the energy increasing transitions. While it is not immediately obvious, this is actually consistent with the scattering operator derived in the Background chapter.

- It is consistent with scattering with a phonon mode at $\mathbf{q}\eta$ which is in thermal equilibrium with the lattice:
  - Transitions associated with phonon creation— which decrease carrier energy by $\hbar\omega_{\mathbf{q}\eta}$— are proportional to $n_{\mathbf{q}\eta} + 1$.
  - Transitions associated with phonon destruction— which increase carrier energy by $\hbar\omega_{\mathbf{q}\eta}$— are proportional to $n_{\mathbf{q}\eta}$.
  - The lattice temperature Bose-Einstein distribution has the property that $n_{\text{B-E}} + 1 = n_{\text{B-E}} e^{\frac{\hbar\omega_{\mathbf{q}\eta}}{kT_L}}$. Therefore the first rate is greater than the second rate by

a factor $e^{\frac{\hbar\omega_{\mathbf{q}\eta}}{kT_L}}$.

- It is consistent with scattering involving a transition between a pair of partner electron states $|\mathbf{p}_A\mu_A\rangle$ and $|\mathbf{p}_B\mu_B\rangle$ separated by an energy $\Delta\varepsilon = \varepsilon_{\mathbf{p}_B\mu_B} - \varepsilon_{\mathbf{p}_A\mu_A}$ which are in thermal equilibrium with the lattice:

  - Transitions in which $|\mathbf{p}_B\mu_B\rangle$ is the final partner state and $|\mathbf{p}_A\mu_A\rangle$ is the initial partner state— which decrease the primary carrier energy by $\Delta\varepsilon$— are associated with a transition rate proportional to $f_{\text{F-D}}(\varepsilon_{\mathbf{p}_A\mu_A})\big(1 - f_{\text{F-D}}(\varepsilon_{\mathbf{p}_B\mu_B})\big)$.

  - Transitions in which $|\mathbf{p}_A\mu_A\rangle$ is the final partner state and $|\mathbf{p}_B\mu_B\rangle$ is the initial partner state— which increase the primary carrier energy by $\Delta\varepsilon$— are associated with a transition rate proportional to $f_{\text{F-D}}(\varepsilon_{\mathbf{p}_B\mu_B})\big(1 - f_{\text{F-D}}(\varepsilon_{\mathbf{p}_A\mu_A})\big)$.

  - The lattice temperature Fermi–Dirac distribution has the property that $1 - f_{\text{F-D}}(\varepsilon) = e^{\frac{\varepsilon}{kT_L}} f_{\text{F-D}}(\varepsilon)$. Accordingly the first rate is proportional to $f_{\text{F-D}}(\varepsilon_{\mathbf{p}_A\mu_A}) f_{\text{F-D}}(\varepsilon_{\mathbf{p}_B\mu_B}) e^{\frac{\varepsilon_{\mathbf{p}_B\mu_B}}{kT_L}}$, and the second rate is proportional to $f_{\text{F-D}}(\varepsilon_{\mathbf{p}_A\mu_A}) f_{\text{F-D}}(\varepsilon_{\mathbf{p}_B\mu_B}) e^{\frac{\varepsilon_{\mathbf{p}_A\mu_A}}{kT_L}}$. Therefore the first rate is greater to the second rate by a factor $e^{\frac{\Delta\varepsilon}{kT_L}}$.

This Boltzmann factor difference in the rate of scattering transitions which decrease (primary) carrier energy and scattering transitions which increase (primary) carrier energy is a fundamental fact of scattering with a system of partner bodies in thermal equilibrium. This is described in more detail in the following aside.

Aside: On Scattering with Partners in Thermal Equilibrium

Suppose we have a set A of g_A carrier states with kinetic energy ε_A which has an occupation rate f_A, and we have a set B of g_B carrier states with kinetic energy ε_B which has an occupation rate f_B.

Let us suppose these carriers scatter with a bath of partner states which is in thermal equilibrium at a temperature T_L, and let us suppose some of these scattering events are able to create transitions from states in A to states in B, and vice versa.

The number of carriers in A is $n_A = g_A f_A$, and the number of carriers in B is $n_B = g_B f_B$. The rate carriers are scattered from A to B can be defined as follows:

$$\left(\frac{\partial n_B}{\partial t}\right)_{A\to B} = -\left(\frac{\partial n_A}{\partial t}\right)_{A\to B} = \overbrace{g_A f_A}^{\text{Occupied } A \text{ states}} \overbrace{g_B(1-f_B)}^{\text{Empty } B \text{ states}} \overbrace{\mathcal{T}_{A\to B}}^{\text{Transition rate}} . \tag{4.7}$$

Here $\mathcal{T}_{A\to B}$ is the transition rate from A to B per empty final state at B. Similarly, the rate carriers are scattered from B to A can be defined as follows

$$\left(\frac{\partial n_A}{\partial t}\right)_{B\to A} = -\left(\frac{\partial n_B}{\partial t}\right)_{B\to A} = \overbrace{g_B f_B}^{\text{Occupied } B \text{ states}} \overbrace{g_A(1-f_A)}^{\text{Empty } A \text{ states}} \overbrace{\mathcal{T}_{B\to A}}^{\text{Transition rate}} . \tag{4.8}$$

Here $\mathcal{T}_{B\to A}$ is the transition rate from B to A per empty final state at A. We are interested in which of these two rates is larger, which can be expressed via the ratio of the two rates:

$$\begin{aligned}\left|\left(\frac{\partial n_A}{\partial t}\right)_{A\to B} \Big/ \left(\frac{\partial n_A}{\partial t}\right)_{B\to A}\right| &= \left|\left(\frac{\partial n_B}{\partial t}\right)_{A\to B} \Big/ \left(\frac{\partial n_B}{\partial t}\right)_{B\to A}\right| \\ &= \frac{f_A(1-f_B)\mathcal{T}_{A\to B}}{f_B(1-f_A)\mathcal{T}_{B\to A}}.\end{aligned} \tag{4.9}$$

When the carriers in A and B are in thermal equilibrium with a temperature T_L, the ratio of these rates must be equal 1, due to detailed balance. By definition of the lattice temperature Fermi function, in this case when the carriers at A and B are at temperature T_L, we have the following:

$$\frac{f_A}{1-f_A} = e^{-\frac{\varepsilon_A}{kT_L}}, \tag{4.10a}$$

$$\frac{f_B}{1-f_B} = e^{-\frac{\varepsilon_B}{kT_L}}. \tag{4.10b}$$

As such, in thermal equilibrium, we must have that the rate of the transition rates per density of final states is as follows:

$$\frac{\mathcal{T}_{A\to B}}{\mathcal{T}_{B\to A}} = \frac{e^{\frac{\varepsilon_A}{kT_L}}}{e^{\frac{\varepsilon_B}{kT_L}}}. \tag{4.11}$$

This ratio of transition rates is not affected by the occupation rate of A or B, and therefore eq. (4.11) must also be true when A and B are *not* in thermal equilibrium. Therefore, if we attribute an EFFECTIVE CHEMICAL POTENTIAL to a state X by $\mu_X^{\text{eff}} = \varepsilon_X + kT_L \ln\left(\frac{f_X}{1-f_X}\right)$, we can re-express eq. (4.9) in terms of these effective chemical potentials:

$$\left| \left(\frac{\partial n_A}{\partial t}\right)_{A\to B} \Big/ \left(\frac{\partial n_A}{\partial t}\right)_{B\to A} \right| = e^{\frac{\mu_A^{\text{eff}} - \mu_B^{\text{eff}}}{kT_L}}. \tag{4.12}$$

The above ratio of rates of particle transfer is greater than 1 when $\mu_A > \mu_B$ and is less than 1 when $\mu_A < \mu_B$. As such, between any two carrier states, the net rate of particle transfer stimulated by scattering with partners in thermal equilibrium is such that net carriers transfer is always from the state of higher effective chemical potential to the state of lower effective chemical potential. The important point is that this is a completely general result, and does not presume the carrier distribution to be at, or near, equilibrium. The only thermal equilibrium required is that the *scattering partner* distribution is in thermal equilibrium at kT_L.

4.5.2 Chemical Potential Equalization by Total Energy Diffusion

Unlike the case for the full Boltzmann transport equation where only the explicit scattering term is a many-to-one mapping of the distribution function, in the Elastically Constrained Equilibrium model *both* the pure inelastic scattering operator *and* the diffusion at constant total energy are many-to-one, entropy increasing functionals, since the latter includes the *elastic* component of scattering. As a result, *both* components contribute toward irreversibly driving the energy distributions toward a subspace of the set of all possible energy distributions.

Diffusion at constant total energy always acts to equalize the occupation rate at constant total energy of distribution function at neighbouring positions. This can be converted into an equalization of effective chemical potential by rewriting the effective chemical potential as a function of *total* energy as opposed to *kinetic* energy. We refer to this as the *thermodynamic form* of the effective chemical potential. That is, if $f_H(H, \mathbf{r}, t)$ is the distribution function as a function of *total energy* H, rather than kinetic energy ε, the thermodynamic effective chemical potential μ_H^{eff} is defined as follows:

$$\mu_H^{\text{eff}} = H + kT_L \ln\left(\frac{f_H(H, \mathbf{r}, t)}{1 - f_H(H, \mathbf{r}, t)}\right). \tag{4.13}$$

According to this definition, the distribution function as a function of total energy is defined as follows:

$$f_H = \frac{1}{1 + e^{\frac{H - \mu_H^{\text{eff}}(H, \mathbf{r}, t)}{kT_L}}}. \tag{4.14}$$

4.5.3 Elastically and Inelastically Connected States

Both the inelastic scattering operator and the diffusion at constant total energy act to equalize the thermodynamic effective chemical potential between states. The difference is that the inelastic scattering operator acts to equalize the thermodynamic effective chemical potential between states at the same position but different energies, whereas diffusion at constant total energy acts to equalize the thermodymamic effective chemical potential between states at the same total energy but at neighbouring positions. The cru-

cial point to understanding the interplay between these two equilibriating forces is that, if we are in the far-from-lattice-temperature regime, diffusion at constant total energy must be *much more efficient* at equalizing thermodynamic effective chemical potential than the inelastic scattering operator is. If this were not the case, then the distribution at each point in space would be in thermal equilibrium with the lattice.

Accordingly, the inelastic scattering operator can only be expected to approximately achieve the local equalization[14] of the effective chemical potential different energy states in the far-from-thermal-equilibrium regime *if and only if diffusion at constant total energy is not primarily acting to disrupt this local equalization of effective chemical potential*. Because of this fact, it is useful to define two different types of states. We define the ELASTICALLY CONNECTED STATES as any state which has a positive kinetic energy path[15] to a point where the local electron population is in thermal equilibrium with the lattice[16]— that is, to a point where the local electron population is defined by a single effective chemical potential. Conversely, we define INELASTICALLY CONNECTED STATES as all remaining states. We note that it is simple to show that at any point $\mathbf{r}$, if a state at an energy ε is an elastically connected state, then *all states are higher energies* at the same position are also elastically connected states. Accordingly, we can define two regions of the total energy axis at any given point: those associated with elastically connected states, and those associated with inelastically connected states. Such a partitioning is illustrated schematically in Fig. 4.3.

As indicated earlier, we can neatly seperate near-equilibrium and non-equilibrium transport by the relative efficiency of the thermodynamic effective chemical potential equalization by pure inelastic scattering and by diffusion at constant total energy. In devices where pure inelastic scattering is more efficient at equalizing thermodynamic effective chemical potential, the effective chemical potential is a single valued function of position, and the transport is defined to be near equilibrium. In devices where diffusion at constant total energy is more efficient at equalizing the thermodynamic effective chem-

[14]By "local equalization", we mean equalization at a single position in space.

[15]By "has a positive kinetic energy path", we simply mean to say that if there is point that is separated from the lattice temperature distribution by a potential barrier which cannot be detoured around, then it is not elastically connected.

[16]This "point where the local electron population is in thermal equilibrium with the lattice" is roughly equivalent to the high energy terminal from which electrons originate. The only difference is that if there is near-equilibrium transport near the terminal, this point can move to lower total energy.

ical potential, elastically connected states will tend to have a much higher chemical potential than inelastically connected states. This is true for all but a small energy region of states near the crossover energy between elastically and inelastically connected states.

We note that the only influence that drives the distribution function away from a constant chemical potential at energies *well above* the crossover region[17] is the energy dependence of the diffusion parameter. We expect that this second-order perturbing force is not close to being strong enough to overcome the thermalizing effect of acoustic phonons, since in distributions which are already near lattice temperature thermal equilibrium, the relaxation to a lattice temperature equilibrium occurs on the time-scale of the acoustic phonon scattering time. As a result, we expect that the distribution function asymptotically approaches some maximum effective chemical potential at high energies. This constant effective chemical potential part of the distribution is referred to as the THERMAL TAIL. As a result of this fact, it is only the inelastically connected and crossover regions that may have distribution function forms that are not determined by the lattice temperature. Put another way, *it is only the inelastically connected and crossover regions that are far-from-lattice-temperature-equilibrium*. We note that while the crossover region will move upwards in total energy in non-equilibrium— owing to the fact that the net transfer of carriers toward lower chemical potential states by the inelastic scattering operator will generally be *faster* when the difference in effective chemical potential occurs over small energy changes, since most inelastic scattering events involve small energy changes[18]— this movement will be stabilized and largely prevented by elastic diffusion, as illustrated schematically in Fig. 4.3.

We note that in the case where the scattering operator is dominated by very narrow bands of energy transition that are large compared to the thermal energy, the distribution function below the tail will have gaps, associated with inelastically connected states that are very weakly coupled to the most heavily occupied elastically connected states. In silicon however, the most frequent scattering processes have ~ 25meV or less, meaning that any oscillation in the energy distribution occurs can be neglected. This leaves the problem of defining the overall shape as a large-scale function of energy. We

[17] If $\langle|\Delta\varepsilon|\rangle$ is the average magnitude of energy exchange in a scattering event, "well above the crossover region" is defined as being $\sim 3\langle|\Delta\varepsilon|\rangle$ above the crossover region.

[18] If this were not the case, we would be in the near equilibrium regime.

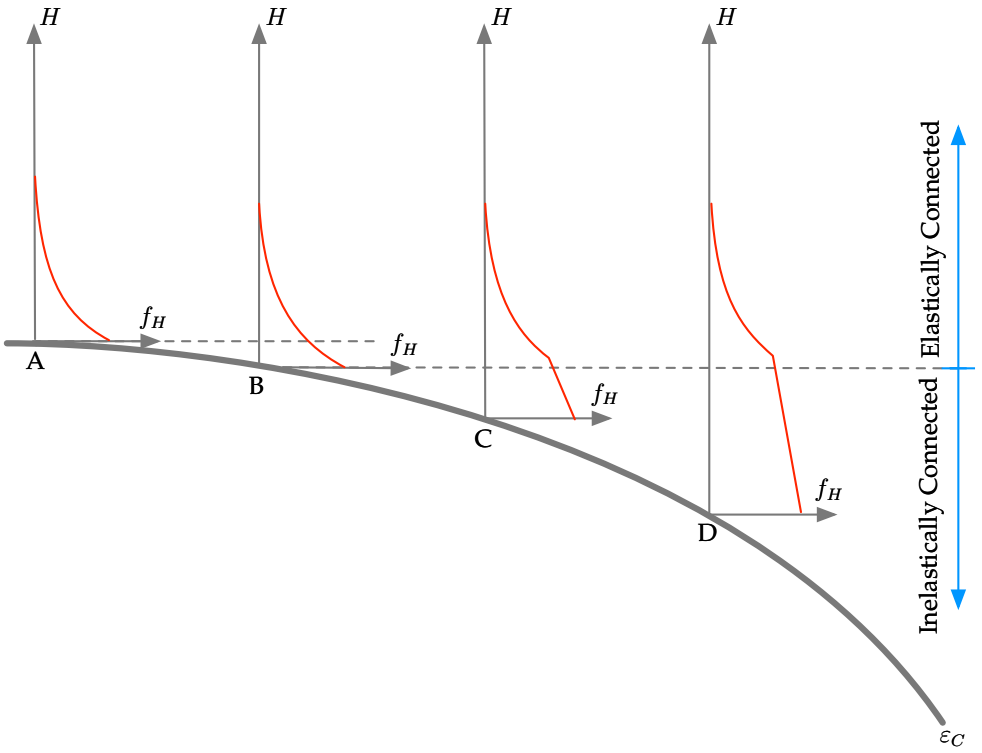


Figure 4.3: Schematic of dynamics, where ε_C is the conduction band minimum and f_H is a schematic depiction of the energy distribution function as a function of total energy. At point A, we have a distribution in thermal equilibrium, which defines the elastically connected states for the adjacent point B. At point B, the elastic diffusion is inefficient enough that the rate of inelastic scattering is sufficient to equilibrate the chemical potential of the inelastically connected states with the elastically connected states. Accordingly the elastically connected states are redefined at the bottom of the conduction band at point B. At point C, elastic diffusion now equilibrates the chemical potential of the elastically connected states with the chemical potential at point B faster than the inelastic scattering can equalize the chemical potential of the inelastic states at point C with the elastically connected states at point C. As a result, the elastically connected states are only in chemical equilibrium with one another, excepting a small crossover region that loses particle density. This crossover region has to be limited to a small region because the decrease in particle density results in an increase in the particle flux at point B, thus stabilizing it. These same dynamics occur at point D, with the crossover region becoming a little larger, and the tail in chemical equilibrium starting a little higher. The number of inelastically connected states grows even faster, and the gap between the chemical potential in the tail and the chemical potential at the bottom of the conduction band grows even larger.

make the assumption that, in the case where there are no "gaps" of weakly coupled states, the distribution simply maximizes entropy, subject to the local maximum effective chemical potential, local energy density and local particle density. This turns out to be equivalent to the assumption that the crossover region and inelastically connected states maximize entropy subject to *their* local energy density and local particle density. We refer to this maximum entropy distribution function as a CHEMICALLY-CONSTRAINED QUASI-EQUILIBRIUM. In Appendix G, we prove that such a distribution f_{CCE} is given by the following form:

$$f_{\text{CCE}} = \begin{cases} \frac{1}{e^{\alpha_\varepsilon + \beta\varepsilon} + 1} & \text{for } \varepsilon < \varepsilon^*, \\ \frac{1}{e^{\frac{\varepsilon - \mu_\varepsilon^{\max}}{kT_L}} + 1} & \text{for } \varepsilon \geq \varepsilon^*. \end{cases} \tag{4.15}$$

Here α_ε and β are uniquely defined by the constraints on the particle density and energy density, $\mu_\varepsilon^{\max}$ is the maximum effective chemical potential and ε^* is the KNEE ENERGY, which is implicitly defined by demanding f_{CCE} is continuous:

$$\alpha_\varepsilon + \beta\varepsilon^* = \frac{\varepsilon^* - \mu_\varepsilon^{\max}}{kT_L} \tag{4.16}$$

Accordingly, we now have a simple ansatz for the electron subpopulation injected from the high energy "source" terminal into a device with highly inhomogeneous fields, which is subject to scattering with partners which are in thermal equilibrium with the lattice. This is precisely the subpopulation which is *not* described by ordinary forms of equilibrium. When combined with the subpopulation of electrons in thermal equilibrium with one another due to electron-electron scattering, and the subpopulation of electrons which are near effective chemical equilibrium with the low energy "drain" terminal, we can write down an ansatz for the *entire electron population* in the innately inhomogeneous regime. This ansatz forms the basis of the THREE QUASI-EQUILIBRIA TRANSPORT model which is derived in the Results II chapter.

Chapter 5

Results I: Elastically-Constrained Transport

5.1 Introduction

In the background chapter we derived a valid model of semiclassical non-equilibrium transport: a Boltzmann transport equation, subject to a full band structure and scattering operator. The aim of this thesis as a whole is to derive a valid model of innately inhomogeneous semiclassical transport that is *much faster to solve*.

In the theoretical framework, we argued that the complexity of the transport equation is naturally linked to the size of the domain, or degrees of freedom, of the distribution function. This is because the more degrees of freedom the distribution function has, the larger the set of continuity equations the transport equation needs to express in order to be closed. Therefore, if we reduce the domain of the distribution function in a theoretically sound manner, we can naturally reduce the complexity of the transport equation in a theoretically sound manner.

The question then becomes how to reduce the domain size of the distribution function in a theoretically sound manner. We showed that the *reproducible* parts of the Hamiltonian,

associated with non-scattering terms in the Boltzmann equation, are of no assistance in this regard. This was shown to be a consequence of Liouville's theorem. The only terms *mathematically capable* of squeezing the distribution function into some subdomain of the general domain are the *non-reproducible* parts of the Hamiltonian, associated with the scattering operator.

Our next aim then was to find what the smaller function space the full scattering operator might act to squeeze the general distribution function into in non-equilibrium transport. One of the things we noticed is that semiclassical non-equilibrium transport is characterized by frequent phonon scattering, since the external field has to be approximately uniform on the length scale of decoherence. In phonon scattering, crystal momentum is relaxed efficiently, but energy is not. We argued that the efficient crystal momentum relaxation of scattering drives the distribution efficiently toward a distribution function constrained by point symmetry, and on top of this the inefficient relaxation of energy results in a distribution function which is constrained to have the same occupation rate in all states at the same energy. In classical statistical mechanics terms, this is a distribution function representing an ensemble of particles where all microstates *consistent with a given energy distribution function* are equally probable. Accordingly, we referred to this distribution as an *elastically-constrained quasi-equilibrium*.

Our aim in this chapter is to incorporate this tendency of the scattering operator to squeeze the distribution function toward an elastically-constrained equilibrium into our model of non-equilibrium transport. In doing so, we will derive the ELASTICALLY CONSTRAINED QUASI-EQUILIBRIUM MODEL of transport: a theoretically sound simplification of the Boltzmann transport equation that is valid for semiclassical non-equilibrium transport that involves frequent low-energy collisions.

This chapter consists of two major sections. In the first section, we derive the generic form of the elastically constrained equilibrium model. In the second section, we describe in detail how to calculate the energy-dependent transport parameters from the scattering operator and band structure described in the background chapter.

5.2 The Generic Elastically Constrained Transport Model

5.2.1 Overview

In a semiclassical non-equilibrium transport system, carriers frequently scatter with partners which exchange only a *small* amount of carrier energy. Accordingly, energy relaxation requires many collisions and therefore passes through many intermediate states on its way back to equilibrium. Thus *we cannot* approximate this relaxation of the distribution back to *thermal* equilibrium using a relaxation time approximation. On the other hand, the crystal momentum exchanged in these scattering events is often *large*, typically of the same order of magnitude as the crystal momentum of the carrier. As such, crystal momentum is randomized in very few (~ 1) collisions. As shown in the theoretical framework, this allows us to assume that the distribution relaxes toward an ELASTICALLY CONSTRAINED QUASI-EQUILIBRIUM in a relaxation time that is similar to the scattering time. In this section, we build a model of collision-dominated non-equilibrium electron transport based on this observation.

The content of this section can be summarized as follows.

- We separate out the purely elastic component of scattering, by arguing the real scattering operator can be approximated by the sum of a purely elastic scattering operator and a purely inelastic scattering operator.

- We argue that in the collision-dominated regime, the relaxation time approximation can be used to describe the *purely elastic* component of scattering. This is *not* the same as the highly dubious approach of using the relaxation time approximation to describe *inelastic* scattering processes.

- We separate the Boltzmann transport equation into an energy dependent component, an antisymmetric component and a symmetric perturbation to the energy dependent component.

- We show that the antisymmetric component is a perturbation to the energy dependent component that is first order in relaxation time, and that the symmetric component is second order in relaxation time. Accordingly, the perturbations to the first order approximation to the antisymmetric component are *third order* in the relaxation time.

- We derive a drift-diffusion model for particle flux *at single energy*, which requires only an accurate estimate for the energy dependent component and the antisymmetric component. This model is derived by taking an appropriate weighted integrals of the Boltzmann transport equation.

5.2.2 The Elastically-Constrained Scattering Operator

The aim of this subsection is to simplify the scattering operator by incorporating the tendency of scattering to relax the distribution toward an elastically constrained equilibrium. Since this is only *part* of the effect of scattering, our first step is to separate out the part of the scattering operator that is associated with only this effect. This requires we make an approximation, since the natural decomposition of the scattering operator is in terms of different carrier–partner scattering terms, each of which contributes to crystal momentum relaxation.

The approximation we will make is, at an abstract level, widespread in physics. It is the *decoupling of phenomena at different timescales*. The basic idea is that when we have two coupled phenomena, but one has a characteristic timescale that is very short compared to the other, we approximate the faster phenomena as occurring while the slower phenomena is held as static.[1] In our case, the two coupled phenomena we are interested in are the relaxation to an elastically constrained equilibrium, and the relaxation to thermal equilibrium.

It is perhaps useful at this stage to provide a concrete example that explicably displays these phenomena occurring at different timescales. The most explicit such example is where there are no spatial gradients in occupation rate, since then there are no fluxes

[1] A classic example of this type of approximation is the Born–Oppenheimer approximation [26].

to distort the effect of scattering on the distribution function. Accordingly, imagine if at $t = 0$, we place a spatially-uniform arbitrary distribution function $f(\mathbf{k}\nu, 0)$ that has an average energy of about ~ 1eV, into a zero-field collision-dominated device that has a uniform temperature lattice. We expect that, after some short period of time comparable to the average collision time $\tau_1 \sim \langle \tau_C \rangle$, the distribution will have been driven toward an elastically constrained equilibrium. That is, we expect the following:

$$f(\mathbf{k}\nu, t) = f_\varepsilon(\varepsilon_{\mathbf{k}\nu}, t) \qquad \text{for } t \geq \tau_1 \sim \langle \tau_C \rangle \,. \tag{5.1}$$

This is our fast relaxation phenomena, and we investigated in detail the underlying microscopic physics that leads to such a distribution. It is not until a much later time, $\tau_2 \sim 10\tau_1$, that the distribution will have reached thermal equilibrium with the lattice:

$$f(\mathbf{k}\nu, t) = f_{F-D}(\varepsilon_{\mathbf{k}\nu}) \qquad \text{for } t \geq \tau_2 \sim 10\tau_1. \tag{5.2}$$

This therefore is our slow relaxation phenomena. The general technique of decoupling phenomena at different timescales suggests we can decouple scattering into two components. The component that leads to elastically constrained equilibrium, while the relaxation toward thermal equilibrium is static, and the part that leads to thermal equilibrium assuming the distribution is already in elastically constrained equilibrium. We refer to the former as the PURELY ELASTIC scattering term, and the latter as the PURELY INELASTIC scattering term.

To achieve this technically, we will express the inelastic scattering term in terms of an inelastic scattering operator that acts only on the *energy* distribution function, which we define as the average occupation rate at energy ε:

$$f_\varepsilon(\varepsilon, \dots) = \frac{\sum_\nu \int_{\mathrm{BZ}} f(\mathbf{k}\nu, \dots)\delta(\varepsilon_{\mathbf{k}\nu} - \varepsilon)\mathrm{d}\mathbf{k}}{\sum_\nu \int_{\mathrm{BZ}} \delta(\varepsilon_{\mathbf{k}\nu} - \varepsilon)\mathrm{d}\mathbf{k}}. \tag{5.3}$$

We have, thus far, glossed over the phenomena of creation and annihilation. We assume creation and annihilation processes typically occur at the timescales longer than the momentum relaxation timescale, and so we also construct fictitious pure creation and pure annihilation operators that also only operate on the energy dependent distribution function. Accordingly, we will approximate the scattering term as the sum of *four* fictitious "pure" scattering terms:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}(\mathbf{k}\nu) = \left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{elastic}}(\mathbf{k}\nu) + \left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{inelastic}}(\varepsilon_{\mathbf{k}\nu}) + \left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{creation}}(\varepsilon_{\mathbf{k}\nu}) + \left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{annihilation}}(\varepsilon_{\mathbf{k}\nu}). \tag{5.4}$$

We define each pure term type as being generated from a corresponding scattering operator $S_{\text{pure type}}$ as follows, where the square brackets indicate the distribution function that the scattering operator acts on.

- The elastic scattering term:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{elastic}}(\mathbf{k}\nu) = -\frac{f(\mathbf{k}\nu,\dots) - f_{\varepsilon}(\varepsilon_{\mathbf{k}\nu},\dots)}{\tau_{\text{relax}}(\mathbf{k}\nu,\dots)}.$$

- The inelastic scattering term:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{inelastic}}(\varepsilon_{\mathbf{k}\nu}) = S_{\text{inelastic}}(\varepsilon^{\text{i}};\varepsilon^{\text{f}})[f_{\varepsilon}].$$

- The particle creation term:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{creation}}(\varepsilon_{\mathbf{k}\nu}) = S_{\text{creation}}(\varepsilon^{\text{i}};\varepsilon^{\text{f}},\varepsilon_{e}^{\text{f}},\varepsilon_{h}^{\text{f}})[f_{\varepsilon}].$$

- The particle annihilation term:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}^{\text{annihilation}}(\varepsilon_{\mathbf{k}\nu}) = S_{\text{annihilation}}(\varepsilon^{\text{i}},\varepsilon_{h}^{\text{i}})[f_{\varepsilon}].$$

The assumptions underlying the very simple, relaxation time form of the elastic scattering operator have been given in the theoretical framework. If our assumptions made in the theoretical framework are true, the relaxation time at the state is equal to the scattering time. We do not make this substitution on the basis that if, for instance, crystal momentum is *nearly but not totally relaxed* in a scattering time, we can phenomenologically correct for this by using a relaxation time that is slightly longer than the scattering time.

5.2.3 Simplifying via the Elastically-Constrained Scattering Operator

Having simplified the scattering operator, we now wish to find a transport equation for the distribution function. We begin with the Boltzmann transport equation:

$$\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{\text{scat}} - \mathbf{v}\cdot\nabla_{\mathbf{r}} f - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f. \tag{5.5}$$

As with any function, we can separate the electron distribution function into its symmetric and antisymmetric parts with respect to inversions in $\mathbf{k}$) $f = f_S + f_A$, and we can also separate the scattering term into its symmetric and antisymmetric parts[2]:

$$\frac{\partial f_S}{\partial t} + \frac{\partial f_A}{\partial t} = \left(\frac{\partial f_S}{\partial t}\right)_{\text{scat}} + \left(\frac{\partial f_A}{\partial t}\right)_{\text{scat}} - \mathbf{v}\cdot\nabla_{\mathbf{r}} f_S - \mathbf{v}\cdot\nabla_{\mathbf{r}} f_A - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_S + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A. \tag{5.6}$$

Separating the symmetric and antisymmetric terms, we form the following two coupled equations:

$$\frac{\partial f_S}{\partial t} = \left(\frac{\partial f_S}{\partial t}\right)_{\text{scat}} - \mathbf{v}\cdot\nabla_{\mathbf{r}} f_A - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A, \tag{5.7a}$$

$$\frac{\partial f_A}{\partial t} = \left(\frac{\partial f_A}{\partial t}\right)_{\text{scat}} - \mathbf{v}\cdot\nabla_{\mathbf{r}} f_S - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_S. \tag{5.7b}$$

We note that in general $f_S = f_\varepsilon + f_{S/\varepsilon}$, where $f_{S/\varepsilon}$ is symmetric with respect to inversions of $\mathbf{k}$, but is associated with zero total density when integrated over local states in a thin shell of constant energy. Accordingly, we can separate the symmetric component into two parts in a similar manner, leading to three coupled equations, where $(\ldots)_\varepsilon$ represents the energy dependent component of the enclosed function, and $(\ldots)_{S/\varepsilon}$ represents the symmetric perturbation component of the enclosed function:

[2]By this we mean that $\left(\frac{\partial f_S}{\partial t}\right)_{\text{scat}}$ should be interpreted as the symmetric component of scattering term, not necessarily the action of the scattering operator on the symmetric component of the distribution function. That is, at this stage we are treating the scattering term as an ordinary function that can be split into antisymmetric and symmetric components.

$$\frac{\partial f_\varepsilon}{\partial t} = \left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}} - \left(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_A + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A\right)_{\varepsilon}, \tag{5.8a}$$

$$\frac{\partial f_{S/\varepsilon}}{\partial t} = \left(\frac{\partial f_{S/\varepsilon}}{\partial t}\right)_{\text{scat}} - \left(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_A + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A\right)_{S/\varepsilon}, \tag{5.8b}$$

$$\frac{\partial f_A}{\partial t} = \left(\frac{\partial f_A}{\partial t}\right)_{\text{scat}} - \mathbf{v}\cdot\nabla_{\mathbf{r}} f_S - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_S. \tag{5.8c}$$

So far we have made no approximations. The equations in eq. (5.8) are in every way equivalent to the full Boltzmann transport equation. At this point, however, we insert our approximation for the scattering operator. According to this approximation, $\left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}} = S[f_\varepsilon]$, $\left(\frac{\partial f_{S/\varepsilon}}{\partial t}\right)_{\text{scat}} = -\frac{f_{S/\varepsilon}}{\tau_{\text{relax}}}$ and $\left(\frac{\partial f_A}{\partial t}\right)_{\text{scat}} = -\frac{f_A}{\tau_{\text{relax}}}$. This leads to the fundamental transport equations of the Elastically-Constrained Equilibrium regime:

$$\frac{\partial f_\varepsilon}{\partial t} = S[f_\varepsilon] - \left(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_A + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A\right)_{\varepsilon}, \tag{5.9a}$$

$$\frac{\partial f_{S/\varepsilon}}{\partial t} = -\frac{f_{S/\varepsilon}}{\tau_{\text{relax}}(\mathbf{k}\nu,\mathbf{r})} - \left(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_A + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A\right)_{S/\varepsilon}, \tag{5.9b}$$

$$\frac{\partial f_A}{\partial t} = -\frac{f_A}{\tau_{\text{relax}}(\mathbf{k}\nu,\mathbf{r})} - \mathbf{v}\cdot\nabla_{\mathbf{r}} f_S - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_S. \tag{5.9c}$$

The elastic relaxation time is on the order of picoseconds, which is almost always much faster than the characteristic time for the rate of change of boundary conditions in the device.[3] Accordingly, we can assume that the antisymmetric and symmetric perturbation components are in steady-state in a typical device. This leads to the following direct expressions for the antisymmetric and symmetric components:

$$f_{S/\varepsilon} = \tau_{\text{relax}}(\mathbf{k}\nu,\mathbf{r})\left(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_A + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_A\right)_{S/\varepsilon}, \tag{5.10a}$$

$$f_A = -\tau_{\text{relax}}(\mathbf{k}\nu,\mathbf{r})\left(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_S + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_S\right). \tag{5.10b}$$

This set of equations can be iteratively solved using perturbation theory, beginning with

[3] In the DAMOCLES model, plasmon scattering is modelled implicitly by updating the Poisson's equation significantly faster than the plasma frequency, which is of the order of *femtoseconds*. Thus in the elastically-constrained model we need to model plasmon scattering *explicitly* as another scattering mechanism analogous to the other scattering mechanisms. We leave this as a problem for future work.

the approximation that $f_S = f_\varepsilon$. The first order approximation to the distribution function is the following:

$$f_A^1 = -\tau_{\text{relax}}(\mathbf{k}\nu, \mathbf{r}) \left(\mathbf{v} \cdot \nabla_{\mathbf{r}} f_\varepsilon + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f_\varepsilon \right). \tag{5.11}$$

We note that, if we interpret f_ε as a function of $(\varepsilon, \mathbf{r}, t)$ rather than a function of $(\mathbf{k}, \mathbf{r}, t)$, we can use the chain rule to re-express $\nabla_{\mathbf{k}} f_\varepsilon$ as $\nabla_{\mathbf{k}} \varepsilon \frac{\partial f_\varepsilon}{\partial \varepsilon}$. Since $\nabla_{\mathbf{k}} \varepsilon = \hbar \mathbf{v}$, this leads to the following expression:

$$f_A^1 = -\tau_{\text{relax}} \mathbf{v} \cdot \left(\nabla_{\mathbf{r}} f_\varepsilon + \mathbf{F} \frac{\partial f_\varepsilon}{\partial \varepsilon} \right). \tag{5.12}$$

This equation becomes even simpler if we express the energy distribution function as a function of total energy f_H. The proper application of the chain rule in this case is slightly more subtle, so this time we will take care when changing derivatives. Technically what we mean when we change variables is that $f_\varepsilon = f_H \circ (H, \mathbf{r}, t)$, where $(H, \mathbf{r}, t)$ is a vector valued function of the *argument of* f_ε— in this case $(\varepsilon, \mathbf{r}, t)$— and "$\circ$" is the composition operator. Conversely, we can express $f_H = f_\varepsilon \circ (\varepsilon, \mathbf{r}, t)$, where $(\varepsilon, \mathbf{r}, t)$ is a vector valued function of $(H, \mathbf{r}, t)$. Let us use the latter form, and expand $\nabla_{\mathbf{r}} f_H$ using the full multivariate chain rule. For clarity, we write $(\ldots)_{x,y}$ to mean the partial derivative where x, y is constant:

$$\begin{aligned}
(\nabla_{\mathbf{r}} f_H)_{H,t} &= (\nabla_{\mathbf{r}} \varepsilon)_{H,t} \left(\frac{\partial f_\varepsilon}{\partial \varepsilon} \right)_{\mathbf{r},t} + (\nabla_{\mathbf{r}} \mathbf{r})_{H,t} (\nabla_{\mathbf{r}} f_\varepsilon)_{\varepsilon,t} + (\nabla_{\mathbf{r}} t)_{H,t} \left(\frac{\partial f_\varepsilon}{\partial t} \right)_{\varepsilon,\mathbf{r}} \\
&= \Big(\nabla_{\mathbf{r}} \left(H - V(\mathbf{r}, t) \right) \Big)_{H,t} \left(\frac{\partial f_\varepsilon}{\partial \varepsilon} \right)_{\mathbf{r},t} + (\nabla_{\mathbf{r}} f_\varepsilon)_{\varepsilon,t}.
\end{aligned} \tag{5.13}$$

Here $V(\mathbf{r}, t)$ is the potential energy. Removing the subscripts, and noting that the force $\mathbf{F}$ is equal to $\mathbf{F} = -\nabla_{\mathbf{r}} V(\mathbf{r}, t)$, we have the expression for the transformation of the spatial gradient:

$$\nabla_{\mathbf{r}} f_H = \mathbf{F} \frac{\partial f_\varepsilon}{\partial \varepsilon} + \nabla_{\mathbf{r}} f_\varepsilon. \tag{5.14}$$

Accordingly, we can rewrite the first order approximation to the antisymmetric distribution function given in eq. (5.12) as the following wonderfully simple expression:

$$f_A^1 = -\tau_{\text{relax}} \mathbf{v} \cdot \nabla_{\mathbf{r}} f_H. \tag{5.15}$$

If we wanted to, we can substitute this expression into eq. (5.10a) in order to find an estimate for the symmetric perturbation to the distribution function. This will be proportional to τ_{relax}^2. We can then in turn substitute this expression back into eq. (5.10b) in order to find an improved estimate for the antisymmetric distribution, which will differ from eq. (5.15) by an absolute term proportional to τ_{relax}^3, or equivalently a relative term proportional to τ_{relax}^2. If we refer to r as a rough characterization of the relative magnitude of the first order antisymmetric distribution to the energy distribution, we expect the absolute error in the first order of approximation to the antisymmetric distribution to be proportional to r^3 of the energy distribution, and the relative error to be proportional to r^2. Thus, this first order approximation is definitely accurate if $r^2 \ll 1$, and in some circumstances it will be sufficient if $r^3 \ll 1$. Outside these regimes, it is necessary to calculate higher order perturbations.

We will focus the rest of this thesis on the first order elastically constrained equilibrium model, which is defined by the following two transport equations:

$$\frac{\partial f_\varepsilon}{\partial t} = S[f_\varepsilon] - \left(\mathbf{v} \cdot \nabla_{\mathbf{r}} f_A + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f_A \right)_\varepsilon, \tag{5.16a}$$

$$f_A = -\tau_{\text{relax}} \mathbf{v} \cdot \nabla_{\mathbf{r}} f_H. \tag{5.16b}$$

5.2.4 The Energy-Dependent Particle Continuity Formulation

One of the most useful properties of the Boltzmann transport equation is that we can derive continuity equations for smaller sets of information by evaluating weighted integrals of the Boltzmann transport equation over momentum space.

Since the electron state in the elastically constrained equilibrium model is uniquely defined by the position and energy dependent particle density, the most intuitive way to think about the elastically constrained equilibrium transport equations is in terms of the continuity of particle density at each position and energy. We can derive this set of con-

tinuity equations from the Boltzmann transport equation by integrating over constant energy surfaces. In order to do this, it is convenient to define the CONSTANT ENERGY SURFACE FUNCTIONAL, $\sigma^{\varepsilon}(\varepsilon)$, as follows:

$$\sigma^{\varepsilon}(\varepsilon)\left[X\right] = \Gamma \sum_{\nu} \int_{BZ} \delta(\varepsilon_{\mathbf{k}\nu} - \varepsilon) X \mathrm{d}\mathbf{k}. \tag{5.17}$$

Here X is a some arbitrary function. The factor Γ refers to the density of states, per unit volume, in $\mathbf{k}$ space multiplied by the degeneracy of the bands indexed by ν. For spin-degenerate bands of silicon, this is equal to $\frac{1}{4\pi^3}$.

The Constant Energy Surface functional is linear, on the basis of the linearity of the integration operator. In addition, it possesses the following behaviour with the differential operators in the Boltzmann transport equation:

$$\sigma^{\varepsilon}(\varepsilon)\left[\frac{\partial X}{\partial t}\right] = \frac{\partial}{\partial t}\sigma^{\varepsilon}(\varepsilon)\left[X\right], \tag{5.18a}$$

$$\sigma^{\varepsilon}(\varepsilon)\left[\nabla_{\mathbf{r}} X\right] = \nabla_{\mathbf{r}}\sigma^{\varepsilon}(\varepsilon)\left[X\right], \tag{5.18b}$$

$$\sigma^{\varepsilon}(\varepsilon)\left[\nabla_{\mathbf{k}} X\right] = \frac{\partial}{\partial \varepsilon}\sigma^{\varepsilon}(\varepsilon)\left[X \nabla_{\mathbf{k}} \varepsilon\right]. \tag{5.18c}$$

The first identity assumes that the bandstructure is a constant function of time, which may be broken if the bandstructure model is coupled to the Schrödinger equation. The second identity assumes that the bandstructure is independent of position, which may be broken if the device is heterogeneous. The final identity is due to a subtle multidimensional form of integration by parts. We prove this non-trivial piece of calculus in detail in an aside found at the end of this section.

In addition, we can define the following terms.

- The Density of States:

$$D(\varepsilon) := \sigma^{\varepsilon}(\varepsilon)\left[1\right]. \tag{5.19a}$$

- The Energy-Dependent Particle Density:

$$n^{\varepsilon}(\varepsilon, \mathbf{r}, t) := \sigma^{\varepsilon}(\varepsilon)\left[f\right] = \sigma^{\varepsilon}\left[f_{\varepsilon}\right] = D(\varepsilon) f_{\varepsilon}. \tag{5.19b}$$

- The Energy-Dependent Particle Flux:

$$\mathbf{j}^{\varepsilon}(\varepsilon, \mathbf{r}, t) := \sigma^{\varepsilon}(\varepsilon)\left[\mathbf{v}f\right] = \sigma^{\varepsilon}\left[\mathbf{v}f_A\right]. \tag{5.19c}$$

- The Energy-Dependent Density Scattering Term:

$$\left(\frac{\partial n^{\varepsilon}}{\partial t}\right)_{\text{scat}} := \sigma^{\varepsilon}(\varepsilon)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}\right] = D(\varepsilon)\left(\frac{\partial f_{\varepsilon}}{\partial t}\right)_{\text{scat}}. \tag{5.19d}$$

We can apply the constant energy surface functional to both sides of the Boltzmann transport equation. Using the definitions given above, this results in the following equation:

$$\sigma^{\varepsilon}(\varepsilon)\left[\frac{\partial f}{\partial t}\right] = \sigma^{\varepsilon}(\varepsilon)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}} - \mathbf{v}\cdot\nabla_{\mathbf{r}}f - \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}f\right]. \tag{5.20}$$

Applying the linearity of the Constant Energy Surface functional, we have the following:

$$\sigma^{\varepsilon}(\varepsilon)\left[\frac{\partial f}{\partial t}\right] = \sigma^{\varepsilon}(\varepsilon)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}\right] - \sigma^{\varepsilon}(\varepsilon)\left[\mathbf{v}\cdot\nabla_{\mathbf{r}}f\right] - \frac{\mathbf{F}}{\hbar}\cdot\sigma^{\varepsilon}(\varepsilon)\left[\nabla_{\mathbf{k}}f\right]. \tag{5.21}$$

Applying the differential identities of (5.18), leads to the following:

$$\frac{\partial}{\partial t}\sigma^{\varepsilon}(\varepsilon)\left[f\right] = \sigma^{\varepsilon}(\varepsilon)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}\right] - \nabla_{\mathbf{r}}\cdot\sigma^{\varepsilon}(\varepsilon)\left[\mathbf{v}f\right] - \mathbf{F}\cdot\frac{\partial}{\partial\varepsilon}\sigma^{\varepsilon}(\varepsilon)\left[\mathbf{v}f\right]. \tag{5.22}$$

Here we have used the assumption that the spatial divergence commutes with the Constant Energy Surface functional in the same way as the spatial gradient, and that $\mathbf{v}\cdot\nabla_{\mathbf{r}}f = \nabla_{\mathbf{r}}\cdot(f\mathbf{v})$, using the assumption again that the bandstructure is position-independent. In addition, we have used the identity that $\nabla_{\mathbf{k}}\varepsilon = \hbar\mathbf{v}$. Finally, substituting the definitions in (5.19) leads to the Energy-Dependent Particle Density Continuity Equation:

$$\frac{\partial n^{\varepsilon}}{\partial t} = \left(\frac{\partial n^{\varepsilon}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}}\cdot\mathbf{j}^{\varepsilon} - \mathbf{F}\cdot\frac{\partial\mathbf{j}^{\varepsilon}}{\partial\varepsilon}. \tag{5.23}$$

In a similar fashion to eq. (5.15) we can make this more elegant by defining the particle flux as a function of total energy H, rather than as a function of kinetic energy ε. This

leads to the following simple, intuitive expression:

$$\frac{\partial n^{\varepsilon}}{\partial t} = \left(\frac{\partial n^{\varepsilon}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}^{H}. \tag{5.24}$$

It is important to note that so far in this section, the only assumption we have used is that the bandstructure is not a function of position or time. In order to derive this expression, we have made no assumptions concerning the scattering operator or the form of the expression for the energy dependent particle flux. We now turn to determining an expression for the energy dependent particle flux, which does require using the assumptions we have previously made about scattering and the size of perturbations. In this thesis we will use the first-order approximation for the antisymmetric scattering derived earlier. To derive an expression for the particle flux, we simply multiple both sides of eq. (5.15) by the velocity $\mathbf{v}$, and apply the Constant Energy Surface functional to both sides. The result is the following:

$$\sigma^{\varepsilon}(\varepsilon)\left[\mathbf{v} f_A\right] = \sigma^{\varepsilon}(\varepsilon)\left[-\tau_{\text{relax}}\mathbf{v}(\mathbf{v}\cdot\nabla_{\mathbf{r}} f_H)\right]. \tag{5.25}$$

The term on the L.H.S. is the energy-dependent particle flux, which can be defined either as a function of kinetic energy or total energy. For this equation we will use the total energy form, since this is the form used in eq. (5.24). On the R.H.S. we notice that the term $\nabla_{\mathbf{r}} f_H$ can be moved to the outside of the Constant Energy Surface functional, on the basis that it is a single valued function at a single energy. In order to achieve this we make use of the tensor product $\otimes$, which can be used in order to "delay" a dot product operation; that is, $A(B \cdot C) = (A \otimes B) \cdot C$. This leads to the following:

$$\mathbf{j}^{H} = -D(\varepsilon)\mathcal{D}^{\varepsilon} \cdot \nabla_{\mathbf{r}} f_H. \tag{5.26}$$

Here we define the termEnergy-Dependent Diffusion Tensor $\mathcal{D}^{\varepsilon}$ as follows:

$$\mathcal{D}^{\varepsilon}(\varepsilon, \mathbf{r}) = \frac{\sigma^{\varepsilon}(\varepsilon)\left[\tau_{\text{relax}}\mathbf{v}\otimes\mathbf{v}\right]}{\sigma^{\varepsilon}(\varepsilon)[1]}. \tag{5.27}$$

Thus, in a device where the distribution relaxes to an Elastically Constrained Quasi-Equilibrium in a given relaxation time, to first order the particle flux at each energy is determined by *pure diffusion* which is driven by the spatial gradient in occupation rate at constant total energy. This is a simple, elegant and theoretically sound model of

semiclassical non-equilibrium electron transport.

Aside: Non-Trivial Calculus

We wish to prove the following identity:

$$\sigma^{\varepsilon}(\varepsilon)\left[\nabla_{\mathbf{k}}X\right] = \frac{\partial}{\partial\varepsilon}\sigma^{\varepsilon}(\varepsilon)\left[X\nabla_{\mathbf{k}}\varepsilon\right]. \tag{5.28}$$

Or in explicit form:

$$\Gamma\sum_{\nu}\int_{\mathrm{BZ}}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k} = \frac{\partial}{\partial\varepsilon}\Gamma\sum_{\nu}\int_{\mathrm{BZ}}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)X\nabla_{\mathbf{k}}\varepsilon\mathrm{d}\mathbf{k}. \tag{5.29}$$

This identity relies on a kind of multidimensional integration by parts. In order to derive this identity, we will express the delta function on the left hand side in as an infinitesimal volume integral:

$$\int_{\mathrm{BZ}}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k} = \lim_{\delta\varepsilon\to 0}\frac{Q}{\delta\varepsilon}\int_{\Omega^{\nu}(\varepsilon,\varepsilon+\delta\varepsilon)}\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k}. \tag{5.30}$$

Here $\Omega^{\nu}(\varepsilon,\varepsilon+\delta\varepsilon)$ is the volume of $\mathbf{k}$ space in which states in the ν band have a kinetic energy in $(\varepsilon,\varepsilon+\delta\varepsilon)$, and Q is some weight factor. We can use the divergence theorem for gradients to rewrite the integral on the R.H.S. of eq. (5.30) above in terms of the surface integrals that are infinitesimally close to one another:

$$\lim_{\delta\varepsilon\to 0}\int_{\Omega^{\nu}(\varepsilon,\varepsilon+\delta\varepsilon)}\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k} = \lim_{\delta\varepsilon\to 0}\int_{\partial\Omega^{\nu}(\varepsilon+\delta\varepsilon)}X\mathbf{S}_{\mathbf{k}} - \int_{\partial\Omega^{\nu}(\varepsilon)}X\mathbf{S}_{\mathbf{k}}.$$

Here $\delta\Omega(\varepsilon)$ is the $2-$D space of surface of states with kinetic energy ε, and $\mathbf{dS}_{\mathbf{k}}$ is an infinitesimal surface element vector, which points normal the the surface.

According to the fundamental theorem of calculus, we can express the R.H.S. of eq. (5.31) as a definite integral of a derivative:

$$\lim_{\delta\varepsilon\to 0}\int_{\Omega^\nu(\varepsilon,\varepsilon+\delta\varepsilon)}\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k}=\lim_{\delta\varepsilon\to 0}\int_{\varepsilon}^{\varepsilon+\delta\varepsilon}\frac{\partial}{\partial\varepsilon'}\left(\int_{\partial\Omega^\nu(\varepsilon')}X\mathbf{dS}_{\mathbf{k}}\right)\mathrm{d}\varepsilon'. \tag{5.31}$$

The variation of the surface integral with respect to ε' is a constant to first order between ε and $\varepsilon+\delta\varepsilon$. Therefore we can take $\frac{\partial}{\partial\varepsilon'}=\frac{\partial}{\partial\varepsilon}$. This leads to the following:

$$\lim_{\delta\varepsilon\to 0}\int_{\Omega^\nu(\varepsilon,\varepsilon+\delta\varepsilon)}\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k}=\lim_{\delta\varepsilon\to 0}\frac{\partial}{\partial\varepsilon}\int_{\varepsilon}^{\varepsilon+\delta\varepsilon}\int_{\partial\Omega^\nu(\varepsilon')}X\mathbf{dS}_{\mathbf{k}}\mathrm{d}\varepsilon'. \tag{5.32}$$

We now wish to convert the right hand side back to an integral over $\Omega^\nu(\varepsilon,\varepsilon+\delta\varepsilon)$. This means converting the element $\mathrm{d}\varepsilon$ to an element $\mathbf{dk}_{\text{norm}}$, an infinitisimal crystal momentum element which points normal to the constant energy surface. It is clear that we must have the following:

$$\nabla_{\mathbf{k}}\varepsilon\cdot\mathbf{dk}_{\text{norm}}=\mathrm{d}\varepsilon. \tag{5.33}$$

Accordingly, we now have the following identity:

$$\int_{\Omega^\nu(\varepsilon,\varepsilon+\delta\varepsilon)}\nabla_{\mathbf{k}}X\mathrm{d}\mathbf{k}=\frac{\partial}{\partial\varepsilon}\int_{\Omega^\nu(\varepsilon,\varepsilon+\delta\varepsilon)}X\nabla_{\mathbf{k}}\varepsilon\mathrm{d}\mathbf{k}. \tag{5.34}$$

We can now complete the proof of eq. (5.29) by multiplyeq. (5.34) by $\frac{Q}{\delta\varepsilon}$, returning them to delta functions, and then multiplying both sides of the resulting expression by Γ and summing over all states.

5.3 The Energy-Dependent Transport Parameters in Silicon

5.3.1 Overview

We have constructed expressions for the full electron–partner scattering operators, and for the band structures of phonons and electrons. This is microscopic information necessary to model the transport of an electron–partner distribution function. However, as discussed earlier, we are not interested in the transport of the partner system; we are only interested in the transport of the *electron* distribution. Accordingly, we can reduce the full electron–partner scattering operators to the part of each scattering operator that acts on the electron state. We do this by integrating over all possible initial and final partner states, for a single partner type:

$$S_{\text{par}}(\mathbf{k}\nu;\mathbf{k}'\nu') = \sum_{\mathfrak{s}_{\text{par}}}\sum_{\mathfrak{s}'_{\text{par}}} S_{\text{par}}(\mathbf{k}\nu,\mathfrak{s}_{\text{par}};\mathbf{k}'\nu',\mathfrak{s}'_{\text{par}}). \tag{5.35}$$

Here $S_{\text{par}}(\mathbf{k}\nu;\mathbf{k}'\nu')$ is the rate of transition from $\mathbf{k}\nu$ to $\mathbf{k}'\nu'$ per occupied initial state $\mathbf{k}\nu$, per unoccupied final state $\mathbf{k}'\nu'$ due to interaction with the partners of type "par". The sum over $\mathfrak{s}_{\text{par}}$ is the sum over occupied initial partner states, and the sum over $\mathfrak{s}'_{\text{par}}$ is the sum over available final partner states.

The reduced scattering operators given above, together with the band structure, are the parameters necessary to understand the precise transport of an arbitrary electron distribution function $f(\mathbf{k},\mathbf{r},t)$. However, as discussed in the last section, in elastically constrained non-equilibrium transport we are not concerned with the transport of an *arbitrary* distribution function, we are concerned with the transport of a distribution function that is a small antisymmetric perturbation from a non-degenerate, energy-dependent distribution function $f_\varepsilon(\varepsilon,\mathbf{r},t)$. Using a relaxation time approximation for the antisymmetric part of the scattering operator, we have shown how to derive a closed transport equation for the symmetric part of the distribution function. The parameters required for this closed transport equation for the energy-dependent symmetric distribution function, which is valid in the collision-dominated regime, are simpler than those

required for the Boltzmann transport equation, which requires the full band structure and the reduced scattering operator. Specifically, we only require the following set of salient transport parameters that are only associated with the kinetic energy of electrons.

- The Density of States:

$$D(\varepsilon) = \Gamma \sum_{\nu} \int_{\mathrm{BZ}} \delta(\varepsilon - \varepsilon_{\mathbf{k}\nu}) \mathrm{d}\mathbf{k}. \tag{5.36}$$

- The Energy-Dependent Diffusion Tensor:

$$\mathcal{D}^{\varepsilon}(\varepsilon, \mathbf{r}) = \frac{\Gamma \sum_{\nu} \int_{\mathrm{BZ}} \tau_{\mathrm{relax}}(\mathbf{k}\nu, \mathbf{r}) \mathbf{v}(\mathbf{k}\nu) \otimes \mathbf{v}(\mathbf{k}\nu) \delta(\varepsilon - \varepsilon_{\mathbf{k}\nu}) \mathrm{d}\mathbf{k}}{\Gamma \sum_{\nu} \int_{\mathrm{BZ}} \delta(\varepsilon - \varepsilon_{\mathbf{k}\nu}) \mathrm{d}\mathbf{k}}. \tag{5.37}$$

- The Conservative[4] Inelastic Scattering Operator:

$$S(\varepsilon_i; \varepsilon_f) = \frac{V^2 \Gamma^2 \sum_{\nu} \sum_{\nu'} \int_{\mathrm{BZ}} \int_{\mathrm{BZ}} S(\mathbf{k}\nu; \mathbf{k}'\nu') \delta(\varepsilon_{\mathbf{k}\nu} - \varepsilon_i) \delta(\varepsilon_{\mathbf{k}'\nu'} - \varepsilon_f) \mathrm{d}\mathbf{k} \mathrm{d}\mathbf{k}'}{V \Gamma \sum_{\nu} \int_{\mathrm{BZ}} \delta(\varepsilon_{\mathbf{k}\nu} - \varepsilon_i) \mathrm{d}\mathbf{k}}. \tag{5.38}$$

- The Energy-Dependent Creation/Annihilation Operator[5]:

$$C(\varepsilon_i^e; \varepsilon_f^e, \varepsilon_f^h, \varepsilon_f^{e2}),$$
$$A(\varepsilon_i^e, \varepsilon_i^h).$$

We will now attempt to find *directly computable* expressions for the contribution to the salient transport parameters for each scattering partner type.

For dopants, scattering events are elastic and conservative, therefore dopant scattering only contributes to the diffusion tensor. For phonons, scattering events are inelastic and conservative, therefore phonon scattering contributes to the diffusion tensor and the conservative inelastic scattering operator. For impact ionization, scattering events are non-conservative, therefore these scattering events contribute to the diffusion tensor

[4] In this thesis, we use the term *conservative* to mean that the scattering type conserves the number of particles.

[5] We omit further details here, since they are complex and irrelevant. The only detail of this operator that ends up being relevant is the rate of electron creation/destruction.

and the creation operator. Finally for electron scattering partners, scattering events are inelastic and conservative, but are not expected to have a net contribution toward the elastic relaxation time, therefore electron–electron scattering contributes to the inelastic scattering operator.

We will equate the elastic relaxation time with the velocity relaxation time, in order to phenomenologically account for the fact that the crystal moment and velocity may not be perfectly randomized in a single scattering time. We use the velocity relaxation rate as it is the velocity relaxation rate that is correlated with the distance electrons travel in a relaxation time. A simple approximation for the velocity relaxation rate due to a scattering partner is the following:

$$\frac{1}{\tau^v_{\text{par}}} = \frac{1}{\langle|\mathbf{v}|\rangle_\varepsilon}\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{par}}. \tag{5.39}$$

Here $\langle|\mathbf{v}|\rangle_\varepsilon$ is the average speed at constant energy, and $\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{par}}$ is the average rate of change of speed of the distribution due to scattering with the partners.

5.3.2 Dopant Scattering

The only salient scattering parameter that is effected by ionized dopant scattering is the diffusion tensor, via the velocity relaxation rate. The aim of this subsection is therefore to write down an expression for the dopants' contribution to the velocity relaxation rate.

We begin by integrating out the dopant states, by writing a special case of eq. (5.35):

$$S_{\text{dop}}(\mathbf{k}\nu;\mathbf{k}'\nu') = \sum_{\mathbf{R}_{\text{dop}}}\sum_{\mathbf{R}'_{\text{dop}}} S_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}'\nu',\mathbf{R}_{\text{dop}}). \tag{5.40}$$

There are two facts mentioned in the full scattering operator section we can use to greatly simplify eq. (5.40). First, we have argued it is safe to assume that the state of the dopant is not changed by carrier scattering. As such, the sum over available final partner states always contains only one non-zero term. Second, we have shown that the carrier–dopant scattering rate is independent of the precise initial position of the ion.

As such, the sum over all initial states is simply the product of the carrier–dopant scattering rate over a single initial carrier state and the number of initial states. These two facts lead to the following simplification:

$$S_{\text{dop}}(\mathbf{k}\nu;\mathbf{k}'\nu') = N_{\text{dop}} V S_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}'\nu',\mathbf{R}_{\text{dop}}). \tag{5.41}$$

The rate of change of speed due to dopant scattering is defined by the following integral:

$$\begin{aligned}\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{dop}}(\mathbf{k}\nu) &= V\Gamma\sum_{\nu'}\int_{\text{BZ}}|\mathbf{v}(\mathbf{k}'\nu')-\mathbf{v}(\mathbf{k}\nu)|S_{\text{dop}}(\mathbf{k}\nu;\mathbf{k}'\nu')\mathrm{d}\mathbf{k}' \\ &= N_{\text{dop}}V^2\Gamma\sum_{\nu'}\int_{\text{BZ}}|\mathbf{v}(\mathbf{k}'\nu')-\mathbf{v}(\mathbf{k}\nu)|S_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}'\nu',\mathbf{R}_{\text{dop}})\mathrm{d}\mathbf{k}'.\end{aligned} \tag{5.42}$$

Before we try to calculate this integral, we remind the reader that according to Fermi's golden rule, all the scattering operators we deal with can be expressed as the following product:

$$S_{\text{par}}(\mathbf{k}\nu,\mathfrak{s}_{\text{par}};\mathbf{k}'\nu',\mathfrak{s}'_{\text{par}}) = s_{\text{par}}(\mathbf{k}\nu,\mathfrak{s}_{\text{par}};\mathbf{k}'\nu',\mathfrak{s}'_{\text{par}})\delta\big(\varepsilon_{\mathbf{k}'\nu'}+\varepsilon_{\mathfrak{s}'_{\text{par}}}-\varepsilon_{\mathbf{k}\nu}-\varepsilon_{\mathfrak{s}_{\text{par}}}\big). \tag{5.43}$$

Here $s_{\text{par}}(\mathbf{k}\nu,\mathfrak{s}_{\text{par}};\mathbf{k}'\nu',\mathfrak{s}'_{\text{par}}) = \frac{2\pi}{\hbar}\left|\left\langle \mathbf{k}'\nu',\mathfrak{s}'_{\text{par}}\right|\hat{\mathcal{H}}_{\text{car–par}}\left|\mathbf{k}\,\nu,\ \mathfrak{s}_{\text{par}}\right\rangle\right|^2$. Accordingly, in the integral over the Brillouin zone of final states we rewrite eq. (5.42) so that the delta function is explicit:

$$\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{unit-dop}}(\mathbf{k}\nu,\beta_S) = V^2\Gamma\sum_{\nu'}\int_{\text{BZ}}|\mathbf{v}(\mathbf{k}'\nu')-\mathbf{v}(\mathbf{k}\nu)|s_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}'\nu',\mathbf{R}_{\text{dop}})\delta\big(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon_{\mathbf{k}\nu}\big)\mathrm{d}\mathbf{k}'. \tag{5.44}$$

The point to takeaway from this is that scattering operators are in general *infinitely sparse*. The scattering rate between any two random electron states is almost certainly zero, because it is almost certain that these two random electron states do not satisfy the zeroth-order energy conservation condition. This is important, because it means any naïve attempt to numerically integral eq. (5.44) using a classic trapezium type method will almost certainly return zero.

Accordingly, to correctly compute an integral such as that of eq. (5.44), we need to first

locate the $2-$D subspaces of final states that conserve energy, and only then can we numerically integrate in a classical manner over the subspace. In order to do this, we use an simple algorithm based on the approach taken by Fischetti and Laux [16], and Gilat and Raubenheimer [88].

Aside: Integrating Over an Energy Surface

We begin by calculating the band structure from empirical pseudopotentials on a cubic mesh irreducible Brillouin zone. We then use inverse point symmetry operations, and a spline type interpolation to produce a continuous band structure over the entire Brillouin zone that possesses the correct point symmetry.[6] We then construct a "fine" cubic mesh of the Brillouin zone. Using the continuous band structure, we can compile a list of the following quantities for each of the cubes in the fine mesh.

- The kinetic energy at the centre of the cube, for each band.
- The velocity at the centre of the cube, for each band.
- The maximum and minimum kinetic energy inside the cube, for each band, assuming the velocity at the centre is constant throughout the cube.

In addition, we collect a group of fine mesh cubes, say $8 \times 8 \times 8$, into a "rough" mesh cube. A sensible rule of thumb is to have about as many

[6] The more theoretically effective way to do this is to interpolate using a basis set of functions that possesses the correct symmetries, such as those suggested by Monkhorst and Pack [89]. We use the method suggested instead for convenience. Well-tested, easy-to-use cubic or cubic spline interpolation packages exist in programs like MATLAB that are especially efficient on cubic grids. The output of such packages will possess a point symmetry if its fed data with that same point symmetry. Perfect translational symmetry is not possible with these packages, but all that really matters is that the interpolation *inside the first Brillouin zone* has the shape associated with the translational symmetry of the bandstructure. This can be achieved by simply adding points outside the first Brillouin zone to the interpolated data set.

fine mesh cubes in a rough mesh cube, as there is rough mesh cubes in the Brillouin zone. For the rough mesh cube, we compile a list of only the maximum and minimum kinetic energy inside the cube, for each band, from the list of fine mesh cube data.

For convenience, we will use $\mathbf{k}^{i}_{\text{rough}}$ to refer to the *set* of $\mathbf{k}$ points that defines the volume of i^{th} rough mesh cube, and $\mathbf{k}^{i,j}_{\text{fine}}$ to define the *set* of points that define the volume of the j^{th} fine mesh cube in the i^{th} rough mesh cube. Similarly, we use $\varepsilon_{\mathbf{k}^{i}_{\text{rough}}\nu}$ to refer to the *set* of kinetic energies in the ν band of the i^{th} rough cube, and $\varepsilon_{\mathbf{k}^{i,j}_{\text{fine}}\nu}$ to refer to the the *set* of kinetic energies in the ν band of the $(i,j)^{\text{th}}$ fine mesh cube. Since we are viewing the band structure as a continuous function, these energy sets are simply the closed intervals between the minimum and maximum kinetic energy for each band.

We will now use this notation to rewrite a specific scattering integral, given in eq. (5.44), as follows:

$$\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{unit-dop}}(\mathbf{k}\nu,\beta_S) = V^2\sum_{\nu'}\sum_{i,j}\Gamma\int_{\mathbf{k}^{i,j}_{\text{fine}}}|\mathbf{v}(\mathbf{k}'\nu')-\mathbf{v}(\mathbf{k}\nu)|s_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}'\nu',\mathbf{R}_{\text{dop}})\delta\big(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon_{\mathbf{k}\nu}\big)\mathrm{d}\mathbf{k}'. \tag{5.45}$$

We can now remove all i from the sum iff $\varepsilon_{\mathbf{k}^{i}_{\text{rough}}\nu'}\cap\varepsilon_{\mathbf{k}\nu}=\emptyset$. For the remaining i, we can remove all j from the sum iff $\varepsilon_{\mathbf{k}^{i,j}_{\text{fine}}\nu'}\cap\varepsilon_{\mathbf{k}\nu}=\emptyset$. These relations can be checked quickly from using the precompiled list of quantities associated with each cube.[7]

Additionally, we make an important approximation. We assume that the fine mesh cubes are sufficiently small, that both the velocity and the matrix element in the fine mesh cube can be approximated by velocity and the matrix element at the centre of the fine mesh cube $\mathbf{k}^{i,j}_{\text{centre}}$. As a result of this approximation we can bring these terms outside the fine mesh cube integral, and as a result of the earlier elimination we can perform the sum over a filtered of (i,j), which we represent by a dash over the Sigma:

[7] We note that, in the general case, this elimination step is slightly more complicated as the target energy for the kinetic energy is itself a function of $\mathbf{k}'$, rather than being a constant such as $\varepsilon_{\mathbf{k}\nu}$.

$$\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{unit-dop}}(\mathbf{k}\nu,\beta_S) = V^2\sum_{\nu'}\sum_{i,j}{}' |\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu') - \mathbf{v}(\mathbf{k}\nu)| s_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}^{i,j}_{\text{centre}}\nu',\mathbf{R}_{\text{dop}}) \times\Gamma\int_{\mathbf{k}^{i,j}_{\text{fine}}}\delta\big(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon_{\mathbf{k}\nu}\big)\mathrm{d}\mathbf{k}'. \tag{5.46}$$

We note that the definition of the density of states (per unit volume), in band ν', in the $(i,j)^{\text{th}}$ mesh cube, is the following:

$$D^{i,j}_{\nu'}(\varepsilon) = \Gamma\int_{\mathbf{k}^{i,j}_{\text{fine}}}\delta(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon)\mathrm{d}\mathbf{k}'. \tag{5.47}$$

As such, we can rewrite eq. (5.46) in terms of eq. (5.47)[8]:

$$\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{unit-dop}}(\mathbf{k}\nu,\beta_S) = V^2\sum_{\nu'}\sum_{i,j}{}' |\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu') - \mathbf{v}(\mathbf{k}\nu)| s_{\text{dop}}(\mathbf{k}\nu,\mathbf{R}_{\text{dop}};\mathbf{k}^{i,j}_{\text{centre}}\nu',\mathbf{R}_{\text{dop}}) D^{i,j}_{\nu'}(\varepsilon_{\mathbf{k}\nu}). \tag{5.48}$$

We note that eq. (5.48) is directly computable so long as we find a computable expression for $D^{i,j}_{\nu'}(\varepsilon)$. Accordingly we turn our attention to finding such an expression. The intrinsic density of states per volume of $\mathbf{k}$ space, for a spin-degenerate band is Γ. To convert this to a density of states per interval of ε space, we simply need to multiply this by the conversion factor between (infinitesimal) intervals of ε space, and volumes of $\mathbf{k}$ space within a fine mesh cube.

We remind the reader that we have made the assumption that the velocity in a fine mesh cube is assumed to be constant and equal to $\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu')$. This is equivalent to the assumption that the band structure inside the fine mesh cube simply consists of planes of equal energy normal to the velocity. Suppose $A^{i,j}(\varepsilon)$ is the cross sectional area of the plane of states in the ν'

[8] We note once again, that in the general case, this substitution is more complicated as the target energy is a function of $\mathbf{k}'$ rather than being a constant such as $\varepsilon_{\mathbf{k}\nu}$. For phonon scattering, the target energy varies sufficiently slowly over a fine mesh cube that we can approximate it as constant for fixed (i,j) and therefore the analysis is similar. For electron–electron scattering, the target energy varies sufficiently quickly over a fine mesh cube that the analysis is changed qualitatively.

band with energy ε, in the fine mesh cube at i, j and $\frac{\partial k_\perp}{\partial \varepsilon}$ is the rate of change of $\mathbf{k} \cdot \hat{\mathbf{n}}$ per change in energy. The density of states in the fine mesh cube for the ν band is therefore written as follows:

$$\begin{aligned} D_{\nu'}^{i,j}(\varepsilon) &= \Gamma A_{\nu'}^{i,j}(\varepsilon) \frac{\partial k_\perp}{\partial \varepsilon} \\ &= \Gamma A_{\nu'}^{i,j}(\varepsilon) \frac{1}{\hbar \left| \mathbf{v}(\mathbf{k}_{\text{centre}}^{i,j}) \right|}. \end{aligned} \tag{5.49}$$

The only remaining unknown in eq. (5.49) is the cross sectional area of the intersection of an arbitrary plane and cube. This can be expressed in closed analytic form, as is given in the paper of Gilat and Raubenheimer. Accordingly, we now wish to sketch the conceptual underpinnings of this closed analytic form.

We begin by choosing a sensible coordinate system. It is natural to choose a coordinate system in which the origin is at $\mathbf{k}_{\text{centre}}^{i,j}$, and to choose a perpendicular coordinate system aligned with the edges of the cube. Suppose we call the axes 1, 2 and 3. We note that we still possess some freedom in our coordinate system, since 1 could refer to one of six directions. To specify our coordinate system completely, suppose we express the velocity direction our coordinate system as follows:

$$\frac{\mathbf{v}(\mathbf{k}_{\text{centre}}^{i,j} \nu')}{\left| \mathbf{v}(\mathbf{k}_{\text{centre}}^{i,j} \nu') \right|} = (\hat{v}_1, \hat{v}_2, \hat{v}_3). \tag{5.50}$$

We choose the coordinate system for which $\hat{v}_1 \geq \hat{v}_2 \geq \hat{v}_3 \geq 0$. It is most natural to find an expression for the cross sectional area in terms of the direction of the normal to the plane (i.e. $\hat{\mathbf{v}}$), and the distance the plane is from the centre of the cube, $\Delta k_\perp$. We will refer to this function as $A(\Delta k_\perp)$.[9]. We note that $A(-\Delta k_\perp) = A(\Delta k_\perp)$ so we can assume $\Delta k_\perp \geq 0$ without loss of generality.

The plane can be considered to partition the corners of the cube into two sets. The shape of the cross sectional area is determined by how many

[9]To find the cross sectional area in terms of energy, we simply need to write down the composition function $A_{\nu'}^{i,j}(\varepsilon) = A(\Delta k_\perp) \circ \Delta k_{\perp,\nu'}^{i,j}(\varepsilon)$

edges connect the two sets. The number of such edges is uniquely determined by the shape of made by the smaller set of corners. The possibilities are as follows.

1 corner: The plane cuts three edges; the cross section is a triangle.

2 corners: The plane cuts four edges; the cross section is a trapezium.

3 corners: The plane cuts five edges; the cross section is a pentagon (four parallel sides).

4 corners, square: The plane cuts four edges; the cross section is a parallelogram.

4 corners, tetrahedron: The plane cuts six edges; the cross section is a hexagon (six parallel sides).

When $\Delta k_\perp = 0$, the set of corners in bisected into two sets of four. As $\Delta k_\perp$ increases, the shape of the cross section will change every time the plane intersects a corner. As such, we can expect then that $A(\Delta k_\perp)$ is a piecewise function that changes each time the plane intersects a corner. Suppose the cube is of side length $2b$. The 8 corners are then at $(\pm b, \pm b, \pm b)$, where each $\pm$ sign can vary independently. According to elementary geometry, the corners intersect the plane at the following values of $\Delta k_\perp^{\text{critical}}$, where once again each $\pm$ sign can vary independently:

$$\begin{aligned} \Delta k_\perp^{\text{critical}} &= (\hat{v}_1, \hat{v}_2, \hat{v}_3) \cdot (\pm b, \pm b, \pm b) \\ &= b(\pm \hat{v}_1 \pm \hat{v}_2 \pm \hat{v}_3). \end{aligned} \tag{5.51}$$

It is useful to label order these $\Delta k_\perp^{\text{critical}}$ from lowest to highest. We label any negative critical value using a negative index since these are not in the domain of $\Delta k_\perp$, and any positive critical value using a positive index:

$$
\begin{aligned}
\Delta k_{\perp}^{-4} &= b(-\hat{v}_1 - \hat{v}_2 - \hat{v}_3), \\
\Delta k_{\perp}^{-3} &= b(-\hat{v}_1 - \hat{v}_2 + \hat{v}_3), \\
\Delta k_{\perp}^{-2} &= b(-\hat{v}_1 + \hat{v}_2 - \hat{v}_3), \\
\Delta k_{\perp}^{-1} &= \min\Big[b(\hat{v}_1 - \hat{v}_2 - \hat{v}_3), b(-\hat{v}_1 + \hat{v}_2 + \hat{v}_3)\Big], \\
\Delta k_{\perp}^{1} &= \max\Big[b(\hat{v}_1 - \hat{v}_2 - \hat{v}_3), b(-\hat{v}_1 + \hat{v}_2 + \hat{v}_3)\Big], \\
\Delta k_{\perp}^{2} &= b(\hat{v}_1 - \hat{v}_2 + \hat{v}_3), \\
\Delta k_{\perp}^{3} &= b(\hat{v}_1 + \hat{v}_2 - \hat{v}_3), \\
\Delta k_{\perp}^{4} &= b(\hat{v}_1 + \hat{v}_2 + \hat{v}_3).
\end{aligned}
$$

We note that the smallest positive critical point has the value $\Delta k_{\perp}^{1} = |b(\hat{v}_1 - \hat{v}_2 - \hat{v}_3)|$ unambiguously. What is ambiguous is the *position* of the corner of the cube that is intersected at this value of Δk. If $\hat{v}_1 > \hat{v}_2 + \hat{v}_3$, then the corner intersected at $\Delta k_{\perp}^{1}$ is coplanar with the corners intersected at higher $\Delta k_{\perp}^{\text{critical}}$; as such, below $\Delta k_{\perp}^{1}$ the plane cuts four edges and the cross section is a parallelogram. If $\hat{v}_1 < \hat{v}_2 + \hat{v}_3$, then the associated corner intersected at $\Delta k_{\perp}^{1}$ is not coplanar with the corners intersected at higher $\Delta k_{\perp}^{\text{critical}}$; as such, below $\Delta k_{\perp}^{1}$ the plane cuts six edges and the cross-section is a hexagon.

Deriving the precise formula for the calculation of the cross sectional areas of the various shapes is tedious. It is derived by viewing each shape as a combination of parallelograms and triangles. Since this is conceptually simple, we will simply write down the piecewise formula for the cross-sectional area:

$$A(\Delta k_\perp) = \begin{cases} \frac{4b^2}{\hat{v}_1} & (\Delta k_\perp^1 \geq \Delta k_\perp \geq 0) \wedge (\hat{v}_1 \geq \hat{v}_2 + \hat{v}_3), \\ \frac{2b^2(\hat{v}_1\hat{v}_2+\hat{v}_2\hat{v}_3+\hat{v}_3\hat{v}_1)-(\Delta k_\perp)^2-b^2}{\hat{v}_1\hat{v}_2\hat{v}_3} & (\Delta k_\perp^1 \geq \Delta k_\perp \geq 0) \wedge (\hat{v}_1 \leq \hat{v}_2 + \hat{v}_3), \\ \frac{2b^2(\hat{v}_1\hat{v}_2+3\hat{v}_2\hat{v}_3+\hat{v}_3\hat{v}_1)+b\Delta k_\perp(\hat{v}_1-\hat{v}_2-\hat{v}_3)-\frac{1}{2}(\Delta k_\perp)^2-\frac{1}{2}b^2}{\hat{v}_1\hat{v}_2\hat{v}_3} & \Delta k_\perp^2 \geq \Delta k_\perp \geq \Delta k_\perp^1, \\ \frac{2b^2\hat{v}_3(\hat{v}_1+\hat{v}_2)-2b\Delta k_\perp\hat{v}_3}{\hat{v}_1\hat{v}_2\hat{v}_3} & \Delta k_\perp^3 \geq \Delta k_\perp \geq \Delta k_\perp^2, \\ \frac{(b(\hat{v}_1+\hat{v}_2+\hat{v}_3)-\Delta k_\perp)^2}{2\hat{v}_1\hat{v}_2\hat{v}_3} & \Delta k_\perp^4 \geq \Delta k_\perp \geq \Delta k_\perp^3, \\ 0 & \Delta k_\perp \geq \Delta k_\perp^4. \end{cases} \tag{5.52}$$

We note that $\Delta k_\perp(\varepsilon) = \frac{\varepsilon - \varepsilon_{\mathbf{k}^{i,j}_{\text{centre}}\nu'}}{\hbar|\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}})|}$. As a result, the density of states at ε for the ν' band, in the i,j fine mesh cube is given by the following analytic function:

$$D^{i,j}_{\nu'}(\varepsilon) = \Gamma \frac{1}{\hbar\left|\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu')\right|} A\left(\frac{\varepsilon - \varepsilon_{\mathbf{k}^{i,j}_{\text{centre}}\nu'}}{\hbar|\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu')|}\right). \tag{5.53}$$

As such, we can now compute eq. (5.48) directly, and calculate $\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{unit-dop}}(\mathbf{k}\nu, \beta_S)$ on a mesh of $\mathbf{k}\nu$ and β_S values. For scattering rates $\frac{1}{\tau_X}$ that depend on the screening wavevector, rather than tabulating $\frac{1}{\tau_X}$ directly, we will generally tabulate $\frac{\beta_S^4}{\tau_X}$, as this function is typically very slowly varying as function of β.

5.3.3 Phonon Scattering

The scattering components affected by phonon scattering are energy dependent diffusion tensor via the velocity relaxation rate, and the conservative inelastic scattering operator. The aim of this section is therefore to derive the contribution to these terms.

The first step to deriving the contribution to both terms is to find the reduced scattering operator associated with all phonon transitions. We will find its is more convenient to treat phonons in different bands as different scattering partners, and therefore to find the electron scattering operator with all phonon transitions in a single band, η. The initial state and final for each phonon mode in that band are characterized by an initial and final occupation numbers, $n_{\mathbf{q}\eta}$ and $n'_{\mathbf{q}\eta}$. The electron scattering operator associated with all phonon transitions in a band is then the following special case of eq. (5.35):

$$S^{\eta}_{\text{pho}}(\mathbf{k}\nu;\mathbf{k}'\nu') = \sum_{\mathbf{q}} \sum_{n'_{\mathbf{q}\eta}} S_{\text{pho}}(\mathbf{k}\nu, n_{\mathbf{q}\eta};\mathbf{k}'\nu', n'_{\mathbf{q}\eta}). \tag{5.54}$$

Here the initial occupation number $n_{\mathbf{q}\eta}$ can be calculated from the analytic approximation to the phonon band structure and from the assumption that the phonons are in thermal equilibrium. We argued in the derivation of the full scattering operator that the full scattering operator is zero unless except for phonon state transition that annihilate or create a single phonon. Therefore we eliminate every term in the sum over final phonon states, except for these phonon annihilation and creation transitions:

$$S^{\eta}_{\text{pho}}(\mathbf{k}\nu;\mathbf{k}'\nu') = \sum_{\mathbf{q}} \Big(S_{\text{pho}}(\mathbf{k}\nu, n_{\mathbf{q}\eta};\mathbf{k}'\nu', n_{\mathbf{q}\eta}-1) + S_{\text{pho}}(\mathbf{k}\nu, n_{\mathbf{q}\eta};\mathbf{k}'\nu', n_{\mathbf{q}\eta}+1) \Big). \tag{5.55}$$

We also note that we have argued that the scattering operator is necessarily zero if crystal momentum not conserved modulo $\mathbf{G}$. For phonon annihilation transitions, we can therefore eliminate every term in the sum over $\mathbf{q}$ except the unique crystal momentum $\mathbf{q} = [\mathbf{k}' - \mathbf{k}] \mod \mathbf{G}$.[10] For phonon creation transitions, we can similarly eliminate every term in the sum over $\mathbf{q}$ except and a unique crystal momentum conserving mode $\mathbf{q}$ specified by $\mathbf{q} = [\mathbf{k} - \mathbf{k}'] \mod \mathbf{G}$:

$$S^{\eta}_{\text{pho}}(\mathbf{k}\nu,\mathbf{k}'\nu') = \sum_{\pm} s_{\text{pho}}(\mathbf{k}\nu, n_{\pm[\mathbf{k}'-\mathbf{k}]^{\mathbf{G}}\eta};\mathbf{k}'\nu', n_{\pm[\mathbf{k}'-\mathbf{k}]^{\mathbf{G}}\eta} \mp 1)\delta(\varepsilon_{\mathbf{k}'\nu'} - \varepsilon_{\mathbf{k}\nu} \mp \hbar\omega_{\pm[\mathbf{k}'-\mathbf{k}]^{\mathbf{G}}\eta}). \tag{5.56}$$

Here all the $\pm$ are simultaneously fixed, we use the unusual summation index to indi-

[10] We use the notation "$[\mathbf{x}] \mod \mathbf{G}$"— or "$[\mathbf{x}]^{\mathbf{G}}$" for short— to mean that a reciprocal lattice vector $\mathbf{G}'$ is added to $\mathbf{x}$ so that $\mathbf{x} + \mathbf{G}'$ is in the (Gamma centered) Brillouin zone.

cate that the two resulting possibilities are added together. We note that in an inversion symmetric material such as Silicon, we have that $\hbar\omega_{\mathbf{q}\eta} = \hbar\omega_{-\mathbf{q}\eta}$, and so the energy of the created and annihilated phonon is identical. The scattering rate is therefore zero unless $|\varepsilon_{\mathbf{k}'\nu'} - \varepsilon_{\mathbf{k}\nu}| = \hbar\omega_{[\mathbf{k}'-\mathbf{k}]\mathrm{G}\eta}$. Thus, for a fixed initial state $\mathbf{k}\nu$, the rate of scattering is zero except for all final states not on a special energy conserving $2-\mathrm{D}$ subspace of the Brillouin zone.

We are interested in the following two integrals of the single particle scattering operator due to phonons. Firstly, the rate of speed change due to phonons, $\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{pho}}$:

$$\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{pho}}(\mathbf{k}\nu) = V\Gamma\sum_{\nu'}\int_{\text{BZ}}\left|\mathbf{v}(\mathbf{k}'\nu') - \mathbf{v}(\mathbf{k}\nu)\right|S_{\text{pho}}(\mathbf{k}\nu,\mathbf{k}'\nu')\mathrm{d}\mathbf{k}'. \tag{5.57}$$

Secondly the average rate of transfer from an occupied state at ε_i to a state at ε_f due to phonons, $S_{\text{pho}}(\varepsilon_i;\varepsilon_f)$:

$$S_{\text{pho}}(\varepsilon_i;\varepsilon_f) = \frac{V^2\Gamma^2\sum_\nu\sum_{\nu'}\int_{BZ}\int_{BZ}S_{\text{pho}}(\mathbf{k}\nu;\mathbf{k}'\nu')\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon_i)\delta(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon_f)\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{k}'}{V\Gamma\sum_\nu\int_{BZ}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon_i)\mathrm{d}\mathbf{k}}. \tag{5.58}$$

The integration of $S_{\text{pho}}(\varepsilon_i;\varepsilon_f)$ is a *line* integral rather than a *surface integral*, however if we set the interval of allowed finite energies to be a finite interval, then the result is a thin surface "ribbon" that we can integrate using a similar algorithm to that used to determine the dopant density. The issue with this approach is that it requires a very fine mesh of the Brillouin zone to do accurately. An alternative approach is to assume a functional form for the inelastic scattering operator that can be specified by a few moments of the operator.

We note that the phonon band structure specified by Fischetti and Laux has a very delta function in the density of states at the maximum energy associated with each band. This is a crude reflection of a real phenomena in silicon, where the phonon bands near the edge of the Brillouin zone are very flat. Because of this, we can view approximate the inelastic scattering operator due to each band of phonons as possessing only two final energies associated with $\varepsilon_i \pm \hbar\omega^\eta_{\text{max}}$, or only one final energy when $\varepsilon_i - \hbar\omega^\eta_{\text{max}} < 0$. We can then force this simple inelastic scattering operator to have the correct rate of scattering with η band phonons, $\frac{1}{\tau^\eta_{pho}(\varepsilon_i)}$ and the correct average energy change in η band phonon scattering events, $\left\langle\Delta\varepsilon_{\text{pho}}\right\rangle^\eta(\varepsilon_i)$, by writing it as follows:

$$S_{\text{pho}}(\varepsilon_i;\varepsilon_f) = \sum_{\eta} \frac{1}{\tau^{\eta}_{\text{pho}}(\varepsilon_i)} \times \begin{cases} r^{\eta}\delta\big(\varepsilon_f - \varepsilon_i + \hbar\omega^{\eta}_{\text{max}}\big) + (1-r^{\eta})\delta\big(\varepsilon_f - \varepsilon_i - \hbar\omega^{\eta}_{\text{max}}\big) & \text{for } \varepsilon_i - \hbar\omega^{\eta}_{\text{max}} > 0, \\ \delta\big(\varepsilon_f - \varepsilon_i + \big\langle\Delta\varepsilon_{\text{pho}}\big\rangle^{\eta}(\varepsilon_i)\big) & \text{for } \varepsilon_i - \hbar\omega^{\eta}_{\text{max}} \leq 0. \end{cases} \tag{5.59}$$

Here the parameter r^{η} is the fraction of phonon creation events to phonon scattering events that ensures the average energy is correct:

$$r^{\eta} = \frac{\big\langle\Delta\varepsilon_{\text{pho}}\big\rangle^{\eta}}{2\hbar\omega^{\eta}_{\text{max}}} + \frac{1}{2}. \tag{5.60}$$

The average rate of scattering with η phonons as a function of electron energy, $\frac{1}{\tau^{\eta}_{pho}(\varepsilon)}$, is defined in the following manner:

$$\frac{1}{\tau^{\eta}_{\text{pho}}(\varepsilon)} = \frac{V\Gamma\sum_{\nu}\int_{\text{BZ}}\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\mathrm{d}\mathbf{k}}{V\Gamma\sum_{\nu}\int_{\text{BZ}}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\mathrm{d}\mathbf{k}}. \tag{5.61}$$

While the average energy of electrons due in an η band phonon scattering event, $\big\langle\Delta\varepsilon_{\text{pho}}\big\rangle^{\eta}(\varepsilon)$, is defined by in the following manner:

$$\big\langle\Delta\varepsilon_{\text{pho}}\big\rangle^{\eta}(\varepsilon) = \frac{V\Gamma\sum_{\nu}\int_{\text{BZ}}\big(\frac{\partial\varepsilon}{\partial t}\big)^{\eta}_{\text{pho}}(\mathbf{k}\nu)\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\mathrm{d}\mathbf{k}}{V\Gamma\sum_{\nu}\int_{\text{BZ}}\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\mathrm{d}\mathbf{k}}. \tag{5.62}$$

The important thing to notice about these two equations, is that so long as the functions $\big(\frac{\partial\varepsilon}{\partial t}\big)^{\eta}_{\text{pho}}(\mathbf{k}\nu)$ and $\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}$ are known, and can be approximated as being constant within a fine mesh cube, then the equations in both these integrals can be computed using exactly the same numerical method used to compute eq. (5.44). If we use this approach, we are led to the following computable expressions for these terms:

$$\big\langle\Delta\varepsilon_{\text{pho}}\big\rangle^{\eta}(\varepsilon) = \frac{\sum_{\nu}\sum'_{i,j}\big(\frac{\partial\varepsilon}{\partial t}\big)^{\eta}_{\text{pho}}(\mathbf{k}^{i,j}_{\text{centre}}\nu)D^{i,j}_{\nu}(\varepsilon)}{\sum_{\nu}\sum'_{i,j}\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}^{i,j}_{\text{centre}}\nu)}D^{i,j}_{\nu}(\varepsilon)}, \tag{5.63a}$$

$$\frac{1}{\tau^{\eta}_{\text{pho}}(\varepsilon)} = \frac{\sum_{\nu}\sum'_{i,j}\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}^{i,j}_{\text{centre}}\nu)}D^{i,j}_{\nu}(\varepsilon)}{\sum_{\nu}\sum'_{i,j}D^{i,j}_{\nu}(\varepsilon)}. \tag{5.63b}$$

We have thus reduced the problem of computing $S(\varepsilon_i,\varepsilon_f)$, to the problem of finding

the function $\left(\frac{\partial \varepsilon}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu)$ and the function $\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}$. The function $\left(\frac{\partial \varepsilon}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu)$ is the rate of energy change of an electron at $\mathbf{k}\nu$ due to interactions with η band phonons, and is defined as follows:

$$\left(\frac{\partial \varepsilon}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu) = V\Gamma \sum_{\nu'} \int_{\text{BZ}} \left(\varepsilon_{\mathbf{k}'\nu'} - \varepsilon_{\mathbf{k}\nu}\right) S^{\eta}_{\text{pho}}(\mathbf{k}\nu, \mathbf{k}'\nu') \mathrm{d}\mathbf{k}'. \tag{5.64}$$

The function $\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}$ is the rate an electron at $\mathbf{k}\nu$ scatters with η band phonons, and is defined as follows:

$$\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)} = V\Gamma \sum_{\nu'} \int_{\text{BZ}} S^{\eta}_{\text{pho}}(\mathbf{k}\nu, \mathbf{k}'\nu') \mathrm{d}\mathbf{k}'. \tag{5.65}$$

When combined with eq. (5.57), we therefore reduced the salient information required from the scattering operator, to *three* integrals of $S^{\eta}_{\text{pho}}(\mathbf{k}\nu, \mathbf{k}'\nu')$. We will now derive directly computable expressions for these integrals.

Our method is essentially identical to the computation of eq. (5.44) with a minor adjustment to allow for the fact that the instead of a single constant target energy for the delta function, there are two target energies $\varepsilon_{\text{target}} = \varepsilon_{\mathbf{k}\nu} \pm \hbar\omega_{([\mathbf{k}'-\mathbf{k}]^{\mathbf{G}})\eta}$, which are functions of the Brillouin zone. The adjustment is minor because these target energies vary slowly enough that we are able to safely assume they are constant over a fine mesh cube.

The main adjustment concerns the elimination of rough cubes. For each rough cube, we determine the point on the i^{th} rough cube closest to $\mathbf{k}$ (mod $\mathbf{G}$) , and the point on the i^{th} rough cube farthest from $\mathbf{k}$ (mod $\mathbf{G}$). From this, we can quickly determine the minimum and maximum crystal momentum between $\mathbf{k}$ and a point in the rough cube:

$$\mathbf{q}^{i}_{\text{min}}(\mathbf{k}) = \min\Big(|[\mathbf{k} - \mathbf{k}^{i}_{\text{rough}}]^{\mathbf{G}}|\Big), \tag{5.66a}$$

$$\mathbf{q}^{i}_{\text{max}}(\mathbf{k}) = \max\Big(|[\mathbf{k} - \mathbf{k}^{i}_{\text{rough}}]^{\mathbf{G}}|\Big). \tag{5.66b}$$

As discussed in the section deriving the full scattering operator, we assume phonon energy is a monotonically increasing function of the magnitude of the phonon wavevector. Thus, by computing the phonon energies associated with the wavevectors in eq. (5.66), we can compute the range of phonon energies. *If* the range of final state electron en-

ergies does not overlap the initial energy plus or minus the range of phonon energies, *then* we can eliminate that rough cube from the summation. That is, we eliminate all i for which the following statement is true:

$$\left[\varepsilon^{\min}_{\mathbf{k}^i_{\text{rough}},\nu}, \varepsilon^{\max}_{\mathbf{k}^i_{\text{rough}},\nu}\right] \cap \varepsilon_{\mathbf{k}\nu} \pm \left[\hbar\omega_{\mathbf{q}^i_{\min}(\mathbf{k})\eta}, \hbar\omega_{\max \mathbf{q}^i_{\max}(\mathbf{k})\eta}\right] = \emptyset. \tag{5.67}$$

We note that this is a necessary but not sufficient condition for energy conservation. That is, if a rough mesh cube makes it through the above filter, this is not sufficient to guarantee there is energy and momentum conserving transition within that cube, as the phonon corresponding to the correct energy change may have the wrong crystal momentum for the transition.

For the rough cubes that pass the filter, we evaluate the fine mesh. In contrast to the rough mesh, we make the assumption that the phonon energy associated with a transition to *any* point in the $(i,j)^{\text{th}}$ fine mesh cube is equal to $\pm\hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}},\nu}$, where $\mathbf{q}^{i,j}_{\text{centre}}$ is the change in crystal momentum associated with a transition from an initial state $\mathbf{k}$ to the *centre* of the $(i,j)^{\text{th}}$ fine mesh cube. Accordingly, we can eliminate the j^{th} fine mesh cube of the i^{th} rough mesh cube if the following statement is true:

$$\varepsilon_{\mathbf{k}\nu} \pm \hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}},\nu} \notin \left[\varepsilon^{\min}_{\mathbf{k}^i_{\text{fine}},\nu}, \varepsilon^{\max}_{\mathbf{k}^i_{\text{fine}},\nu}\right]. \tag{5.68}$$

We can now write down a directly computable expression for $\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu)$, the rate of velocity relaxation due to scattering with η band phonons for an electron at $\mathbf{k}\nu$ using the same method as was used for eq. (5.42). Namely:

1. We partitioning the integral into the sum of the integrals over the cubes in the fine mesh that are not filtered out by the two elimination steps.

2. We assume that the velocity and matrix element are constant within a unit cell and equal to their values at the centre of the fine mesh cube. Accordingly we can move all except the delta function outside the integral over a fine mesh cube.

3. Given the assumption that the target energies for the delta function are constant over a fine mesh cube, we relate the remaining integral to the density of states

of electrons at the target energies, in the band ν', in the $(i,j)^{\text{th}}$ fine mesh cube. As discussed earlier, this density of states has an analytic expression due to the assumption that velocity is constant within a fine mesh cube.

$$\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu) = \sum_{\nu'}\sum_{i,j}{}' \left|\mathbf{v}(\mathbf{k}^{i.j}_{\text{centre}}\nu') - \mathbf{v}(\mathbf{k}\nu)\right| s^{\eta}_{\text{pho}}(\mathbf{k}\nu;\mathbf{k}^{i.j}_{\text{centre}}\nu') \times \left(D^{i,j}_{\nu'}\left(\varepsilon_{\mathbf{k}\nu}+\hbar\omega_{\mathbf{q}\eta}\right) + D^{i,j}_{\nu'}\left(\varepsilon_{\mathbf{k}\nu}-\hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}}\eta}\right)\right). \tag{5.69}$$

This method is also sufficient to write a directly computable expression for $\left(\frac{\partial\varepsilon}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu)$, the rate of energy change for an electron in state $\mathbf{k}\nu$ due to interactions with the phonons in the band η. The energy change can be brought outside the integral because it is equivalent to plus or minus the phonon energy, which is constant over a fine mesh cube according to the third assumption:

$$\left(\frac{\partial\varepsilon}{\partial t}\right)^{\eta}_{\text{pho}}(\mathbf{k}\nu) = \sum_{\nu'}\sum_{i,j}{}' \hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}}\eta} s^{\eta}_{\text{pho}}(\mathbf{k}\nu;\mathbf{k}^{i.j}_{\text{centre}}\nu') \times \left(D^{i,j}_{\nu'}\left(\varepsilon_{\mathbf{k}\nu}+\hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}}\eta}\right) - D^{i,j}_{\nu'}\left(\varepsilon_{\mathbf{k}\nu}-\hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}}\eta}\right)\right). \tag{5.70}$$

Finally, it is obvious that this method is sufficient to write a directly computable expression for $\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}$, the rate that an electron at state $\mathbf{k}\nu$ scatters with phonons in the band η:

$$\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)} = \sum_{\nu'}\sum_{i,j}{}' s^{\eta}_{\text{pho}}(\mathbf{k}\nu;\mathbf{k}^{i.j}_{\text{centre}}\nu') \times \left(D^{i,j}_{\nu'}\left(\varepsilon_{\mathbf{k}\nu}+\hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}}\eta}\right) + D^{i,j}_{\nu'}\left(\varepsilon_{\mathbf{k}\nu}-\hbar\omega_{\mathbf{q}^{i,j}_{\text{centre}}\eta}\right)\right). \tag{5.71}$$

We have thus written down directly computable expressions for all the universal scattering parameters associated with phonons that are relevant to our transport model.

5.3.4 Electron–Electron Scattering

The scattering component affected by electron–electron scattering is the conservative inelastic scattering operator, which can be defined as follows:

$$S_{\text{ee}}(\varepsilon_i;\varepsilon_f;\beta_S) = \frac{V^2\Gamma^2\sum_{\nu,\nu'}\int_{BZ}\int_{BZ} S_{\text{ee}}(\mathbf{k}\nu;\mathbf{k}'\nu';\beta_S)\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon_i)\delta(\varepsilon_{\mathbf{k}'\nu'}-\varepsilon_f)\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{k}'}{V\Gamma\sum_{\nu}\int_{BZ}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon_i)\mathrm{d}\mathbf{k}}. \quad (5.72)$$

Here $S_{\text{ee}}(\mathbf{k}\nu;\mathbf{k}'\nu';\beta_S)$ is the single particle electron–electron scattering operator, which is defined from the two particle electron–electron scattering operator as follows:

$$S_{\text{ee}}(\mathbf{k}\nu;\mathbf{k}'\nu';\beta_S) = V\Gamma\sum_{\mu,\mu'}\int_{\text{BZ}} S_{\text{ee}}(\mathbf{k}\nu,\mathbf{p}\mu;\mathbf{k}'\nu',\mathbf{p}'\mu';\beta_S)f(\mathbf{p}\mu,\mathbf{r},t)\big(1-f(\mathbf{p}'\mu',\mathbf{r},t)\big)\mathrm{d}\mathbf{p}. \quad (5.73)$$

The single particle form of the inelastic electron–electron scattering operator is not particularly useful as it is infeasible to precompute and tabulate. The reason is that, unlike for other scattering types where the distribution of scattering partners is fixed, the distribution of partner electrons dynamically changes in a device, as it is the distribution of electrons. As such, it is better to keep the inelastic scattering operator due to electrons expressed in terms of a two-particle operator. Thus the quantity we are wish to define is $S(\varepsilon_i,\varepsilon_i^{\text{par}};\varepsilon_f,\varepsilon_f^{\text{par}};\beta_S)$.

In order to determine the non-zero terms in scattering operator, we need to integrate over the $1-$D manifold of final states that is consistent with total energy conservation, total crystal momentum pseudoconservation, and the fact that one particle possess ε_f. Using the algorithm we have described, this can be done by integrating over a series of $2-$D "ribbons" surfaces between $(\varepsilon_f,\varepsilon_f+\Delta\varepsilon_f)$, but this is inefficient. We avoided this route in the phonon subsection by assuming a particular form for the distribution of final energies which could be characterized by a simpler integral. We will take the same approach here.

For electron–electron scattering, we will make a similar assumption to that made by Fischetti and Laux for impact ionization: namely we will assume that the probability of a pair of final state energies $(\varepsilon_f,\varepsilon_f^{\text{par}})$ is proportional to the simultaneous density of states $D(\varepsilon_f)D(\varepsilon_f^{\text{par}})$. This allows us define the full inelastic two-particle scattering operator from the total scattering rate for a pair of electrons of energy $(\varepsilon_i,\varepsilon_i^{\text{par}})$, by noting that the integral of the two-particle scattering operator over all final state energies must to be equal to the latter:

$$S_{\text{ee}}\left(\varepsilon_i, \varepsilon_i^{\text{par}}; \varepsilon_f, \varepsilon_f^{\text{par}}; \beta_S\right) = \left(\int_0^{\varepsilon_i+\varepsilon_i^{\text{par}}} D(\varepsilon_f) D(\varepsilon_i + \varepsilon_i^{\text{par}} - \varepsilon_f) \mathrm{d}\varepsilon_f\right)^{-1} \times \frac{1}{\tau_{\text{ee}}(\varepsilon_i, \varepsilon_i^{\text{par}}; \beta_S)} \delta(\varepsilon_f + \varepsilon_f^{\text{par}} - \varepsilon_i - \varepsilon_i^{\text{par}}) D(\varepsilon_f) D(\varepsilon_f^{\text{par}}). \tag{5.74}$$

Technically, the total scattering rate for a pair of electrons of energy $(\varepsilon_i, \varepsilon_i^{\text{par}})$, $\frac{1}{\tau_{\text{ee}}(\varepsilon_i, \varepsilon_i^{\text{par}}; \beta_S)}$, is defined the average scattering rate per unit energy, between electrons with energy ε_i and $\varepsilon_i^{\text{par}}$, per unit partner electron density. Specifically, it is the quantity that when weighted by the energy dependent electron density, and integrated over all energy states, yields to the electron–electron scattering rate for electrons at ε_i:

$$\frac{1}{\tau_{\text{ee}}}(\varepsilon_i, \beta_S) = \int_0^{\infty} \frac{1}{\tau_{\text{ee}}(\varepsilon_i, \varepsilon_i^{\text{par}}; \beta_S)} D(\varepsilon_i^{\text{par}}) f_{\varepsilon}(\varepsilon_i^{\text{par}}) \mathrm{d}\varepsilon_i^{\text{par}}. \tag{5.75}$$

We can now calculate $\frac{1}{\tau_{\text{ee}}(\varepsilon_i, \varepsilon_i^{\text{par}}; \beta_S)}$ using a similar technique to that which we have used for calculating the other energy dependent scattering rates for other particles. We begin by defining the scattering rate between an electron at $\mathbf{k}\nu$ and a density of electrons at ε_{par}, $\frac{1}{\tau_{\text{ee}}(\mathbf{k}\nu; \varepsilon_{\text{par}}; \beta_S)}$[11]:

$$\frac{1}{\tau_{\text{ee}}(\mathbf{k}\nu; \varepsilon_{\text{par}}; \beta_S)} = \frac{V^3 \Gamma^2 \sum_{\nu',\mu,\mu'} \int_{\text{BZ}} \int_{\text{BZ}} s_{\text{ee}}(\mathbf{k}\nu, \mathbf{p}\mu; \mathbf{k}'\nu', [\mathbf{k}+\mathbf{p}-\mathbf{k}']^{\mathbf{G}}\mu'; \beta_S) \delta(\Delta\varepsilon_{\text{total}}) \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{k}'}{V\Gamma \sum_{\mu'} \int_{\text{BZ}} \delta(\varepsilon_{\mathbf{p}\mu} - \varepsilon_{\text{par}}) \mathrm{d}\mathbf{p}}. \tag{5.76}$$

Here we have already separated out the crystal momentum and energy conserving delta functions from the scattering operator, and thus we represent the remaining part of the scattering operator using a lower case s. The scattering operator used here is the full electron–electron scattering operator discussed in the background chapter. The denominator is simply $VD(\varepsilon_{\text{par}})$. We can calculate the numerator using a similar discretization technique to that which was used to calculate the phonon scattering operator. There are two major differences. The first major difference is that this time we have a double integral. The second major difference is that this time rough cells are not useful, as there is no simple predefined relationship between the change in electron crystal momentum

[11]The extra volume factor V in the numerator comes from the fact that we have defined $\tau_{\text{ee}}(\varepsilon_i, \varepsilon_i^{\text{par}}; \beta_S)$ in eq. (5.75) to be already have its factor of V incorporated.

and the "target energy" for a transition: it depends on the initial state of the partner electron. As a result of these facts, we perform the integration over *two* sets of *fine* cells directly: one associated with integral over initial partner states indexed by l, and one associated with the integral over final electron states indexed by m. We can express this integral as follows:

$$\frac{1}{\tau_{\text{ee}}(\mathbf{k}\nu, \varepsilon_{\text{par}}; \beta_S)} = V^2 \sum_{\nu',\mu,\mu'} {\sum_{l,m}}' \left(w_\mu^l s_{\text{ee}}(\mathbf{k}\nu, \mathbf{p}^l_{\text{centre}}\mu; \mathbf{k}^m_{\text{centre}}\nu', [\mathbf{k} + \mathbf{p}^l_{\text{centre}} - \mathbf{k}^m_{\text{centre}}]^{\mathbf{G}}\mu'; \beta_S) \times \Gamma \int_{\mathbf{k}^m_{\text{fine}}} \delta(\varepsilon_{\mathbf{k}'\nu'} + \varepsilon_{[\mathbf{k}+\mathbf{p}^l_{\text{centre}}-\mathbf{k}']^{\mathbf{G}}\mu'} - \varepsilon_{\mathbf{k}\nu} - \varepsilon_{\text{par}}) \mathrm{d}\mathbf{k}' \right). \tag{5.77}$$

Here the weight faction w_μ^l are the fraction of total partner density that is expected to reside in the the l^{th} fine cell. If $\varepsilon_{\text{par}} = 0$, $w_\mu^l = \frac{1}{6}$ for the six valid fine cells. Otherwise it is given as follows as the fraction of the density of states that lies in the fine cell:

$$w_\mu^l = \frac{D_\mu^l(\varepsilon_{\text{par}})}{D(\varepsilon_{\text{par}})}. \tag{5.78}$$

Since we do not have a simple expression for the target energy in the m^{th} fine cell, we cannot yet simplify the expression for the appropriate density of states. The dash over the sum indicates that the sum over l is restricted to those fine cells that intersect the nominal partner energy ε_{par}, and the sum over m is restricted to cells that might conserve total energy. To find the set of valid m, we use the following process. We restrict ourselves to $\mathbf{k}$ that point to the centre of fine mesh points, as this means that $\mathbf{p}'(\mathbf{k}') = [\mathbf{k} + \mathbf{p}^l_{\text{centre}} - \mathbf{k}']^{\mathbf{G}}$ defines a fine mesh cube for the locus of points at $\mathbf{k}^m_{\text{fine}}$. Accordingly, using the maximum and minimum kinetic energies that occur in the fine mesh cube $\mathbf{k}^m_{\text{fine}}$, and the fine mesh cube $\mathbf{p}'(\mathbf{k}^{i,j}_{\text{fine}})$, we can *eliminate* all m from the summation for condition is true:

$$\varepsilon_{\mathbf{k}^m_{\text{fine}},\nu'} \cap \left(\varepsilon_{\mathbf{k}\nu} + \varepsilon_{\text{par}} - \varepsilon_{\mathbf{p}'(\mathbf{k}^m_{\text{fine}})\mu'} \right) = \emptyset. \tag{5.79}$$

This filtering of the set of fine mesh cubes cannot be done as efficiently as was the case for phonons. In order to be able to use a rough mesh, we would require that $\mathbf{k}$ and $\mathbf{p}^l_{\text{centre}}$ is aligned with the centre of the rough mesh, which is unreasonably restrictive.

For the set of fine mesh cubes that are can plausibly conserve energy, we will use the same assumptions as for phonons: namely that velocity is constant in a fine mesh cube and equal to the value of velocity at the centre. However, this time we are not interested in the density of states inside the cube; that is, we are not interested in the planes of states in the fine mesh cubes for which $\varepsilon_{\mathbf{k}'\nu'}$ is constant. Instead, we are interested in the planes of states in the cube for which $\varepsilon_{\mathbf{k}'\nu'} + \varepsilon_{[\mathbf{k}+\mathbf{p}^l_{\text{centre}}-\mathbf{k}']^{\mathbf{G}}\mu'}$ is constant, which will not generally be perpendicular to the velocity at $\mathbf{k}^{i,j}_{\text{centre}}\nu'$.

According to the assumption of constant velocity inside a mesh cube, the individual final state energies can be expressed as follows:

$$\varepsilon_{\mathbf{k}'\nu'} = \varepsilon_{\mathbf{k}^m_{\text{centre}},\nu'} + \hbar\mathbf{v}(\mathbf{k}^m_{\text{centre}}\nu')\cdot(\mathbf{k}' - \mathbf{k}^{i,j}_{\text{centre}}) \tag{5.80a}$$

$$\varepsilon_{\mathbf{p}'(\mathbf{k}')\mu'} = \varepsilon_{\mathbf{p}'(\mathbf{k}^m_{\text{centre}})\mu'} - \hbar\mathbf{v}\big(\mathbf{p}'(\mathbf{k}^m_{\text{centre}})\mu'\big)\cdot(\mathbf{k}' - \mathbf{k}^{i,j}_{\text{centre}}). \tag{5.80b}$$

Here $\mathbf{p}'(\mathbf{k}') = [\mathbf{k} + \mathbf{p}^l_{\text{centre}} - \mathbf{k}']^{\mathbf{G}}$. Therefore combining the equations, we can create a *effective band structure* inside the m^{th} fine mesh cube for the *total energy*:

$$\varepsilon^{\text{total}}_{\mu'}(\mathbf{k}'\nu') = \Big(\varepsilon_{\mathbf{k}^m_{\text{centre}},\nu'} + \varepsilon_{\mathbf{p}'(\mathbf{k}^m_{\text{centre}})\mu'}\Big) + \hbar\Big(\mathbf{v}(\mathbf{k}^m_{\text{centre}}\nu') - \mathbf{v}\big(\mathbf{p}'(\mathbf{k}^m_{\text{centre}})\mu'\big)\Big)\cdot(\mathbf{k}' - \mathbf{k}^{i,j}_{\text{centre}}). \tag{5.81}$$

Here the first term in large parentheses is the total energy of the centre of the cube $\varepsilon^{\text{total}}_{\mu'}(\mathbf{k}^m_{\text{centre}}\nu')$, and the second term in large parentheses EFFECTIVE VELOCITY for total energy, $\mathbf{v}^{\text{eff}}_{\mu'}(\mathbf{k}^m_{\text{centre}}\nu')$. This transformation is extremely useful because the target *total energy* for delta function in eq. (5.77) is the constant. Accordingly, we have now transformed the problem into one in which we can use the same method as for dopants. We can express the scattering rate per unit partner electron density, $\frac{1}{\tau_{ee_{\text{unit}}}}$, as follows:

$$\frac{1}{\tau_{ee}(\mathbf{k}\nu, \varepsilon_{\text{par}}; \beta_S)} = V^2 \sum_{\nu',\mu,\mu'} {\sum_{l,m}}' w^l_\mu s_{\text{ee}}(\mathbf{k}\nu, \mathbf{p}^l_{\text{centre}}\mu; \mathbf{k}^m_{\text{centre}}\nu', [\mathbf{k} + \mathbf{p}^l_{\text{centre}} - \mathbf{k}^m_{\text{centre}}]^{\mathbf{G}}\mu') D^m_{\nu',\mu'}(\varepsilon_{\mathbf{k}\nu} + \varepsilon_{\text{par}}). \tag{5.82}$$

Here $D^m_{\nu'\mu'}$ is the density of states of the effective band structure, when final electrons are in the m^{th} fine mesh cell and the band ν', and final partners are in band μ':

$$D^m_{\nu',\mu'}(\varepsilon) = \Gamma \frac{1}{\hbar\big|\mathbf{v}^{\text{eff}}_{\mu'}(\mathbf{k}^m_{\text{centre}}\nu')\big|} A\left(\frac{\varepsilon - \varepsilon^{\text{total}}_{\mu'}(\mathbf{k}^m_{\text{centre}}\nu')}{\hbar|\mathbf{v}^{\text{eff}}_{\mu'}(\mathbf{k}^m_{\text{centre}}\nu')|}\right). \tag{5.83}$$

We have now constructed an algorithm for the calculation of the scattering rate between an electron at $\mathbf{k}\nu$ and a set of partners at energy ε_{par}. In order to turn this into a calculation for the rate of scattering over kinetic energy, we simply integrate the resulting expression over constant total energy. This can be determined in an analogous manner to eq. (5.61):

$$\frac{1}{\tau_{ee}(\varepsilon,\varepsilon_{\text{par}};\beta_S)} = \frac{V\Gamma\sum_\nu \int_{\text{BZ}} \frac{1}{\tau_{ee}(\mathbf{k}\nu,\varepsilon_{\text{par}};\beta_S)}\delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\mathrm{d}\mathbf{k}}{V\Gamma\sum_\nu \int_{\text{BZ}} \delta(\varepsilon_{\mathbf{k}\nu}-\varepsilon)\mathrm{d}\mathbf{k}}. \tag{5.84}$$

We have now constructed an algorithm that can in principle calculate the average scattering rate between electrons at ε and partner electrons at ε_{par}, and by extension with eq. (5.86), the two-electron inelastic scattering operator.

The problem with this algorithm is that, compared to the algorithm for calculating the inelastic scattering operator for phonons, it is extremely computationally intensive. This is not simply because the scattering rates need to be calculated at a range of values of the partner energy and screening wavevector. The main problem is that calculation of the scattering rate *at a single* partner energy and screening wavevector is much more expensive to calculate than scattering rate due to an entire band of phonons. The reason is as follows. For phonon transitions, either the initial or final crystal momentum of the phonon state is zero. For a given change in electron crystal momentum $\Delta\mathbf{k}$, there is therefore only two phonon transitions per band that are consistent with this. In contrast, for electron–electron, there is no similar requirement for either the initial or the final partner state to have zero momentum, and so there is a $2-\text{D}$ space of different possible electron transitions— associated with the different possible initial states of the partner— that are consistent a single change in the crystal momentum of the electron of $\Delta\mathbf{k}$. Therefore, the two processes that take order of magnitude the same amount of time to compute is the average scattering rate between an electron at a given energy and an entire band of phonons, and the scattering rate between an electron at a given energy and a single initial (electron) partner state $(\mathbf{p}\nu)$.

In order to make the calculation of the two-particle inelastic electron–electron scattering operator complete a reasonable amount of time, we may need to simplify the distribution of partners. In order to do this we need to make assumption on the functional form for the pair scattering rate $\frac{1}{\tau_{\text{ee}}(\varepsilon,\varepsilon_{\text{par}},\beta_S)}$. One obvious assumption such assumption we can

make is that it only depends on the total energy:

$$\frac{1}{\tau_{\text{ee}}(\varepsilon, \varepsilon_{\text{par}}, \beta_S)} = \frac{1}{\tau_{\text{ee}}^{\text{total}}(\varepsilon + \varepsilon_{\text{par}}, \beta_S)}. \tag{5.85}$$

The justification for this assumption is that the number of possible final states is also only a function of total energy, *if we ignore momentum conservation*. Using this assumption, it is now sufficient to calculate the pair scattering rate for only the six partners at $\varepsilon_{\text{par}} = 0$, or one can average the scattering time $\frac{1}{\tau_{\text{ee}}(\varepsilon_{\text{total}} - \varepsilon_{\mathbf{p}\mu}, \mathbf{p}\mu; \beta_S)}$ across a representative sample of scattering partner states scattered throughout the Brillouin zone. [12]

This line of reasoning can be taken even be further to argue that the two particle inelastic scattering operator can be written in terms of a single phenomenological coupling constant $\tau_{\text{ee}}^{\text{cons}}$, in the following manner:

$$S_{\text{ee}}\left(\varepsilon_i, \varepsilon_i^{\text{par}}; \varepsilon_f, \varepsilon_f^{\text{par}}; \beta_S\right) = \frac{1}{\tau_{\text{ee}}^{\text{cons}}} D(\varepsilon_f) D(\varepsilon_f^{\text{par}}) \delta(\varepsilon_f + \varepsilon_f^{\text{par}} - \varepsilon_i - \varepsilon_i^{\text{par}}). \tag{5.86}$$

This coupling constant can then be estimated either empirically, or again by calculating the electron–electron scattering operator at a representative sample of initial electron states $\mathbf{k}\nu$ and initial partner electrons states $\mathbf{p}\mu$:

$$\frac{1}{\tau_{\text{ee}}^{\text{cons}}} = \left(\int_0^{\varepsilon_{\mathbf{k}\nu} + \varepsilon_{\mathbf{p}\mu}} D(\varepsilon_f) D(\varepsilon_i + \varepsilon_i^{\text{par}} - \varepsilon_f) \mathrm{d}\varepsilon_f \right)^{-1} \times \frac{1}{\tau_{\text{ee}}(\mathbf{k}\nu, \mathbf{p}\mu; \beta_S)}. \tag{5.87}$$

In order to partially account for the error induced by this assumption, one can even make the coupling constant a function of the *average energy* of the electron distribution by changing the representative sample accordingly.

5.3.5 Impact Ionization

The scattering components affected by impact ionization is the velocity relaxation time and the electron creation operator.

[12] This is calculated from eq. (5.82) by setting the weight function w_μ^l at to 1 at $\mathbf{p}^l = \mathbf{p}$, and to 0 everywhere else.

We assume that impact ionization scattering events are totally velocity relaxing, on the basis that while crystal momentum is pseudo-conserved it is partly distributed to a particle of opposite charge, destroying the net current often associated with a net crystal momentum. Therefore the velocity relaxation rate associated with impact ionization is the same as the impact ionization rate given in the background chapter:

$$\begin{aligned}\frac{1}{\tau_{ii}^{v}}(\varepsilon_i) &= \frac{1}{\tau_{ii}}(\varepsilon_i) \\ &= \sum_{i=1} \theta\left(\varepsilon - \varepsilon_i^{\text{thr}}\right) P_i \left(\frac{\varepsilon - \varepsilon_i^{\text{thr}}}{\varepsilon_i^{\text{thr}}}\right)^2 .\end{aligned} \tag{5.88}$$

For the energy-dependent creation operator, we note that the expression given in the background for the impact ionization scattering operator is already energy-independent. Therefore in order to determine the purely inelastic scattering operator we simply need to multiply it by the density of final three-particle states, which is product of three single-particle density of states:

$$\begin{aligned}S(\varepsilon_i;\varepsilon_f,\varepsilon_e,\varepsilon_h) = &\left(\int_0^{\varepsilon-\varepsilon_{\text{gap}}}\int_0^{\varepsilon_i-\varepsilon_{\text{gap}}-\varepsilon_e} D_{\text{con}}(\varepsilon_e)D_{\text{val}}(\varepsilon_h)D_{\text{car}}(\varepsilon_{\mathbf{k}\nu}-\varepsilon_{\text{gap}}-\varepsilon_e-\varepsilon_h)\mathrm{d}\varepsilon_h\mathrm{d}\varepsilon_e\right)^{-1} \\ &\times \frac{1}{\tau_{ii}(\varepsilon_{\mathbf{k}\nu})}\delta(\varepsilon_f+\varepsilon_e+\varepsilon_h+\varepsilon_{\text{gap}}-\varepsilon_i)D(\varepsilon_f)D(\varepsilon_e)D_{\text{val}}(\varepsilon_h).\end{aligned} \tag{5.89}$$

5.3.6 The Energy Dependent Diffusion Tensor

We now need to write down an computable expression for the energy dependent diffusion tensor:

$$\mathcal{D}^{\varepsilon}(\varepsilon) = \frac{\Gamma\sum_\nu \int_{\text{BZ}} \tau^{v}_{\text{relax}}(\mathbf{k}\nu)\mathbf{v}(\mathbf{k}\nu)\otimes\mathbf{v}(\mathbf{k}\nu)\delta(\varepsilon-\varepsilon_{\mathbf{k}\nu})\mathrm{d}\mathbf{k}}{\Gamma\sum_\nu\int_{\text{BZ}}\delta(\varepsilon-\varepsilon_{\mathbf{k}\nu})\mathrm{d}\mathbf{k}}. \tag{5.90}$$

The first thing we require is an expression for the velocity relaxation time *as a function of* $\mathbf{k}\nu$. If $|v|(\varepsilon)$ is the average speed as a function of energy, then the $\mathbf{k}\nu$ dependent expression for the velocity relaxation rate is as follows:

$$\frac{1}{\tau^{v}_{\text{par}}}(\mathbf{k}\nu) = \frac{1}{\langle|\mathbf{v}|\rangle_\varepsilon}\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{par}}(\mathbf{k}\nu). \tag{5.91}$$

We make the assumption that impact ionization events are perfectly velocity relaxing, and that electron–electron scattering has no net velocity relaxation rate since these collisions pseudoconserve crystal momentum. This leads to the following expression for the total velocity relaxation time:

$$\begin{aligned}\frac{1}{\tau^v}(\mathbf{k}\nu,\beta_S) &= N_{\text{dop}}\frac{1}{\langle|\mathbf{v}|\rangle_\varepsilon}\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{unit-dop}}(\mathbf{k}\nu,\beta_S)+\left(\frac{1}{\langle|\mathbf{v}|\rangle_\varepsilon}\left(\frac{\partial|\mathbf{v}|}{\partial t}\right)_{\text{pho}}(\mathbf{k}\nu)+\frac{1}{\tau_{ii}(\mathbf{k}\nu)}\right)\\ &= \frac{N_{\text{dop}}}{\tau^v_{\text{unit-dop}}(\mathbf{k}\nu,\beta_S)}+\frac{1}{\tau^v_{\text{non-dop}}}.\end{aligned} \tag{5.92}$$

We wish to tabulate the energy-dependent diffusion coefficient. We could tabulate it entirely, in terms of ε, N_{dop}, and β_S. However, we can avoid tabulating in terms of N_{dop} if we make the simple approximation that we can tabulate the diffusion coefficients for dopants and non-dopants separately, and can combine them to find a total diffusion constant as we would velocity relaxation times to find a total relaxation time:

$$\begin{aligned}\mathcal{D}^\varepsilon(\varepsilon) &= \left(\frac{1}{\mathcal{D}^\varepsilon_{\text{dop}}(\varepsilon,\beta_S,N_{\text{dop}})}+\frac{1}{\mathcal{D}^\varepsilon_{\text{non-dop}}(\varepsilon)}\right)^{-1}\\ &= \left(\frac{N_{\text{dop}}}{\mathcal{D}^\varepsilon_{\text{unit-dop}}(\varepsilon,\beta_S)}+\frac{1}{\mathcal{D}^\varepsilon_{\text{non-dop}}(\varepsilon)}\right)^{-1}.\end{aligned} \tag{5.93}$$

We can now write down a computable expression for both $\mathcal{D}^\varepsilon(\varepsilon)$, using the same technique as was used for the calculation of eq. (5.44) since we are interested in a constant energy surface:

$$\mathcal{D}^\varepsilon_{\text{unit-dop}}(\varepsilon,\beta_S) = \frac{1}{D(\varepsilon)}\sum_\nu{\sum_{i,j}}' \tau^v_{\text{unit/non-dop}}(\mathbf{k}^{i,j}_{\text{centre}}\nu,\beta_S)\,\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu)\otimes\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu)\,D^{i,j}_\nu(\varepsilon), \tag{5.94}$$

$$\mathcal{D}^\varepsilon_{\text{non-dop}}(\varepsilon) = \frac{1}{D(\varepsilon)}\sum_\nu{\sum_{i,j}}' \tau^v_{\text{unit/non-dop}}(\mathbf{k}^{i,j}_{\text{centre}}\nu)\,\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu)\otimes\mathbf{v}(\mathbf{k}^{i,j}_{\text{centre}}\nu)\,D^{i,j}_\nu(\varepsilon). \tag{5.95}$$

Here the density of states is given by the following obviously computable expression:

$$D(\varepsilon) = \sum_\nu{\sum_{i,j}}' D^{i,j}_\nu(\varepsilon). \tag{5.96}$$

Aside: Exploiting Brillouin Zone Point Symmetry

The band structure has the following point symmetry relations:

$$\varepsilon(k_x, k_y, k_x) = \varepsilon(k_a, k_b, k_c), \tag{5.97}$$

where
$$\begin{aligned} k_a &\in \{k_x, -k_x, k_y, -k_y, k_z, -k_z\}, \\ k_b &\in \{k_x, -k_x, k_y, -k_y, k_z, -k_z\} \backslash \{k_a, -k_a\}, \\ k_c &\in \{k_x, -k_x, k_y, -k_y, k_z, -k_z\} \backslash \{k_a, -k_a, k_b, -k_b\}. \end{aligned}$$

Since there are 6 choices for k_a, 4 choices for k_b, and 2 choices for k_c, the band structure can be reproduced from a wedge in the Brillouin zone that is $\frac{1}{48}^{\text{th}}$ the size of the Brillouin zone. A natural way to define such a wedge is to enforce the following:

$$k_a \geq k_b \geq k_c \geq 0. \tag{5.98}$$

It is worth understanding the extent to which these point symmetries can be used to speed up computation.

Put simply, integrals over the Brillouin zone can be reduced to an irreducible Brillouin zone so long as the integrand has the same point symmetries as the Brillouin zone. This is *not* true for scattering or bandstructure functions that depend on multiple points on the Brillouin zone such as $S^{\eta}_{\text{pho}}(\mathbf{k}\nu, \mathbf{k}'\nu')$, since the second point breaks the symmetry, but is true for scattering or bandstructure functions that depend explicitly only on one point in the Brillouin zone, such as $\frac{1}{\tau^{\eta}_{\text{pho}}(\mathbf{k}\nu)}$.

For scalar functions, this transformation of the integral is simple. The integral is simply performed over the irreducible wedge and multiplied by the number of symmetries— in the case of silicon this is 48. For non-scalar functions, such tensorial term in the diffusion coefficient $\mathbf{v}(\mathbf{k}\nu) \otimes \mathbf{v}(\mathbf{k}\nu)$, the

transformation is more subtle.

Suppose that $\mathbf{v}(k_x, k_y, k_z) = \mathbf{T}\mathbf{v}(k_a, k_b, k_c)$ where $\mathbf{T}$ is a linear transformation. We can then write down an expression for the transfomation for the tensor $\mathbf{v}(k_x, k_y, k_z) \otimes \mathbf{k}(k_x, k_y, k_z)$ as follows:

$$\begin{aligned}\mathbf{v}(k_x, k_y, k_z) \otimes \mathbf{v}(k_x, k_y, k_z) &= (\mathbf{T}\mathbf{v}(k_a, k_b, k_c)) \otimes (\mathbf{v}(k_a, k_b, k_c)) \\ &= (\mathbf{T} \otimes \mathbf{T})\, \mathbf{v}(k_a, k_b, k_c) \otimes \mathbf{v}(k_a, k_b, k_c). \quad (5.99)\end{aligned}$$

If we suppose that $\mathbf{T}$ is a different constant for each wedge, then we simply need to find the multiply the integral over the irreducible wedge, by the sum of the 48 different transformation vectors. So we simply need to find an expression for $\mathbf{T}$. We start by writing down an expression for $v_x(k_x, k_y, k_z)$ in terms of $\mathbf{v}(k_a, k_b, k_c)$:

$$\begin{aligned}v_x(k_x, k_y, k_z) &= \frac{\partial \varepsilon(k_x, k_y, k_z)}{\partial k_x} \\ &= \frac{\partial \varepsilon(k_a, k_b, k_c)}{\partial k_x} \\ &= \frac{\partial (k_a, k_b, k_c)}{\partial k_x} \cdot \nabla_{(k_a, k_b, k_c)} \varepsilon(k_a, k_b, k_c) \\ &= \mathbf{T}_x \cdot \mathbf{v}(k_a, k_b, k_c). \quad (5.100)\end{aligned}$$

We can then extend this to $v_y(k_x, k_y, k_z)$ and $v_z(k_x, k_y, k_z)$ to find the following:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_x \\ \mathbf{T}_y \\ \mathbf{T}_z \end{bmatrix} = \begin{bmatrix} \frac{\partial (k_a, k_b, k_c)}{\partial k_x} \\ \frac{\partial (k_a, k_b, k_c)}{\partial k_y} \\ \frac{\partial (k_a, k_b, k_c)}{\partial k_x} \end{bmatrix}. \quad (5.101)$$

Here the definition of (k_a, k_b, k_c) implies that the vectors $\mathbf{T}_x$, $\mathbf{T}_y$, and $\mathbf{T}_z$ must be limited to the following possibilities:

$$\begin{aligned}
&\mathbf{T}_x \in \{(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1),\\
&\mathbf{T}_y \in \{(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)\} \backslash \{\pm \mathbf{T}_x\},\\
&\mathbf{T}_z \in \{(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)\} \backslash \{\pm \mathbf{T}_z, \pm \mathbf{T}_y\}.
\end{aligned} \tag{5.102}$$

Thus a typical example will look like the following:

$$\mathbf{T} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}. \tag{5.103}$$

Which implies that $k_x = -k_b$, $k_y = k_a$, $k_z = -k_c$. Suppose we label the 48 possible transformation tensors by k. The diffusion tensor will then be related to the diffusion tensor on an irreducible wedge in the following manner:

$$\begin{aligned}
\mathcal{D}^{\varepsilon}_{\text{BZ}}(\varepsilon) &= \left(\frac{1}{48} \sum_k \mathbf{T}^k \otimes \mathbf{T}^k \right) \mathcal{D}^{\varepsilon}_{\text{Wedge}}(\varepsilon) \\
&= \begin{bmatrix}
\frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3}
\end{bmatrix} \mathcal{D}^{\varepsilon}_{\text{Wedge}}(\varepsilon) \\
&= \mathbf{I} \left(\frac{1}{3} \mathcal{D}^{\varepsilon,xx}_{\text{Wedge}} + \frac{1}{3} \mathcal{D}^{\varepsilon,yy} + \frac{1}{3} \mathcal{D}^{\varepsilon,zz}_{\text{Wedge}} \right).
\end{aligned} \tag{5.104}$$

That is, no matter what the diffusion tensor we determine on the irreducible wedge of the Brillouin zone, the diffusion tensor over the Brillouin

zone is equivalent to a scalar given by average of its diagonal entries.[13] Thus in any material with dioctahedral symmetry, such as unstrained silicon, the diffusion coefficient must be a simple scalar.

---

[13] We have assumed here that the integral in the denominator of the wedge diffusion coefficient, that is the density of states, has also been restricted to a wedge.

Chapter 6

Results II: Three Quasi-Equilibria Transport

6.1 Introduction

In the first results chapter, we derived a model of semiclassical non-local transport by assuming that the distribution function relaxed to an Elastically Constrained Quasi-Equilibrium on a time-scale similar to the scattering time. The result was a simple model Here the electron state is defined by an energy and position dependent distribution function subject to pure diffusion at constant total energy, and pure inelastic local scattering. The only issue with this model is that it is a couple of orders of magnitude slower to solve than local macroscopic models.

The aim of this chapter is to derive a model of semiclassical non-local transport that can be solved at a speed comparable to existing local macroscopic models. In order to do this, we need to make a stronger assumption than the Elastically Constrained Quasi-Equilibrium assumption. The Elastically Constrained Quasi-Equilibrium is a maximum entropy state subject to a $1-$D scalar field of local constraints— the particle density associated with each energy level. In order to simplify the model, we need to assume the local electron state is defined by a small finite set scalars, i.e. a small finite set $0-$D

scalar fields.

In order to achieve this, we view the electron distribution as the superposition of three seperate quasi-equilibria, each of which is defined as the maximum entropy distribution constrained by a few scalars. The result of this model is an electron state defined by 6 position dependent scalars. If electron–electron scattering between high-energy electrons is not strong, then one of the quasi-equilibria can be neglected and the resulting electron state can be defined by 4 position dependent scalars.

As a result of the small amount of information required to define the electron state, the Three Quasi-Equilibria model can in principle be solved in a similar amount of time as it takes to solve as a local macroscopic model. However, in the same way that the Elastically-Constrained Equilibria transport model only provides device solutions more efficiently than the full Boltzmann transport equation if the *energy dependant* transport parameters are *precomputed and stored in a look-up table*, similarly the Three Quasi-Equilibria transport model only provides device solutions more efficiently than the Elastically-Constrained Equilibria transport model if the *ansatz-parameter dependent* transport parameters are *precomputed and stored in a look-up table*. Thus the speed increase of the Three Quasi-Equilibria model comes at the one-time upfront CPU cost and persistent memory and cost of calculating and storing a large look-up table of ansatz-parameter dependent transport parameters.

We divide this chapter into three major sections. In the first section, we describe the Three Quasi-Equilibria ansatz. In the second section, we derive an unclosed equation of motion for the ansatz parameters. In the third section, we close the equation of motion by analytically relating the undefined terms to a large set of *macroscopic transport parameters*. These macroscopic transport parameters are our precomputed quantities, and are defined in terms of integrals of each component quasi-equilibrium over the bandstructure and scattering operator.

6.2 The Three Quasi-Equilibria Ansatz

6.2.1 Overview

The purpose of this section is firstly to define the class of devices that the Three Quasi-Equilibria transport model applies to, secondly to describe how the carriers in the device will be represented, and thirdly to describe how the ansatz evolves in response to inelastic scattering. We can summarize the content of this section as follows.

- We restrict ourselves to investigating semiclassical electron transport in source-drain devices Here inelastic scattering is broadband, and all states are elastically connected to the source and/or drain.

- We separate the electron population into three populations: a "cold" electron population— primarily associated with electrons injected from the drain; a "stunted" population— primarily associated with electrons injected from the source; and a "warm" electron population— primarily associated with stunted electrons that have scattered with each other.

- We assume that the cold electron population is a small perturbation from a lattice temperature Fermi-Dirac distribution.

- We assume that the warm electron population is a small perturbation from a heated Maxwell-Boltzmann distribution.

- We assume that the stunted electron population is a small perturbation from the non-degenerate maximum entropy distribution associated with a given density, energy density, and maximum chemical potential. This is the most idiosyncratic assumption in our model.

- We describe the qualitative effect of inelastic scattering on these populations.

6.2.2 The Limitations on Device Geometry

Beyond the assumptions of the Elastically-Constrained Equilibrium transport model, the Three Quasi-Equilibria transport model described in this chapter is intrinsically limited to semiclassical devices in which inelastic scattering processes are broad spectrum relative to the mean energy exchange.[1] In the interest of presenting the critical kernel of this model as explicitly as possible, we furthermore assume the device is a homogeneous "source-drain" type device, and the transport is monopolar, and that there are no *classically trapped* states. Finally, for convenience, we assume the homogeneous material is silicon, and that the monopolar carriers are conduction electrons. A device meets these constraints if charge transport is dominated by

1. the movement of electrons,

2. from a single "source" terminal to a single "drain" terminal,

3. via states that are elastically connected at least one of these terminals

4. through a homogeneous piece of silicon,

5. that is of a characteristic length much larger than the decoherence length of the electron.

The source is defined as the terminal in contact with the reservoir of electrons which has a higher chemical potential and the drain as the terminal in contact with the reservoir of electrons which has the lower chemical potential. We note that there may be other terminals in such a device— for example, gates— but we make the restriction that charges injected into the device from these additional terminals can be assumed *not to occupy the same physical volume as the charges that are injected by the source or drain*. As such, the charges from other terminals can only have long-range interactions with source and drain electrons that are mediated by Poisson's equation. In contrast, source and drain electrons may have short-range interactions with one another.

[1] That is, inelastic scattering is *does not* consist of sharply defined energy transitions.

6.2.3 The Three Quasi-Equilibrium Populations

We partition the carriers in our device into three populations:

"Stunted" electrons A population of electrons that is assumed to be an antisymmetric perturbation from a chemical potential constrained quasi-equilibrium, an atypical constrained quasi-equilibria that has been defined in the theoretical framework.

"Warm" electrons A population of electrons that is assumed to be an antisymmetric from an internal equilibrium with one another that is generally hotter than the lattice.

"Cold" electrons A population of electrons that are assumed to be an antisymmetric perturbation from thermal equilibrium with the lattice.

For the various electron populations, we will make mixed degeneracy assumptions. We will assume that degenerate statistics apply for the cold electron population, but we will assume that non-degenerate statistics apply for both the warm and the stunted electron population. We note that this assumption of non-degeneracy for these populations means that we ignore the often significant reduction in the empty density of states by the cold electron population. This assumption is justified using the argument that in typical non-equilibrium transport scenarios, the vast majority of occupied warm and stunted electron states are at energies much higher than fermi-level of the cold electron population. These degeneracy assumptions are actually usually not most problematic at the overlap between the stunted electron distribution and the cold electron distribution, but rather in the places Here the stunted electron distribution near lattice temperature equilibrium near the source. As will be seen later, these degeneracy effects can be roughly accounted for in a simple manner.

In the absence of electron–electron scattering, an electron which enters the device from the source terminal would be considered to belong to the stunted electron population, an electron which enters the drain terminal would be considered to belong to the cold electron population, and the warm electron population would be empty. The particle transfers induced by electron–electron scattering are discussed in the next section.

As shown in the theoretical framework, the non-degenerate limit of the distribution function associated with a chemically constrained equilibria is given by following the three parameter function:

$$f_\varepsilon = \begin{cases} e^{-(\alpha^\varepsilon + \beta\varepsilon)} & \varepsilon \leq \varepsilon^*, \\ e^{-\frac{\varepsilon - \mu^\varepsilon}{kT_L}} & \varepsilon \geq \varepsilon^*. \end{cases} \tag{6.1a}$$

This is a three parameter functions since α^ε can be defined in terms of ε^*, β, and μ^ε by noting that the maximum entropy distribution is continuous at ε^*:

$$\alpha^\varepsilon = -\frac{\mu^\varepsilon}{kT_L} + \varepsilon^* \left(\frac{1}{kT_L} - \beta \right). \tag{6.1b}$$

We refer to α^ε SLOTBOOM ALPHA, μ^ε as the SLOTBOOM CHEMICAL POTENTIAL, and ε^* as the KNEE ENERGY. The prefix "Slotboom" for α^ε and μ^ε, refers to the fact that these are terms relate to kinetic energies ε rather than total energies H. If we write the stunted distribution ansatz in terms of a total energy H,[2] we can do so as follows:

$$f_H = \begin{cases} e^{-(\alpha^H + \beta H)} & H \leq H^*, \\ e^{-\frac{H - \mu^H}{kT_L}} & H \geq H^*. \end{cases} \tag{6.1c}$$

Here continuity implies the following relation:

$$\alpha^H = -\frac{\mu^H}{kT_L} + H^* \left(\frac{1}{kT_L} - \beta \right). \tag{6.1d}$$

In this case we use the prefix "thermodynamic" for the relevant variables, and so refer to α^H as THERMODYNAMIC ALPHA, and μ^H as the THERMODYNAMIC CHEMICAL POTENTIAL. For the same reason that f_ε and f_H were both useful forms of the energy dependent distribution function, the Slotboom forms and the thermodynamic forms are both useful forms of the stunted electron distribution parameters. The mapping from one form to another is trivial if one knows the local potential, so one should consider them to be different expressions of the same set of variables.

The parameter β is the same in both eq. (6.1) and eq. (6.2), so while we could use either

[2]By *kinetic energy* ε, we simple mean the difference between total energy H, and the potential energy of the conduction band minima ε_C, so $H = \varepsilon + \varepsilon_C$.

prefix, we will generally refer to it as termthermodynamic beta. Thermodynamic beta is most intuitively thought of in terms of the corresponding "bulk" temperature:

$$T_{\text{bulk}} = \frac{k}{\beta}. \tag{6.1e}$$

The reason we prefer to use the thermodynamic beta formulation rather than temperature to describe the energy dependence of the stunted distribution function below the knee energy, is that it is possible for the bulk temperature to be *infinite* or even *negative*. Accordingly, it makes more sense store transport parameters in terms of thermodynamic beta rather than the temperature, which instead of crossing from positive infinity to negative infinity, infinitesimally similar distribution functions, simply crosses from above zero to below zero. As a side note, this is also explains why we refer to warm electron population as "warm" rather than "hot"— because typically it is the bulk of the stunted electron distribution that contains the hottest electrons.

Note that we have omitted any qualification in these names related to the fact that they refer only to the stunted electron population, and we have also omitted the qualification that the chemical potential is in fact a *maximum* chemical potential for the stunted distribution. This is a convenient notational choice owing to the fact that the derivation of the stunted electron transport equations is the least trivial, and thus we spend the most time discussing it. Accordingly it makes sense for variables to the stunted distribution by default and require qualifications only when referring to other populations. So we refer to the stunted electron distribution by f, rather than by f^{stunted}, and instead refer to the total electron distribution function is denoted with a superscript "total". The ansatz for the two forms of the warm electron energy distribution is given as follows:

$$f_{\varepsilon}^{\text{warm}} = e^{-\left(\alpha_{\text{warm}}^{\varepsilon} + \frac{\varepsilon}{kT_{\text{warm}}}\right)}, \tag{6.2a}$$

$$f_{H}^{\text{warm}} = e^{-\left(\alpha_{\text{warm}}^{H} + \frac{H}{kT_{\text{warm}}}\right)}. \tag{6.2b}$$

Here were use thermodynamic alpha formulation again in order to avoid confusion with the chemical potential, which we would like to have always measured relative to the *lattice temperature*. Finally the ansatz for the two forms of the cold electron energy distribution is given as follows:

$$f_\varepsilon^{\text{cold}} = \frac{1}{e^{\frac{\varepsilon - \mu_{\text{cold}}^\varepsilon}{kT_L}} + 1}, \tag{6.3a}$$

$$f_H^{\text{cold}} = \frac{1}{e^{\frac{H - \mu_{\text{cold}}^H}{kT_L}} + 1}. \tag{6.3b}$$

And our ansatz for the two forms of the total energy distribution function is as follows:

$$f_\varepsilon^{\text{total}} = f_\varepsilon + f_\varepsilon^{\text{warm}} + f_\varepsilon^{\text{cold}}, \tag{6.4a}$$

$$f_H^{\text{total}} = f_H + f_H^{\text{warm}} + f_H^{\text{cold}}. \tag{6.4b}$$

6.2.4 The Effect of Scattering on the Populations

We define CARRIER–LATTICE scattering as all scattering that is not carrier-carrier scattering. In the case of DAMOCLES scattering operator for silicon described in the background chapter this includes all phonon scattering, dopant scattering and impact ionization processes.

We assume that an electron subject to carrier–lattice scattering, the *population type* is conserved. Accordingly, the purely inelastic effect of these scattering types is only to change the CHARACTERISTIC DENSITIES associated with each population: that is, the set of densities sufficient to characterize the ansatz for the population. The particle density is sufficient to characterize the cold electron population, for the warm electron we also require the energy density, and finally for the stunted population we require *particle density above the knee energy* in addition to the other densities.

In the deep sub-micrometer scale of devices we are interested in, the strength of doping is sufficiently strong that it is typically unreasonable to assume the electron–electron scattering is negligible. As such, any model which does not incorporate these effects is bound to be unphysical.

When electron–electron scattering is strong, it qualitatively changes the pure inelastic scattering operator. For an input distribution of consisting of a delta function at an en-

ergy ε, in one electron–electron scattering time the distribution very broadly distributed between 0 and 2ε. Most importantly, unlike carrier–lattice scattering, it does not generally preserve the maximum chemical potential. This is problematic for the ansatz of the cold electron population and the stunted electron population, both of which assume the existence of a maximum chemical potential.

We assume that the effect of electron–electron on the pure-inelastic scattering operator can be accounted for by *moving* electrons between our three quasi-equilibrium distributions. The precise movement that occurs depends on the initial populations of the electrons involved in scattering. Accordingly, we need to break down the various possible short-range electron–electron interactions according to the pair of populations interacting. In general, n populations can combine in $\frac{n(n+1)}{2}$ ways, implying that three populations combine in six ways. These are as follows.

Stunted—Stunted Electron Scattering This will result both electrons no longer being in limited by the local maximum chemical potential for the stunted population. The concept of a maximum chemical potential was derived on the assumption that while stunted electrons are not in thermal equilibrium themselves, they only interact inelastically with scattering partners at thermal equilibrium. It was proved that in this scenario, scattering leads to the maximum chemical potential decreasing monotonically. Accordingly, the electrons states after stunted—stunted electron scattering will not be well described by the stunted electron ansatz. We make the assumption that both electrons involved in such an interaction are moved into the warm electron population. Note, this will not quite be true, since the resulting distribution will in fact have a sharp cutoff at twice the knee-energy, but we consider this fact of minimal importance, since the higher the cutoff energy relative to the average energy, the less of an influence it has on the shape of the distribution function.

Warm—Warm Electron Scattering This will not effect the validity of the warm electron ansatz, and in fact will only make it more accurate.

Cold—Cold Electron Scattering This will not effect the validity of the cold electron ansatz these will not act to drive the cold electrons away from thermal equilib-

rium. Therefore it is completely physically justified to ignore cold—cold electron–electron scattering.

Stunted—Warm Electron Scattering This removes an electron from the stunted population, and adds it to the warm electron population.

Stunted—Cold Electron Scattering This is an interaction between stunted electrons and a set of partner bodies that are in thermal equilibrium. The assumption that stunted electrons only scatter with partner bodies in a lattice temperature thermal equilibrium is maintained, and accordingly the concept of a maximum chemical potential for the stunted distribution still holds. Thus when a cold electron interacts with a stunted electron, an electron is removed from the cold population and added to the stunted population.

Warm—Cold Electron Scattering This removes the electron from the cold population, and adds it to the warm electron population.

Finally, we assume that the population transfer effects associated with plasmon scattering— the *long-range* scattering between a single electron and the collective state of the electrons— is always the same: the particle is removed from its current population and a particle is added to the warm electron distribution.[3]

We note that we assume that moving an electron from population A to population B involves decreasing the total energy of population A and increasing the total energy of population B by the average energy of A, and decreasing the population number of A by one and increasing the population number of B by one. This assumption neglects the fact that the average energy of electrons that scatter may not be the same as the average energy of the electrons in the population. If it is found that there is important differences between these two averages, this error can be removed simply by precomputing both types of average energy.

[3]Note that we do not describe how to calculate the plasmon scattering rate in this thesis, as we have been using the DAMOCLES scattering operator, and in the DAMOCLES program scattering is modelled implicitly by updating Poisson's equation at a very high rate. However, this implicit model of plasmon scattering is not compatible with either the Elastically Constrained Transport model or the Three Quasi-Equilibria Model. We have left defining the plasmon scattering operator as a problem for future work.

Thus we have described our model of the purely inelastic, population changing effects of electron–electron scattering. We further assume that these *are the only effects* electron–electron scattering has. That is, we assume that the elastic relaxation rate of electron–electron scattering is zero, since such scattering pseudoconserves the total crystal momentum in the electron distribution.[4] This is an assumption, since in a complex band structure, pseudoconserving total crystal momentum is not the same as conserving total velocity except at low energies. Accordingly, a strictly more accurate method to model the effects of electron–electron scattering would be to assume that the electron–electron scattering forms a *drifted* quasi-equilibrium that conserves total crystal momentum. It is not obvious however, how such a model— which is considerably more complex— would result in systematic changes to the particle flux distribution in energy *apart from the caused indirectly by the inelastic changes to electrons*.

6.3 Generic Equations of Motion for the Ansatz Parameters

6.3.1 Overview

The purpose of this section is to formulate an equation of motion[5] for the ansatz parameters. The equations of motion derived in this section contain more unknowns variables than equations, and are therefore *open*. The section can be summarized as follows.

- We relate the ansatz parameters to characteristic densities
- We derive continuity equations— or equations of motion— for these characteristic densities by taking weighted integrals of the Boltzmann transport equation.
- We relate the time derivative in the characteristic densities to the time derivative in the ansatz parameters.

[4] By "pseudoconserves", we simply mean conserves modulo a reciprocal lattice vector.
[5] By *equation of motion* we simply mean an equation that defines the time derivative in terms of the

6.3.2 The Continuity Equations for the Characteristic Densities

The Stunted Electron Population

Our ansatz is that the stunted electron distribution equation is an antisymmetric perturbation from an energy distribution characterized by *only three scalar quantities* at *each* point in space and time. There is a large choice of sets of scalar quantities we can use. For instance we can use any set of three independent quantities the equations presented in eq. (6.1). However, for the same reason that the most intuitive continuity equations associated with the Elastically Constrained Transport model involved the *densities* that characterize the energy dependent distribution function, the most intuitive continuity equations associated with the stunted electron transport equations involve the *densities* that characterize the distribution function.

The stunted electron distribution function is formally the maximum entropy distribution subject to a local particle density, local energy density and local maximum chemical potential. We describe rate of change in particle and energy density directly, by deriving a continuity equation for these quantities. We describe the rate of change in the maximum chemical potential *indirectly*, by finding a continuity equation for the rate of change in density above the knee energy. Since the knee energy is a function of position and time, and the continuity equations involve derivatives with respect to position and time that do not intend the knee energy to change, there is a high chance of confusion if we use the term *knee energy* to describe both the attribute of the ansatz, and the cutoff energy of the continuity equation. To avoid this confusion, will refer to the knee energy in the context of a continuity equation as the CUTOFF ENERGY ε_{cut}.

The continuity equations we derive for these characteristic densities are physically intuitive. Each states that, in a small volume, the rate of change of a conserved quantity— either particles, energy, or particles above the cutoff energy— is equal to the net rate of flow of the conserved quantity into the volume, plus the rate of change of the conserved quantity due to scattering in the volume, plus the rate of change of the conserved

instantaneous state. In this case, an equation that defines the time derivative of the ansatz parameters in terms of the instantaneous state of a $3-$D field of ansatz parameters.

quantity due to the local electric field. Accordingly, it would arguably be legitimate to simply state these continuity equations without derivation. We take the longer route of describing how they are derived from the Boltzmann transport equation simply for the sake of rigor and completeness.

In order to derive these equations, we need to take appropriate weighted integrals of the Boltzmann Transport Equation. To assist us in this regard, we define the m^{th}-order energy density functional, or m^{th}-order RHO FUNCTIONAL $\rho^m(\varepsilon_{\text{min}}, \varepsilon_{\text{max}})$ as follows:

$$\begin{aligned}\rho^m(\varepsilon_{\text{min}}, \varepsilon_{\text{max}})[X] &= \Gamma \sum_\nu \int_{\text{BZ}} \Big(\theta(\varepsilon_{\mathbf{k}\nu} - \varepsilon_1)\theta(\varepsilon_2 - \varepsilon_{\mathbf{k}\nu})\Big)\varepsilon_{\mathbf{k}\nu}^m X \mathrm{d}\mathbf{k} \\ &= \int_{\varepsilon_{\text{min}}}^{\varepsilon_{\text{max}}} \varepsilon^m \sigma^\varepsilon(\varepsilon)[X] \mathrm{d}\varepsilon. \end{aligned} \tag{6.5}$$

Here $\theta(x)$ is the Heaviside step function and $\sigma^\varepsilon(\varepsilon)[X]$ is the Constant Energy Surface functional defined in the first results chapter. The m^{th}-order Rho Functional therefore measures the integral of an arbitrary function between two constant energy surfaces in the Brillouin Zone, weighted by the m^{th} power of the kinetic energy.

Let us apply the Rho Functional to both sides of the Boltzmann transport equation:

$$\rho^m(\varepsilon_{\text{min}}, \varepsilon_{\text{max}}) \left[\frac{\partial f}{\partial t}\right] = \rho^m(\varepsilon_{\text{min}}, \varepsilon_{\text{max}}) \left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}} - \mathbf{v} \cdot \nabla_{\mathbf{r}} f - \frac{\mathbf{F}}{\hbar} \nabla_{\mathbf{k}} f\right]. \tag{6.6}$$

Using eq. (6.5), the definition of the Rho Functional in terms of the Constant Energy Surface functional, and eq. (5.23), the expression from the first results chapter associated with the application of the Constant Energy Surface functional on the Boltzmann transport equation, we have the following:

$$\int_{\varepsilon_{\text{min}}}^{\varepsilon_{\text{max}}} \frac{\partial n^\varepsilon}{\partial t} \mathrm{d}\varepsilon = \int_{\varepsilon_{\text{min}}}^{\varepsilon_{\text{max}}} \varepsilon^m \left(\left(\frac{\partial n^\varepsilon}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}^\varepsilon - \mathbf{F} \cdot \frac{\partial \mathbf{j}^\varepsilon}{\partial \varepsilon}\right) \mathrm{d}\varepsilon. \tag{6.7}$$

For clarity, we substitute back in the relation $n^\varepsilon = D(\varepsilon) f_\varepsilon$ and $\left(\frac{\partial n^\varepsilon}{\partial t}\right)_{\text{scat}} = D(\varepsilon) \left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}}$ of eq. (5.19). The temporal and spatial derivatives are taken with respect to constant energy and so can be taken outside the integral. Finally, we can re-express the derivative of particle flux per kinetic energy using integration by parts. This leads to the following

expression, Here we have labelled all the terms according to their role in the continuity of a density (which we have named ρ):

$$\frac{\partial}{\partial t}\overbrace{\int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \varepsilon^m D(\varepsilon) f_\varepsilon \mathrm{d}\varepsilon}^{\rho} = \overbrace{\int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \varepsilon^m D(\varepsilon) \left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}} \mathrm{d}\varepsilon}^{\left(\frac{\partial \rho}{\partial t}\right)_{\text{scat}}} - \nabla_{\mathbf{r}} \cdot \overbrace{\int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \varepsilon^m \mathbf{j}^\varepsilon \mathrm{d}\varepsilon}^{\mathbf{J}_\rho} \overbrace{- \mathbf{F} \cdot \left(\varepsilon^m \mathbf{j}^\varepsilon \Big|_{\varepsilon_{\min}}^{\varepsilon_{\max}} - m \int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \varepsilon^{m-1} \mathbf{j}^\varepsilon \mathrm{d}\varepsilon \right)}^{\left(\frac{\partial \rho}{\partial t}\right)_{\text{field}}}. \tag{6.8}$$

The specific densities we are interested in the particle density n, energy density by w and particle density above the cutoff energy by $n_{\varepsilon>\varepsilon_{\text{cut}}}$. These are defined by the following values of the Rho Functional:

$$n(\mathbf{r},t) = \rho^0(0,\infty)[f_\varepsilon] \quad = \int_0^\infty D(\varepsilon) f_\varepsilon \mathrm{d}\varepsilon, \tag{6.9a}$$

$$w(\mathbf{r},t) = \rho^1(0,\infty)[f_\varepsilon] \quad = \int_0^\infty \varepsilon D(\varepsilon) f_\varepsilon \mathrm{d}\varepsilon, \tag{6.9b}$$

$$n_{\varepsilon>\varepsilon_{\text{cut}}}(\varepsilon_{\text{cut}},\mathbf{r},t) = \rho^0(\varepsilon_{\text{cut}},\infty)[f_\varepsilon] = \int_{\varepsilon_{\text{cut}}}^\infty D(\varepsilon) f_\varepsilon \mathrm{d}\varepsilon. \tag{6.9c}$$

The corresponding rate of change of the characteristic densities due to scattering is defined as follows:

$$\left(\frac{\partial n}{\partial t}\right)_{\text{scat}}(\mathbf{r},t) = \rho^0(0,\infty)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}\right] \quad = \int_0^\infty D(\varepsilon)\left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}} \mathrm{d}\varepsilon, \tag{6.10a}$$

$$\left(\frac{\partial w}{\partial t}\right)_{\text{scat}}(\mathbf{r},t) = \rho^1(0,\infty)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}\right] \quad = \int_0^\infty \varepsilon D(\varepsilon)\left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}} \mathrm{d}\varepsilon, \tag{6.10b}$$

$$\left(\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}\right)_{\text{scat}}(\varepsilon_{\text{cut}},\mathbf{r},t) = \rho^0(\varepsilon_{\text{cut}},\infty)\left[\left(\frac{\partial f}{\partial t}\right)_{\text{scat}}\right] = \int_{\varepsilon_{\text{cut}}}^\infty D(\varepsilon)\left(\frac{\partial f_\varepsilon}{\partial t}\right)_{\text{scat}} \mathrm{d}\varepsilon. \tag{6.10c}$$

While the corresponding characteristic fluxes are defines as follows:

$$\mathbf{j}(\mathbf{r},t) = \rho^0(0,\infty)[\mathbf{v}f] \quad = \int_0^\infty \mathbf{j}^\varepsilon \mathrm{d}\varepsilon, \tag{6.11a}$$

$$\mathbf{S}(\mathbf{r},t) = \rho^1(0,\infty)[\mathbf{v}f] \quad = \int_0^\infty \varepsilon\mathbf{j}^\varepsilon \mathrm{d}\varepsilon, \tag{6.11b}$$

$$\mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}}(\varepsilon_{\text{cut}},\mathbf{r},t) = \rho^0(\varepsilon_{\text{cut}},\infty)[\mathbf{v}f] = \int_{\varepsilon_{\text{cut}}}^\infty \mathbf{j}^\varepsilon \mathrm{d}\varepsilon. \tag{6.11c}$$

For the term associated with the rate of change in the characteristic densities due to the field, we note the following:

- $\varepsilon^m \mathbf{j}^\varepsilon(\varepsilon,\mathbf{r},t)|^{\varepsilon=0} = 0$, due to the velocity of states being zero at zero energy.
- $\varepsilon^m \mathbf{j}^\varepsilon(\varepsilon,\mathbf{r},t)|^{\varepsilon=\infty} = 0$, due to energy distribution tending exponentially to zero at infinite energy, which dominates the tendency of the ε^m to diverge to infinity for positive values of m.

On the basis of these equalities, we can derive the following:

$$\begin{aligned}\left(\frac{\partial n}{\partial t}\right)_{\text{field}}(\mathbf{r},t) = \rho^0(0,\infty)\left[\frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f\right] \quad &= -\mathbf{F}\cdot\left(\varepsilon^0\mathbf{j}^\varepsilon\Big|_0^\infty - 0\int_0^\infty \varepsilon^{-1}\mathbf{j}^\varepsilon \mathrm{d}\varepsilon\right)\\ &= 0,\end{aligned} \tag{6.12a}$$

$$\begin{aligned}\left(\frac{\partial w}{\partial t}\right)_{\text{field}}(\mathbf{r},t) = \rho^1(0,\infty)\left[\frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f\right] \quad &= -\mathbf{F}\cdot\left(\varepsilon^1\mathbf{j}^\varepsilon\Big|_0^\infty - \int_0^\infty \varepsilon^0\mathbf{j}^\varepsilon \mathrm{d}\varepsilon\right)\\ &= \mathbf{F}\cdot\mathbf{j}(\mathbf{r},t),\end{aligned} \tag{6.12b}$$

$$\begin{aligned}\left(\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}\right)_{\text{field}}(\varepsilon_{\text{cut}},\mathbf{r},t) = \rho^0(\varepsilon_{\text{cut}},\infty)\left[\frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f\right] &= -\mathbf{F}\cdot\left(\varepsilon^0\mathbf{j}^\varepsilon\Big|_{\varepsilon_{\text{cut}}}^\infty - 0\int_0^\infty \varepsilon^{-1}\mathbf{j}^\varepsilon \mathrm{d}\varepsilon\right)\\ &= \mathbf{F}\cdot\mathbf{j}^\varepsilon(\varepsilon_{\text{cut}},\mathbf{r},t).\end{aligned} \tag{6.12c}$$

These relations are intuitively easy to understand. Eq. (6.12a) says that the local electric field does not create to create particles. Eq. (6.12b) states that the component of the particle flux that is flowing parallel to the external force will drive an increase in the kinetic energy. Eq. (6.12c) says that the component of the particle flux per kinetic energy at the cutoff energy that is flowing parallel to the external force will drive particles from below the cutoff energy to above the cutoff energy.

Accordingly, the continuity equation for the particle density is defined as following familiar equation:

$$\boxed{\frac{\partial n}{\partial t} = \left(\frac{\partial n}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}.} \tag{6.13a}$$

The continuity for the energy density is defined as the following expression:

$$\boxed{\frac{\partial w}{\partial t} = \left(\frac{\partial w}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{S} + \mathbf{F} \cdot \mathbf{j}.} \tag{6.13b}$$

Finally, we not that if we express the cutoff for the particle flux in terms of a *total* energy, the term in eq. (6.12c) disappears since the local field does not cause any electrons to change their total energy. Accordingly, as we show formally in an aside at the end of this section, the continuity equation for the density above cutoff energy can be defined as follows:

$$\boxed{\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t} = \left(\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}_{H>H_{\text{cut}}}.} \tag{6.13c}$$

The Warm And Cold Electron Populations

In deriving the continuity equations for the characteristic densities of the stunted electron population, we have simultaneously derived continuity equations for the warm and cold electron populations. This is because the characteristic densities of the warm and cold populations are subsets of the characteristic densities on the stunted electron distribution.

For the warm electron distribution, the characteristic densities are the warm electron particle density and the warm electron energy density. The continuity equation for the warm particle density is a simple copy of eq. (6.13):

$$\boxed{\frac{\partial n_{\text{warm}}}{\partial t} = \left(\frac{\partial n_{\text{warm}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}_{\text{warm}}.} \tag{6.14a}$$

The continuity for the energy density is a simply copy of eq. (6.14):

$$\boxed{\frac{\partial w_{\text{warm}}}{\partial t} = \left(\frac{\partial w_{\text{warm}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{S}_{\text{warm}} + \mathbf{F} \cdot \mathbf{j}_{\text{warm}}.} \tag{6.14b}$$

For the cold electron distribution, the characteristic density is the cold electron particle density:

$$\boxed{\frac{\partial n_{\text{cold}}}{\partial t} = \left(\frac{\partial n_{\text{cold}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}_{\text{cold}}.} \tag{6.15}$$

The terms in the warm and cold continuity equations are defined by simply substituting the stunted electron ansatz f in eq. (6.9), eq. (6.10), and eq. (6.11) with the warm electron ansatz f^{warm} or the cold electron ansatz f^{cold}.

Aside: Simplifying the Tail Density Continuity Equation

The continuity equation for the particle density above ε_{cut} can be derived by substituting $m = 0$, $\varepsilon_{\text{max}} = \infty$, and $\varepsilon_{\text{min}} = \varepsilon_{\text{cut}}$ into eq. (**??**). This leads to the following relation:

$$\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t} = \left(\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}} + \mathbf{F} \cdot \mathbf{j}^{\varepsilon}(\varepsilon_{\text{cut}}, \dots). \tag{6.16}$$

Suppose we define $\mathbf{j}_{H>H_{\text{cut}}}(H_{\text{cut}}, \mathbf{r}, t)$ to be total energy version of $\mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}}(\varepsilon_{\text{cut}}, \mathbf{r}, t)$. Using a similar approach to eq. (5.13), we can express the total energy form of the as following composition of functions, Here $(\varepsilon_{\text{cut}}, \mathbf{r}, t)$ is a vector valued function of $(H_{\text{cut}}, \mathbf{r}, t)$:

$$\mathbf{j}_{H>H_{\text{cut}}} = \mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}} \circ (\varepsilon_{\text{cut}}, \mathbf{r}, t). \tag{6.17}$$

We can now express the spatial derivative in $\mathbf{j}_{H>H_{\text{cut}}}$ using the chain rule,

Here the first factor in the terms on the RHS refers to partial derivatives in the coordinate system $(\varepsilon_{\text{cut}}, \mathbf{r}, t)$, while the second factor refers to partial derivatives in the coordinate system $(H_{\text{cut}}, \mathbf{r}, t)$:

$$\nabla_{\mathbf{r}} \cdot \mathbf{j}_{H>H_{\text{cut}}} = \left(\frac{\partial \mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial \varepsilon_{\text{cut}}} \right) \cdot (\nabla_{\mathbf{r}} \varepsilon_{\text{cut}}) + (\nabla_{\mathbf{r}} \cdot \mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}}) \left(\frac{\partial \mathbf{r}}{\partial \mathbf{r}} \right). \tag{6.18}$$

The term $\frac{\partial \mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial \varepsilon_{\text{cut}}} = -\mathbf{j}^{\varepsilon}(\varepsilon_{\text{cut}}, \dots)$ while $\nabla_{\mathbf{r}} \varepsilon_{\text{cut}} = \nabla_{\mathbf{r}}(H_{\text{cut}} - \varepsilon_C) = \mathbf{F}$. Substituting these relations, and multiplying by -1 leads to the following expression:

$$\nabla_{\mathbf{r}} \cdot \mathbf{j}_{H>H_{\text{cut}}} = \nabla_{\mathbf{r}} \cdot \mathbf{j}_{\varepsilon>\varepsilon_{\text{cut}}} - \mathbf{F} \cdot \mathbf{j}^{\varepsilon}(\varepsilon_{\text{cut}}, \dots). \tag{6.19}$$

This can clearly be substituted into eq. (6.16) in order to yield eq. (6.14).

6.3.3 Converting the Characteristic Continuity Equations Into Equations of Motion for Ansatz Parameters

The Stunted Electron Distribution

In time-varying analysis the continuity equations will produce a non-zero rate of change for the characteristic densities. If the characteristic densities are the fundamental variables, then it is obvious how to update the state of the fundamental variables in a given time step. For this reason, characteristic densities are typically used as the fundamental variables in time-varying analysis in most transport models.

However in the Three Quasi-Equilibria model we propose, using characteristic densities as the fundamental variables is quite problematic, since it is awkward to precompute

and store the transport parameters as functions of the functions of the characteristic densities. It is far more natural to store these parameters as functions of the Slotboom version of the ansatz parameters. Accordingly, in this subsection we convert the rate of change of characteristic densities into a rate of change of ansatz parameters. Accordingly, we will make the ansatz parameters the fundamental variables of our model *even in time-varying analysis*. As such, in this subsection we describe how to determine equations of motion for the ansatz parameters from equations of motion— a.k.a. continuity equations— for the characteristic densities.

Suppose we have an expression for each of the characteristic densities in terms of the ansatz parameters. By using the multivariate chain rule, we can define a relationship between the rate of change in the characteristic densities and the rate of change in the ansatz parameters:

$$\begin{bmatrix} \frac{\partial n}{\partial t} \\ \frac{\partial w}{\partial t} \\ \frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial n}{\partial \mu^{\varepsilon}} & \frac{\partial n}{\partial \beta} & \frac{\partial n}{\partial \varepsilon^*} \\ \frac{\partial w}{\partial \mu^{\varepsilon}} & \frac{\partial w}{\partial \beta} & \frac{\partial w}{\partial \varepsilon^*} \\ \frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial \mu^{\varepsilon}} & \frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial \beta} & \frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial \varepsilon^*} \end{bmatrix} \begin{bmatrix} \frac{\partial \mu^{\varepsilon}}{\partial t} \\ \frac{\partial \beta}{\partial t} \\ \frac{\partial \varepsilon^*}{\partial t} \end{bmatrix}. \tag{6.20}$$

Here once again we use ε_{cut} only to emphasize the fact that the cutoff energy does not change as the knee energy ε^* varies. This partial derivative is the most interesting of the terms in the matrix, because it is not immediately clear that it is well-defined. However in Fig. 6.1, we argue that it is in fact well-defined, and is equal to zero. Additionally, we note that at constant ε^* and μ^{ε}, a variation in β will not effect the density above ε^*. Accordingly, two of the terms on the bottom row of the matrix in eq. (6.20) are zero:

$$\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial \varepsilon^*} = \frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial \beta} = 0. \tag{6.21}$$

Because of this, we will not begin by inverting the whole matrix in eq. (6.20). Instead, we note that the equation associated with the third row of eq. (6.20) describes a one-to-one relationship between $\frac{\partial \mu^{\varepsilon}}{\partial t}$ and $\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}$ that can be inverted separately. This inversion leads to a simple expression for the rate of change of the chemical potential:

$$\boxed{\frac{\partial \mu^{\varepsilon}}{\partial t} = \left(\frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial \mu^{\varepsilon}}\right)^{-1} \frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}.} \tag{6.22}$$

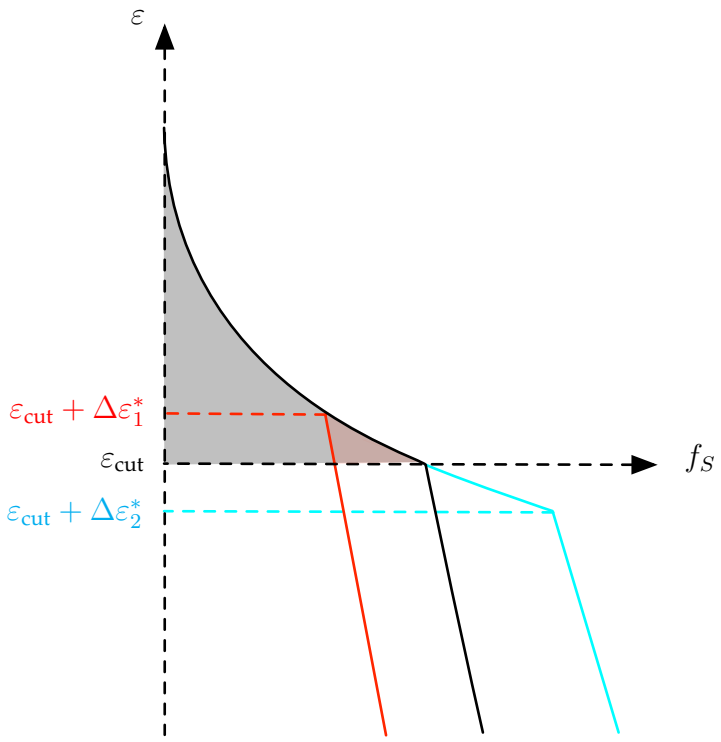


Figure 6.1: An illustration of the effect varying ε^* on the distribution function and the density above ε_{cut}. The black plot shows the case without variation, Here $\varepsilon^* - \varepsilon_{\text{cut}} = 0$. The red plot shows the case Here $\varepsilon^* - \varepsilon_{\text{cut}} = \Delta\varepsilon_1^* > 0$. The cyan plot shows Here the case Here $\varepsilon^* - \varepsilon_{\text{cut}} = \Delta\varepsilon_2^* < 0$. The black line overlaps the red plot and the cyan plot after the intersection. We are interested in the variation of particle density above ε_{cut}, with respect to variations in ε^*. The particle density above ε_{cut} is a weighted integral of the plot above ε_{cut}. The density above ε_{cut} associated with zero variation is represented shaded area. The density above ε_{cut} is unchanged by negative variation, since the plot is the same above ε_{cut}. The density above ε_{cut} is reduced by the weighted area of the slightly red shaded region for positive variation. However, for small $\Delta\varepsilon_1^*$, both the height and the average width of this region are proportional to $\Delta\varepsilon_1^*$. Therefore there is zero first order variation in the density above ε_{cut} with respect to infinitesimal positive variations. The partial derivative of the density above ε_{cut} with respect to ε^* is therefore well-defined as zero.

Having already used the third row from eq. (6.20), we can remove this from the equation, and having determined $\frac{\partial \mu^\varepsilon}{\partial t}$, we can remove this from the vector of unknowns. The updated form of the multivariate chain rule is then expressed as follows:

$$\begin{bmatrix} \frac{\partial n}{\partial t} \\ \frac{\partial w}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial n}{\partial \beta} & \frac{\partial n}{\partial \varepsilon^*} \\ \frac{\partial w}{\partial \beta} & \frac{\partial w}{\partial \varepsilon^*} \end{bmatrix} \begin{bmatrix} \frac{\partial \beta}{\partial t} \\ \frac{\partial \varepsilon^*}{\partial t} \end{bmatrix} + \begin{bmatrix} \frac{\partial \mu^\varepsilon}{\partial t} \frac{\partial n}{\partial \mu^\varepsilon} \\ \frac{\partial \mu^\varepsilon}{\partial t} \frac{\partial w}{\partial \mu^\varepsilon} \end{bmatrix}. \tag{6.22a}$$

Since all entries of the matrix are typically non-zero, we will proceed to invert this matrix equation in the standard manner so that the vector of unknowns the subject:

$$\begin{bmatrix} \frac{\partial \beta}{\partial t} \\ \frac{\partial \varepsilon^*}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial n}{\partial \beta} & \frac{\partial n}{\partial \varepsilon^*} \\ \frac{\partial w}{\partial \beta} & \frac{\partial w}{\partial \varepsilon^*} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial n}{\partial t} - \frac{\partial \mu^\varepsilon}{\partial t} \frac{\partial n}{\partial \mu^\varepsilon} \\ \frac{\partial w}{\partial t} - \frac{\partial \mu^\varepsilon}{\partial t} \frac{\partial w}{\partial \mu^\varepsilon} \end{bmatrix}. \tag{6.22b}$$

It is as this point we note that the partial derivatives of the densities with respect to μ^ε have a particularly simple form. The reason is as follows. The maximum chemical potential μ^ε only effects a scale factor of the occupation function, via an exponential relation. Densities are weighted integrals of the occupation function, and the dependence on the scale factor can always be brought outside this weighted integral. The derivative of the scale factor with respect to μ^ε is proportional to the scale factor. Accordingly, the derivative of the density with respect to μ^ε is proportional to the density:

$$\frac{\partial n}{\partial \mu^\varepsilon} = \frac{1}{kT_L} n, \tag{6.22c}$$

$$\frac{\partial w}{\partial \mu^\varepsilon} = \frac{1}{kT_L} w, \tag{6.22d}$$

$$\frac{\partial n_{\varepsilon > \varepsilon^*}}{\partial \mu^\varepsilon} = \frac{1}{kT_L} n_{\varepsilon > \varepsilon^*}. \tag{6.22e}$$

We can now eliminate all references to μ^ε from the eq. (6.23), by substituting in eq. (6.22) and eq. (6.23). The result is the following:

$$\boxed{\begin{bmatrix} \frac{\partial \beta}{\partial t} \\ \frac{\partial \varepsilon^*}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial n}{\partial \beta} & \frac{\partial n}{\partial \varepsilon^*} \\ \frac{\partial w}{\partial \beta} & \frac{\partial w}{\partial \varepsilon^*} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial n}{\partial t} - \frac{n}{n_{\varepsilon > \varepsilon^*}} \frac{\partial n_{\varepsilon > \varepsilon_{\text{cut}}}}{\partial t} \\ \frac{\partial w}{\partial t} - \frac{w}{n_{\varepsilon > \varepsilon^*}} \frac{\partial n_{\varepsilon > \varepsilon_{\text{cut}}}}{\partial t} \end{bmatrix}.} \tag{6.23}$$

The above equation is fairly simple to understand. The change in $n_{\varepsilon > \varepsilon_{\text{cut}}}$ is caused purely

by a change in μ^{ε}, which means it is caused purely by a change in the scale factor for the ansatz. Accordingly, the fractional change in $n_{\varepsilon>\varepsilon_{\text{cut}}}$ is equal to the fractional change of n and w that can be attributed to a change in μ^{ε}. If we remove this change from the rate of change of particle and energy density, then the remaining the changes in particle and energy density must purely be due to changes in β and ε^{*}.

If the terms on the right hand side of eq. (6.23) are known, then solving this equation is trivial.[6] Thus determining the rate of change in the ansatz parameters is a trivial extension of the problem of determining the rate of change in the characteristic fluxes.

The Warm and Cold Electron Populations

For the warm electron density, we have according to the chain rule the following expression for the rate of change of the characteristic densities:

$$\begin{bmatrix} \frac{\partial n_{\text{warm}}}{\partial t} \\ \frac{\partial w_{\text{warm}}}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial n_{\text{warm}}}{\partial \alpha^{\varepsilon}_{\text{warm}}} & \frac{\partial n_{\text{warm}}}{\partial T_{\text{warm}}} \\ \frac{\partial w_{\text{warm}}}{\partial \alpha^{\varepsilon}_{\text{warm}}} & \frac{\partial w_{\text{warm}}}{\partial T_{\text{warm}}} \end{bmatrix} \begin{bmatrix} \frac{\partial \alpha^{\varepsilon}_{\text{warm}}}{\partial t} \\ \frac{\partial T_{\text{warm}}}{\partial t} \end{bmatrix}. \tag{6.24}$$

On the basis of the simple energy-independent scaling behaviour of $\alpha^{\varepsilon}_{\text{warm}}$, we have that $\frac{\partial n_{\text{warm}}}{\partial \alpha^{\varepsilon}_{\text{warm}}} = -n$ and $\frac{\partial w_{\text{warm}}}{\partial \alpha^{\varepsilon}_{\text{warm}}} = -w$. Substituting in these relations, are rearranging eq. (6.24) in order to make the ansatz parameters the subject, we reach the following:

$$\boxed{\begin{bmatrix} \frac{\partial \alpha^{\varepsilon}_{\text{warm}}}{\partial t} \\ \frac{\partial T_{\text{warm}}}{\partial t} \end{bmatrix} = \begin{bmatrix} -n_{\text{warm}} & \frac{\partial n_{\text{warm}}}{\partial T_{\text{warm}}} \\ -w_{\text{warm}} & \frac{\partial w_{\text{warm}}}{\partial T_{\text{warm}}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial n_{\text{warm}}}{\partial t} \\ \frac{\partial w_{\text{warm}}}{\partial t} \end{bmatrix}.} \tag{6.25}$$

Finally for the cold electron population, the chain rule leads to the simple relation:

$$\frac{\partial n_{\text{cold}}}{\partial t} = \frac{\partial n_{\text{cold}}}{\partial \mu^{\varepsilon}_{\text{cold}}} \frac{\partial \mu^{\varepsilon}_{\text{cold}}}{\partial t}. \tag{6.26}$$

The cold electron population *does not* show simple energy-independent scaling behaviour with $\mu^{\varepsilon}_{\text{cold}}$. Rearranging to make the ansatz parameters the subject, we have the following:

[6] For instance, there is a simple analytic formula for the inverse of a general 2×2 matrix.

$$\boxed{\frac{\partial \mu^{\varepsilon}_{\text{cold}}}{\partial t} = \left(\frac{\partial n_{\text{cold}}}{\partial \mu^{\varepsilon}_{\text{cold}}}\right)^{-1} \frac{\partial n_{\text{cold}}}{\partial t}.} \tag{6.27}$$

6.4 Closing the Ansatz Parameter Equation of Motion

6.4.1 Overview

The purpose of this section is to close the equation of motion for the ansatz parameters derived in the last section. This is essentially achieved by taking every term in the equation of motion that is not defined in terms of ansatz parameters and defining it in terms of ansatz parameters. In order to do this efficiently, we make heavy use of precomputed macroscopic transport parameters, which are defined as integrals of the scattering operator and bandstructure weighted by the energy component of the ansatz for each of the three quasi-equilibrium populations. We summarise this section as follows.

- We express the characteristic particle densities as the product of a tabulated macroscopic average of the density of states and an analytic integral of the distribution function.

- We express the characteristic energy densities as the product of a tabulated macroscopic average energy and the characteristic particle density.

- We express the partial derivatives of these functions using the product rule, and tabulated values of the partial derivatives of the macroscopic average density of states and the average energies.

- We express the characteristic fluxes as the sum of a fluxes each due to spatial gradient in each thermodynamic ansatz parameter

- We express the flux due to the gradient in each thermodynamic ansatz parameter

as being proportional to the product of an tabulated macroscopic average diffusion coefficient, particle density, and spatial gradient in the ansatz parameter.

- We express the characteristic rate of change in density due to scattering following the population dynamics described in the first section of this chapter. We express these scattering rates in terms of small set of macroscopic average scattering parameters associated with each scattering partner type.

6.4.2 The Characteristic Densities as a Function of Ansatz Parameters

The Stunted Electron Population

The aim of this subsection is to describe a method of quickly relating the ansatz parameters to the characteristic densities and their partial derivatives. In the case of the stunted population the characteristic densities are defined as follows:

$$n = \int_0^\infty f_\varepsilon D(\varepsilon) \mathrm{d}\varepsilon, \tag{6.28a}$$

$$w = \int_0^\infty \varepsilon f_\varepsilon D(\varepsilon) \mathrm{d}\varepsilon, \tag{6.28b}$$

$$n_{\varepsilon>\varepsilon^*} = \int_{\varepsilon^*}^\infty f_\varepsilon D(\varepsilon) \mathrm{d}\varepsilon. \tag{6.28c}$$

These integrals need to be calculated numerically since we are not assuming an analytically integrable form for the density of states $D(\varepsilon)$. It is unacceptably slow to numerically compute these integrals at runtime, therefore we will make use of precomputation and tabulation. However, rather than precomputing and tabulating the mapping of ansatz parameters and the densities directly, we will search for intermediate quantities to precompute and tabulate that are analytically related to the densities and possess the following qualities: the quantities should be described by the minimum number of variables possible, and the quantities should vary as slowly as possible.

In order to minimize the number of variables, we eliminate the dependence of any tabulated quantity on μ^{ε}, by noting that this term is associated with a uniform change in scale of the entire distribution function. having already achieved the same for the characteristic densities and characteristic fluxes, the scale of the entire distribution.

In order to minimize the rate of change of the variable, we note that the densities also have a natural exponential dependence on β and ε^*. For the energy density, we can remove this natural dependence by simply tabulating the average energy as a function of β and ε^*, reducing the problem to the problem of removing the natural exponential dependence of the particle density. In order to remove this natural dependence for the particle density, we tabulate the actual particle density to what the particle density would be if the density of states was equal to unity at all energies. This quantity is both relatively slowly varying, and has a natural interpretation as a macroscopic measure of the density of states. Due to the fact that we have a seperate continuity equation for the tail, it is sensible to write down seperate our expressions for the density of states measure and average energy for the "hot bulk" of the distribution comprised of below the knee energy, and for the "thermal tail" of the distribution that has kinetic energy greater than ε^*. In order to be able to quickly determine the partial derivatives needed to convert the continuity equations into equations of motion for the ansatz parameters, we also need to precompute and tabulate the partial derivatives of the density of states measures and average energies.

We define list of precomputed ansatz-parameter dependant density measures in Table 6.1 using the Rho Functional. We note that for functions $X(\varepsilon;\dots)$ that are only implicitly dependent on the crystal momentum and band index via the kinetic energy, the Rho Functional takes the following simple form:

$$\begin{aligned}\rho^m(\varepsilon_{\text{min}},\varepsilon_{\text{max}})[X(\varepsilon;\dots)] &= \int_{\varepsilon_{\text{min}}}^{\varepsilon_{\text{max}}} \sigma^{\varepsilon}(\varepsilon)[1]\varepsilon^m X(\varepsilon;\dots)\mathrm{d}\varepsilon \\ &= \int_{\varepsilon_{\text{min}}}^{\varepsilon_{\text{max}}} D(\varepsilon)\varepsilon^m X(\varepsilon;\dots)\mathrm{d}\varepsilon. \end{aligned} \tag{6.29}$$

By solving the integrals on the denominator analytically, we find that the following analytic mapping between the stunted electron ansatz parameters and the tail and bulk densities:

$D_{\text{tail}}(\varepsilon^*)$	$=$	$\dfrac{\rho^0(\varepsilon^*,\infty)\,[f_\varepsilon]}{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$	Density of states measure for electrons in the thermal tail.
$D_{\text{bulk}}(\varepsilon^*,\beta)$	$=$	$\dfrac{\rho^0(0,\varepsilon^*)\,[f_\varepsilon]}{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$	Density of states measure for electrons in the hot bulk.
$\langle\varepsilon\rangle_{\text{tail}}(\varepsilon^*)$	$=$	$\dfrac{\rho^1(\varepsilon^*,\infty)\,[f_\varepsilon]}{\rho^0(\varepsilon^*,\infty)\,[f_\varepsilon]}$	Average energy of electrons in the thermal tail.
$\langle\varepsilon\rangle_{\text{bulk}}(\varepsilon^*,\beta)$	$=$	$\dfrac{\rho^1(0,\varepsilon^*)\,[f_\varepsilon]}{\rho^0(0,\varepsilon^*)\,[f_\varepsilon]}$	Average energy of electrons in the hot bulk.
$\dfrac{\partial D_{\text{tail}}}{\partial\varepsilon^*}(\varepsilon^*)$	$=$	$\dfrac{\partial}{\partial\varepsilon^*}\dfrac{\rho^0(\varepsilon^*,\infty)\,[f_\varepsilon]}{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$	Rate of change of density of states measure with respect to ε^* for electrons in the thermal tail.
$\dfrac{\partial D_{\text{bulk}}}{\partial\varepsilon^*}(\varepsilon^*,\beta)$	$=$	$\dfrac{\partial}{\partial\varepsilon^*}\dfrac{\rho^0(0,\varepsilon^*)\,[f_\varepsilon]}{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$	Rate of change of density of states measure with respect to ε^* for electrons in the hot bulk.
$\dfrac{\partial D_{\text{bulk}}}{\partial\beta}(\varepsilon^*,\beta)$	$=$	$\dfrac{\partial}{\partial\beta}\dfrac{\rho^0(0,\varepsilon^*)\,[f_\varepsilon]}{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$	Rate of change of density of states measure with respect to β for electrons in the hot bulk.
$\dfrac{\partial\langle\varepsilon\rangle_{\text{bulk}}}{\partial\varepsilon^*}(\varepsilon^*,\beta)$	$=$	$\dfrac{\partial}{\partial\varepsilon^*}\dfrac{\rho^1(0,\varepsilon^*)\,[f_\varepsilon]}{\rho^0(0,\varepsilon^*)\,[f_\varepsilon]}$	Rate of change of average energy with respect to ε^* for electrons in the hot bulk.
$\dfrac{\partial\langle\varepsilon\rangle_{\text{bulk}}}{\partial\beta}(\varepsilon^*,\beta)$	$=$	$\dfrac{\partial}{\partial\beta}\dfrac{\rho^1(0,\varepsilon^*)\,[f_\varepsilon]}{\rho^0(0,\varepsilon^*)\,[f_\varepsilon]}$	Rate of change of average energy with respect to β for electrons in the hot bulk.

Table 6.1: Description of the precomputed and tabulated stunted ansatz parameter dependent density measures. These allow the values of characteristic densities and their partial derivatives to be quickly computed from the ansatz parameters at run-time.

$$n_{\text{tail}} = \left(kT_L e^{-\frac{\varepsilon^*-\mu^\varepsilon}{kT_L}}\right) D_{\text{tail}}, \tag{6.30a}$$

$$n_{\text{bulk}} = \left(\frac{1}{\beta}\left(e^{\beta\varepsilon^*}-1\right)e^{-\frac{\varepsilon^*-\mu^\varepsilon}{kT_L}}\right) D_{\text{bulk}}, \tag{6.30b}$$

$$w_{\text{tail}} = n_{\text{tail}}\,\langle\varepsilon\rangle_{\text{tail}}\,, \tag{6.30c}$$

$$w_{\text{bulk}} = n_{\text{bulk}}\,\langle\varepsilon\rangle_{\text{bulk}}\,. \tag{6.30d}$$

We now can write down expressions for the partial derivatives of these functions with respect to the ansatz parameters. We have already discussed the simple partial derivative of any density with respect to the chemical potential, and the tail densities do not dependent on the bulk thermodynamic beta. The remaining partial derivatives for the tail and bulk particle densities are defined as follows:

$$\frac{\partial n_{\text{tail}}}{\partial \varepsilon^*} = -e^{-\frac{\varepsilon^* - \mu^\varepsilon}{kT_l}} \left(D_{\text{tail}} + kT_L \frac{\partial D_{\text{tail}}}{\partial \varepsilon^*} \right), \tag{6.31a}$$

$$\frac{\partial n_{\text{bulk}}}{\partial \varepsilon^*} = e^{-\frac{\varepsilon^* - \mu^\varepsilon}{kT_l}} \left[\left(-\frac{1}{kT_L \beta} \left(e^{\beta \varepsilon^*} - 1 \right) + e^{\beta \varepsilon^*} \right) D_{\text{bulk}} + \frac{1}{\beta} \left(e^{\beta \varepsilon^*} - 1 \right) \frac{\partial D_{\text{bulk}}}{\partial \varepsilon^*} \right], \tag{6.31b}$$

$$\frac{\partial n_{\text{bulk}}}{\partial \beta} = \frac{1}{\beta} e^{-\frac{\varepsilon^* - \mu^\varepsilon}{kT_L}} \left[\left(-\frac{1}{\beta} \left(e^{\beta \varepsilon^*} - 1 \right) + \varepsilon^* e^{\beta \varepsilon^*} \right) D_{\text{bulk}} + \left(e^{\beta \varepsilon^*} - 1 \right) \frac{\partial D_{\text{bulk}}}{\partial \beta} \right]. \tag{6.31c}$$

Having expressed the partial derivatives of the tail and bulk densities in terms of the ansatz parameters, it is now trivial to define an expression for the partial derivatives of the energy densities in terms of ansatz parameters:

$$\frac{\partial w_{\text{tail}}}{\partial \varepsilon^*} = \frac{\partial n_{\text{tail}}}{\partial \varepsilon^*} \langle \varepsilon \rangle_{\text{tail}} + n_{\text{tail}} \frac{\partial \langle \varepsilon \rangle_{\text{tail}}}{\partial \varepsilon^*}, \tag{6.32a}$$

$$\frac{\partial w_{\text{bulk}}}{\partial \varepsilon^*} = \frac{\partial n_{\text{bulk}}}{\partial \varepsilon^*} \langle \varepsilon \rangle_{\text{bulk}} + n_{\text{bulk}} \frac{\partial \langle \varepsilon \rangle_{\text{bulk}}}{\partial \varepsilon^*}, \tag{6.32b}$$

$$\frac{\partial w_{\text{bulk}}}{\partial \beta} = \frac{\partial n_{\text{bulk}}}{\partial \beta} \langle \varepsilon \rangle_{\text{bulk}} + n_{\text{bulk}} \frac{\partial \langle \varepsilon \rangle_{\text{bulk}}}{\partial \beta}. \tag{6.32c}$$

The Warm and Cold Electron Populations

A useful relation is that shape of the bulk electron population is identical to the warm population when $\varepsilon^* = \infty$ and $\beta = \frac{1}{kT_{\text{warm}}}$. Accordingly, whenever we have a macroscopic quantity for the bulk of the distribution as a function (ε^*, β), we can always convert it into a corresponding macroscopic quantity for the warm distribution as a function of T_{warm}. This leads to the following relations:

$$D_{\text{warm}}(T_{\text{warm}}) = D_{\text{bulk}}\left(\infty, \frac{1}{kT_{\text{warm}}}\right), \tag{6.33a}$$

$$\langle\varepsilon\rangle_{\text{warm}}(T_{\text{warm}}) = \langle\varepsilon\rangle_{\text{warm}}\left(\infty, \frac{1}{kT_{\text{warm}}}\right), \tag{6.33b}$$

$$\frac{\partial D_{\text{warm}}}{\partial T_{\text{warm}}} = -\frac{1}{kT_{\text{warm}}^2}\frac{\partial D_{\text{bulk}}}{\partial \beta}\left(\infty, \frac{1}{kT_{\text{warm}}}\right), \tag{6.33c}$$

$$\frac{\partial \langle\varepsilon\rangle_{\text{warm}}}{\partial T_{\text{warm}}} = -\frac{1}{kT_{\text{warm}}}\frac{\partial \langle\varepsilon\rangle_{\text{bulk}}}{\partial \beta}\left(\infty, \frac{1}{kT_{\text{warm}}}\right). \tag{6.33d}$$

We can now express the warm particle and energy density in terms of these functions of the ansatz parameters as follows:

$$n_{\text{warm}} = kT_{\text{warm}} e^{-\alpha^{\varepsilon}_{\text{warm}}} D_{\text{warm}}, \tag{6.34a}$$

$$w_{\text{warm}} = n_{\text{warm}} \langle\varepsilon\rangle_{\text{warm}}. \tag{6.34b}$$

And we can express the partial derivatives of these densities with respect to temperature as functions of the ansatz parameters:

$$\frac{\partial n_{\text{warm}}}{\partial T_{\text{warm}}} = e^{-\alpha^{\varepsilon}_{\text{warm}}}\left(kD_{\text{warm}} + kT_L \frac{\partial D_{\text{warm}}}{\partial T_{\text{warm}}}\right), \tag{6.35a}$$

$$\frac{\partial w_{\text{tail}}}{\partial T_{\text{warm}}} = \frac{\partial n_{\text{warm}}}{\partial T_{\text{warm}}}\langle\varepsilon\rangle_{\text{warm}} + n_{\text{warm}}\frac{\partial \langle\varepsilon\rangle_{\text{warm}}}{\partial T_{\text{warm}}}. \tag{6.35b}$$

We now turn to the cold particle density. We note that for the cold population, we need to calculate a new density of states measure, and partial derivative of that density of states measure. These are listed in Table 6.2.

As with the particle densities of the other distributions, we can now express the particle density as the product of an analytic integral of the distribution function and the density of states measure. Since the distribution is now a Fermi-Dirac function, the associated integral is non-trivial:

$D_{\text{cold}}(\mu^{\varepsilon}_{\text{cold}}) = \frac{\rho^0(0,\infty)\left[f_\varepsilon\right]}{\rho^0(0,\infty)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$		Density of states measure for electrons in the cold distribution.
$\frac{\partial D_{\text{cold}}}{\partial \mu^{\varepsilon}_{\text{cold}}}(\mu^{\varepsilon}_{\text{cold}}) = \frac{\partial}{\partial \mu^{\varepsilon}_{\text{cold}}} \frac{\rho^0(0,\infty)\left[f_\varepsilon\right]}{\rho^0(0,\infty)\left[\frac{f_\varepsilon}{D(\varepsilon)}\right]}$		Rate of change of density of states measure with respect to $\mu^{\varepsilon}_{\text{cold}}$ for electrons in the cold population.

Table 6.2: Description of the precomputed and tabulated cold chemical potential dependent density measures. These allow the particle density of the cold population and its derivative to be quickly computed from the cold electron chemical potential at run-time.

$$\begin{aligned} n_{\text{cold}} &= \left(\int_0^\infty f_{\text{cold}} \mathrm{d}\varepsilon\right) D_{\text{cold}} \\ &= \left[\varepsilon - kT_L \ln\left(e^{\frac{\varepsilon-\mu^{\varepsilon}_{\text{cold}}}{kT_L}} + 1\right)\right]\Bigg|_0^\infty D_{\text{cold}} \\ &= \left[\mu^{\varepsilon}_{\text{cold}} + kT_L \ln\left(e^{-\frac{\mu^{\varepsilon}_{\text{cold}}}{kT_L}} + 1\right)\right] D_{\text{cold}}. \end{aligned} \tag{6.36}$$

Here the derivative of the cold distribution with respect to the cold chemical potential can be expressed as follows:

$$\frac{\partial n_{\text{cold}}}{\partial \mu^{\varepsilon}_{\text{cold}}} = \frac{1}{1 + e^{-\frac{\mu^{\varepsilon}_{\text{cold}}}{kT_L}}} D_{\text{cold}} + \frac{n_{\text{cold}}}{D_{\text{cold}}} \frac{\partial D_{\text{cold}}}{\partial \mu^{\varepsilon}_{\text{cold}}}. \tag{6.37}$$

We can now conclude this section, having presented expressions for the characteristic densities as functions of the ansatz parameters, and the partial derivatives of these functions. So long as the quantities listed in Table 6.1 and Table 6.2 are precomputed and stored in a look-up table, these characteristic densities and there partial derivatives are able to be computed easily from the ansatz parameters at run-time.

6.4.3 The Characteristic Fluxes as a Function of Ansatz Parameters

The Stunted Electron Population

The characteristic fluxes for the stunted electron population are determined by the following integrals:

$$\mathbf{j} = \int_0^\infty \mathbf{j}^\varepsilon \mathrm{d}\varepsilon, \tag{6.38a}$$

$$\mathbf{S} = \int_0^\infty \varepsilon \mathbf{j}^\varepsilon \mathrm{d}\varepsilon, \tag{6.38b}$$

$$\mathbf{j}_{H>H_{\text{cut}}} = \int_{H_{\text{cut}}-\varepsilon_C}^\infty \mathbf{j}^\varepsilon \mathrm{d}\varepsilon. \tag{6.38c}$$

Here $\varepsilon_C(\mathbf{r}, t)$ is local the potential energy of the conduction band minimum, and the particle flux per kinetic energy is given by the following:

$$\mathbf{j}^\varepsilon(\varepsilon, \mathbf{r}, t) = D(\varepsilon)\mathcal{D}^\varepsilon \cdot \nabla_{\mathbf{r}} f_H. \tag{6.39}$$

Here the function f_H is assumed to be an explicit function of $(H, \mathbf{r}, t)$. Given the form of our stunted ansatz in eq. (6.2), we wish to express this partial derivative in the case Here the space and time dependence of f_H is replaced with a dependence on the ansatz parameters. While the stunted ansatz only has 3 degrees of freedom— owing to the continuity at H^*— it is simpler if we express the f_H as an explicit function of all 4 degrees of freedom $(\mu^H, H^*, \alpha^H, \beta)$, and simplify the resulting expression by enforce continuity. First, applying the chain rule leads to the following:

$$\nabla_{\mathbf{r}} f_H = \frac{\partial f_H}{\partial H^*}\nabla_{\mathbf{r}} H^* + \frac{\partial f_H}{\partial \mu^H}\nabla_{\mathbf{r}}\mu^H + \frac{\partial f_H}{\partial \alpha^H}\nabla_{\mathbf{r}}\alpha^H + \frac{\partial f_H}{\partial \beta}\nabla_{\mathbf{r}}\beta. \tag{6.40}$$

Substituting eq. (6.40) into expression for the particle flux per kinetic energy eq. (6.39), we have the following:

$$\mathbf{j}^\varepsilon = D(\varepsilon)\mathcal{D}^\varepsilon(\varepsilon; ...) \cdot \left(\nabla_{\mathbf{r}} H^* \frac{\partial f_H}{\partial H^*} + \nabla_{\mathbf{r}}\mu^H \frac{\partial f_H}{\partial \mu^H} + \nabla_{\mathbf{r}}\alpha^H \frac{\partial f_H}{\partial \alpha^H} + \nabla_{\mathbf{r}}\beta \frac{\partial f_H}{\partial \beta} \right). \tag{6.41}$$

The problem with this expression is that the partial derivative $\frac{\partial f_H}{\partial H^*}$ is not well defined due to the gradient discontinuity at that point in the ansatz. However, we are not interested in $\mathbf{j}^\varepsilon$ itself, but in the three weighted integrals of $\mathbf{j}^\varepsilon$ defined by eq. (6.38). We show in an aside at the end of this section, that the weighted integrals of this term can be defined, and that in the case Here f_H is continuous, are equal to zero. The remaining partial derivatives in eq. (6.41) can be expressed analytically:

$$\frac{\partial f_H}{\partial \mu^H} = \begin{cases} 0 & \text{for } H < H^*, \\ \frac{1}{kT_L} f_H & \text{for } H > H^*, \end{cases} \tag{6.42a}$$

$$\frac{\partial f_H}{\partial \alpha^H} = \begin{cases} -f_H & \text{for } H < H^*, \\ 0 & \text{for } H > H^*, \end{cases} \tag{6.42b}$$

$$\frac{\partial f_H}{\partial \beta} = \begin{cases} -H f_H & \text{for } H < H^*, \\ 0 & \text{for } H > H^*. \end{cases} \tag{6.42c}$$

The partial derivatives of eq. (6.42) are simply defined they are defined relative to constant H^*, which was our reason for making H^* an explicit variable. When substituting eq. (6.42) into eq. (6.41), we can change all terms not associated with spatial gradients into their kinetic energy form, and we our final expression for the flux per kinetic energy:

$$\mathbf{j}^\varepsilon(\varepsilon, \mathbf{r}, t) = -D(\varepsilon)\mathcal{D}^\varepsilon(\varepsilon; ...) f_\varepsilon \cdot \begin{cases} \left(\nabla_{\mathbf{r}} \alpha^H + \varepsilon \nabla_{\mathbf{r}} \beta\right) & \text{for } \varepsilon < \varepsilon^*, \\ -\frac{1}{kT_L} \nabla_{\mathbf{r}} \mu^H & \text{for } \varepsilon > \varepsilon^*. \end{cases} \tag{6.43}$$

This is expression is quite simple conceptually. The above flux relationship describes ELASTIC DIFFUSION, or diffusion at constant total energy. Like near-equilibrium transport processes, elastic diffusion is an entropic process driven by gradients in thermodynamic potentials.

We can rewrite our expressions for the characteristic fluxes in eq. (6.38) using the eq. (6.43) and the Rho Functional. This leads to the following:

$$\mathbf{j} = \nabla_{\mathbf{r}}\alpha^H \cdot \rho^0(0,\varepsilon^*)[\mathcal{D}^\varepsilon f_\varepsilon] + \nabla_{\mathbf{r}}\beta \cdot \rho^1(0,\varepsilon^*)[\mathcal{D}^\varepsilon f_\varepsilon] - \frac{1}{kT_L} \cdot \nabla_{\mathbf{r}}\mu^H \rho^0(\varepsilon^*,\infty)[\mathcal{D}^\varepsilon f_\varepsilon], \tag{6.44a}$$

$$\mathbf{S} = \nabla_{\mathbf{r}}\alpha^H \cdot \rho^1(0,\varepsilon^*)[\mathcal{D}^\varepsilon f_\varepsilon] + \nabla_{\mathbf{r}}\beta \cdot \rho^2(0,\varepsilon^*)[\mathcal{D}^\varepsilon f_\varepsilon] - \frac{1}{kT_L} \cdot \nabla_{\mathbf{r}}\mu^H \rho^1(\varepsilon^*,\infty)[\mathcal{D}^\varepsilon f_\varepsilon], \tag{6.44b}$$

$$\mathbf{j}_{H>H_{\text{cut}}} = -\frac{1}{kT_L}\nabla_{\mathbf{r}}\mu^H \rho^1(\varepsilon^*,\infty)[\mathcal{D}^\varepsilon f_\varepsilon] + \int_{\varepsilon_{\text{cut}}}^{\varepsilon^*} \mathbf{j}^\varepsilon \mathrm{d}\varepsilon. \tag{6.44c}$$

Computing the Rho Functional at run-time will make the model unreasonably slow, accordingly we precompute and store the list of macroscopic diffusion coefficients in Table 6.3, that allow us to quickly calculate the Rho Functionals in eq. (6.44).

Once again, when storing a numerical function, it is most efficient to store one that varies as slowly as possible over its domain— thus instead of storing the Rho Functionals in eq. (6.44) directly, we normalize each by an associated particle density, which has the special benefit of being a natural definition of the macroscopic diffusion parameter. It is also most efficient to store a function that depends on as few variables as possible— thus instead of storing the full diffusion coefficient as a function of $(\varepsilon^*, \beta, \beta_S, N_{\text{dop}})$, we store the dopant and non-dopant diffusion coefficients separately and approximate the total macroscopic diffusion coefficient in the following manner:

$$\mathcal{D}^m_{\text{tail/bulk}} = \left(\frac{1}{\mathcal{D}^{\text{m, non-dop}}_{\text{tail/bulk}}} + \frac{N_{\text{dop}}}{\mathcal{D}^{\text{m, dop}}_{\text{tail/bulk}}} \right)^{-1}. \tag{6.45}$$

Here if the macroscopic diffusion coefficients are genuine tensors, the preceding calculation is performed on *entry-wise*. We note that this is sufficient to simplify all terms except the KNEE DISLOCATION FLUX, defined by $\int_{\varepsilon_{\text{cut}}}^{\varepsilon^*} \mathbf{j}^\varepsilon \mathrm{d}\varepsilon$. This term is zero when $\varepsilon_{\text{cut}} = \varepsilon^*$, which is true in the limit that the $\varepsilon_{\text{cut}} = \varepsilon^*$. In a realistic discretization scheme it is however important to account for the dislocation flux. As explained in an aside in the end of this subsection, we incorporate the perturbation due to the knee dislocation flux into the chemical potential driven flux, which in turn becomes the flux driven by the gradient in a *clamped* chemical potential. This leads to the following set of equations:

$\mathcal{D}^{0,(\text{non-})\text{dop}}_{\text{tail}}(\varepsilon^*;\beta_S)$	$= \dfrac{\rho^0(\varepsilon^*,\infty)\left[f_\varepsilon \mathcal{D}^{\varepsilon}_{(\text{non-})\text{dop}}\right]}{\rho^0(\varepsilon^*,\infty)\left[f_\varepsilon\right]}$	Average particle diffusion coefficient for electrons in the thermal tail associated with to (non-)dopants.
$\mathcal{D}^{0,(\text{non-})\text{dop}}_{\text{bulk}}(\varepsilon^*,\beta;\beta_S)$	$= \dfrac{\rho^0(0,\varepsilon^*)\left[f_\varepsilon \mathcal{D}^{\varepsilon}_{(\text{non-})\text{dop}}\right]}{\rho^0(0,\varepsilon^*)\left[f_\varepsilon\right]}$	Average particle diffusion coefficient for electrons in the hot bulk associated with to (non-)dopants.
$\mathcal{D}^{1,(\text{non-})\text{dop}}_{\text{tail}}(\varepsilon^*;\beta_S)$	$= \dfrac{\rho^1(\varepsilon^*,\infty)\left[f_\varepsilon \mathcal{D}^{\varepsilon}_{(\text{non-})\text{dop}}\right]}{\rho^0(\varepsilon^*,\infty)\left[f_\varepsilon\right]}$	Average energy diffusion coefficient for electrons in the thermal tail associated with to (non-)dopants.
$\mathcal{D}^{1,(\text{non-})\text{dop}}_{\text{bulk}}(\varepsilon^*,\beta;\beta_S)$	$= \dfrac{\rho^1(0,\varepsilon^*)\left[f_\varepsilon \mathcal{D}^{\varepsilon}_{(\text{non-})\text{dop}}\right]}{\rho^0(0,\varepsilon^*)\left[f_\varepsilon\right]}$	Average energy diffusion coefficient for electrons in the hot bulk associated with to (non-)dopants.
$\mathcal{D}^{2,(\text{non-})\text{dop}}_{\text{bulk}}(\varepsilon^*,\beta;\beta_S)$	$= \dfrac{\rho^2(0,\varepsilon^*)\left[f_\varepsilon \mathcal{D}^{\varepsilon}_{(\text{non-})\text{dop}}\right]}{\rho^0(0,\varepsilon^*)\left[f_\varepsilon\right]}$	Average square energy diffusion coefficient for electrons in the hot bulk associated with to (non-)dopants.

Table 6.3: Description of the precomputed macroscopic diffusion parameters, which allow the characteristic fluxes of the stunted distribution to be quickly computed at run-time.

$$\boxed{\mathbf{j} = n_{\text{bulk}}\mathcal{D}^0_{\text{bulk}} \cdot \nabla_{\mathbf{r}}\alpha^H + n_{\text{bulk}}\mathcal{D}^1_{\text{bulk}} \cdot \nabla_{\mathbf{r}}\beta - \frac{n_{\text{tail}}}{kT_L}\mathcal{D}^0_{\text{tail}} \cdot \nabla_{\mathbf{r}}\mu^H,} \tag{6.46a}$$

$$\boxed{\mathbf{S} = n_{\text{bulk}}\mathcal{D}^1_{\text{bulk}} \cdot \nabla_{\mathbf{r}}\alpha^H + n_{\text{bulk}}\mathcal{D}^2_{\text{bulk}} \cdot \nabla_{\mathbf{r}}\beta - \frac{n_{\text{tail}}}{kT_L}\mathcal{D}^1_{\text{tail}} \cdot \nabla_{\mathbf{r}}\mu^H,} \tag{6.46b}$$

$$\boxed{\mathbf{j}_{H>H_{\text{cut}}} = \begin{cases} -\frac{n_{\text{tail}}(\varepsilon_{\text{cut}})}{kT_L}\mathcal{D}^0_{\text{tail}}(\varepsilon_{\text{cut}};\dots)\cdot\nabla_{\mathbf{r}}\mu^H & \varepsilon_{\text{cut}} \geq \varepsilon^*, \\ -\frac{n_{\text{tail}}(\varepsilon^*)}{kT_L}\mathcal{D}^0_{\text{tail}}(\varepsilon^*;\dots)\cdot\nabla_{\mathbf{r}}\mu^H + \mathbf{j}^{\varepsilon^*}_{\varepsilon_{\text{cut}}} & \varepsilon_{\text{cut}} < \varepsilon^*. \end{cases}} \tag{6.46c}$$

Here the term $\mathbf{j}^{\varepsilon^*}_{\varepsilon_{\text{cut}}}$ refers to the knee-displacement particle flux, a term which is zero in the continuous limit, but which is non-zero in finite volume discretization scheme. We estimate this term as follows:

$$\mathbf{j}^{\varepsilon^*}_{\varepsilon_{\text{cut}}} = D(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2})\mathcal{D}^{\varepsilon}(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2};\dots)\cdot\left(\frac{1}{\beta}\nabla_{\mathbf{r}}\alpha^H + \frac{1}{\beta^2}(\beta\varepsilon+1)\nabla_{\mathbf{r}}\beta\right)e^{-(\alpha^{\varepsilon}+\beta\varepsilon)}\Big|^{\varepsilon^*}_{\varepsilon_{\text{cut}}}. \tag{6.47}$$

The Warm and Cold Electron Populations

The identity of the bulk population ansatz at $\varepsilon = \infty$ and the warm population ansatz leads to the following definition of the warm macroscopic diffusion coefficients:

$$\mathcal{D}^m_{\text{warm}}(T_{\text{warm}};\dots) = \mathcal{D}^m_{\text{bulk}}(\infty, \frac{1}{kT_{\text{warm}}};\dots). \tag{6.48}$$

And we can define the characteristic fluxes for the warm electron population by simply converting the gradients in β to gradients in T_{warm} using the chain rule:

$$\boxed{\mathbf{j}_{\text{warm}} = n_{\text{warm}}\mathcal{D}^0_{\text{warm}}\cdot\nabla_{\mathbf{r}}\alpha^H_{\text{warm}} - \frac{n_{\text{warm}}}{kT^2_{\text{warm}}}\mathcal{D}^1_{\text{warm}}\cdot\nabla_{\mathbf{r}}T_{\text{warm}},} \tag{6.49a}$$

$$\boxed{\mathbf{S}_{\text{warm}} = n_{\text{warm}}\mathcal{D}^1_{\text{warm}}\cdot\nabla_{\mathbf{r}}\alpha^H_{\text{warm}} - \frac{n_{\text{warm}}}{kT^2_{\text{warm}}}\mathcal{D}^2_{\text{warm}}\cdot\nabla_{\mathbf{r}}T_{\text{warm}}.} \tag{6.49b}$$

The case of the flux relations for the cold electron distribution is more complicated, due to degeneracy. We note that the corresponding form eq. (6.41) for the cold electron population is as follows:

$$\mathbf{j}^{\varepsilon}_{\text{cold}} = -D(\varepsilon)\mathcal{D}^{\varepsilon}\frac{\partial f^{\text{cold}}_H}{\partial\mu^H_{\text{cold}}}\cdot\nabla_{\mathbf{r}}\mu^H_{\text{cold}}. \tag{6.50}$$

Leading to the following definition of the cold electron particle flux:

$$\mathbf{j}_{\text{cold}} = \nabla_{\mathbf{r}}\mu_{\text{cold}} \cdot \rho^0(0,\infty)[\mathcal{D}^{\varepsilon}\frac{\partial f_H^{\text{cold}}}{\partial \mu_{\text{cold}}}]. \tag{6.51}$$

We want our cold electron diffusion coefficient to tend toward $\frac{\rho^0(0,\infty)\left[\mathcal{D}^{\varepsilon} f_{\varepsilon}^{\text{cold}}\right]}{\rho^0(0,\infty)[f_{\varepsilon}^{\text{cold}}]}$ in the non-degenerate limit, and thus define its dopant and non-dopant components in the manner defined in Table 6.4.

Having defined the cold electron diffusion coefficient, we can now define the cold electron particle flux is defined as follows:

$$\boxed{\mathbf{j}_{\text{cold}} = -\frac{n_{\text{cold}}}{kT_L}\nabla_{\mathbf{r}}\mu_{\text{cold}}^{H} \cdot \mathcal{D}_{\text{cold}}^{0}\cdot} \tag{6.52}$$

$\mathcal{D}_{\text{cold}}^{0,\text{(non-)dop}}(\mu_{\text{cold}}^{\varepsilon}, \beta_S)$	$= \frac{\rho^0(0,\infty)\left[-kT_L\frac{\partial f_{\varepsilon}^{\text{cold}}}{\partial \mu_{\text{cold}}^{\varepsilon}}\mathcal{D}_{\text{(non-)dop}}^{\varepsilon}\right]}{\rho^0(0,\infty)\left[f_{\varepsilon}^{\text{cold}}\right]}$	Average particle diffusion coefficient for electrons in the cold electron population due to (non-)dopants.

Table 6.4: Precomputed macroscopic diffusion parameters which allow the characteristic flux term of the cold electron population to be quickly computed at run-time.

Aside: On the Knee-Energy Driven Fluxes

The partial derivative of the function f_H with respect to a change in knee-energy, $\frac{\partial f_H}{\partial H^*}$, is not well-defined near H^*. To understand why, it is useful to define $\Delta f_H(\Delta H^*; H, H^*, \mu, \alpha, \beta)$ to be the change to the distribution function for a *finite* change in H^* equal to ΔH^*, with the all other variables kept constant:

$$\Delta f_H(\Delta H^*;\dots) = \begin{cases} \text{sign}(\Delta H^*)\left(e^{-(\beta H+\alpha^H)} - e^{-\frac{H-\mu^H}{kT_L}}\right) & \text{for } H \in \mathscr{H}_{\Delta H^* \text{ of } H^*}, \\ 0 & \text{for } H \notin \mathscr{H}_{\Delta H^* \text{ of } H^*}. \end{cases} \tag{6.53}$$

Here $\mathscr{H}_{\Delta H^* \text{ of } H^*}$ is the interval $(H^*, H^*+\Delta H^*)$ if ΔH^* is positive, and $(H^*+\Delta H^*, H^*)$ if ΔH^* is negative. The essential point to notice is that there is no linear relationship between the magnitude of ΔH^* and the size of the Δf_H at H as $\Delta H^* \to 0$. Instead, Δf_H has a step change in magnitude when ΔH^* passes a threshold (namely, when it is larger in magnitude than $H - H^*$ and is of the same sign). While the partial derivative can be safely defined to be zero for total energies far from H^*, there is no meaningful interpretation of the partial derivative near H^*. This means that our ansatz is manifestly incompatible with the Boltzmann transport equation near H^*.

However, we are not interested in solving the Boltzmann transport equation per se, but in solving a few specific moments of the Boltzmann transport equation. These moment equations do not require that $\frac{\partial f_H}{\partial H^*}$ is well-defined in the abstract, but only that a few specific weighted integrals of $\frac{\partial f_H}{\partial H^*}$ are well-defined. These integrals have been referred to as the *knee-energy driven fluxes*, and are defined as follows:

$$\mathbf{j}_{\nabla H^*} = \nabla_{\mathbf{r}} H^* \cdot \int_0^\infty D(\varepsilon)\mathcal{D}^\varepsilon \frac{\partial f}{\partial H^*} \mathrm{d}\varepsilon, \tag{6.54a}$$

$$\mathbf{S}_{\nabla H^*} = \nabla_{\mathbf{r}} H^* \cdot \int_0^\infty \varepsilon D(\varepsilon)\mathcal{D}^\varepsilon \frac{\partial f}{\partial H^*} \mathrm{d}\varepsilon, \tag{6.54b}$$

$$\mathbf{j}^{H>H_{\text{cut}}}_{\nabla H*} = \nabla_{\mathbf{r}} H^* \cdot \int_{\varepsilon_{\text{cut}}}^\infty D(\varepsilon)\mathcal{D}^\varepsilon \frac{\partial f}{\partial H^*} \mathrm{d}\varepsilon. \tag{6.54c}$$

If we express eq. (6.54) by first taking the weighted integral of $\frac{\Delta f_H(\Delta H^*)}{\Delta H^*}$, and then taking the limit as $\Delta H^* \to 0$, we obtain the following expression for the knee-energy driven fluxes:

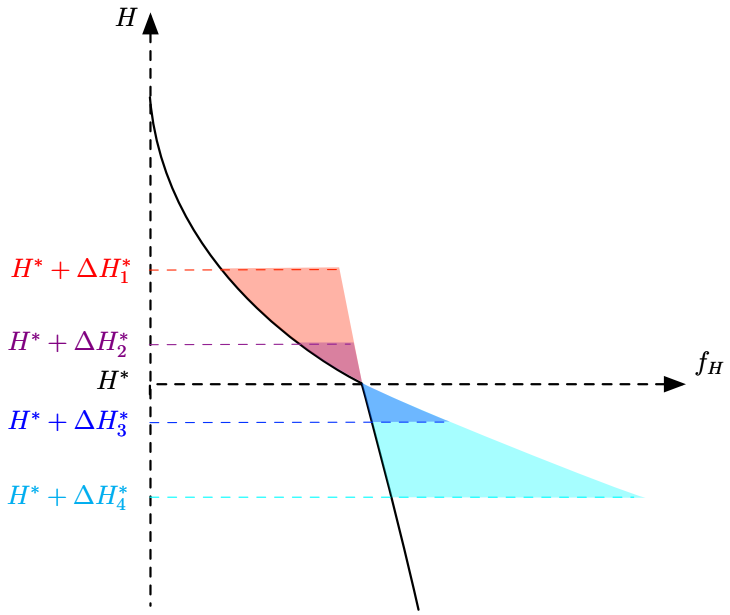


Figure 6.2: An illustration of the effect varying ΔH^* on the weighted integral of $\Delta f_H(\Delta H^*)$. The black plot shows the case without variation. The shaded red areas represent the weighted integral over Δf_H for $\Delta H^* > 0$, and the shaded blue areas represent the weighted integral over Δf_H for $\Delta H* < 0$. The darker shaded regions are associated with smaller magnitude ΔH^*, and their shaded regions sit on top of the lighter shaded regions associated with larger magnitude ΔH^*. We note that as ΔH^* becomes small, both the height and the average width of this region are proportional to ΔH^*. Therefore, the size of these weighted integrals, divided by ΔH^* will itself tend to zero as ΔH^* tends to zero. This fact is not changed by the possibility that the weight function may be zero in some regions, such as the region below some given H_{cut}.

$$\mathbf{j}_{\nabla H^*} = \nabla_{\mathbf{r}} H^* \cdot \left(\lim_{\Delta H^* \to 0} \frac{1}{\Delta H^*} \int_0^\infty D(\varepsilon) D^\varepsilon \Delta f_H(\Delta H^*) \mathrm{d}\varepsilon \right), \quad \text{(6.55a)}$$

$$\mathbf{S}_{\nabla H^*} = \nabla_{\mathbf{r}} H^* \cdot \left(\lim_{\Delta H^* \to 0} \frac{1}{\Delta H^*} \int_0^\infty \varepsilon D(\varepsilon) D^\varepsilon \Delta f_H(\Delta H^*) \mathrm{d}\varepsilon \right), \quad \text{(6.55b)}$$

$$\mathbf{j}_{\nabla H*}^{H>H_{\text{cut}}} = \nabla_{\mathbf{r}} H^* \cdot \left(\lim_{\Delta H^* \to 0} \frac{1}{\Delta H^*} \int_{\varepsilon_{\text{cut}}}^\infty D(\varepsilon) D^\varepsilon \Delta f_H(\Delta H^*) \mathrm{d}\varepsilon \right). \quad \text{(6.55c)}$$

It can be shown that the integrals defined by eq. (6.55) are well-defined and equal to zero. This can be demonstrated by simply substituting eq. (6.53) and performing the integration. As this is quite tedious, we take the approach of demonstrating it graphically in Fig. 6.2.

Aside: On Discretization and the Knee Dislocation

In a simple discretization of the continuity equations, the values of the ansatz parameters will be determined at the roughly the *centre* of a finite volume, Hereas the values associated with the gradients in the continuity equations will be determined on the *on the faces* of a finite volume. The fluxes depend on both the gradients of the ansatz parameters, and the ansatz parameters themselves, therefore in order to determine the fluxes on the faces we need to interpolate the ansatz parameters.

Many interpolation schemes for α^H, β, μ^H, α^H_{warm}, T_{warm} and μ^H_{cold} are valid, and even a simple linear interpolation scheme will lead to good results. The problem is that in any sensible interpolation scheme will result in a dislocation between the cutoff energy and the knee energy, as the cutoff total energy stays constant while the knee total energy is implicitly interpolated. As a result, the "knee dislocation" flux $\int_{\varepsilon_{\text{cut}}}^{\varepsilon^*} \mathbf{j}^\varepsilon \mathrm{d}\varepsilon$ will be non-zero on the faces of the finite volume Here the flux is measured. This is shown schematically in Fig. 6.3.

Accounting for the knee dislocation flux is trivial in the case that $\varepsilon_{\text{cut}} > \varepsilon^*$— instead of calculating the tail flux for a knee energy of ε^*, we calculate the tail flux for a knee energy of ε_{cut}. In the case that $\varepsilon_{\text{cut}} < \varepsilon^*$, the problem is not as trivial, but the flux can easily be estimated in a number of ways. One straight-forward approach is to simply assume that the density of states and energy-dependent diffusion coefficients are constant between ε^* and ε_{cut}. This leads to the following expression for the below knee dislocation flux $\mathbf{j}^{\varepsilon^*}_{\varepsilon_{\text{cut}}}$:

$$\begin{aligned}
\mathbf{j}_{\varepsilon_{\text{cut}}}^{\varepsilon^*} &= \int_{\varepsilon_{\text{cut}}}^{\varepsilon^*} -e^{-\alpha^\varepsilon-\beta\varepsilon} D(\varepsilon)\mathcal{D}^\varepsilon(\varepsilon;...)\cdot\left(\nabla_{\mathbf{r}}\alpha^H+\varepsilon\nabla_{\mathbf{r}}\beta\right)\mathrm{d}\varepsilon \\
&\approx D(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2})\mathcal{D}^\varepsilon(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2};...)\cdot\int_{\varepsilon_{\text{cut}}}^{\varepsilon^*} -e^{-\alpha^\varepsilon-\beta\varepsilon}\left(\nabla_{\mathbf{r}}\alpha^H+\varepsilon\nabla_{\mathbf{r}}\beta\right)\mathrm{d}\varepsilon \\
&= D(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2})\mathcal{D}^\varepsilon(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2};...)\cdot\left(\frac{1}{\beta}\nabla_{\mathbf{r}}\alpha^H+\frac{1}{\beta^2}(\beta\varepsilon+1)\nabla_{\mathbf{r}}\beta\right)e^{-(\alpha^\varepsilon+\beta\varepsilon)}\Big|_{\varepsilon_{\text{cut}}}^{\varepsilon^*}.
\end{aligned}$$

This leads to the following expression for the particle flux above the cutoff energy is the case of non-zero knee displacements:

$$\mathbf{j}_{H>H_{\text{cut}}} = \begin{cases} -\frac{n_{\text{tail}}(\varepsilon_{\text{cut}})}{kT_L}\mathcal{D}^0_{\text{tail}}(\varepsilon_{\text{cut}};\dots)\cdot\nabla_{\mathbf{r}}\mu^H & \varepsilon_{\text{cut}}\geq\varepsilon^*, \\ -\frac{n_{\text{tail}}(\varepsilon^*)}{kT_L}\mathcal{D}^0_{\text{tail}}(\varepsilon^*;\dots)\cdot\nabla_{\mathbf{r}}\mu^H+\mathbf{j}_{\varepsilon_{\text{cut}}}^{\varepsilon^*} & \varepsilon_{\text{cut}}<\varepsilon^*. \end{cases} \tag{6.56}$$

6.4.4 The Characteristic Scattering Terms as a Function of Ansatz Parameters

The Stunted Electron Population

The aim of this subsection is to describe a method of efficiently relating the ansatz parameters to the rates of change of characteristic densities due to scattering— in the case of the stunted electron population, $\left(\frac{\partial n}{\partial t}\right)_{\text{scat}}$, $\left(\frac{\partial w}{\partial t}\right)_{\text{scat}}$ and $\left(\frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial t}\right)_{\text{scat}}$. The starting point for this method is the calculation of the set of kinetic energy dependent, inelastic scattering parameters tabulated in Table 6.5.

We wish to generate *ansatz parameter* dependent inelastic scattering parameters from this set of kinetic-energy dependent inleastic scattering parameters that we can precompute and tabulate. Since the tail of the stunted distribution follows its own continuity equa-

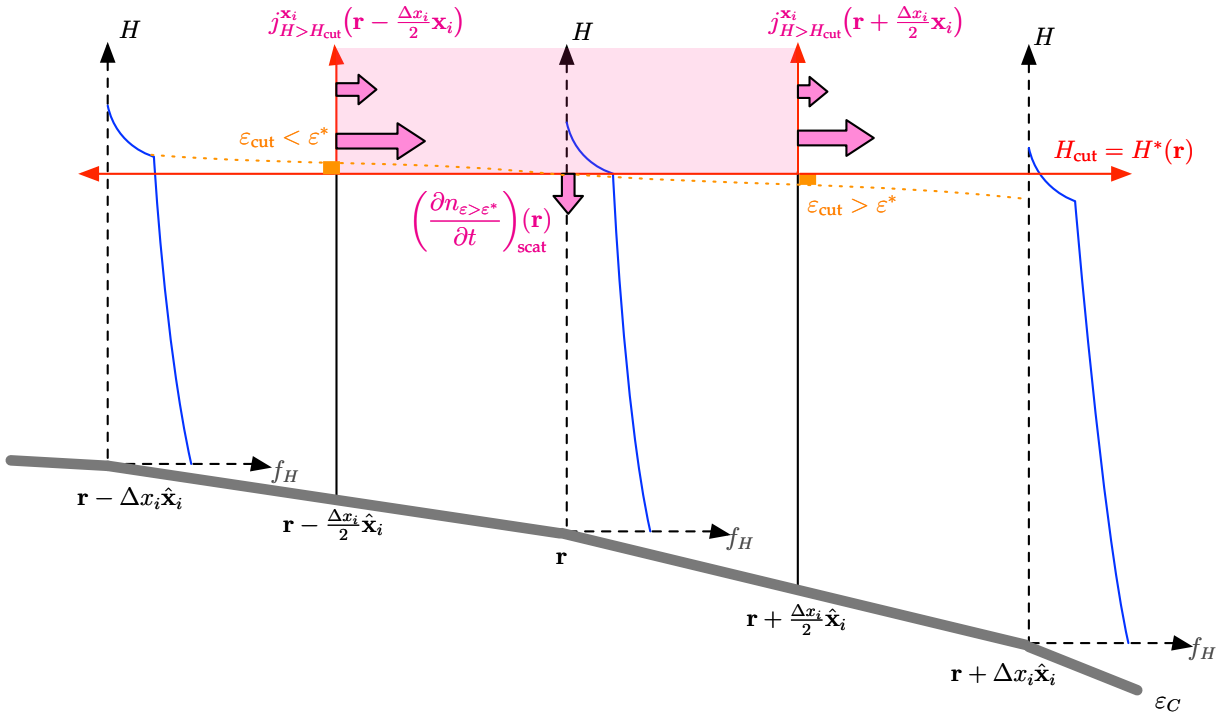


Figure 6.3: A schematic discretization of the continuity equation for the density above a cutoff energy. The global horizontal axis represents position along some direction $\hat{\mathbf{x}}^i$, the local horizontal axis represents symmetric occupation rate, and the vertical axis represents total energy. The potential energy at the bottom of the conduction band is represented by the thick grey line. The vertical black lines enclose the cell associated with $\mathbf{r}$. The shaded pink region represents the particle density above $H_{\text{cut}} = H^*(\mathbf{r})$. The pink arrows represent the particle flows in and out of the shaded pink region region. The dotted orange lines represents the interpolation of the knee energy, which is generally non-linear even in schemes Here the interpolation of α^H, β, and μ^H is linear. The finite "knee displacement" between the local knee energy and the cutoff energy that is typical in a discretization scheme is represented by the orange bar at $\mathbf{r}\pm\frac{\Delta x_i}{2}\hat{\mathbf{x}}_i$.

$\frac{1}{\tau_{\text{pho}}}(\varepsilon_i) = \int_0^\infty S_{\text{pho}}(\varepsilon_i, \varepsilon_f)\mathrm{d}\varepsilon_f$		Average rate of phonon scattering for an electrons at initial kinetic energy ε_i.
$\frac{1}{\tau_{\text{pho}}^{\text{tail-f}}}(\varepsilon_i, \varepsilon_{\text{cut}}) = \int_{\varepsilon_{\text{cut}}}^\infty S_{\text{pho}}(\varepsilon_i, \varepsilon_f)\mathrm{d}\varepsilon_f$		Average rate of phonon scattering for an electrons at initial kinetic energy ε_i that results in a final state kinetic energy ε_{cut}.
$\left\langle \varepsilon_{\text{pho}} \right\rangle (\varepsilon_i) = \tau_{\text{pho}} \int_0^\infty (\varepsilon_i - \varepsilon_f) S_{\text{pho}}(\varepsilon_i, \varepsilon_f)\mathrm{d}\varepsilon_f$		Average energy of phonons emitted from electron states at initial kinetic energy at ε_i. Absorbed phonons contribute negatively to this average.
$\frac{1}{\tau_{ii}}(\varepsilon_i) = \iiint_0^\infty S_{ii}(\varepsilon_i, \varepsilon_f, \varepsilon_f^e, \varepsilon_f^h)\mathrm{d}\varepsilon_f \mathrm{d}\varepsilon_f^e \mathrm{d}\varepsilon_f^h$		Average rate of impact ionization for electron states at initial kinetic energy ε_i.
$\langle \varepsilon_h \rangle (\varepsilon_i) = \tau_{ii} \iiint_0^\infty \varepsilon_f^h S_{ii}(\varepsilon_i, \varepsilon_f, \varepsilon_f^e, \varepsilon_f^h)\mathrm{d}\varepsilon_f \mathrm{d}\varepsilon_f^e \mathrm{d}\varepsilon_f^h$		Average kinetic energy of holes created by the impact ionization of electrons at initial kinetic energy ε_i.

Table 6.5: Table of energy-dependent inelastic scattering parameters that are relevant to the Three Quasi-Equilibria transport model.

tion, we will seperate each ansatz dependent scattering parameters into a component associated with the tail, as this is necessary in most but not all cases.

To aid tabulation, it is crucial that the ansatz parameter dependent scattering functions are slowly varying, and dependent on as few degrees of freedom as possible. Beyond ensuring that the scattering parameter does not depend on extraneous parameters, and is not exponentially increasing with as a function of ansatz parameters, we focus primarily on scattering parameters with a simple physical meaning.

For impact ionization, we simply determine the mean value of the hole energy and the mean impact ionization rate for electron in the tail and the bulk. For phonon scattering, we determine a mean energy relaxation rate for electrons in the bulk and tail, as well as the mean rate that tail electrons enter the bulk, and the mean rate that bulk electrons

enter the tail. These sets of parameters are tabulated in Table 6.6

$\frac{1}{\tau_{ii}^{\text{tail}}}(\varepsilon^*)$	$= \frac{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon}{\tau_{ii}}\right]}{\rho^0(\varepsilon^*,\infty)\left[f_\varepsilon\right]}$	Average impact ionization rate for electrons in the thermal tail.
$\frac{1}{\tau_{ii}^{\text{bulk}}}(\varepsilon^*,\beta)$	$= \frac{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon}{\tau_{ii}}\right]}{\rho^0(0,\varepsilon^*)\left[f_\varepsilon\right]}$	Average impact ionization rate for electrons in the hot bulk.
$\langle\varepsilon_h\rangle^{\text{tail}}(\varepsilon^*)$	$= \frac{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon\langle\varepsilon_h\rangle}{\tau_{ii}}\right]}{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon}{\tau_{ii}}\right]}$	Average kinetic energy of the holes generated by the impact ionization of electrons in the thermal tail.
$\langle\varepsilon_h\rangle^{\text{bulk}}(\varepsilon^*,\beta)$	$= \frac{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon\langle\varepsilon_h\rangle}{\tau_{ii}}\right]}{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon}{\tau_{ii}}\right]}$	Average kinetic energy of the holes generated by the impact ionization of electrons in the hot bulk.
$\frac{1}{\tau_{w,\text{tail}}^{\text{pho}}}(\varepsilon^*)$	$= \frac{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon\langle\varepsilon_{\text{pho}}\rangle}{\tau_{\text{pho}}}\right]}{\rho^1(\varepsilon^*,\infty)\left[f_\varepsilon\right]}$	Energy relaxation rate associated with phonon scattering of electrons in the thermal tail.
$\frac{1}{\tau_{w,\text{bulk}}^{\text{pho}}}(\varepsilon^*,\beta)$	$= \frac{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon\langle\varepsilon_{\text{pho}}\rangle}{\tau_{\text{pho}}}\right]}{\rho^1(0,\varepsilon^*)\left[f_\varepsilon\right]}$	Energy relaxation rate associated with phonon scattering of electrons in the hot bulk
$\frac{1}{\tau_{\text{tail–dest}}^{\text{pho}}}(\varepsilon^*)$	$= \frac{\rho^0(\varepsilon^*,\infty)\left[\frac{f_\varepsilon}{\tau_{\text{pho}}} - \frac{f_\varepsilon}{\tau_{\text{pho}}^{\text{tail-f}}}\right]}{\rho^0(\varepsilon^*,\infty)[f_\varepsilon]}$	Tail electron destruction rate associated with phonon scattering of electrons in the thermal tail.
$\frac{1}{\tau_{\text{tail–prod}}^{\text{pho}}}(\varepsilon^*,\beta)$	$= \frac{\rho^0(0,\varepsilon^*)\left[\frac{f_\varepsilon}{\tau_{\text{pho}}^{\text{tail-f}}}\right]}{\rho^0(0,\varepsilon^*)[f_\varepsilon]}$	Tail electron production rate associated with phonon scattering of electrons in the hot bulk.

Table 6.6: Table of ansatz-parameter dependent scattering functions associated with stunted electron–lattice inelastic scattering. These are to be precomputed and tabulated so that the characteristic scattering terms can be computed quickly at run-time.

For electron–electron scattering, for electrons in the tail and bulk, we calculate the mean

scattering time per unit stunted electron density, per unit warm electron density and per unit cold electron density. Thus there are six precomputed scattering parameters associated with electron–electron scattering. In order to define these scattering parameters, it is useful to first define the scattering rate at a given kinetic energy, per unit stunted electron density, per unit warm electron density and per unit cold electron density:

$$\frac{1}{\tau_{\text{ee-stunted}}}(\varepsilon, \varepsilon^*, \beta; \beta_S) = \frac{\rho^0(0,\infty)\left[\frac{f_\varepsilon}{\tau_{ee}}\right]}{\rho^0(0,\infty)[f_\varepsilon]}, \tag{6.57a}$$

$$\frac{1}{\tau_{\text{ee-warm}}}(\varepsilon, T_{\text{warm}}; \beta_S) = \frac{\rho^0(0,\infty)\left[\frac{f_\varepsilon^{\text{warm}}}{\tau_{ee}}\right]}{\rho^0(0,\infty)[f_\varepsilon^{\text{warm}}]}, \tag{6.57b}$$

$$\frac{1}{\tau_{\text{ee-cold}}}(\varepsilon; \beta_S) = \frac{\rho^0(0,\infty)\left[\frac{1}{\tau_{ee}} e^{-\frac{\varepsilon}{kT_L}}\right]}{\rho^0(0,\infty)\left[e^{-\frac{\varepsilon}{kT_L}}\right]}. \tag{6.57c}$$

Here we have ignored degeneracy effects associated with the cold electron density on the basis that the cold electrons scattering is only physically important in cases Here the electrons scattering with the cold electrons have a much larger average energy than the cold electrons. We now express our macroscopic electron–electron scattering parameters in Table 6.7 in terms of the scattering rates in eq. (6.57).

We will spend the rest of this subsection describing precisely how the rate of change of characteristic densities is related to these macroscopic scattering functions.

We begin with determining the rate of change of electrons in the stunted electron distribution due to scattering. According to our model, a stunted electron can be destroyed by warm electron scattering scattering, or stunted electron scattering which also destroys the stunted partner electron. A stunted electron can be created cold electron scattering or by impact ionization. This leads to the following expression for the rate of change in the particle density of stunted electrons:

$\frac{1}{\tau^{\text{tail}}_{\text{ee-stunted}}}(\varepsilon^*, \beta; \beta_S)$	$= \frac{\rho^0(\varepsilon^*, \infty)\left[\frac{f_\varepsilon}{\tau_{\text{ee-stunted}}}\right]}{\rho^0(\varepsilon^*, \infty)\left[f_\varepsilon\right]}$	Average tail electron scattering rate with stunted electron partners, per unit stunted electron density.
$\frac{1}{\tau^{\text{bulk}}_{\text{ee-stunted}}}(\varepsilon^*, \beta; \beta_S)$	$= \frac{\rho^0(0, \varepsilon^*)\left[\frac{f_\varepsilon}{\tau_{\text{ee-stunted}}}\right]}{\rho^0(0, \varepsilon^*)\left[f_\varepsilon\right]}$	Average bulk electron scattering rate with stunted electron partners, per unit stunted electron density.
$\frac{1}{\tau^{\text{tail}}_{\text{ee-warm}}}(\varepsilon^*, \beta, T_{\text{warm}}; \beta_S)$	$= \frac{\rho^0(\varepsilon^*, \infty)\left[\frac{f_\varepsilon}{\tau_{\text{ee-warm}}}\right]}{\rho^0(\varepsilon^*, \infty)\left[f_\varepsilon\right]}$	Average tail electron scattering rate with warm partners, per unit warm electron density.
$\frac{1}{\tau^{\text{bulk}}_{\text{ee-warm}}}(\varepsilon^*, \beta, T_{\text{warm}}; \beta_S)$	$= \frac{\rho^0(0, \varepsilon^*)\left[\frac{f_\varepsilon}{\tau_{\text{ee-warm}}}\right]}{\rho^0(0, \varepsilon^*)\left[f_\varepsilon\right]}$	Average bulk electron scattering rate with warm partners, per unit warm electron density.
$\frac{1}{\tau^{\text{tail}}_{\text{ee-cold}}}(\varepsilon^*, \beta; \beta_S)$	$= \frac{\rho^0(\varepsilon^*, \infty)\left[\frac{f_\varepsilon}{\tau_{\text{ee-cold}}}\right]}{\rho^0(\varepsilon^*, \infty)\left[f_\varepsilon\right]}$	Average tail electron scattering rate with cold partners, per unit cold electron density.
$\frac{1}{\tau^{\text{bulk}}_{\text{ee-cold}}}(\varepsilon^*, \beta; \beta_S)$	$= \frac{\rho^0(0, \varepsilon^*)\left[\frac{f_\varepsilon}{\tau_{\text{ee-cold}}}\right]}{\rho^0(0, \varepsilon^*)\left[f_\varepsilon\right]}$	Average bulk electron scattering rate with cold partners, per unit cold electron density.

Table 6.7: Table of ansatz-parameter dependent scattering functions associated with stunted electron–electron scattering. These are to be precomputed and tabulated so that the characteristic scattering terms can be computed quickly at run-time. If it is too difficult to store the stunted-warm electron scattering rates, these can be approximated by the stunted-stunted scattering rates since the distributions are expected to have similar average energies.

$$\left(\frac{\partial n}{\partial t}\right)_{\text{scat}} = -\overbrace{2\left(n_{\text{tail}} + n_{\text{bulk}}\right)\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right)}^{\text{stunted+stunted}\to\text{warm+warm}} - \overbrace{n_{\text{warm}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right)}^{\text{stunted+warm}\to\text{warm+warm}}$$
$$+ \overbrace{n_{\text{cold}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-cold}}}\right)}^{\text{stunted+cold}\to\text{stunted+stunted}} + \overbrace{\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{ii}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{ii}}\right)}^{\text{stunted}\to\text{stunted+stunted+hole}} \quad . \tag{6.58}$$

Each of these scattering mechanisms also effects the energy density. In the case of

electron–electron scattering, the electron approximately removes the average energy of the distribution starts and adds it to the distribution it moves to. For impact ionization we expect that each impact ionization event reduces the total kinetic energy of the population by the average energy of the hole produced. In addition to these energy change mechanisms that are mediated by stunted particle creation and annihilation, we also have the energy density change due to phonon scattering, which we express in terms of the tabulated macroscopic energy relaxation time. Accordingly we have the following expression for the rate of change of the energy density:

$$\begin{aligned}\left(\frac{\partial w}{\partial t}\right)_{\text{scat}} = &- \overbrace{2\left(n_{\text{tail}} + n_{\text{bulk}}\right)\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}} + \frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right)}^{\text{stunted+stunted}\to\text{warm+warm}} \overbrace{n_{\text{warm}}\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}} + \frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right)}^{\text{stunted+warm}\to\text{warm+warm}} \\ &+ \overbrace{n_{\text{cold}}\left\langle\varepsilon_{\text{cold}}\right\rangle\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-cold}}}\right)}^{\text{stunted+cold}\to\text{stunted+stunted}} - \overbrace{\left(\frac{n_{\text{tail}}\left\langle\varepsilon_h\right\rangle^{\text{tail}}}{\tau^{\text{tail}}_{ii}} + \frac{n_{\text{bulk}}\left\langle\varepsilon_h\right\rangle^{\text{bulk}}}{\tau^{\text{bulk}}_{ii}}\right)}^{\text{stunted}\to\text{stunted+stunted+hole}} \\ &- \overbrace{\left(\frac{w_{\text{tail}}}{\tau^{\text{pho}}_{w,\text{tail}}} + \frac{w_{\text{bulk}}}{\tau^{\text{pho}}_{w,\text{bulk}}}\right)}^{\text{stunted}\to\text{stunted}}.\end{aligned} \tag{6.59}$$

Finally, for the rate of change of thermal tail density due to scattering, we assume that each assume that each time an electron in the thermal tail undergoes impact ionization or cold electron scattering, the total number of electrons in the thermal tail *reduces* by one. That is, we assume that total number of electrons in the bulk grows by two for each of these processes. Adding in the transfer of electrons between the bulk and tail mediated by phonon scattering, and we are led to the following expression:

$$\begin{aligned}\left(\frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial t}\right)_{\text{scat}} = &- \overbrace{\left(2n_{\text{tail}} + n_{\text{bulk}}\right)\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}}}^{\text{tail+stunted}\to\text{warm+warm}} - \overbrace{n_{\text{warm}}\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}}}^{\text{tail+warm}\to\text{warm+warm}} \\ &- \overbrace{\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{ii}}}^{\text{tail}\to\text{bulk+bulk+hole}} - \overbrace{n_{\text{cold}}\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}}}^{\text{tail+cold}\to\text{bulk+bulk}} - \overbrace{\frac{n_{\text{tail}}}{\tau^{\text{pho}}_{\text{tail—dest}}}}^{\text{tail}\to\text{bulk}} + \overbrace{\frac{n_{\text{bulk}}}{\tau^{\text{pho}}_{\text{tail—prod}}}}^{\text{bulk}\to\text{tail}}.\end{aligned} \tag{6.60}$$

Thus we have now expressed the rate of change in the stunted characteristic densities due to scattering as a function of the ansatz parameters that is efficiently computable at run-time.

The Warm and Cold Electron Populations

The identity of the bulk population ansatz at $\varepsilon = \infty$ and the warm population ansatz leads to the following definition of the average rate warm electrons scatter with cold electrons, per unit cold electron density:

$$\frac{1}{\tau^{\text{warm}}_{\text{ee-cold}}}(T_{\text{warm}};\dots) = \frac{1}{\tau^{\text{stunted}}_{\text{ee-cold}}}(\infty, \frac{1}{kT_{\text{warm}}};\dots). \tag{6.61}$$

The processes that create warm electrons are stunted-stunted electron scattering, warm-stunted electron scattering, warm-cold electron scattering, and warm electron impact ionization, while no processes destroy warm electrons. This leads to the following expression for the rate of change of the warm electron density:

$$\left(\frac{\partial n_{\text{warm}}}{\partial t}\right)_{\text{scat}} = \overbrace{2\left(n_{\text{tail}}+n_{\text{bulk}}\right)\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right)}^{\text{stunted+stunted}\to\text{warm+warm}} + \overbrace{n_{\text{warm}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right)}^{\text{stunted+warm}\to\text{warm+warm}}$$
$$+ \overbrace{\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{ii}}}^{\text{warm}\to\text{warm+warm+hole}} + \overbrace{n_{\text{cold}}\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{\text{ee-cold}}}}^{\text{warm+cold}\to\text{warm+warm}} . \tag{6.62}$$

Using the same argument as before for the stunted electron energy density, we can convert the warm electron particle density into a warm electron energy density by multiplying each term in eq. (6.62) by an appropriate average energy, adding a term that defines the rate of change of density due to phonon scattering. This leads to the following expression:

$$\left(\frac{\partial w_{\text{warm}}}{\partial t}\right)_{\text{scat}} = \overbrace{2\left(n_{\text{tail}}+n_{\text{bulk}}\right)\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}}+\frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right)}^{\text{stunted+stunted}\to\text{warm+warm}} + \overbrace{n_{\text{warm}}\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}}+\frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right)}^{\text{stunted+warm}\to\text{warm+warm}}$$
$$- \overbrace{\frac{n_{\text{warm}}\left\langle\varepsilon_h\right\rangle^{\text{warm}}}{\tau^{\text{warm}}_{ii}}}^{\text{warm}\to\text{warm+warm+hole}} + \overbrace{n_{\text{cold}}\left\langle\varepsilon_{\text{cold}}\right\rangle\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{\text{ee-cold}}}}^{\text{warm+cold}\to\text{warm+warm}} - \overbrace{\frac{w_{\text{warm}}}{\tau^{\text{pho}}_{w,\text{warm}}}}^{\text{stunted}\to\text{stunted}} . \tag{6.63}$$

Finally, for cold electrons we can ignore impact ionization, meaning there are zero creation processes for cold electrons. Instead there is simply the cold electron destruction by scattering with stunted or warm electrons:

$$\left(\frac{\partial n_{\text{cold}}}{\partial t}\right)_{\text{scat}} = -\overbrace{n_{\text{cold}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-cold}}}\right)}^{\text{stunted+cold}\to\text{stunted+stunted}} - \overbrace{n_{\text{cold}}\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{\text{ee-cold}}}}^{\text{warm+cold}\to\text{warm+warm}} . \tag{6.64}$$

Thus we have now expressed the rate of change in the characteristic densities due to scattering as a function of the ansatz parameters, which is efficiently computable at run-time. Having already achieved the same for the characteristic densities and characteristic fluxes, we now derived an expression for the equation of motion of the ansatz parameters that can in principle be computed efficiently at run-time. This is the Three Quasi-Equilibria model, which when coupled with Poisson's equation, can be used to numerically model both time-dependent and steady-state non-local electron transport in silicon devices.

6.5 Summary

In this chapter, we have derived a system of equations for modelling semiclassical electron transport in highly inhomogeneous fields for source-drain devices made of homogeneous silicon. We can summarise this equation system as follows.

We partition the electrons into a stunted electron population of density n, a warm electron population of density n_{warm} and a cold population of density n_{cold}. The external force on these populations resulting from long range Coloumb interactions can be calculating via the Poisson equation:

$$\mathbf{F} = e\nabla_{\mathbf{r}}\phi, \tag{6.65a}$$

$$\nabla_{\mathbf{r}}\cdot(\epsilon\nabla_{\mathbf{r}}\phi) = e(n + n_{\text{warm}} + n_{\text{cold}} - \sum_i N^i_{\text{dop}} Z^i_{\text{dop}}). \tag{6.65b}$$

The distribution of the stunted electron population are defined by the ansatz parameters μ^{ε}, β, and ε^*. The distribution of the warm electron population is defined by the ansatz parameters $\alpha^{\varepsilon}_{\text{warm}}$, and T_{warm} and the distribution of the cold electron population is defined by the ansatz parameter $\mu^{\varepsilon}_{\text{cold}}$. We can define a relationship between the ansatz

parameters and a set of densities that characterize the electron population as follows:

$$n = n_{\text{tail}} + n_{\text{bulk}}, \tag{6.66a}$$

$$w = w_{\text{tail}} + w_{\text{bulk}}, \tag{6.66b}$$

$$n_{\text{tail}} = \left(kT_L e^{-\frac{\varepsilon^* - \mu^\varepsilon}{kT_L}} \right) D_{\text{tail}}, \tag{6.66c}$$

$$n_{\text{warm}} = kT_{\text{warm}} e^{-\alpha^\varepsilon_{\text{warm}}} D_{\text{warm}}, \tag{6.66d}$$

$$w_{\text{warm}} = n_{\text{warm}} \left\langle \varepsilon \right\rangle_{\text{warm}}, \tag{6.66e}$$

$$n_{\text{cold}} = \left[\mu^\varepsilon_{\text{cold}} + kT_L \ln\left(e^{-\frac{\mu^\varepsilon_{\text{cold}}}{kT_L}} + 1 \right) \right] D_{\text{cold}}. \tag{6.66f}$$

Here n_{bulk}, w_{tail} and w_{bulk} are defined as follows:

$$n_{\text{bulk}} = \left(\frac{1}{\beta} \left(e^{\beta\varepsilon^*} - 1 \right) e^{-\frac{\varepsilon^* - \mu^\varepsilon}{kT_L}} \right) D_{\text{bulk}}, \tag{6.67a}$$

$$w_{\text{tail}} = n_{\text{tail}} \left\langle \varepsilon \right\rangle_{\text{tail}}, \tag{6.67b}$$

$$w_{\text{bulk}} = n_{\text{bulk}} \left\langle \varepsilon \right\rangle_{\text{bulk}}. \tag{6.67c}$$

From these relationships, we can define an equation for the rate of change in the ansatz parameters in terms of the rate of change in the characteristic densities:

$$\frac{\partial \mu^\varepsilon}{\partial t} = \frac{kT_L}{n_{\text{tail}}} \frac{\partial n_{\varepsilon > \varepsilon_{\text{cut}}}}{\partial t}, \tag{6.68a}$$

$$\begin{bmatrix} \frac{\partial \beta}{\partial t} \\ \frac{\partial \varepsilon^*}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial n}{\partial \beta} & \frac{\partial n}{\partial \varepsilon^*} \\ \frac{\partial w}{\partial \beta} & \frac{\partial w}{\partial \varepsilon^*} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial n}{\partial t} - \frac{n}{n_{\text{tail}}} \frac{\partial n_{\varepsilon > \varepsilon_{\text{cut}}}}{\partial t} \\ \frac{\partial w}{\partial t} - \frac{w}{n_{\text{tail}}} \frac{\partial n_{\varepsilon > \varepsilon_{\text{cut}}}}{\partial t} \end{bmatrix}, \tag{6.68b}$$

$$\begin{bmatrix} \frac{\partial \alpha^\varepsilon_{\text{warm}}}{\partial t} \\ \frac{\partial T_{\text{warm}}}{\partial t} \end{bmatrix} = \begin{bmatrix} -n_{\text{warm}} & \frac{\partial n_{\text{warm}}}{\partial T_{\text{warm}}} \\ -w_{\text{warm}} & \frac{\partial w_{\text{warm}}}{\partial T_{\text{warm}}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial n_{\text{warm}}}{\partial t} \\ \frac{\partial w_{\text{warm}}}{\partial t} \end{bmatrix}, \tag{6.68c}$$

$$\frac{\partial \mu^\varepsilon_{\text{cold}}}{\partial t} = \left(\frac{\partial n_{\text{cold}}}{\partial \mu^\varepsilon_{\text{cold}}} \right)^{-1} \frac{\partial n_{\text{cold}}}{\partial t}. \tag{6.68d}$$

We can express the rate of change in the characteristic densities in terms via macroscopic continuity equations:

$$\frac{\partial n}{\partial t} = \left(\frac{\partial n}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}, \tag{6.69a}$$

$$\frac{\partial w}{\partial t} = \left(\frac{\partial w}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{S} + \mathbf{F} \cdot \mathbf{j}, \tag{6.69b}$$

$$\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t} = \left(\frac{\partial n_{\varepsilon>\varepsilon_{\text{cut}}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}_{H>H_{\text{cut}}}, \tag{6.69c}$$

$$\frac{\partial n_{\text{warm}}}{\partial t} = \left(\frac{\partial n_{\text{warm}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}_{\text{warm}}, \tag{6.69d}$$

$$\frac{\partial w_{\text{warm}}}{\partial t} = \left(\frac{\partial w_{\text{warm}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{S}_{\text{warm}} + \mathbf{F} \cdot \mathbf{j}_{\text{warm}}, \tag{6.69e}$$

$$\frac{\partial n_{\text{cold}}}{\partial t} = \left(\frac{\partial n_{\text{cold}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \mathbf{j}_{\text{cold}}. \tag{6.69f}$$

In these macroscopic continuity equations, we can express the characteristic fluxes in terms of ansatz parameters as follows:

$$\mathbf{j} = n_{\text{bulk}} \mathcal{D}^0_{\text{bulk}} \cdot \nabla_{\mathbf{r}} \alpha^H + n_{\text{bulk}} \mathcal{D}^1_{\text{bulk}} \cdot \nabla_{\mathbf{r}} \beta - \frac{n_{\text{tail}}}{kT_L} \mathcal{D}^0_{\text{tail}} \cdot \nabla_{\mathbf{r}} \mu^H, \tag{6.70a}$$

$$\mathbf{S} = n_{\text{bulk}} \mathcal{D}^1_{\text{bulk}} \cdot \nabla_{\mathbf{r}} \alpha^H + n_{\text{bulk}} \mathcal{D}^2_{\text{bulk}} \cdot \nabla_{\mathbf{r}} \beta - \frac{n_{\text{tail}}}{kT_L} \mathcal{D}^1_{\text{tail}} \cdot \nabla_{\mathbf{r}} \mu^H, \tag{6.70b}$$

$$\mathbf{j}_{H>H_{\text{cut}}} = \begin{cases} -\frac{n_{\text{tail}}(\varepsilon_{\text{cut}})}{kT_L} \mathcal{D}^0_{\text{tail}}(\varepsilon_{\text{cut}}; \dots) \cdot \nabla_{\mathbf{r}} \mu^H & \varepsilon_{\text{cut}} \geq \varepsilon^*, \\ -\frac{n_{\text{tail}}(\varepsilon^*)}{kT_L} \mathcal{D}^0_{\text{tail}}(\varepsilon^*; \dots) \cdot \nabla_{\mathbf{r}} \mu^H + \mathbf{j}^{\varepsilon^*}_{\varepsilon_{\text{cut}}} & \varepsilon_{\text{cut}} < \varepsilon^*, \end{cases} \tag{6.70c}$$

$$\mathbf{j}_{\text{warm}} = n_{\text{warm}} \mathcal{D}^0_{\text{warm}} \cdot \nabla_{\mathbf{r}} \alpha^H_{\text{warm}} - \frac{n_{\text{warm}}}{kT^2_{\text{warm}}} \mathcal{D}^1_{\text{warm}} \cdot \nabla_{\mathbf{r}} T_{\text{warm}}, \tag{6.70d}$$

$$\mathbf{S}_{\text{warm}} = n_{\text{warm}} \mathcal{D}^1_{\text{warm}} \cdot \nabla_{\mathbf{r}} \alpha^H_{\text{warm}} - \frac{n_{\text{warm}}}{kT^2_{\text{warm}}} \mathcal{D}^2_{\text{warm}} \cdot \nabla_{\mathbf{r}} T_{\text{warm}}, \tag{6.70e}$$

$$\mathbf{j}_{\text{cold}} = -\frac{n_{\text{cold}}}{kT_L} \nabla_{\mathbf{r}} \mu^H_{\text{cold}} \cdot \mathcal{D}^0_{\text{cold}}. \tag{6.70f}$$

Here the thermodynamic versions of the ansatz parameters and knee dislocation particle flux is defined as follows:

$$\alpha^H = \frac{\varepsilon^* - \mu^\varepsilon}{kT_L} - \beta(\varepsilon^* - e\phi), \tag{6.71a}$$

$$\mu^H = \mu^\varepsilon - e\phi, \tag{6.71b}$$

$$\alpha^H_{\text{warm}} = \alpha^\varepsilon_{\text{warm}} + \beta e\phi, \tag{6.71c}$$

$$\mu^H_{\text{cold}} = \mu^\varepsilon_{\text{cold}} - e\phi, \tag{6.71d}$$

$$\mathbf{j}^{\varepsilon^*}_{\varepsilon_{\text{cut}}} = D(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2})\mathcal{D}^\varepsilon(\tfrac{\varepsilon^*+\varepsilon_{\text{cut}}}{2};...)\cdot\left(\frac{1}{\beta}\nabla_{\mathbf{r}}\alpha^H + \frac{1}{\beta^2}(\beta\varepsilon+1)\nabla_{\mathbf{r}}\beta\right)e^{-(\alpha^\varepsilon+\beta\varepsilon)}\Big|^{\varepsilon^*}_{\varepsilon_{\text{cut}}}. \tag{6.71e}$$

And finally we can express the rate of change in the characteristic densities due to scattering in terms of ansatz parameters as follows:

$$\begin{aligned}\left(\frac{\partial n}{\partial t}\right)_{\text{scat}} = &-2\left(n_{\text{tail}}+n_{\text{bulk}}\right)\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right)-n_{\text{warm}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right)\\ &+n_{\text{cold}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-cold}}}\right)+\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{ii}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{ii}}\right),\end{aligned} \tag{6.72a}$$

$$\begin{aligned}\left(\frac{\partial w}{\partial t}\right)_{\text{scat}} = &-2\left(n_{\text{tail}}+n_{\text{bulk}}\right)\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}}+\frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right)n_{\text{warm}}\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}}+\frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right)\\ &+n_{\text{cold}}\left\langle\varepsilon_{\text{cold}}\right\rangle\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}}+\frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-cold}}}\right)-\left(\frac{n_{\text{tail}}\left\langle\varepsilon_h\right\rangle^{\text{tail}}}{\tau^{\text{tail}}_{ii}}+\frac{n_{\text{bulk}}\left\langle\varepsilon_h\right\rangle^{\text{bulk}}}{\tau^{\text{bulk}}_{ii}}\right)\\ &-\left(\frac{w_{\text{tail}}}{\tau^{\text{pho}}_{w,\text{tail}}}+\frac{w_{\text{bulk}}}{\tau^{\text{pho}}_{w,\text{bulk}}}\right),\end{aligned} \tag{6.72b}$$

$$\begin{aligned}\left(\frac{\partial n_{\varepsilon>\varepsilon^*}}{\partial t}\right)_{\text{scat}} = &-\left(2n_{\text{tail}}+n_{\text{bulk}}\right)\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}}\quad -n_{\text{warm}}\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}}\\ &-\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{ii}}-n_{\text{cold}}\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}}-\frac{n_{\text{tail}}}{\tau^{\text{pho}}_{\text{tail—dest}}}+\frac{n_{\text{bulk}}}{\tau^{\text{pho}}_{\text{tail—prod}}},\end{aligned} \tag{6.72c}$$

$$\left(\frac{\partial n_{\text{warm}}}{\partial t}\right)_{\text{scat}} = 2\left(n_{\text{tail}} + n_{\text{bulk}}\right)\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right) + n_{\text{warm}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right) + \frac{n_{\text{warm}}}{\tau^{\text{warm}}_{ii}} + n_{\text{cold}}\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{\text{ee-cold}}}, \tag{6.72d}$$

$$\left(\frac{\partial w_{\text{warm}}}{\partial t}\right)_{\text{scat}} = 2\left(n_{\text{tail}} + n_{\text{bulk}}\right)\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-stunted}}} + \frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-stunted}}}\right) + n_{\text{warm}}\left(\frac{w_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-warm}}} + \frac{w_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-warm}}}\right) - \frac{n_{\text{warm}}\left\langle\varepsilon_h\right\rangle^{\text{warm}}}{\tau^{\text{warm}}_{ii}} + n_{\text{cold}}\left\langle\varepsilon_{\text{cold}}\right\rangle\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{\text{ee-cold}}} - \frac{w_{\text{warm}}}{\tau^{\text{pho}}_{w,\text{warm}}}, \tag{6.72e}$$

$$\left(\frac{\partial n_{\text{cold}}}{\partial t}\right)_{\text{scat}} = -n_{\text{cold}}\left(\frac{n_{\text{tail}}}{\tau^{\text{tail}}_{\text{ee-cold}}} + \frac{n_{\text{bulk}}}{\tau^{\text{bulk}}_{\text{ee-cold}}}\right) - n_{\text{cold}}\frac{n_{\text{warm}}}{\tau^{\text{warm}}_{\text{ee-cold}}}. \tag{6.72f}$$

The result of this chain of logic is that we have derived a closed equation-of-motion for the stunted electron ansatz parameters. This efficient solution of this equation of motion relies on the pre-computation and tabulation of the many macroscopic transport parameters as functions of the ansatz variables.

The precomputation of these macroscopic transport parameters is a rather substantial computational task, and the tabulation will require several gigabytes of storage until good analytic approximations for these parameters are found, but this precomputation and tabulation will save an enormous amount of computational effort at run-time. This is because according to the assumptions of our model, these macroscopic transport parameters are universal to all semiclassical source-drain unstrained silicon devices. Thus if these macroscopic transport parameters are tabulated, we can model any device geometry in this class without having to either make strong assumptions about scattering operator or band structure, and without having to recompute numerically any of the integrals required for closure.

If we precompute the macroscopic transport parameters, the time taken to solve our Three Quasi-Equilibria Model in any device geometry that belongs to the specified class should be comparable to a more conventional analytically closed macroscopic model. However, unlike many macroscopic models, the accuracy and robustness of the model

proposed will not be intrinsically limited due unphysical assumptions concerning band-structure and scattering. The accuracy and robustness of the model proposed is limited only by the accuracy of the ansatz. Another way to state this is that the accuracy and robustness of the model proposed will approach the accuracy and robustness of a detailed Monte-Carlo simulation to the extent that the proposed ansatz reflects reality. If the ansatz we propose is unreasonable, then the effort we have taken to incorporate the precise microscopic details that are used in accurate Monte-Carlo simulations will achieve *almost nothing* in terms of improving accuracy and robustness.

However, if the ansatz we have proposed is reasonable, then the Three Quasi-Equilibria model of non-equilibrium transport we have proposed will have a *speed* comparable to a conventional macroscopic model, and an *accuracy* and *robustness* comparable to a detailed Monte Carlo simulation. In the view of this author, this benefit easily justifies the upfront computational effort and the storage costs associated with tabulating the macroscopic transport parameters.

Chapter 7

Discussion

7.1 Introduction

In this thesis, we have derived two models of semiclassical electron transport– the ELASTICALLY CONSTRAINED TRANSPORT MODEL and the THREE EQUILIBRIA MODEL.

Our derivation of the Elastically Constrained Transport model makes four major assumptions in addition to the ordinary semiclassical assumptions, each of which we have explained and justified in the previous chapters.

Assumption 1: Scattering is the combination of a purely elastic relaxation time, and a purely inelastic scattering operator that non-locally connects energy levels.

Assumption 2: Electron-electron scattering does not contribute to the elastic relaxation time.

Assumption 3: The antisymmetric perturbation from an elastically-constrained equilibrium is small.

Assumption 4: The symmetric perturbation from elastically constrained equilibrium is

negligible.

Our derivation of the Three Quasi-Equilibrium model of assumes the validity of the Elastically Constrained Transport model, and makes the following additional four major assumptions.

Assumption 5: The energy distribution for electrons injected from a non-drain terminal subject to scattering with lattice temperature scattering partners is a STUNTED EQUILIBRIUM, defined as the maximum entropy distribution subject to a local particle density, energy density and maximum chemical potential.

Assumption 6: Degeneracy effects in subpopulations of electrons other than the drain electron population can be ignored.

Assumption 7: Electron-electron scattering between electrons in a stunted equilibrium distribution instantly relax into an internal thermal equilibrium energy distribution.

Assumption 8: The vast majority of electron transport occurs between a single source terminal and a single drain terminal.

In this discussion chapter, we will briefly discuss opportunities to make these assumptions closer to the pure semiclassical transport assumptions described in the background section, and then also . We discuss this in two major sections. In the first section, we discuss possibilities for making these models closer to the model of pure semiclassical electron transport in silicon. In the second section, we discuss the possibilities for adapting these models in the more complex regime of transport that is common in modern devices.

To roughly quantify the difficulty of the extensions we discuss, we will use the following "difficulty score" scale to describe the level of effort required to make the improvements proposed.

Level 1: Trivial extension with no new insight or theory required.

Level 2: Seemingly fixable without much new insight or theory required.

Level 3: Open question as to whether problem is fixable, considerable research needed in order generate new insights and theory.

Level 4: Seemingly unfixable.

To roughly quantify the urgency of the extensions we discuss, we will use the following "urgency score".

Level 1: The model does not need this extension in order to become a powerful, practical tool.

Level 2: The power and usefulness of model is noticeably limited without this extension.

Level 3: Essential for model to be practically useful.

7.2 Simplifying the Homogeneous Semiclassical Electron State

We begin by analyzing the major assumptions of the Elastically Constrained Transport model.

Assumption 1: It is the position of this thesis that the separation of the scattering operator into the sum of an elastic relaxation time and a purely inelastic scattering operator between energy levels is fundamentally consistent with the semiclassical assumptions. We do not believe that any important physical effects are missed by this assumption. One interesting way to test this hypothesis would be to perform an ensemble Monte Carlo simulation of electron transport using this formulation of the scattering operator and compare the results to a DAMOCLES ensemble Monte Carlo simulation. No improvements are needed to this basic fundamental assumption.

However the calculation for actual values of the elastic relaxation time and pure inelastic scattering operator could be improved. The expressions derived in the background chapter for the screening wavevector and overlap integral are dubious, and need improvement. Most of this work has already been done by the authors of the DAMOCLES model [6]. In addition it would be optimal to include the effect of plasmons in the inelastic scattering operator, as including the effect of plasma oscillations implicitly by solving a time-dependent Poisson equation is inappropriate for the model we have proposed. *Difficulty:* $1-2$. *Urgency:* $2-3$.

Assumption 2: Improving upon the assumption that electron–electron scattering events do not effect the elastic relaxation time, while at the same time ensuring that the expression for the energy-dependent diffusion coefficient is feasible to store in a look-up table requires fundamental new insight. In addition, it is unclear that improving upon this assumption will make any practically relevant changes to the models predictions. *Difficulty:* $3-4$. *Urgency:* 1.

Assumption 3: The assumption that the antisymmetric distribution is small is probably the largest weakness of the Elastically Constrained Transport model, since when gradients in energy dependent occupation rate are very large it will not be true. However, it is important to remember that because the inelastic component of relaxation is generally relatively slow compared to the elastic component, in steady-state the gradients in the energy dependent component in a large field than would be the case if the inelastic component of scattering was highly efficient. That said, it is the view of this thesis that the assumption that the antisymmetric component is small is not necessarily fundamental to the elastically constrained equilibrium. One can easily imagine a model in which the particle flux at constant energy does *not* scale linearly with increasing gradient in total energy distribution function, but *saturates* as the gradient becomes very large. The importance of this saturation will only be clear once the current formulation of the Elastically Constrained Transport model is tested. *Difficulty:* 3. *Urgency:* 1.

Assumption 4: When the gradient of the antisymmetric distribution at constant total energy is large, there is a generating force for a symmetric perturbation to the elastically constrained distribution function. Dmitruk et al. [21] have already shown how to incorporate this term using a highly idealized scaling argument, but more research is

needed to understand the role of this term in the regime where the antisymmetric distribution properly understand it's role in the regime where the antisymmetric distribution function has a similar scale to the energy distribution function. *Difficulty:* 3. *Urgency:* 1.

We now move on analyzing the assumptions of the Three Quasi-Equilibrium model.

Assumption 5: The assumption that the electrons which enter a non-drain terminal have the energy distribution of associated with a "stunted equilibrium" in a material with a broadband inelastic scattering operator such as silicon is central to the simplicity of the Three Quasi-Equilibrium model. The arguments for this distribution being maximum entropy falls apart if there is a strong correlation between the energy state of an Bloch particle in the distribution and the time it spends in the device. This will occur if there is a large classical potential well that trap Bloch particles below a certain energy. In devices with such wells, one need to add an additional "trapped source electron" population to the distribution function, as such a population of trapped electrons will always undergo a sufficient number of inelastic collisions in order to reach thermal equilibrium with the lattice. *Difficulty:* $1-2$. *Urgency:* 1

A more obvious error in the stunted equilibrium distribution is however the shape of the distribution near the knee energy, as the shape of the distribution discontinuity in the distribution is likely to be softened by both constant energy diffusion and the effects of very low-energy inelastic scattering events. The corrections to the shape over such a small energy scale however, are not likely to be important. Another important question is the consistency of the ansatz with low-density homogeneous field data. In a low-density homogeneous field, the knee energy can be expected to diverge to infinity so long as spatial diffusion is faster than energy diffusion. The ansatz is only consistent with homogeneous field data so long as low-density homogeneous field data is thermal. This is approximately, but not precisely true [6]. While the ansatz is not expected to be precisely accurate, it is expected to roughly accurate, which is what ultimately matters in order to be able to correctly predict the future values of ansatz parameters. It is difficult to see how to generally improve without sacrificing the simplicity of the transport model. *Difficulty:* $3-4$. *Urgency:* 1.

Assumption 6: The electrons that enter the device through a non-drain terminal will

typically end up spread thinly throughout across a broadband of energy levels, and therefore ignoring the effects of degeneracy on non-drain electrons is expected to be theoretically sound in most positions in the device. The only position where this is potentially not true is very near the injection terminal, as in this position electrons are not spread thinly throughout a broadband of energies. In order to fix this problem, we suggest the following "hack". If the ansatz occupation rate is ever greater than 1, then β should be set to 0, and maximum chemical potential set such that the knee energy has an occupation rate of 1. This will exhibit the approximate transport characteristics of a Fermi-Dirac distribution without having to complicate the model by formally incorporating degeneracy. *Difficulty:* $1-2$. *Urgency:* $1-2$.

Assumption 7: The basis of this approximation is the well known fact that electron-electron scattering are very efficient at relaxing a distribution of electrons toward an internal equilibrium, because the scattering operator associated with such collisions is strongly coupled to most states in the Brillouin zone, since there are so many surfaces that total energy and crystal momentum. The difficulty is that when two electrons scatter, they are only strongly coupled to electron energy states that have less than the total of the two electrons involved. This means that, since electron states more than a few thermal energies above the knee energy are essentially empty, scattering between electrons in a stunted equilibrium will never result in a significant occupation of electrons in states with energies higher than double the knee-energy. The problem with incorporating this into the model is that, in cases where electron-electron scattering is strong, one needs to incorporate more and more warmed distributions with more and more thresholds. Furthermore, unless these additional thresholds are very near the band gap and thus the thresholds for impact ionization it is unlikely they will have any significant effect the physics of transport. Thus it is possible to extend the model to correct for this approximation, but it adds considerable complexity, and is only expected to have a physically significant effect on the transport model when the source-drain bias is in a small range of values very close to half the band gap voltage. *Difficulty:* $2-3$. *Urgency:* 1.

Assumption 8: Many important devices can be viewed as being dominated by the transport between a single source and single drain terminal with all other electrons– such as electrons associated with a "gate" terminal– only affecting this transport via long range

effects. However this restriction does limit the flexibility of the model. Adding electron populations associated with other terminals is expected to be relatively trivial, as no fundamentally new mechanics need to be described. The resulting model will still be a three equilibria model, in the sense that transport will always involve three *types* of equilibrium, even if there are multiple populations of each type. *Difficulty:* $1-2$. *Urgency:* 2.

Finally, we note that the problem of discretizing these models is relatively trivial since the models are underlying partial differential equations are never hyperbolic.

7.3 Beyond a Homogeneous Semiclassical Electron State

Hole Transport: Incorporating an accurate model of hole transport is critical for the flexibility of the device model. *Difficulty:* $1-2$. *Urgency:* 3.

First-Order Quantum Corrections: First order quantum corrections can be included by replacing the external potential with an *effective* external potential which is a weighted average over the decoherence length scale. Such corrections are simple to implement and have physically significant effects in many devices of interest. *Difficulty:* $1-2$. *Urgency:* $2-3$.

Dynamic Coupling To Schrödingers Equation: Some modern Monte Carlo simulators are coupled dynamically to Schrödingers equation [90]. Attempting to do the same with the models described in this thesis is deeply problematic because the models rely heavily on being able to *precompute and store* integrals of the distribution function and scattering operator that are very expensive to compute at run-time. If this is not possible, because the bandstructure and scattering operator associated with a device are not known ahead of run-time, then the models in this thesis lose their advantage over Monte Carlo simulation. *Difficulty:* $3-4$. *Urgency:* 1.

Non-Silicon Materials and Heterojunctions: While silicon is the backbone of the semiconductor industry, the ability to model transport accurately in other materials and

across heterojunctions is becoming more and more important. The Elastically-Constrained Equilibrium Model is expected can be expected to be valid in any material where acoustic scattering is dominant, and the Three Quasi-Equilibrium Model is expected to be valid in any material where the inelastic scattering operator is broadband. Describe the transport of these models across heterojunctions is a slightly more complicated issue, that will require some new insights but will be based on similar physics that has been used in ensemble Monte Carlo simulations of heterojunctions [6]. *Difficulty:* $2-3$. *Urgency:* $2-3$.

Dynamic Coupling to Lattice Temperature: The dynamic coupling of both models to a dynamic *uniform* lattice temperature is trivial. However, the dynamic coupling of the models to a dynamic *non-uniform* lattice temperature is an open research problem for electron transport models in general [91]. It is likely that a non-uniform lattice temperature is fundamentally incompatible with the Three Quasi-Equilibrium model, as the concept of a local maximum chemical potential becomes much more dubious when lattice temperature changes from position to position. *Difficulty:* $3-4$. *Urgency:* 2.

Chapter 8

Conclusion

In this thesis we have described a theoretically sound way to simplify the semiclassical model of electron transport in silicon to a model which depends only on a $4-$D or $3-$D electron state. This had already been achieved in cases where the electron distribution is near local dynamic equilibrium with the local field, but it had not yet been achieved in the innately inhomogeneous regime where the models proposed in this thesis are valid.

The fundamental reason for being able to simplify the semiclassical model of electron transport is that the model itself relied on the assumption that the field is relatively constant over a decoherence length. In silicon, the primary decoherence mechanism is acoustic phonon scattering and so in order for the semiclassical regime to be valid the external field needs to be relatively constant over the mean free path between scattering events. This fact is often missed by authors who argue that one can model ballistic transport semiclassically.

The fact that the semiclassical model is only valid in silicon when the external field is relatively constant over the mean free path between acoustic phonon scattering events means that the electron distribution in the Boltzmann transport equation is *not* free to be an arbitrary $6-$D function. There is an acoustic phonon state associated with every crystal momentum that has a very small amount of energy. Therefore in a relaxation time that is similar to the acoustic scattering time, the distribution is driven efficiently

toward a state in which all crystal momentum states associated with the same energy distribution function are equally probable. We refer to this as relaxation to an ELASTICALLY CONSTRAINED QUASI-EQUILIBRIUM. Due to the semiclassical assumptions, the external field is relatively constant on the length scale associated with this ELASTIC RELAXATION TIME, and therefore the electron distribution function is confined to be a simple local field perturbation to an energy dependent distribution function. This is the ELASTICALLY CONSTRAINED TRANSPORT MODEL, which is expected to be valid in essentially the same semiclassical conditions as the Boltzmann transport equation requires.[1] The difference is that rather than an $6-$D electron state, the Elastically Constrained Transport model is associated with a $4-$D electron state, which diffuses at constant total energy as a result of the elastic relaxation time approximation, and is subject to a complex, purely inelastic, scattering operator. Thus it is much less computationally intensive to solve.

We further showed that the semiclassical model of electron transport in silicon can be further simplified into a model with a $3-$D electron state by analysing the boundary conditions the Elastically Constrained Transport model is typically subject to in a typical silicon device. All silicon devices are typically subject to terminals that supply electrons at a given chemical potential. The electrons that enter the device via the lowest chemical potential "drain" channel will generally remain in thermal equilibrium unless they scatter with non-drain electrons. The electrons which enter via a non-drain terminal will not generally be in thermal equilibrium, but due to the mechanics of the Elastically Constrained Transport model, the *highest energy* states of such a population of electrons will remain in thermal equilibrium with one another unless they scatter with non-drain electrons.

We argue— on the basis of an inelastic scattering operator which is broadband, and on the basis that diffusion at constant total energy erases the detailed patterns of energy-dependent particle creation and destruction caused by the inelastic scattering operator— that the energy distribution of electrons associated with electrons from a non-drain terminal that have not scattered with non-drain electrons can be approximated as being a STUNTED EQUILIBRIUM DISTRIBUTION defined as the maximum entropy distribution sub-

[1] Our derivation of the elastically constrained model assumes that the antisymmetric part of the distribution function is small, resulting in a particle flux at constant energy that is proportional to the gradient in the occupation rate at constant total energy. There is however, no fundamental reason we cannot attempt to remove this assumption and model the *saturation* of the flux that occurs when the antisymmetric distribution is large.

ject to a local particle density, energy density and maximum chemical potential. We argue that the energy distribution of electrons that result from electron–electron scattering between non-drain electrons can be approximated as a heated thermal equilibrium distribution.

For a device consisting of one non-drain "source" terminal and one drain terminal, we can define an ansatz for the energy distribution as the sum of a lattice temperature equilibrium, a stunted equilibrium, and a heated thermal equilibrium. We then derive the THREE QUASI-EQUILIBRIA MODEL of semiclassical electron transport in silicon based on this ansatz. In this model the electron state is defined by five $3-$D scalar fields, making it comparable in speed to other macroscopic models.

Essential to both these models is the precomputation and storage of transport parameters that are often $2-$D and $3-$D functions. In the short term, it is important that accessing the values in the "look-up tables" associated with these functions is well-optimized. In the long term, it would be ideal to find simple analytic forms that accurately recreate the look-up table data so that the models could be closed analytically.

The practical implications of this thesis are significant, as these models have the potential to have a large positive impact on the semiconductor industry. In order for this impact to materialize these models must be transformed into flexible, easy to use computer programs that can be used to model a wide range of devices. This will require significant additional research and development, both within the TCAD industry and in academia.

Aside from the real-world implications, it is expected that this thesis will also be of theoretical interest to many physicists. The models themselves add much new insight into a previously opaque regime of electron transport. Furthermore the theoretical framework used to derive these models shows one path of how new results in the field of non-equilibrium statistical mechanics can be obtained.

Appendix A

Bloch Waves and Bandstructure

A.1 Bloch Waves

We wish to understand the eigenstates available to a single carrier state in an otherwise ideal semiconductor. Before the carrier is added to this ideal semiconductor, each electron was in the ground state of the system and the total electron density would have the same symmetry as the crystal lattice. Therefore the interaction Hamiltonian between the carrier and the ideal semiconductor crystal must have the same symmetry as the underlying crystal lattice. As such, the interaction Hamiltonian must commute with a translation operator $\hat{\mathbf{R}}$, so long as the corresponding translation $\mathbf{R}$ is a real lattice vector; that is, if the primitive lattice vectors are $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$, $\mathbf{R}$ must belong to the following set:

$$\mathbf{R} \in \left\{ n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \;\middle|\; n_1, n_2, n_3 \in \mathbb{Z} \right\}. \tag{A.1}$$

It is a well-known result known as Bloch's Theorem [26] that the eigenstates of a particle described by a Hamiltonian which commutes with lattice translation operator has the form of a BLOCH WAVE or BLOCH STATE. A Bloch wave consists of a plane wave of wavevector $\mathbf{k}$, modified by a periodic complex valued "amplitude" function $u(\mathbf{r})$ which has the same periodicity as the primitive unit cell:

$$\psi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u(\mathbf{r}), \qquad \text{(A.2)}$$
$$\text{where} \quad u(\mathbf{r}) = u(\mathbf{r}+\mathbf{R}).$$

We would like to find a set of variables that uniquely identify our Bloch waves. An obvious choice is to try to make $\mathbf{k}$— the wavevector of the plane wave part— a unique identifier, however we can show unless we specify another degree of freedom this is an ambiguous choice. In order to see this, we make a small detour to describe an important concept: *the reciprocal lattice*.

Aside: The Reciprocal Lattice

Suppose we want to construct a plane wave that *always* fits an integral number of wavelengths between two point separated by *any* lattice vector $\mathbf{R}$. The *smallest* wavevector in each direction which can possibly satisfy this demand must represent a plane wave than can fit exactly 1 wavelength along one primitive lattice vector, while fitting exactly 0 wavelengths along the remaining two. Each such wavevector corresponds to a plane wave which travels in a direction perpendicular to two of the primitive lattice vectors— thus fitting exactly zero wavelengths along these directions— and has a magnitude such that the the plane wave is exactly one wavelength when projected onto the remaining primitive vector. We can construct such a wavevector for each dimension of a $3-$D crystal lattice, which gives rise to three PRIMITIVE WAVEVECTORS $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$:

$$\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\} = \left\{ \frac{2\pi\,(\mathbf{a}_2 \times \mathbf{a}_3)}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)} \;,\; \frac{2\pi\,(\mathbf{a}_3 \times \mathbf{a}_1)}{\mathbf{a}_2 \cdot (\mathbf{a}_3 \times \mathbf{a}_1)} \;,\; \frac{2\pi\,(\mathbf{a}_1 \times \mathbf{a}_2)}{\mathbf{a}_3 \cdot (\mathbf{a}_1 \times \mathbf{a}_2)} \right\}. \qquad \text{(A.3)}$$

We define $\mathbf{G}$ as a member of the *span* of these primitive wavevectors:

$$\mathbf{G} \in \left\{ n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3 \;\middle|\; n_1, n_2, n_3 \in \mathbb{Z} \right\}. \tag{A.4}$$

Since the span of primitive wavevectors defines a lattice in wavevector— or RECIPROCAL— space, we refer to $\mathbf{G}$ as a RECIPROCAL LATTICE VECTOR, and refer to the primitive wavevectors as defining a PRIMITIVE CELL OF THE RECIPROCAL LATTICE.

Any reciprocal lattice vector $\mathbf{G}$ must fit an integral number of wavelengths along each *real space* primitive lattice vector by construction, and vice-versa, every plane wave which fits an integral number of wavelengths along each along each primitive lattice vector is a reciprocal lattice vector $\mathbf{G}$. Since all lattice translation vectors $\mathbf{R}$ are made up of an integral number of primitive vectors, it obviously follows that a plane wave of wavevector $\mathbf{G}$ must also have an integral number of wavelengths fit between two points separated by any $\mathbf{R}$.

The reciprocal lattice is therefore significant primarily because an arbitrary function which is invariant under a real lattice translation— such as $u(\mathbf{r})$— can always be expressed as a unique superposition of plane waves corresponding to reciprocal lattice vectors.

We now return to the problem of demonstrating the ambiguity inherent in attempting to make $\mathbf{k}$— the wavevector for the plane wave part of the Bloch wave— a unique identifier of a Bloch wave/eigenstate. We note that since have defined the reciprocal lattice vectors $\mathbf{G}$ such that an integral number of plane wave wavelengths will fit between two points separated by a lattice vector $\mathbf{R}$, by definition:

$$e^{-i\mathbf{G}\cdot\mathbf{R}} = 1. \tag{A.5}$$

This identity allows us to show a Bloch wave with a plane wave part of wavevector $\mathbf{k}$, is also a Bloch wave with a plane wave part $\mathbf{k}+\mathbf{G}$. All we need to do is change the original periodic function $u(\mathbf{r})$ to different periodic function $u'(\mathbf{r}) = u(\mathbf{r})e^{-i\mathbf{G}\cdot\mathbf{r}}$:

$$\begin{aligned}
\psi_{\mathbf{k}}(\mathbf{r}) &= e^{i\mathbf{k}\cdot\mathbf{r}}u(\mathbf{r}) \\
&= \left(e^{i\mathbf{k}\cdot\mathbf{r}}e^{i\mathbf{G}\cdot\mathbf{r}}\right)\left(e^{-i\mathbf{G}\cdot\mathbf{r}}u(\mathbf{r})\right) \\
&= e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}u'(\mathbf{r}), \\
&= \psi_{\mathbf{k}+\mathbf{G}}(\mathbf{r}), \qquad \text{(A.6)} \\
\text{where} \quad u'(\mathbf{r}+\mathbf{R}) &= e^{-i\mathbf{G}\cdot\mathbf{r}}e^{-i\mathbf{G}\cdot\mathbf{R}}u(\mathbf{r}+\mathbf{R}) \\
&= u'(\mathbf{r}).
\end{aligned}$$

We therefore reach the conclusion that any Bloch wave that is a modulated plane wave of wavevector $\mathbf{k}$ is simultaneously a modulated plane wave of wavevector $\mathbf{k}+\mathbf{G}$, for any reciprocal lattice vector $\mathbf{G}$. This ambiguity can be understood as arising because the complex modulation function $u(\mathbf{r})$ is permitted to contain an undesignated plane wave of any reciprocal lattice wavevector $\mathbf{G}$; as such, the wavevector of the designated plane wave part can only be established modulo $\mathbf{G}$.

There are a number of schemes for rectifying this ambiguity, but in this thesis we adapt the REDUCED ZONE scheme. In the reduced zone scheme, we restrict the designated plane wave part to be the *smallest magnitude wavevector* of the entire set of wavevectors which can be associated with a Bloch wave. The result is that the wavevector $\mathbf{k}$ associated with a Bloch wave is restricted to lie within a particular choice of primitive cell for the reciprocal lattice, known as the FIRST BRILLIOUN ZONE[1], which is made of the locus of reciprocal space points that are closer to the $\mathbf{k}=0$ reciprocal lattice point than any other reciprocal lattice point.

The restriction of $\mathbf{k}$ to the first Brillioun zone means there will be in general an infinite number of different energy Bloch eigenstates associated with each wavevector $\mathbf{k}$. To demonstrate this, imagine we approach the "empty lattice" limit where the strength of periodic potential approaches zero. The energy eigenstates of Bloch waves must become the energy eigenstates of free electron plane waves with no restrictions on

[1]Or simply, the BRILLOUIN ZONE.

wavevector. As such, for each Bloch wave which has a wavevector $\mathbf{k}$ that falls within the first Brillouin zone, there must be an infinite number of possible modulation function $u(\mathbf{r}) = e^{i\mathbf{G}\cdot\mathbf{r}}$ defined for every possible reciprocal lattice vector $\mathbf{G}$, each of which is an energy eigenstate. Therefore, we add a second variable ν, an integer called the BAND INDEX, which enumerates all the Bloch functions associated with $\mathbf{k}$ in the order of increasing energy eigenvalue. Using these two variables, we can now uniquely index a Bloch wave eigenstate.

$$\psi_{\mathbf{k}\nu}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}\nu}(\mathbf{r}), \tag{A.7}$$
$$\text{where} \quad \mathbf{k} \in \left\{ \mathbf{k} \,\middle|\, |\mathbf{k}| \leq |\mathbf{k}+\mathbf{G}| \quad \forall \mathbf{G} \right\}.$$

A.2 The Bandstructure

The energy of all the Bloch wave eigenstates corresponding to a single carrier in an idealized crystal— $\varepsilon_{\mathbf{k}\nu}$— is a function commonly known as the BANDSTRUCTURE. The calculation of the bandstructure from first principles is a non-trivial, active field of research. We will not attempt to illuminate this field of research as the empirical pseudopotential calculation of bandstructure is sufficient for the ends of this thesis, and this is already discussed in Section 2.4 of the Background chapter and in Appendix C.

Instead we will briefly discuss the considerations that result in the near ubiquitous presentation of bandstructures as plots over a $1-$D domain, as this can be a source of confusion. From the discussion so far, one would expect that the plot of eigenvalues will inherently require the $3-$D domain of the entire Brillouin zone. The common $1-$D presentation of bandstructure results *firstly* from a general information preserving, symmetry-based reduction of the $3-$D Brillioun zone to a much smaller $3-$D subvolume; followed *secondly* by much less general argument that the most important bandstructure information often occurs along certain high symmetry lines.

The first argument stems from noticing that the derivation of Bloch's Thereom only required the *translation* symmetry properties of the crystal lattice, while the crystal lattice always has additional *point* symmetries defined by certain reflections, inversions and rotations. These additional symmetries force the energy eigenstates of symmetry

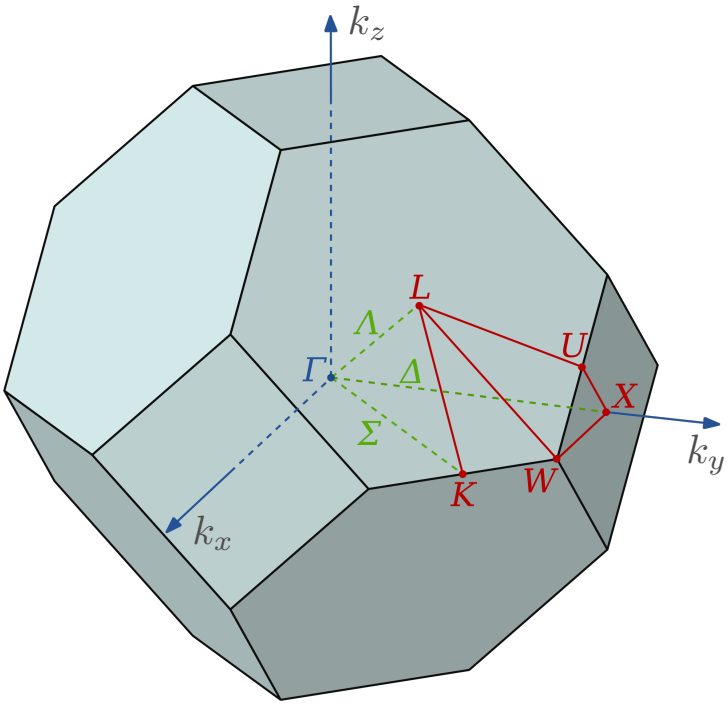


Figure A.1: (***Public Domain***) The Brillouin zone (grey) and irreducible Brillouin zone (green and red) of a face-centred cubic (FCC) crystal lattice. It is worth noting that while the reciprocal lattice of a simple cubic (SC) crystal is also SC, the reciprocal lattice of a FCC crystal transforms to a base-centred cubic (BCC) lattice in reciprocal space. The high symmetry points and edges of the irreducible Brillouin zone are labeled in a manner fitting with standard convention. Notably, the Brillouin zone of silicon is this shape.

related points in the Brillouin zone to be equal. As such the bandstructure is only a unique function of a small wedge— often only $\sim 5\%$ by volume— of the Brillouin zone known as the IRREDUCIBLE BRILLOUIN ZONE, which can be unfolded using the symmetry operations to define the energy eigenvalues of states at any point of the Brillouin zone. An example showing the irreducible wedge of the face-centred cubic Brillouin zone is shown in Fig. A.1.

The reduction by symmetry does not reduce the number of dimensions of the Brillouin zone. If we wish to have a plot of eigenvalues which can be communicated quantitatively on a $2-$D page, we can only plot the eigenstates $1-$D path through the irreducible wedge. An arbitrary path through the irreducible Brillouin zone will generally yield insufficient information to approximately reconstruct the $3-$D Brillouin zone. The primary problem is that the derivatives of the bandstructure in directions *other* than the direction of the path is completely unknown. However, if we carefully select a path

along a line of high symmetry, the gradient of the bandstructure— $\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}$— may have restrictions enforced by its symmetry which allow us to we can infer additional information it from the directional derivative along the path.

The most common and important example of the symmetry restriction of the bandstructure gradient occurs due to reflection symmetry. If there is a plane of reflection symmetry in reciprocal space, $\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}$— if it is well-defined— must lie in that plane, since the values of bandstructure on either side of the plane must be equal. If two planes of reflection meet at an edge, then $\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}$— if it is well-defined— must be collinear to that edge, since it must lie in both planes which form the edge. By the same argument, if three planes of reflection symmetry meet at a vertex, then the vertex must be a local extrema of the bandstructure, and $\nabla_{\mathbf{k}}\varepsilon_{\mathbf{k}\nu}$— if it is well-defined— must equal zero.

We note that if these planes of reflection symmetry exist, they *must* be certain faces of the irreducible wedge, as if they were inside the irreducible wedge, the wedge could be further reduced by symmetry. As an obvious corollary, the edges and vertices where these reflection planes meet will also be certain edges and vertices of the irreducible wedge. A simple $2-$D example of an irreducible wedge which has edges with high reflection symmetry is shown in Fig. A.2.

A $1-$D path between the high symmetry vertices of the irreducible Brillioun zone via the high symmetry edges of the wedge, will therefore contain an unusual concentration of the local extrema in the bandstructure, but it is not clear that these $1-$D plots yield all the information required to model a given phenomenon. In traditional transport calculations, the critical information the bandstructure contains is that required to recreate a harmonic approximation to the *absolute* maximum of the valence band and *absolute* minimum of the conduction band. This is the critical information in traditional transport modelling, because such a recreation will yield the effective mass tensor of the lowest energy carriers of either type, and in approximately thermal distributions of carriers characteristic of traditional transport models, these low energy carriers are overwhelmingly more common than higher energy carriers with different effective mass tensors. It is by no means clear that this $1-$D plot will contain this information, since there is nothing argued so far to stop the *global* extrema from occurring somewhere outside the high symmetry path.

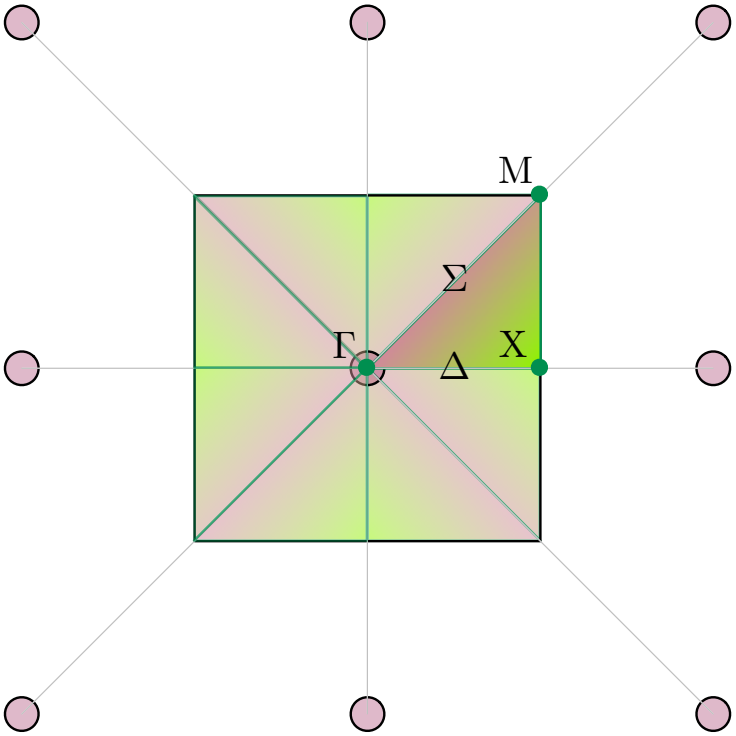


Figure A.2: A small portion of the reciprocal lattice of a $2-$D simple cubic (SC) lattice. The Brillouin zone is shown by the square, and the darkened triangle is an irreducible wedge. The gray lines are lines of reflection symmetry. The entire Brillouin zone can be reproduced by reflecting the irreducible wedge. The high symmetry points and edges are labeled in a manner fitting with standard convention. The green-red colouring schematically indicates the dependance of the states energy eigenvalue on $\mathbf{k}$ for a single band of a fictitious material.

However, in a manner consistent with bandstructures based on highly simplified Hamiltonians, extrema in the bandstructure across many materials *do* tend to occur most frequently when forced to by symmetry restriction at a high reflection symmetry vertex, or when the gradient is forced by symmetry to be collinear to a high symmetry edge. If we "get lucky", and the valence band maximum and conduction band minimum occur at high symmetry vertices, then such absolute extrema will be sufficiently characterized to create a harmonic approximation. This is because if extrema occur at the intersection vertex of three reflection planes, they simultaneously occur at the intersection of three reflection edges, and a plot of the bandstructure along a path following these high symmetry edges will characterize the rate of change of the bandstructure gradients in three dimensions. Additionally when studying the optical properties of semiconductors, we have similar fortune and find that $1-$D plots often contain the critical information required for understanding many optical phenomenon, as these are typically dominated by local and global extrema in the bandstructure.

Indeed, the $1-$D bandstructure plot along high symmetry edges— the style of bandstructure presentation so ubiquitous in the literature— often contains so much useful information it is easy to slip into believing it is equivalent to the full bandstructure. But in far-from-equilibrium transport carriers are often distributed throughout the whole Brillouin zone, and not simply concentrated around a few high symmetry vertices and edges. Therefore, when we refer to full bandstructure in *this* thesis, we are always referring to a $3-$D full bandstructure calculation, and not the $1-$D plot often used in its place.

Appendix B

Newton's Law for Bloch States

B.1 The Expected Value of the Lattice Translation Operator

We follow the derivation of Kroemer [92]. We wish to examine the effect of a uniform force $\mathbf{F}$, on the single carrier Bloch eigenstates $|\mathbf{k}\nu\rangle$ of an ideal crystal. The total single carrier Hamiltonian $\hat{H}$, is the Hamiltonian of a single carrier in an ideal crystal $\hat{H}_0$, plus the position-dependent potential energy associated with the uniform force $\hat{V}_{\text{external}} = \mathbf{F} \cdot \hat{\mathbf{r}}$:

$$\hat{H} = \hat{H}_0 + \mathbf{F} \cdot \hat{\mathbf{r}}. \tag{B.1}$$

We wish to find a *non-perturbative* solution to this problem. To do so, we take the slightly non-obvious step of examining the rate of change of the expected value of a lattice vector translation operator $\hat{\mathbf{R}}$ associated with a lattice vector $\mathbf{R}$. It is a basic result of quantum mechanics that the rate of change of the expected value of an operator $\hat{A}$ is related to the Hamiltonian $\hat{H}$ as follows:

$$\frac{\partial}{\partial t}\left\langle \hat{A} \right\rangle = \frac{i}{\hbar}\left\langle \left[\hat{H}, \hat{A}\right] \right\rangle. \tag{B.2}$$

Therefore, the rate of change of the expected value of the translation operator $\hat{\mathbf{R}}$ is given by eq. (B.3):

$$\frac{\partial}{\partial t}\left\langle \hat{\mathbf{R}} \right\rangle = \frac{i}{\hbar}\left\langle \left[\hat{H}_0, \hat{\mathbf{R}}\right] + \mathbf{F}\cdot\left[\hat{\mathbf{r}}, \hat{\mathbf{R}}\right] \right\rangle. \tag{B.3}$$

The lattice vector translation operator commutes with the unperturbed Hamitonian, since the infinite, ideal crystal is identical before and after such a shift:

$$\left[\hat{H}_0, \hat{\mathbf{R}}\right] = 0. \tag{B.4}$$

On the other hand, to find the commutator $\left[\hat{\mathbf{r}}, \hat{\mathbf{R}}\right]$ we examine the effect of the operator combination $\hat{\mathbf{r}}\hat{\mathbf{R}}$, on an arbitrary wavefunction $|\psi\rangle$. Supposing this wavefunction can be expressed in the position basis as $|\psi\rangle = \int_{\mathbb{R}^3} \psi(\mathbf{r})\,|\mathbf{r}\rangle\,\mathrm{d}\mathbf{r}$, we can derive the following identity:

$$\begin{aligned}
\hat{\mathbf{r}}\hat{\mathbf{R}}\,|\psi\rangle &= \hat{\mathbf{r}}\hat{\mathbf{R}}\int_{\mathbb{R}^3}\psi(\mathbf{r})\,|\mathbf{r}\rangle\,\mathrm{d}\mathbf{r} && \text{by definition of } |\psi\rangle,\\
&= \hat{\mathbf{r}}\int_{\mathbb{R}^3}\psi(\mathbf{r})\,|\mathbf{r}+\mathbf{R}\rangle\,\mathrm{d}\mathbf{r} && \text{by definition of } \hat{\mathbf{R}},\\
&= \int_{\mathbb{R}^3}(\mathbf{r}+\mathbf{R})\psi(\mathbf{r})\,|\mathbf{r}+\mathbf{R}\rangle\,\mathrm{d}\mathbf{r} && \text{by definition of } \hat{\mathbf{r}},\\
&= \hat{\mathbf{R}}\int_{\mathbb{R}^3}(\mathbf{r}+\mathbf{R})\psi(\mathbf{r})\,|\mathbf{r}\rangle\,\mathrm{d}\mathbf{r} && \text{by definition of } \hat{\mathbf{R}},\\
&= \hat{\mathbf{R}}\hat{\mathbf{r}}\,|\psi\rangle + \mathbf{R}\hat{\mathbf{R}}\,|\psi\rangle && \text{by definition of } \hat{\mathbf{r}} \text{ and } |\psi\rangle.
\end{aligned} \tag{B.5}$$

From this identity, it is clear that the commutator $\left[\hat{\mathbf{r}}, \hat{\mathbf{R}}\right]$ is defined as follows:

$$\left[\hat{\mathbf{r}}, \hat{\mathbf{R}}\right] = \mathbf{R}\hat{\mathbf{R}}. \tag{B.6}$$

By substituting eq. (B.4) and eq. (B.6) into eq. (B.3), we arrive at the equation of motion for the expected value of the translation operator:

$$\frac{\partial}{\partial t}\left\langle \hat{\mathbf{R}} \right\rangle = \frac{i}{\hbar}\mathbf{F}\cdot\mathbf{R}\left\langle \hat{\mathbf{R}} \right\rangle. \tag{B.7}$$

The interesting thing about this equation is the expected value of given translation op-

erator can simply be viewed here as some unknown scalar function of time and lattice vector $\mathbf{R}$. The solutions to such an equation are therefore a standard result of first-order differential equation theory, and have the following form:

$$\left\langle \hat{\mathbf{R}} \right\rangle = R_0 e^{i\boldsymbol{\xi}(t)\cdot\mathbf{R}}, \tag{B.8a}$$

$$\text{where} \qquad \frac{\partial \boldsymbol{\xi}}{\partial t} = \frac{1}{\hbar}\mathbf{F}. \tag{B.8b}$$

The arbitrary constants of this equation are $\boldsymbol{\xi}_0$— the value of the vector $\boldsymbol{\xi}$ at time 0— and R_0— the magnitude of the expected value of $\hat{\mathbf{R}}$. Since the differential equation is linear, any linear combination of solutions of this form is also a solution to the differential equation.

B.2 The Equation of Motion of a Pure Bloch State

We note that there is an expected value of $\hat{\mathbf{R}}$ associated with *any* arbitrary mixed quantum state, and the corresponding time dependence of this expected value must be governed by eq. (B.7). To begin with, we examine the special case where the arbitrary mixed quantum state is a pure Bloch state $|\mathbf{k}\nu\rangle$ at time $t = 0$. We note that, for a Bloch wave $|\mathbf{k}\nu\rangle$, the only difference between the wavefunction at a point $\mathbf{r}$ and at a point $\mathbf{r} + \mathbf{R}$ is due to the plane-wave part, which will differ by a factor $e^{i\mathbf{k}\cdot\mathbf{R}}$, for any position $\mathbf{r}$. This implies that $|\mathbf{k}\nu\rangle$ *is an eigenstate* of the operator $\hat{\mathbf{R}}$, with associated eigenvalue $e^{i\mathbf{k}\cdot\mathbf{R}}$:

$$\hat{\mathbf{R}} \left|\mathbf{k}\nu\right\rangle = e^{i\mathbf{k}\cdot\mathbf{R}} \left|\mathbf{k}\nu\right\rangle. \tag{B.9}$$

Therefore the expected value of the translation operator for an over an arbitrary Bloch state $|\mathbf{k}\nu\rangle$ is as follows:

$$\begin{aligned} \left\langle \hat{\mathbf{R}} \right\rangle^{\mathbf{k}\nu} &= \left\langle \mathbf{k}\nu \right| \hat{\mathbf{R}} \left| \mathbf{k}\nu \right\rangle \\ &= e^{i\mathbf{k}\cdot\mathbf{R}} \end{aligned} \tag{B.10}$$

The time variation of this expected value in the presence of an external field $\mathbf{F}$ must be

given by some linear combination of the solutions in eq. (B.8). The appropriate solution is $R_0 = 1$ and $\boldsymbol{\xi}(0) = \mathbf{k}$. By the definition of $\boldsymbol{\xi}(t)$, eq. (B.8) implies NEWTONS LAW FOR BLOCH STATES must hold for $\mathbf{k}$:

$$\mathbf{F} = \hbar \frac{\partial \mathbf{k}}{\partial t}. \tag{B.11}$$

B.3 The Equation of Motion an Arbitrary Mixed State

It is instructive to also examine the expected value of the translation operator associated with an arbitrary mixed state ρ at time $t = 0$. We suppose ρ is defined by $\rho = \sum_i p_i \left|\psi_i\right\rangle$, where p_i is the real classical probability of a being in the arbitrary state $\left|\psi_i\right\rangle$, which is itself a superposition of Bloch states $\left|\psi_i\right\rangle = \sum_\nu \int_{\mathbb{R}^3} \psi^i_{\mathbf{k}\nu} \left|\mathbf{k}\nu\right\rangle \mathrm{d}\mathbf{k}$. The expected value of the translation operator associated with the arbitrary mixed state ρ is as follows:

$$\begin{aligned}
\left\langle \hat{\mathbf{R}} \right\rangle^{\rho} &= \sum_i p_i \left\langle \psi_i \right| \hat{\mathbf{R}} \left| \psi_i \right\rangle \\
&= \sum_{i,\nu,\nu'} \int_{\mathrm{BZ}} \int_{\mathrm{BZ}} p_i \psi^{i*}_{\mathbf{k}\nu} \psi^i_{\mathbf{k}'\nu'} \left\langle \mathbf{k}\nu \right| \hat{\mathbf{R}} \left| \mathbf{k}'\nu' \right\rangle \mathrm{d}\mathbf{k}\mathrm{d}\mathbf{k}' \\
&= \sum_{i,\nu} \int_{\mathrm{BZ}} p_i \left|\psi^i_{\mathbf{k}\nu}\right|^2 \left\langle \mathbf{k}\nu \right| \hat{\mathbf{R}} \left| \mathbf{k}\nu \right\rangle \mathrm{d}\mathbf{k} \\
&= \sum_{\nu} \int_{\mathrm{BZ}} p_{\mathbf{k}\nu} \left\langle \hat{\mathbf{R}} \right\rangle^{\mathbf{k}\nu} \mathrm{d}\mathbf{k}, \qquad \text{(B.12)} \\
\text{where} \quad p_{\mathbf{k}\nu} &= \sum_i p_i |\psi^i_{\mathbf{k}\nu}|^2
\end{aligned}$$

The point of eq. (B.12) is to show no matter what quantum mechanical state a carrier is in, the expected value of the translation operator is the same as the expected value of the translation operator for a *purely classical* statistical distribution of *pure* Bloch states. The time evolution of the expected value of the translation operator in turn implies that an arbitrary quantum superposition of Bloch states must evolve in time as a classical statistical distribution of Bloch states, where the crystal momenta of each follows Newtons law for Bloch states, eq. (B.11). This is one of the important facts that allows us to treat the local electron state as being a *classical statistical distribution* of *pure* Bloch states,

as opposed to needing to treat the local electron state as an arbitrary mixed state ρ. Since we followed a non-perturbative treatment, this fact is true *no matter how large the uniform external force is*.

The only novel effect that very large fields can have on Bloch states is to increase the rate of transition from one band ν to another ν' *at the same crystal momentum value* to the point where transitions from the valence band to the conduction band become a non-negligible source of carriers. This process is known as ZENER TUNNELLING. We will not investigate it in detail in this thesis, since it requires that the change in potential energy across a decoherence length ($\sim 2-3$ nm) due to the external force is larger than the band gap ($\sim 1eV$). Since this condition will not generally be true in highly inhomogeneous transport, this is a special subregime that is not part of the "kernel" model of highly inhomogeneous transport that this thesis attempts to describe.

Appendix C

The Pseudopotential Band Structure

C.1 A Local Pseudopotential in Terms of Reciprocal Lattice Vectors

Let us assume that there exists an accurate, universal, slowly-varying local pseudopotential $V_{\text{pseudo}}(\mathbf{r})$. The form of that local pseudopotential will often be restricted by symmetries related to the positioning of the multiple basis ions *within a unit cell*. To incorporate this symmetry, we view the crystal as the superposition of n centred, *single basis ion* crystals, where n is the number of basis ions in each unit cell. If $\boldsymbol{\tau}_j$ is the basis vector of the jth basis ion, the total pseudopotential $V_{\text{pseudo}}(\mathbf{r})$ is expressed in terms of centred, single basis component pseudopotentials $V^j_{\text{pseudo}}(\mathbf{r})$ as follows:

$$V_{\text{pseudo}}(\mathbf{r}) = \sum_j^n V^j_{\text{pseudo}}(\mathbf{r} - \boldsymbol{\tau}_j). \tag{C.1}$$

We wish to rewrite both the total pseudopotential $V_{\text{pseudo}}(\mathbf{r})$ and the component pseudopotentials $V^j_{\text{pseudo}}(\mathbf{r})$ as Fourier series, on the basis that they are both periodic with respect to lattice translation vectors $\mathbf{R}$:

$$V_{\text{pseudo}}(\mathbf{r}) = \sum_{\mathbf{G}} \tilde{V}_{\text{pseudo}}(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{C.2a}$$

$$\text{and} \qquad V^j_{\text{pseudo}}(\mathbf{r}) = \sum_{\mathbf{G}} \tilde{V}^j_{\text{pseudo}}(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}}. \tag{C.2b}$$

If we multiply either of these equations by $e^{-i\mathbf{G}'\cdot\mathbf{r}}$ and integrate over a unit cell Ω, all terms on the RHS will integrate to zero except where $\mathbf{G} = \mathbf{G}'$. All other terms vanish because $\mathbf{G}' - \mathbf{G}$ is a non-zero reciprocal lattice vector, and non-zero reciprocal lattice vector plane waves integrate to zero over a unit cell. Thus the total and component Fourier series coefficients, $\tilde{V}_{\text{pseudo}}(\mathbf{G})$ and $\tilde{V}^j_{\text{pseudo}}(\mathbf{G})$ respectively, are given by the following expressions:

$$\tilde{V}_{\text{pseudo}}(\mathbf{G}) = \frac{1}{\Omega}\int_{\Omega} V_{\text{pseudo}}(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}, \tag{C.3a}$$

$$\text{and} \qquad \tilde{V}^j_{\text{pseudo}}(\mathbf{G}) = \frac{1}{\Omega}\int_{\Omega} V^j_{\text{pseudo}}(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}. \tag{C.3b}$$

We wish to express the Fourier coefficients of the total pseudopotential $\tilde{V}_{\text{pseudo}}(\mathbf{G})$ in terms of the Fourier coefficients of the component pseudopotential $\tilde{V}^j_{\text{pseudo}}(\mathbf{G})$. We begin by substituting eq. (C.1) into eq. (C.3a):

$$\begin{aligned}
\tilde{V}_{\text{pseudo}}(\mathbf{G}) &= \frac{1}{\Omega}\int_{\Omega} V_{\text{pseudo}}(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} \\
&= \frac{1}{\Omega}\int_{\Omega} \sum_{j=1}^{n} V^j_{\text{pseudo}}(\mathbf{r} - \boldsymbol{\tau}_j) e^{-i\mathbf{G}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} \\
&= \sum_{j=1}^{n} \int_{\Omega} V^j_{\text{pseudo}}(\mathbf{r} - \boldsymbol{\tau}_j) e^{-i\mathbf{G}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}.
\end{aligned} \tag{C.4}$$

For each j, we can substitute the integral over $\mathbf{r}$ for an integral over $\mathbf{r}' = \mathbf{r} - \boldsymbol{\tau}_j$. Due to the periodicity of the function, Ω can be kept the same. By substituting eq. (C.3b), we can express the total pseudopotential Fourier coefficients $\tilde{V}_{\text{pseudo}}(\mathbf{G})$ in terms of the centred, single basis component pseudopotential Fourier coefficients $\tilde{V}^j_{\text{pseudo}}(\mathbf{G})$:

$$\begin{aligned}\tilde{V}_{\text{pseudo}}(\mathbf{G}) &= \sum_{j=1}^{n} \int_{\Omega} V^{j}_{\text{pseudo}}(\mathbf{r}') e^{-i\mathbf{G}\cdot\mathbf{r}'} e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_j} \mathrm{d}\mathbf{r}' \\ &= \sum_{j=1}^{n} e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_j} \int_{\Omega} V^{j}_{\text{pseudo}}(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} \mathrm{d}\mathbf{r} \\ &= \sum_{j=1}^{n} e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_j} \tilde{V}^{j}_{\text{pseudo}}(\mathbf{G}). \end{aligned} \tag{C.5}$$

This expression is useful as it allows us to easily determine if the distribution of basis ions within a unit cell forces particular Fourier series coefficients of the total pseudopotential to vanish due to symmetry.

C.2 Zinc Blende and Diamond Reciprocal Lattices

We now wish to write down the smallest reciprocal lattice vectors for two technologically important lattice types: the ZINC BLENDE and DIAMOND lattices. These lattices are superpositions of face-centred cubic (FCC) single atom crystals. The reciprocal lattice of a FCC of lattice parameter a is a base-centered cubic (BCC) with a lattice parameter $\frac{4\pi}{a}$. Thus the nearest neighbour reciprocal lattice vectors— the lowest spatial frequencies— are as follows, in units of $\frac{2\pi}{a}$:

$$
\begin{aligned}
\mathbf{G}_0 &= (0,0,0) && \text{1 centre lattice point}\\
\mathbf{G}_3 &= (\pm1,\pm1,\pm1) && \text{8 vertices of unit cell}\\
\mathbf{G}_4 &= (\pm2,0,0),(0,\pm2,0),(0,0,\pm2) && \text{6 centres through faces}\\
\mathbf{G}_8 &= (\pm2,\pm2,0),(\pm2,0,\pm2),(0,\pm2,\pm2) && \text{12 centres through edges}\\
\mathbf{G}_{11} &= (\pm3,\pm1,\pm1),(\pm1,\pm3,\pm1),(\pm1,\pm1,\pm3) && \text{24 vertices of adjacent cells}\\
\mathbf{G}_{12} &= (\pm2,\pm2,\pm2) && \text{8 centres through edges}\\
\mathbf{G}_{16} &= (\pm4,0,0),(0,\pm4,0),(0,0,\pm4) && \text{6 centres through faces}\\
\mathbf{G}_{19} &= (\pm1,\pm3,\pm3),(\pm3,\pm1,\pm3),(\pm1,\pm3,\pm3) && \text{24 vertices...}\\
\mathbf{G}_{20} &= (\pm2,\pm4,0),(\pm4,\pm2,0),\ldots && \text{24 centres...}\\
\mathbf{G}_{24} &= (\pm2,\pm2,\pm4),(\pm2,\pm4,\pm2),(\pm4,\pm2,\pm2) && \text{24 centres...}\\
&\vdots && \qquad\qquad (C.6)
\end{aligned}
$$

Shown above are the reciprocal lattice vectors for the 10 sets of nearest neighbours to the gamma point, indexed according to the *square of their distance from the gamma point in units of* $\frac{2\pi}{a}$. These 10 sets of nearest neighbours define a total of 137 unique plane waves. Since the total local pseudopotential is slowly varying, we seek to approximate both the local pseudopotential and the pseudoeigenstates of the corresponding pseudo-Hamiltonian in terms of the plane waves associated with these relatively small reciprocal lattice vectors.

The first point to notice is that the pseudopotential associated with $\mathbf{G}_0$ is fairly arbitary, since it adds a uniform energy shift to every state. The second point to notice is that since the total pseudopotential must have the point symmetry of the crystal, which *typically* means that the Fourier coefficients of the total pseudopotential can only be a function of the *size* of the reciprocal lattice vector $|\mathbf{G}|$, and cannot be a function of its direction, since *usually* small reciprocal lattice vectors of the same magnitude differ only by a point symmetry operation.[1] The third point to notice is that, in the case of diamond, the Fourier coefficients associated with some of the reciprocal lattice vectors

[1]This is not guaranteed. It is perfectly possible to have two valid reciprocal lattice vectors of the same length that do not differ by a point symmetry operation. It is simply that at small reciprocal lattice vectors, this is unlikely.

must vanish.

To understand this third point, we note that with a two-atom basis crystal, the symmetry restrictions are most obvious if we place the lattice point at the midpoint between the two interlaced crystals. In the case of zinc blende and diamond, this means that our basis vectors are $\boldsymbol{\tau}_1 = -\boldsymbol{\tau}_2 = \boldsymbol{\tau} = \frac{a}{8}(1,1,1)$. Therefore the Fourier coefficients of the total pseudopotential can be expressed in the following manner:

$$\begin{aligned}
\tilde{V}_{\text{pseudo}}(\mathbf{G}) &= \sum_{j=1}^{n} e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_j}\tilde{V}^{j}_{\text{pseudo}}(\mathbf{G}) \\
&= \tilde{V}^{1}_{\text{pseudo}}(\mathbf{G})e^{-i\mathbf{G}\cdot\boldsymbol{\tau}} + \tilde{V}^{2}_{\text{pseudo}}(\mathbf{G})e^{i\mathbf{G}\cdot\boldsymbol{\tau}} \\
&= \cos(\mathbf{G}\cdot\boldsymbol{\tau})\left(\tilde{V}^{1}_{\text{pseudo}}(\mathbf{G}) + \tilde{V}^{2}_{\text{pseudo}}(\mathbf{G})\right) - i\sin(\mathbf{G}\cdot\boldsymbol{\tau})\left(\tilde{V}^{1}_{\text{pseudo}}(\mathbf{G}) - \tilde{V}^{2}_{\text{pseudo}}(\mathbf{G})\right) \\
&= \cos(\mathbf{G}\cdot\boldsymbol{\tau})\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}) + i\sin(\mathbf{G}\cdot\boldsymbol{\tau})\tilde{V}^{\text{anti}}_{\text{pseudo}}(\mathbf{G}),
\end{aligned} \tag{C.7}$$

where we have $\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G})$ and $\tilde{V}^{\text{anti}}_{\text{pseudo}}(\mathbf{G})$ defined as follows:

$$\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}) = \tilde{V}^{1}_{\text{pseudo}}(\mathbf{G}) + \tilde{V}^{2}_{\text{pseudo}}(\mathbf{G}) \tag{C.8a}$$

$$\tilde{V}^{\text{anti}}_{\text{pseudo}}(\mathbf{G}) = \tilde{V}^{1}_{\text{pseudo}}(\mathbf{G}) - \tilde{V}^{2}_{\text{pseudo}}(\mathbf{G}). \tag{C.8b}$$

In a diamond lattice— such as that of silicon— the antisymmetric pseudopotential coefficient $\tilde{V}^{\text{anti}}_{\text{pseudo}}$ is zero, since $\tilde{V}^{1}_{\text{pseudo}}(\mathbf{G}) = \tilde{V}^{2}_{\text{pseudo}}(\mathbf{G})$. According to eq. (C.7), this implies that the Fourier series coefficient vanishes for any for at any reciprocal lattice vector where $\mathbf{G}\cdot\boldsymbol{\tau} = \frac{2n+1}{2\pi}$, which is true for any reciprocal lattice vector $\mathbf{G}$ in which the sum of vector coordinates $G_x + G_y + G_z$— in units of $\frac{2\pi}{a}$— is equal to $4m+2$ where m is some integer. This is clearly true for $\mathbf{G}_4$ where the sum of coordinates is ± 2, and for $\mathbf{G}_{12}$ and $\mathbf{G}_{20}$, where the sum of coordinates is ± 2 or ± 6. Accordingly, the Fourier coefficients at these reciprocal lattice vectors vanish. Putting these results together, we can express a slowly varying pseudopotential in terms of Fourier coefficients for $|\mathbf{G}|^2 \in \{3, 8, 11, 16, 19, 24, ...\}$, where the coefficients should be functions only[2] of $|\mathbf{G}|$ and should eventually become rapidly decreasing as $|\mathbf{G}|$ increases.

[2]Again, unless there are two reciprocal lattice vectors of the same magnitude that cannot be mapped onto one another by an octahedral point symmetry operation.

Given these considerations, it is not too surprising that according to Chelikowsky and Cohen [39], there is a fairly accurate local pseudopotential for silicon that can be characterized by the following Fourier coefficients:

$$\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}) = \begin{cases} -0.2241\,\text{Rydbergs} & \text{for } |\mathbf{G}|^2 = 3 \\ 0.0551\,\text{Rydbergs} & \text{for } |\mathbf{G}|^2 = 8 \\ 0.0724\,\text{Rydbergs} & \text{for } |\mathbf{G}|^2 = 11 \\ 0 & \text{otherwise.} \end{cases} \tag{C.9}$$

C.3 Solving the Pseudopotential Hamiltonian

We will now attempt to solve the time-independent Schrödinger equation for pseudoeigenstates:

$$\varepsilon_{\mathbf{k}\nu}\varphi^{\text{pseudo}}_{\mathbf{k}\nu} = \hat{H}_{\text{pseudo}}\varphi^{\text{pseudo}}_{\mathbf{k}\nu}$$

Here $\varphi^{\text{pseudo}}_{\mathbf{k}\nu}$ is the pseudoeigenstate associated with a single crystal momentum $\mathbf{k}$ in the Brillouin zone that is indexed by ν, and $\hat{H}_{\text{pseudo}}$ is the pseudo-Hamiltonian which is defined as the sum of the kinetic energy operator and the pseudopotential operator. Having explicitly defined the pseudopotential in the position basis, we can now write down an explicit expression for this equation in the position basis:

$$\varepsilon_{\mathbf{k}\nu}\varphi^{\text{pseudo}}_{\mathbf{k}\nu}(\mathbf{r}) = \left(-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}} + \sum_{\mathbf{G}} \cos(\mathbf{G}\cdot\boldsymbol{\tau})\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G})e^{\,i\mathbf{G}\cdot\mathbf{r}} \right) \varphi^{\text{pseudo}}_{\mathbf{k}\nu}(\mathbf{r}). \tag{C.10}$$

Since the pseudoeigenstates are Bloch waves, we can express them in the following manner:

$$\begin{aligned} \varphi^{\text{pseudo}}_{\mathbf{k}\nu}(\mathbf{r}) &= e^{\,i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}\nu}(\mathbf{r}) \\ &= e^{\,i\mathbf{k}\cdot\mathbf{r}}\sum_{\mathbf{G}} \tilde{u}_{\mathbf{k}\nu}(\mathbf{G})e^{\,i\mathbf{G}\cdot\mathbf{r}}. \end{aligned} \tag{C.11}$$

Thus, by substituting eq. (C.11) into eq. (C.10), we have the following result:

$$0 = \left(-\frac{\hbar^2}{2m_e}\nabla_{\mathbf{r}}^2 + \sum_{\mathbf{G}} \cos(\mathbf{G}\cdot\boldsymbol{\tau})\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}) e^{\,i\mathbf{G}\cdot\mathbf{r}} - \varepsilon_{\mathbf{k}\nu} \right) \sum_{\mathbf{G}'} \tilde{u}_{\mathbf{k}\nu}(\mathbf{G}') e^{\,i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}}. \tag{C.12}$$

We can expand now expand the brackets, and rewrite the gradient operator in terms of its eigenvalues, since $\nabla_{\mathbf{r}} e^{\,i\mathbf{k}\cdot\mathbf{r}} = i\mathbf{k}e^{\,i\mathbf{k}\cdot\mathbf{r}}$:

$$0 = \sum_{\mathbf{G}'} \tilde{u}_{\mathbf{k}\nu}(\mathbf{G}') \left(\frac{\hbar^2(\mathbf{k}+\mathbf{G}')^2}{2m_e} - \varepsilon_{\mathbf{k}\nu} \right) e^{\,i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}} + \sum_{\mathbf{G}}{}' \sum_{\mathbf{G}} \cos(\mathbf{G}\cdot\boldsymbol{\tau})\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}) e^{\,i(\mathbf{k}+\mathbf{G}'+\mathbf{G})\cdot\mathbf{r}}. \tag{C.13}$$

We notice that we can eliminate one of the summations in each term by multiplying both sides of the equation by $\frac{1}{\Omega} e^{\,-i(\mathbf{k}+\mathbf{G}'')\cdot\mathbf{r}}$ and integrating both sides over a unit cell Ω. The key here is that, after the multiplication, the only $\mathbf{r}$ dependence of any term in the summation is via a phase factor $e^{\,i\mathbf{G}'''\cdot\mathbf{r}}$, for some reciprocal lattice vector $\mathbf{G}'''$. Therefore, after the integral over a unit cell, every term vanishes unless this characteristic reciprocal lattice vector $\mathbf{G}'''$ is zero. This implies that every term in the first summation vanishes except $\mathbf{G} = \mathbf{G}''$, and every term in the second summation vanishes except $\mathbf{G} = \mathbf{G}'' - \mathbf{G}'$:

$$0 = \tilde{u}_{\mathbf{k}\nu}(\mathbf{G}'') \left(\frac{\hbar^2(\mathbf{k}+\mathbf{G}'')^2}{2m_e} - \varepsilon_{\mathbf{k}\nu} \right) + \sum_{\mathbf{G}'} \cos\left((\mathbf{G}''-\mathbf{G}')\cdot\boldsymbol{\tau}\right)\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}''-\mathbf{G}')\tilde{u}_{\mathbf{k}\nu}(\mathbf{G}'). \tag{C.14}$$

We can now rename $\mathbf{G}''$ as $\mathbf{G}$, and move the energy term to the LHS:

$$\varepsilon_{\mathbf{k}\nu}\tilde{u}_{\mathbf{k}\nu}(\mathbf{G}) = \tilde{u}_{\mathbf{k}\nu}(\mathbf{G})\frac{\hbar^2(\mathbf{k}+\mathbf{G})^2}{2m_e} + \sum_{\mathbf{G}'} \cos\left((\mathbf{G}-\mathbf{G}')\cdot\boldsymbol{\tau}\right)\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}-\mathbf{G}')\tilde{u}_{\mathbf{k}\nu}(\mathbf{G}'). \tag{C.15}$$

Thus, we have taken the time-independent Schrödinger equation for the pseudoeigenstates, and transformed it into an equation for the Fourier coefficients of the periodic part of the pseudoeigenstates at $\mathbf{k}$. Given that the pseudopotential is zero at reciprocal lattice vectors for any reciprocal lattice vector where $|\mathbf{G}|^2 > 11$, we assume that any low energy eigenstate of the corresponding time-independent Schödinger equation will be only slightly faster varying than this, and as such can be expressed in terms of a Fourier

series over the 137 reciprocal lattice vectors we introduced earlier in eq. (C.6). As such, suppose we index this set of vectors such we have $\mathbf{G}^i$ defined for $i = 1, \ldots, 137$, and write $\tilde{\mathbf{u}}_{\mathbf{k}\nu} = \left[\tilde{u}_{\mathbf{k}\nu}(\mathbf{G}^1), \ldots, \tilde{u}_{\mathbf{k}\nu}(\mathbf{G}^{137})\right]$. We can then write down an eigenvalue equation for $\tilde{\mathbf{u}}_{\mathbf{k}\nu}(\mathbf{G})$:

$$\varepsilon_{\mathbf{k}\nu}\tilde{\mathbf{u}}_{\mathbf{k}\nu} = \hat{\mathcal{H}}^{\mathbf{k}}_{\text{pseudo}}\tilde{\mathbf{u}}_{\mathbf{k}\nu}. \tag{C.16}$$

Here the 137×137 matrix $\hat{\mathcal{H}}^{\mathbf{k}}_{\text{pseudo}}$ is defined in the following manner:

$$\begin{aligned}\left[\hat{\mathcal{H}}^{\mathbf{k}}_{\text{pseudo}}\right]^{ii} &= \frac{\hbar^2(\mathbf{k}+\mathbf{G}^i)^2}{2m_e} + \tilde{V}^{\text{sym}}_{\text{pseudo}}(0) \\ \left[\hat{\mathcal{H}}^{\mathbf{k}}_{\text{pseudo}}\right]^{ij} &= \cos\left((\mathbf{G}^i - \mathbf{G}^j)\cdot\boldsymbol{\tau}\right)\tilde{V}^{\text{sym}}_{\text{pseudo}}(\mathbf{G}^i - \mathbf{G}^j)\end{aligned} \tag{C.17}$$

Where $\tilde{V}^{\text{sym}}_{\text{pseudo}}(0) = 0$ according to Chelikowsky and Cohen [39]. Thus we can find a set of ~ 137 pseudoeigenvalues by solving the following matrix equation for any $\mathbf{k}$ in the Brillouin zone that interests us:

$$0 = \left(\hat{\mathcal{H}}^{\mathbf{k}}_{\text{pseudo}} - \varepsilon_{\mathbf{k}\nu}\hat{I}\right)\tilde{\mathbf{u}}_{\mathbf{k}\nu}. \tag{C.18}$$

By construction of the local empirical pseudopotential, the $\sim 8-10$ pseudoeigenvalues with the lowest energy will approximate the real eigenvalues of the valence band and the lowest energy conductions bands.

C.4 Exploiting Symmetry in the Brillouin Zone

Much like the translational symmetry of a crystal Hamiltonian ensures that energy eigenvalues separated by a reciprocal lattice vector are degenerate, the point group symmetries of a Hamiltonian ensures there are additional degeneracies in reciprocal space. The point symmetry of the reciprocal space of a Zinc-Blende or Diamond lattice is the same as the 48 point symmetries of a cube, known as the OCTAHEDRAL GROUP. Namely, the energy eigenvalues do not does not change if we invert any subset of the reciprocal lattice vectors, or if we permute the reciprocal lattice vectors. If we use a coordinate system rectilinear with the reciprocal lattice, this can be expressed in the following manner:

$$E(k_x, k_y, k_z) = E(|k_x|, |k_y|, |k_z|), \tag{C.19a}$$

$$\text{and} \qquad E(k_x, k_y, k_z) = E(k_l, k_m, k_n), \tag{C.19b}$$

$$\text{where} \qquad l \in \{x, y, z\}, \quad m \in \{x, y, z\} \backslash \{l\}, \quad n \in \{x, y, z\} \backslash \{l, m\}.$$

The first condition clearly allows us to reduce the reciprocal space to the positive octant of the coordinate system. Additionally, the six permutation relations described in the second condition allows us to reduce the positive octant of the reciprocal space to the wedge in which $k_x \geq k_y \geq k_z$; that is, if we apply the permutation symmetry to this wedge, it is clear that *any* coordinate in the positive octant must belong to some wedge, and that either the coordinates will belong interior of a unique wedge, or will belong to they belong to the border of a wedge. As such, it is clear that this wedge of the Brillouin zone is irreducible.

The most succinct definition of the interior of the irreducible of the wedge is that is the locus of points k_x, k_y, k_z such that the following two conditions are both true:

$$\frac{2\pi}{a} > k_x > k_y > k_z > 0 \qquad \text{due to nearest neighbours and 48 point symmetries,} \tag{C.20a}$$

$$\text{and} \qquad \frac{3\pi}{2a} > (k_x + k_y + k_z) \qquad \text{due to next nearest neighbours.} \tag{C.20b}$$

Thus, the most efficient way to find the energy eigenvalues of the whole Brillouin zone of silicon is to:

1. Solve eq. (C.18) for each $\mathbf{k}$ point in a mesh of the irreducible Brillouin Zone.

2. Use the point symmetry permutations to generate a mesh of eigenvalues on the whole Brillouin zone.

3. Use an interpolation scheme to generate a full band structure $\varepsilon(\mathbf{k}\nu)$.

Here steps 2 and 3 can be inverted if it is more convenient. Alternatively, if generating an interpolation from data on a cubic mesh is more convenient, more symmetry operations can be used in step 2 to generate of mesh of eigenvalues on, for instance, a cube of side length $\frac{4\pi}{a}$.

Appendix D

Fermi's Golden Rule

D.1 A General State in a Perturbed Hamiltonian

We present in this chapter a very typical derivation of Fermi's golden rule [93]. Suppose we have a Hamiltonian that can be expressed as the combination of a soluble Hamiltonian $\hat{H}_0$ and a small perturbing Hamiltonian $\hat{H}'$:

$$\hat{H} = \hat{H}_0 + \hat{H}'. \tag{D.1}$$

Suppose $\Phi_j^0(\mathbf{r}, t) = \phi_j^0(\mathbf{r}) e^{-\frac{i}{\hbar}\varepsilon_j^0 t}$ are time-dependent eigenfunctions of the *unperturbed* Hamiltonian. The time dependence of a general pure state $\Psi^0(\mathbf{r}, t)$ in the *unperturbed* Hamiltonian $\hat{H}^0$ will simply be determined by the superposition of the time dependence of unperturbed eigenfunctions:

$$\Psi^0(\mathbf{r}, t) = \sum_j a_j^0 \phi_j^0(\mathbf{r}) e^{-\frac{i}{\hbar}\varepsilon_j^0 t}. \tag{D.2}$$

Here a_j^0 is a constant, which is subject to the restriction that $\sum_j |a_j^0|^2 = 1$. On the other hand, the time dependence of a general pure state $\Psi(\mathbf{r}, t)$ in the *full* Hamiltonian $\hat{H}$ can be described simply by demanding that the coefficients a_j are not constants, but are instead functions of time:

$$\Psi(\mathbf{r},t) = \sum_j a_j(t)\phi_j^0(\mathbf{r})e^{-\frac{i}{\hbar}\varepsilon_j^0 t}. \tag{D.3}$$

D.2 An Equation of Motion for $a_k(t)$

The time dependence of a particular coefficient $a_k(t)$ of the general wavefunction Ψ will be determined by the time-dependent Schrödinger equation associated with the full Hamiltonian:

$$i\hbar\frac{\partial\Psi}{\partial t} = \hat{H}_0\Psi + \hat{H}'\Psi. \tag{D.4}$$

In order to determine the equation of motion for $a_k(t)$ from this equation, we begin by substituting eq. (D.3) into eq. (D.4), and expanding the time derivative using the product rule:

$$\sum_j \left(a_j(t)\varepsilon_j^0 + i\hbar\frac{\mathrm{d}a_j(t)}{\mathrm{d}t}\right)\phi_j^0(\mathbf{r})e^{-\frac{i}{\hbar}\varepsilon_j^0 t} = \sum_j a_j(t)\left(\varepsilon_j^0 + \hat{H}'\right)\phi_j^0(\mathbf{r})e^{-\frac{i}{\hbar}\varepsilon_j^0 t}. \tag{D.5}$$

Notice that we can subtract $\sum_j a_j(t)\varepsilon_j^0\phi_j^0(\mathbf{r})e^{-\frac{i}{\hbar}\varepsilon_j^0 t}$ from both sides of eq. (D.5), leading to the following equation:

$$i\hbar\sum_j \frac{\mathrm{d}a_j(t)}{\mathrm{d}t}\phi_j^0(\mathbf{r})e^{-\frac{i}{\hbar}\varepsilon_j^0 t} = \sum_j a_j(t)\hat{H}'\phi_j^0(\mathbf{r})e^{-\frac{i}{\hbar}\varepsilon_j^0 t}. \tag{D.6}$$

If we multiply both sides by $\phi_k^{0*}(\mathbf{r})$ and integrate over all space[1], we have the following expression:

$$i\hbar\sum_j \frac{\mathrm{d}a_j(t)}{\mathrm{d}t}e^{-\frac{i}{\hbar}\varepsilon_j^0 t}\int_{\mathbb{R}^3}\phi_k^{0*}(\mathbf{r})\phi_j^0(\mathbf{r})\mathrm{d}\mathbf{r} = \sum_j a_j(t)e^{-\frac{i}{\hbar}\varepsilon_j^0 t}\int_{\mathbb{R}^3}\phi_k^{0*}(\mathbf{r})\hat{H}'\phi_j^0(\mathbf{r})\mathrm{d}\mathbf{r}. \tag{D.7}$$

This development is useful because if we assume the eigenfunctions of the unperturbed wavefunctions are normalized, the integral on the LHS simplifies to δ_{kj}, since the normalized eigenstates always form an orthonormal set of basis functions. Multiplying

[1] A unit cell is sufficient if in cases where Bloch's theorem applies.

both sides by $-\frac{i}{\hbar}e^{\frac{i}{\hbar}\varepsilon_k^0 t}$, we find an equation of motion for $a_k(t)$:

$$\frac{\mathrm{d}a_k(t)}{\mathrm{d}t} = -\frac{i}{\hbar}\sum_j a_j(t)H'_{kj}e^{\frac{i}{\hbar}(\varepsilon_k^0-\varepsilon_j^0)t}, \tag{D.8}$$

$$\text{where} \qquad H'_{kj} = \int_{\mathbb{R}^3}\phi_k^{0*}(\mathbf{r})\hat{H}'\phi_j^0(\mathbf{r})\mathrm{d}\mathbf{r}.$$

In order to solve this equation and find $a_k(t)$, we integrate both sides of the equation from $t' = 0$ to $t' = t$:

$$\int_0^t \frac{\mathrm{d}a_k(t')}{\mathrm{d}t'}\mathrm{d}t' = -\frac{i}{\hbar}\sum_j \int_0^t a_j(t')H'_{kj}e^{i\omega_{kj}t'}\mathrm{d}t', \tag{D.9}$$

$$\text{where} \qquad \omega_{kj} = \frac{\varepsilon_k^0 - \varepsilon_j^0}{\hbar}.$$

This can be in principle be solved by applying the fundamental theorem of calculus:

$$a_k(t) = a_k(0) - \frac{i}{\hbar}\sum_j \int_0^t a_j(t')H'_{kj}e^{i\omega_{kj}t'}\mathrm{d}t'. \tag{D.10}$$

D.3 Transition Rate Between Two Unpertubed Eigenstates

In this section we are interested in the transition rate from one unperturbed eigenstate state to another. Accordingly, we assume the initial state is a pure unperturbed eigenstate:

$$a_j(0) = \begin{cases} 1 & j = j' \\ 0 & j \neq j'. \end{cases} \tag{D.11}$$

In this section we are also interested in the case where the perturbation Hamiltonian $\hat{H}'$ is *weak* and can be associated with a *single driving frequency* ω', such that it can be $\hat{H}'$ can be expressed as follows:

$$\hat{H}' = \hat{\mathcal{H}}'(\mathbf{r})e^{-i\omega' t}. \tag{D.12}$$

.

The *maximum* rate probability amplitude can be transferred per unit time from the j^{th} unperturbed eigenstate to the k^{th} unperturbed eigenstate is equal to $\frac{|\mathcal{H}'_{kj}|}{\hbar}$. Accordingly, if we restrict ourselves to a time scale such that $t \ll \frac{\hbar}{|\mathcal{H}'_{kj}|}$, $a_j(t)$ must be roughly constant:

$$a_j(t) \approx a_j(0) \qquad \text{for} \quad t \ll \frac{\hbar}{|\mathcal{H}'_{kj}|}. \tag{D.13}$$

We note that since we are assuming $\mathcal{H}'$ is small compared to H_0, the time scale on which this approximation is true will generally be much *larger* than the timescale associated with the period of the unperturbed eigenstates, and so *within this timescale we will observe many unperturbed eigenstate oscillations*.

For $k \neq j'$, by substituting eq. (D.13), eq. (D.11), and eq. (D.12) into eq. (D.10), we have the following:

$$\begin{aligned} a_k(t) &\approx -\frac{i}{\hbar}\hat{\mathcal{H}}'_{kj'} \int_0^t e^{\,i(\omega_{kj'}-\omega')t'}\,\mathrm{d}t' \\ &= \frac{\mathcal{H}'_{kj'}}{\hbar(\omega_{kj'}-\omega')}\left(e^{\,i(\omega_{kj'}-\omega')t}-1\right). \end{aligned} \tag{D.14}$$

This leads to a transition probability as a function of time, $|a_k(t)|^2$, defined as follows:

$$\begin{aligned} |a_k(t)|^2 &\approx \frac{|\mathcal{H}'_{kj'}|^2}{\hbar^2(\omega_{kj'}-\omega')^2}\left(e^{\,i(\omega_{kj'}-\omega')t}-1\right)\left(e^{\,-i(\omega_{kj'}-\omega')t}-1\right) \\ &= 2\frac{|\mathcal{H}'_{kj'}|^2}{\hbar^2(\omega_{kj'}-\omega')^2}\Big(1-\cos\big((\omega_{kj'}-\omega')t\big)\Big) \\ &= 4\frac{|\mathcal{H}'_{kj'}|^2}{\hbar^2(\omega_{kj'}-\omega')^2}\sin^2\left(\frac{\omega_{kj'}-\omega'}{2}t\right). \end{aligned} \tag{D.15}$$

This is a basic result of first order perturbation theory. There is a resonance in the amplitude of the transition probability as a function of time when the driving frequency ω' is equal to the difference in frequency between the j^{th} and j'^{th} eigenstates, at which point probability amplitude is initially transferred at the maximum rate, resulting in $|a_k(t)|^2 = \frac{|\mathcal{H}'_{kj'}|^2}{\hbar^2}t^2$. In the non-resonance case, a small fraction of the initial probability oscillates between the states at the BEAT— or *difference from resonance*— FREQUENCY.

From the second to final line of eq. (D.15), we can easily derive the initial transition rate

from the initial j'^{th} unperturbed eigenstate to the final k^{th} unpeturbed eigenstate:

$$S_{j'\to k} = \left.\frac{\mathrm{d}|a_k|^2}{\mathrm{d}t}\right|^0 = 0. \tag{D.16}$$

This result may be somewhat jarring initially. Transitions will obviously eventually occur between individual eigenstates of the unperturbed Hamiltonian— as eq. (D.15) shows— but it is a *second-order rather than a first-order* effect of increasing time. This is caused by the simple fact that initial rate of *probability amplitude* transfer is linear in time, and therefore initial *probability* transfer is quadratic in time. As we will show, the only way to have a finite initial transition rate out from an initial eigenstate is if there is a *continuum* of final states available.

D.4 Transition Rate Into a Continuum of Final States

Suppose we are interested in the probability that the single frequency perturbation Hamiltonian drives initial unperturbed eigenstate i into a *quasicontinuum* of final unperturbed eigenstates f, such that there is $D_f(\varepsilon_f^0)$ states per unit energy at eigenstate energy ε_f^0. Before we address this problem, we note that to find the transition probability to a *countable* set of final unperturbed eigenstates— f_{count}— we would simply calculate the following:

$$P_{i\to f_{\text{count}}}(t) = \sum_{k\in f_{\text{count}}} |a_k(t)|^2. \tag{D.17}$$

If $a_{\varepsilon_f^0}(t)$ is the amplitude associated with unperturbed eigenstates of energy ε_f^0, the analogous expression for the transition probability into the quasicontinuum of final unperturbed eigenstates— f— is the following:

$$P_{i\to f}(t) = \int_0^\infty D_f(\varepsilon_f^0)\,|a_{\varepsilon_f^0}(t)|^2\,\mathrm{d}\varepsilon_f^0. \tag{D.18}$$

Substituting eq. (D.15) and the relation $\omega_{fi} = \omega_f^0 - \omega_i^0$ into eq. (D.18), we find the following:

$$P_{i\to f}(t) = \frac{4}{\hbar}|\hat{\mathcal{H}}'_{fi}|^2 \int_0^\infty D_f(\hbar\omega_f^0)\frac{\sin^2\left(\frac{\omega_f^0-\omega_i^0-\omega'}{2}t\right)}{(\omega_f^0-\omega_i^0-\omega')^2}\mathrm{d}\omega_f^0. \tag{D.19}$$

We make the substitution $\theta = \frac{t}{2}(\omega_f^0 - \omega_i^0 - \omega')$, which implies that $\frac{2}{t}\mathrm{d}\theta = \mathrm{d}\omega_f^0$, and the limits change from $0 \to \infty$, into $-\frac{t}{2}(\omega_i^0 + \omega') \to \infty$.

$$P_{i\to f}(t) = \frac{2t}{\hbar}|\hat{\mathcal{H}}'_{fi}|^2 \int_{-\frac{t}{2}(\omega_i^0+\omega')}^\infty D_f(\hbar\omega_f^0(\theta))\frac{\sin^2\theta}{\theta^2}\mathrm{d}\theta, \tag{D.20}$$

$$\text{where} \qquad \omega_f^0(\theta) = \frac{2\theta}{t} + \omega_i^0 + \omega'.$$

We note that $\sim 90\%$ of the integral of $\frac{sin^2\theta}{\theta^2}$ is between $(-\pi, \pi)$. We note that we are *not* interested in the very rapid tiny oscillations of $P_{i\to f}(t)$ over times periods comparable to the unperturbed eigenvalue periods, but are instead interested in the transition probability at much longer times. Accordingly, we can assume firstly that the bottom limit of the integral is much less than $-\pi$, and can accordingly be approximated as $-\infty$. We can assume secondly that the total variation of the final eigenvalue frequency, ω_f^0, over the domain $(-\pi, \pi)$, which is equal to $\Delta\omega_f^0 = \frac{4\pi}{t}$, is negligible so that $\Delta\omega_f^0 << \omega_f^0$. Accordingly, we can assume $\omega_f^0(\theta) \approx \omega_f^0(0)$. These assumptions lead to the following expression:

$$\begin{aligned} P_{i\to f}(t) &\approx \frac{2t}{\hbar}D_f(\hbar\omega_i^0 + \hbar\omega')|\hat{\mathcal{H}}'_{fi}\int_{-\infty}^\infty \frac{\sin^2\theta}{\theta^2}\mathrm{d}\theta \\ &= \frac{2\pi t}{\hbar}D_f(\varepsilon_i^0 + \hbar\omega')|\hat{\mathcal{H}}'_{fi}|^2. \end{aligned} \tag{D.21}$$

The rate of transition to final states, which is defined as $S_{i\to f} = \frac{\mathrm{d}P_{i\to f}}{\mathrm{d}t}$, is then given by FERMI'S GOLDEN RULE:

$$S_{i\to f} = \frac{2\pi}{\hbar}D_f(\varepsilon_i^0 + \hbar\omega')|\hat{\mathcal{H}}'_{fi}|^2. \tag{D.22}$$

Appendix E

The Strain Tensor

E.1 The Deformation Tensor

Suppose each point of matter in a continuous body is labelled by an initial position vector $\mathbf{r}$, and we deform the body so that the each point of matter in the body initially at $\mathbf{r}$ is now at a final position vector $\mathbf{r}'(\mathbf{r})$. We define the DISPLACEMENT VECTOR FIELD $\mathbf{u}(\mathbf{r})$ as the displacement of the final position vector from the initial position vector, as a function of initial position vector:

$$\mathbf{u}(\mathbf{r}) = \mathbf{r}'(\mathbf{r}) - \mathbf{r}. \tag{E.1}$$

We note that, in many cases, the most important physical effects of a deformation are not associated with how much points are displaced in an *absolute* sense, but are instead associated with the *relative* change in displacements between points. For example, a large uniform shift of the entire body— while associated with a large displacement vector field— does not change the distances between points and as such will not be associated with any physical effects *unless the environment the entire body is in is non-uniform*. We will assume this is not the case.

A displacement vector field which does not change the relative displacement between

points is equivalent to a displacement field in which $\frac{\mathrm{d}\mathbf{r}'}{\mathrm{d}\mathbf{r}}$— the rate of change of final position $\mathbf{r}'$ relative to small changes in initial position $\mathbf{r}$— is equal to unity. The *deviation of* $\frac{\mathrm{d}\mathbf{r}'}{\mathrm{d}\mathbf{r}}$ *from unity*— $\mathbf{d}(\mathbf{r})$— is therefore generally of more direct physical interest than the displacement vector field $\mathbf{u}(\mathbf{r})$ when we examine bodies in uniform environments:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{r}'}{\mathrm{d}\mathbf{r}} &= \frac{\mathbf{r}'(\mathbf{r}+\delta\mathbf{r}) - \mathbf{r}'(\mathbf{r})}{\delta\mathbf{r}} \\ &= \frac{\mathbf{r}+\delta\mathbf{r}+\mathbf{u}(\mathbf{r}+\delta\mathbf{r}) - \mathbf{r} - \mathbf{u}(\mathbf{r})}{\delta\mathbf{r}} \\ &= 1 + \mathbf{d}(\mathbf{r}), \qquad \text{(E.2a)}\end{aligned}$$

$$\text{where} \qquad \mathbf{d}(\mathbf{r}) = \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\mathbf{r}}. \qquad \text{(E.2b)}$$

We refer to $\mathbf{d}(\mathbf{r})$ as the DEFORMATION TENSOR FIELD, and note that it measures how quickly *each* displacement vector component changes in *any* given direction.

E.2 Strain and the Small Deformation Limit

We can often simplify further. We note that many important physical effects are not necessarily associated with changes in the *diplacements* between *all* points, but more simply with changes in the *distance* between *nearby* points. Accordingly the derivative we are most often interested in is $\frac{|\partial\mathbf{r}'|}{|\partial\mathbf{r}|}$, which can be expressed as follows:

$$\begin{aligned}\frac{|\partial\mathbf{r}'|}{|\partial\mathbf{r}|} &= \frac{|\mathbf{r}'(\mathbf{r}+\delta\mathbf{r}) - \mathbf{r}'(\mathbf{r})|}{|\delta\mathbf{r}|} \\ &= \frac{|(1+\mathbf{d})\delta\mathbf{r}|}{|\delta\mathbf{r}|} \\ &= |(1+\mathbf{d})\widehat{\delta\mathbf{r}}| \\ &= \sqrt{1 + 2(\mathbf{d}\widehat{\delta\mathbf{r}})\cdot\widehat{\delta\mathbf{r}} + (\mathbf{d}\widehat{\delta\mathbf{r}})\cdot(\mathbf{d}\widehat{\delta\mathbf{r}})}, \qquad \text{(E.3)}\end{aligned}$$

$$\text{where} \qquad \widehat{\delta\mathbf{r}} = \frac{\delta\mathbf{r}}{|\delta\mathbf{r}|}.$$

Here $\widehat{\delta\mathbf{r}}$ is a unit vector in the direction of $\delta\mathbf{r}$. If assume that the material being is only *slightly deformed* we can make further simplifications. In the small deformation limit,

each entry of $\mathbf{d}$ is small compared to unity, which implies the following:

$$1 \gg 2(\mathbf{d}\widehat{\delta\mathbf{r}}) \cdot \widehat{\delta\mathbf{r}} \gg (\mathbf{d}\widehat{\delta\mathbf{r}}) \cdot (\mathbf{d}\widehat{\delta\mathbf{r}}). \tag{E.4}$$

Combining eq. (E.3) with the small deformation limit in eq. (E.4), and the first order approximation $\sqrt{1+x} \approx 1 + \frac{1}{2}x$, we find the following expression:

$$\frac{|\partial\mathbf{r}'|}{|\partial\mathbf{r}|} \approx 1 + (\mathbf{d}\widehat{\delta\mathbf{r}}) \cdot \widehat{\delta\mathbf{r}}. \tag{E.5}$$

We note that it is only the *symmetric part* of $\mathbf{d}$ that contributes, since for *any* antisymmetric tensor $\mathbf{A}$, and *any* vector $\mathbf{v}$, we have the following relationship:

$$\begin{aligned}
(\mathbf{A}\mathbf{v}) \cdot \mathbf{v} &= \sum_{i,j} A_{ij} v_i v_j \\
&= -\sum_{i,j} A_{ji} v_j v_i \qquad \text{since } A_{ij} = -A_{ji} \text{ and } v_i v_j = v_j v_i, \\
&= -(\mathbf{A}\mathbf{v}) \cdot \mathbf{v} \qquad \text{by definition of } (\mathbf{A}\mathbf{v}) \cdot \mathbf{v}, \\
&= 0.
\end{aligned} \tag{E.6}$$

For this reason, we express $\mathbf{d}$ as the sum of a symmetric tensor $\mathbf{e}$ and an antisymmetric tensor $\mathbf{f}$, with entries defined as follows:

$$e_{ij} = \frac{1}{2}\left(d_{ij} + d_{ji}\right) \tag{E.7a}$$

$$f_{ij} = \frac{1}{2}\left(d_{ij} - d_{ji}\right). \tag{E.7b}$$

This allows us to rewrite eq. (E.5) in terms of the symmetric deformation tensor $\mathbf{e}$:

$$\frac{|\partial\mathbf{r}'|}{|\partial\mathbf{r}|} = 1 + (\mathbf{e}\widehat{\delta\mathbf{r}}) \cdot \widehat{\delta\mathbf{r}}. \tag{E.8}$$

We remind the reader that the above expression is *only true* in the limit that the deformation is small. Outside this limit, one must use the full expression for the relative distance derivative $\frac{|\partial\mathbf{r}'|}{|\partial\mathbf{r}|}$ given in eq. (E.3), which includes a contribution from the antisymmetric deformation component. However the small deformation limit is widely used. Accordingly the symmetric deformation tensor $\mathbf{e}$ is typically referred to as the STRAIN TENSOR,

owing to the fact that it is *typically* the part of the deformation tensor associated with the changes in distance between nearby points. Similarly, the antisymmetric tensor $\mathbf{f}$ is sometimes referred to as the LOCAL ROTATION TENSOR, since its first order effect is a simple rotation of the local coordinate system about the point where $\delta\mathbf{r} = 0$.

E.3 Decomposing the Strain Tensor

There are a number of useful ways to further decompose the strain tensor, each of yield an important insight. One classic decomposition is to view the strain as the *sum* of a purely diagonal NORMAL STRAIN TENSOR— which describes the simple length dilation factor along each coordinate— and purely off-diagonal SHEAR STRAIN TENSOR— which describe the *lateral* displacement factors as one moves along each coordinate. Another classic decomposition is to view the strain as the *product* of a basis transformation tensor, and a purely diagonal, normal strain tensor, which emphasizes the fact that there always exists a— possibly non-orthogonal— coordinate system in which there is *no shear strain*. Finally, the decomposition we are most interested in is to view the strain as the sum of a MEAN DILATION TENSOR— which replaces the length dilation factor along each coordinate with the mean length dilation factor— and a STRAIN DEVIATION TENSOR— which adjusts the length dilation factors to their correct values and adds shear strain. We can express this decomposition as follows:

$$\mathbf{e} = \overbrace{\frac{e_{xx} + e_{yy} + e_{zz}}{3}\mathbf{I}}^{\text{Mean Dilation Tensor}} + \overbrace{\left(\mathbf{e} - \frac{e_{xx} + e_{yy} + e_{zz}}{3}\mathbf{I}\right)}^{\text{Strain Deviation Tensor}}. \tag{E.9}$$

The idea behind this decomposition is very simply that it puts the strain trace information into a single subtensor— the mean dilation tensor— and bundles *all information that is not the strain trace* into the other subtensor— the strain deviation tensor. This is especially useful in cases *where we care most about the strain trace* and care much less about the other components of strain, such as in this thesis where we care most about the *volume dilation* effect of strain.

We can show that the volume dilation is equal to the trace of the strain tensor in the small

deformation limit as follows. First, we note that a rectilinear cuboid volume between $\mathbf{r}$ and $\mathbf{r} + \delta\mathbf{r}$ will deform into a parallelepiped between $\mathbf{r}'$ and $\mathbf{r}' + \delta\mathbf{r}'$. The volume of this parallelepiped is determined by the *projection* of the edges of the parallelepiped back on to the cartesian coordinates. This in turn is determined *only* by the diagonal components of the strain tensor. We can therefore express the rate of change of volume due to the small deformation as follows:

$$\begin{aligned}\frac{\mathrm{d}V'}{\mathrm{d}V} &= \frac{(1+e_{xx})\delta x(1+e_{yy})\delta y(1+e_{zz})\delta z}{\delta x\delta y\delta z} \\ &= \frac{\Big(1+e_{xx}+e_{yy}+e_{zz}+\mathcal{O}(\mathbf{e}^2)\Big)\delta x\delta y\delta z}{\delta x\delta y\delta z}] \\ &\approx 1+\mathrm{Tr}(\mathbf{e}).\end{aligned} \tag{E.10}$$

Accordingly, in the small deformation limit the volume dilation is defined by the trace of the strain tensor. Since the strain deviation tensor has zero trace by definition, it does not contribute to the volume dilation. Since most of the changes to bandstructure energy in this thesis are mediated by volume dilation, the physical effect of the strain deviation tensor is largely neglected in this thesis.

Appendix F

A General Macroscopic Continuity Equation

F.1 A Generic Continuity Equation

In this appendix we derive a general macroscopic continuity equation for the quantity measured by an arbitrary function of crystal momentum $\boldsymbol{\zeta}(\mathbf{k}\nu)$. We begin by noting that the average value of $\boldsymbol{\zeta}(\mathbf{k}\nu)$ for carriers at $(\mathbf{r}, t)$ is as follows:

$$\langle \boldsymbol{\zeta} \rangle (\mathbf{r}, t) = \frac{1}{n(\mathbf{r}, t)} \sum_{\nu} \int_{\mathrm{BZ}} f(\mathbf{k}\nu, \mathbf{r}, t) \boldsymbol{\zeta}(\mathbf{k}\nu) \mathrm{d}\mathbf{k}, \tag{F.1}$$

$$\text{where} \qquad n(\mathbf{r}, t) = \sum_{\nu} \int_{\mathrm{BZ}} f(\mathbf{k}\nu, \mathbf{r}, t) \mathrm{d}\mathbf{k}.$$

As usual, the subscript "BZ" implies the integral is over the Brillouin zone, and the term $n(\mathbf{r}, t)$ is the density of carriers. We can define the *density* of the quantity measured by $\boldsymbol{\zeta}(\mathbf{k}\nu)$ as the product of the average value of $\boldsymbol{\zeta}(\mathbf{k}\nu)$ at a point and the local carrier density:

$$\rho_{\boldsymbol{\zeta}}(\mathbf{r}, t) = n(\mathbf{r}, t) \langle \boldsymbol{\zeta} \rangle (\mathbf{r}, t). \tag{F.2}$$

Note that since carrier density and the distribution function are both scalar, $\langle \boldsymbol{\zeta} \rangle$ and $\rho_{\boldsymbol{\zeta}}$

must be the same kind of geometric object as $\zeta(\mathbf{k}\nu)$; that is, all must be scalar, vector, or tensor fields. The macroscopic continuity equation for the quantity measured by $\zeta(\mathbf{k}\nu)$ is an equation for the rate of change of the arbitrary density— $\frac{\partial \rho_\zeta}{\partial t}$.[1] We can intuitively partition this rate of change into two distinguishable parts: one part is the net rate that the arbitrary quantity flows into a unit volume due to the movement of particles— $\left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{net flow in}}$ — the other part is the net rate the arbitrary quantity is produced inside a unit volume— $\left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{net production}}$:

$$\frac{\partial \rho_\zeta}{\partial t} = \left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{net flow in}} + \left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{net production}}. \tag{F.3}$$

The net rate the arbitrary quantity flows into a volume can be defined in terms of a transport flux of the arbitrary quantity— $\boldsymbol{\Phi}_\zeta(\mathbf{r}, t)$. We represent the transport flux using a boldface variable to acknowledge that in a $m-$D position space, it is a geometric object with m times more degrees of freedom than the density $\rho_\zeta(\mathbf{r}, t)$.[2] Using this transport flux and definition of divergence, we can express eq. (F.3) in familiar form of a generic continuity equation:

$$\frac{\partial \rho_\zeta}{\partial t} = \left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{net production}} - \nabla_{\mathbf{r}} \cdot \boldsymbol{\Phi}_\zeta. \tag{F.4}$$

F.2 The Semiclassical Transport Flux

The transport flux $\boldsymbol{\Phi}_\zeta$ of some arbitrary tensor quantity ζ is geometric object defined such that $\delta A \hat{\mathbf{n}} \cdot \boldsymbol{\Phi}_\zeta$ will yield the rate the arbitrary quantity crosses an oriented flat surface of area δA and unit normal vector $\hat{\mathbf{n}}$. This quantity is commonly understood when the quantity transported is a simple scalar, and it generalizes simply to the case of vectors and tensor by noting that the projection along each dimension of the basis is a scalar field that has an associated transport flux vector.

In order to determine an expression for the transport flux, it helps to first notice that expectation of the rate at which *carriers themselves* cross the infinitesimal surface defines

[1] This rate must also be the same type of geometric object as $\zeta(\mathbf{k}\nu)$.

[2] By this we mean the following: when ρ_ζ is a scalar field, $\boldsymbol{\Phi}_\zeta$ is an $m \times 1$ or $1 \times m$ vector field; when ρ_ζ is an $n \times 1$ vector field, $\boldsymbol{\Phi}_\zeta$ is an $n \times m$ or $nm \times 1$ tensor field; et cetera.

$\boldsymbol{\phi}_1$, or the transport flux associated with $\zeta(\mathbf{k}\nu) = 1$:

$$\delta A \hat{\mathbf{n}} \cdot \boldsymbol{\phi}_1 = \delta A \sum_\nu \int_{\mathrm{BZ}} (\hat{\mathbf{n}} \cdot \mathbf{v})\, f \mathrm{d}\mathbf{k}. \tag{F.5}$$

The rate at which an arbitrary quantity crosses the infinitesimal surface must simply be defined by multiplying the particle density crossing at each $\mathbf{k}\nu$ by the function $\zeta(\mathbf{k}\nu)$:

$$\delta A \hat{\mathbf{n}} \cdot \boldsymbol{\phi}_\zeta = \delta A \sum_\nu \int_{\mathrm{BZ}} (\hat{\mathbf{n}} \cdot \mathbf{v})\, \zeta f \mathrm{d}\mathbf{k}. \tag{F.6}$$

We are now in a position to define transport flux directly, but in order to do so, we must make use of the TENSOR PRODUCT. The tensor product is a way of combining two vectors into a product without any reduction in the total number of degrees of freedom. The tensor product is primarily useful to us because it allows us to "wait" to preform a dot product operation if it is not convenient to do so, by using the relation $(\mathbf{a} \cdot \mathbf{b})\mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \otimes \mathbf{c})$.[3] In eq. (F.7), this manipulation allows us to take the arbitrary constant normal vector $\hat{\mathbf{n}}$ outside the integral, and so allows us to write down an explicit definition of the arbitrary transport flux:

$$\begin{aligned} \delta A \hat{\mathbf{n}} \cdot \boldsymbol{\phi}_\zeta &= \delta A \hat{\mathbf{n}} \cdot \int_{\mathrm{BZ}} (\mathbf{v} \otimes \boldsymbol{\zeta})\, f \mathrm{d}\mathbf{k}, \\ \text{therefore} \qquad \boldsymbol{\phi}_\zeta &= n \left\langle \mathbf{v} \otimes \boldsymbol{\zeta} \right\rangle . \end{aligned} \tag{F.7}$$

F.3 The Semiclassical Production Rate

We now turn our attention to the net production rate term of eq. (F.4). This is associated with the— possibly negative— rate the arbitrary quantity is produced inside a unit volume *which is not due to the divergence of the arbitrary transport flux*. Since the measure $\zeta(\mathbf{k}\nu)$ is purely a function of $\mathbf{k}\nu$, the contribution to the arbitrary quantity associated with a local fictitious Bloch particle will only change *if the Bloch particle changes eigenstate*.

[3] Technically, the dot operator on the right hand side of this identity no longer represents the dot product, but its generalization— the CONTRACTED TENSOR PRODUCT. We also note that the contracted tensor product is equivalent to a simple matrix multiplication for tensors of order two, and so is often represented without any dot symbol at all.

As discussed in the Background chapter, there are two broad processes which cause local transitions of probability density from one Bloch eigenstate to another: scattering by non-reproducible perturbation force fields, and Bloch law acceleration by reproducibly ordered "external" perturbation force fields:

$$\left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{net production}} = \left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{external}} + \left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{scat}}. \tag{F.8}$$

While we largely ignore field assisted interband tunnelling in this thesis, we note that *if we wanted to incorporated it*, we would incorporate it into the scattering term for convenience. Accordingly, the external force term is always mediated entirely by a deterministic change to $\mathbf{k}$ and is not associated with a production rate of n. Therefore, if we expand ρ_ζ using eq. (F.2), we can move n outside the derivative:

$$\begin{aligned}\left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{external}} &= \left(\frac{\partial n \langle \boldsymbol{\zeta} \rangle}{\partial t}\right)_{\text{external}} \\ &= n \left\langle \left(\frac{\partial \boldsymbol{\zeta}}{\partial t}\right)_{\text{external}} \right\rangle.\end{aligned} \tag{F.9}$$

Since the effect of the external field on $\boldsymbol{\zeta}$ is mediated entirely by the effect of the field on $\mathbf{k}$, we use the chain rule to separate out this dependence. Additionally, we again make use of the tensor product in order to delay the evaluation of the dot product:

$$\begin{aligned}\left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{external}} &= n \left\langle \left(\left(\frac{\partial \mathbf{k}}{\partial t}\right)_{\text{external}} \cdot \nabla_{\mathbf{k}} \right) \boldsymbol{\zeta} \right\rangle \\ &= n \left\langle \left(\frac{\partial \mathbf{k}}{\partial t}\right)_{\text{external}} \cdot (\nabla_{\mathbf{k}} \otimes \boldsymbol{\zeta}) \right\rangle.\end{aligned} \tag{F.10}$$

Under the assumption that we can localize the carriers to a region of the device in which the force is due to an approximately uniform electric field, we can invoke eq. (2.2)—Newton's law for Bloch states— to yield an expression for $\left(\frac{\partial \mathbf{k}}{\partial t}\right)_{\text{external}}$. Gathering this together with we can write down an expression for the production rate of ρ_ζ due to an external force $\mathbf{F}(\mathbf{r}, t)$:

$$\left(\frac{\partial \rho_\zeta}{\partial t}\right)_{\text{external}} = \frac{n}{\hbar} \mathbf{F} \cdot \langle \nabla_{\mathbf{k}} \otimes \boldsymbol{\zeta} \rangle. \tag{F.11}$$

We are now in a position to rewrite the continuity equation given eq. (F.4) for the semiclassical regime described in the Background chapter, by substituting eq. (F.7) and eq. (F.11), to yield the macroscopic continuity equation for the density associated with $\boldsymbol{\zeta}(\mathbf{k}\nu)$:

$$\frac{\partial \rho_{\boldsymbol{\zeta}}}{\partial t} = \left(\frac{\partial \rho_{\boldsymbol{\zeta}}}{\partial t}\right)_{\text{scat}} - \nabla_{\mathbf{r}} \cdot \Big(n \left\langle \mathbf{v} \otimes \boldsymbol{\zeta} \right\rangle \Big) + \frac{n}{\hbar} \mathbf{F} \cdot \left\langle \nabla_{\mathbf{k}} \otimes \boldsymbol{\zeta} \right\rangle . \tag{F.12}$$

Appendix G

Maximizing Entropy When Tail Density is Limited

G.1 Maximum Entropy in Classical Statistical Mechanics

The MAXIMUM ENTROPY OCCUPATION FUNCTION[1] is defined by the most probable occupation rate of any *single carrier* state, under the condition that all accessible many-particle states in a small volume ΔV of our device are assumed to be equally probable. A microstate is ACCESSIBLE if it does not violate the ESTABLISHED CONSTRAINTS, which typically take the form of various known expected values of the carrier system. In this appendix, we investigate the entropy maximization subject to established constraints defined as follows for a small region of volume ΔV in our device:

- the expected number of carriers is $n\Delta V$,
- the expected total kinetic energy is $w\Delta V$, and
- the expected number of carriers with energy greater than ε is less than or equal to

[1] We refer to the distribution function here as an *occupation function* for the simple reason that it does not possess spatial dependence in this appendix.

$n_{\text{lim}}(\varepsilon)\Delta V$, where $n_{\text{lim}}(\varepsilon)$ is a known function of kinetic energy.

It is not obvious how to determine whether a given microstate in the volume ΔV violates the established constraints. However, the fact that the maximum entropy occupation function is the *most probable* single state occupation rate allows a transformation that avoids this problem. Suppose we consider a large volume V, defined to contain $\frac{V}{\Delta V}$ independent copies of the original system. The *most probable* single carrier occupation function in the large volume occurs, by definition, when the single carrier occupation functions in each of the $\frac{V}{\Delta V}$ original systems are the *most probable* single carrier occupation functions. Thus, the maximum entropy occupation function associated with the actual system is identical to the maximum entropy occupation function associated with $\frac{V}{\Delta V}$ independent copies of the system. Accordingly, we will investigate the maximum entropy of the enlarged system of volume V.

If we take $V \to \infty$, the accessible states in the enlarged system become clear. By the law of large numbers, almost all accessible microstates of the enlarged system must have densities that tend toward precisely n, energy densities that tend toward precisely w, and an upper limit on the density of particles above ε that must tend toward precisely $n_{\text{lim}}(\varepsilon)$.

To find the maximum entropy occupation function, it helps to isolate a *single carrier* state from the rest of the distribution, since the occupation function defines the occupation rate of any single carrier state. Suppose we do so, and partition a single carrier state A of kinetic energy ε_A from all other states in the volume V. If we assume that neither filling this subsystem with a carrier, nor not filling this subsystem with a carrier violates the established constraints, then there are two accessible "nanostates" for our single state "nanosystem": occupied and unoccupied.[2]

If the single state is occupied, then the microstates associated with the all other states in the volume must have 1 less carrier, and ε_A less total kinetic energy than if the partitioned single state was unoccupied. As such, if f_A is the occupation rate of state A, the system associated with the remaining particles— which we will refer to as the MAJOR

[2]We use the term MICROSTATE to refer to the precise many particle state of a system expected to have a large number of particles. We use the NANOSTATE to refer to the precise state of a special system— referred to as a NANOSYSTEM— where the only possible microstates are occupied or unoccupied.

SYSTEM— is expected to have f_A less carriers, and $f_A\varepsilon_A$ less energy than if the state A is unoccupied. Accordingly the rate of change of the *total number of particles of the major system*— $N_{\text{not } A}$— and the *total energy of the major system*— $\varepsilon_{\text{not } A}$— with respect to *the occupation rate of state* A— f_A— are given respectively as follows:

$$\frac{\partial N_{\text{not } A}}{\partial f_A} = 1, \tag{G.1a}$$

$$\text{and} \qquad \frac{\partial \varepsilon_{\text{not } A}}{\partial f_A} = \varepsilon_A. \tag{G.1b}$$

A change in the total carriers or total energy of the major system will, in general, change the entropy of this system since it changes the set of possible microstates of the system. Our aim is to find the f_A that maximizes the entropy of the *total* system, which is the sum of the *entropy of the nanosystem*—S_A— and the entropy of the major system—$S_{\text{not}A}$. Since entropy is a concave functional [23], this will occur at the point where the derivative of the total entropy with respect to f_A vanishes, or equivalently at:

$$\frac{\partial S_A}{\partial f_A} = -\frac{\partial S_{\text{not } A}}{\partial f_A} \tag{G.2}$$

We assume that for the major system, the *rate of change of entropy* with respect to a *change in total particle number*, or a *change in total energy* over the range $f_A = [0, 1]$ can be considered *constant*. This is because the total number of particles or the total energy in system of all other states changes by an infinitesimal fraction, since the volume V can be arbitrarily large. We will name the constants describing the rate of change of the *entropy of the major system*— $S_{\text{not } A}$— with respect to the *number of particles in the major system*— $N_{\text{not } A}$— and the *total kinetic energy of the major system*— $\varepsilon_{\text{not } A}$— as α_ε and β respectively:

$$\frac{\partial S_{\text{not } A}}{\partial N_{\text{not } A}} = \alpha_\varepsilon, \tag{G.2a}$$

$$\text{and} \qquad \frac{\partial S_{\text{not } A}}{\partial \varepsilon_{\text{not } A}} = \beta. \tag{G.2b}$$

Accordingly the rate of change of entropy of the major system with respect to a change in the occupation rate of state A— $\frac{\partial S_{\text{not } A}}{\partial f_A}$— is given as follows:

$$\begin{aligned}\frac{\partial S_{\text{not } A}}{\partial f_A} &= \frac{\partial S_{\text{not A}}}{\partial N_{\text{not } A}}\frac{\partial N_{\text{not } A}}{\partial f_A} + \frac{\partial S_{\text{not A}}}{\partial \varepsilon_{\text{not } A}}\frac{\partial \varepsilon_{\text{not } A}}{\partial f_A} \\ &= -\alpha_\varepsilon - \beta\varepsilon_A. \end{aligned} \tag{G.3}$$

To solve eq. (G.2) we now only need to determine the rate of change of entropy of the partitioned nanosystem with respect to a change in f_A. The entropy of such a nanosystem with respect to a change in occupation function can be calculated trivially from first principles. In general, the entropy S associated with a set of microstates— each of which is indexed by an integer i and has a probability of occupation p_i— is given by the following expression[3] [23]:

$$S = \sum_i^M p_i \ln p_i. \tag{G.4}$$

The nanosystem associated with A has two micro/nanostates: a full nanostate associated with probability $p_{\text{full}} = f_A$ and an empty nanostate associated with probability $p_{\text{empty}} = 1 - f_A$. Accordingly, its entropy is defined as follows:

$$S_A = (1 - f_A)\ln(1 - f_A) + f_A \ln f_A. \tag{G.5}$$

From here it is simple to calculate the rate of change of entropy of the nanosystem with respect to the occupation rate:

$$\begin{aligned}\frac{\partial S_A}{\partial f_A} &= -\frac{\partial}{\partial f_A}\Big((1 - f_A)\ln(1 - f_A) + f_A \ln f_A\Big) \\ &= \ln\frac{1 - f_A}{f_A}. \end{aligned} \tag{G.6}$$

As stated earlier, f_A is the maximum entropy occupation function if and only if the rate of change of total entropy with respect to f_A vanishes, which leads to eq. (G.2). By substituting eq. (G.3) and eq. (G.6) into this equation, we can find an expression for the occupation rate of the single carrier state A.

[3]We use the dimensionless form of entropy, which differs from the energetic form of entropy by the Boltzmann factor.

$$\ln \frac{1 - f_A}{f_A} = \alpha_\varepsilon + \beta \varepsilon_A,$$

$$\text{therefore} \qquad f_A = \frac{1}{1 + e^{\alpha_\varepsilon + \beta \varepsilon_A}}. \tag{G.7}$$

This argument has yielded the familiar Fermi-Dirac occupation function, which perhaps would be more easily recognized if we were to rename the constants as $\alpha_\varepsilon = -\frac{\mu_\varepsilon}{kT}$ and $\beta = \frac{1}{kT}$. We will not do so as we wish to retain the emphasis on the fact that these constants are *just the rates of change of the (dimensionless) entropy of a very large system with respect to a small perturbation of the total particle number or the total energy of the system*.

The Fermi-Dirac occupation function has been derived in this case because we have assumed that there are always two accessible "nanostates" associated with a single carrier state A that has been partitioned from all other carrier states, *and as such, we have not yet enforced our condition on the upper limit for the expected number of carriers above an energy* ε. Thus the occupation function we have derived is *not* the maximum entropy occupation function for *any* single carrier state but only for those partitioned single carrier states in which both the occupied state and the unoccupied state are accessible.

G.2 The Maximum Entropy Occupation Function Ansatz

We must alter the Fermi-Dirac expression in order to account for cases where an isolated nanosystem A *does not* have two accessible nanostates, due the upper limit on the expected number of carriers above a given energy. We propose doing this in a simple, intuitive manner. Essentially, we note that if the Fermi-Dirac occupation rate *does not* violate the upper limit on particle density, then it appears to be the maximum entropy rate. And if it is *does* violate the upper limit, then the occupation function should get as close to this ideal of the Fermi-Dirac occupation rate as is possible without breaking the limit.

In order to this notion of "as close as possible", we introduce the occupation function $f_{\text{lim}}(\varepsilon)$, which is defined as being equal to the negative right hand derivative of $n_{\text{lim}}(\varepsilon)$,

divided by the density of states $D(\varepsilon)$:

$$f_{\text{lim}}(\varepsilon) = -\frac{1}{D(\varepsilon)} \frac{\mathrm{d}n_{\text{lim}}}{\mathrm{d}\varepsilon^+}\Big|_\varepsilon \tag{G.8}$$

The relevance of the $f_{\text{lim}}(\varepsilon)$ is more concretely illustrated when $n_{\text{lim}}(\varepsilon)$ is written in terms of $f_{\text{lim}}(\varepsilon)$, which shows it is a kind of partner probability density to the cumulative density described by $n_{\text{lim}}(\varepsilon)$:

$$n_{\text{lim}}(\varepsilon) = \int_\varepsilon^\infty f_{\text{lim}}(\varepsilon') D(\varepsilon') \mathrm{d}\varepsilon' \tag{G.9}$$

It is also useful to define a quantity $n_{\text{spare}}(\varepsilon)$, which is positive if and only if the particle density above ε associated with the maximum entropy distribution function is below limit imposed by $n_{\text{lim}}(\varepsilon)$:

$$n_{\text{spare}}(\varepsilon) = \int_\varepsilon^\infty \Big(f_{\text{lim}}(\varepsilon') - f(\varepsilon')\Big) D(\varepsilon') \mathrm{d}\varepsilon'. \tag{G.10}$$

With this, we can propose a simple ansatz for the maximum entropy occupation function. It is simply equal to the Fermi-Dirac occupation function whenever n_{spare} is positive, and when n_{spare} is not positive, it is equal to *the closest f can get to $f_{\text{F-D}}$*, which is simply the Fermi-Dirac occupation function $f_{\text{F-D}}$ if $f_{\text{F-D}} < f_{\text{lim}}$, and the upper limit occupation function f_{lim} otherwise:

$$f(\varepsilon) = \begin{cases} \frac{1}{1+e^{\alpha_\varepsilon + \beta\varepsilon}} & \text{for } \varepsilon \in \mathfrak{e}_{\text{F-D}} = \Big\{\varepsilon \Big| \Big(n_{\text{spare}}(\varepsilon) > 0\Big) \vee \Big(f_{\text{F-D}} < f_{\text{lim}}\Big)\Big\}, \\ f_{\text{lim}}(\varepsilon) & \text{for } \varepsilon \in \mathfrak{e}_{\text{lim}} = \Big\{\varepsilon \Big| \Big(n_{\text{spare}}(\varepsilon) \leq 0\Big) \wedge \Big(f_{\text{F-D}} \geq f_{\text{lim}}\Big)\Big\}. \end{cases} \tag{G.11}$$

Here $\mathfrak{e}_{\text{F-D}}$ and $\mathfrak{e}_{\text{lim}}$ are names for the domains where the maximum entropy occupation function is equal to the Fermi-Dirac occupation function and the upper limit occupation function respectively. While this is an intuitive ansatz, we cannot be confident that it actually maximizes total entropy subject to its established constraints. The reason is simple: by choosing an occupation rate of f_{lim} at some energy ε, we ensure that on the continuous domain of connected lower energies where $f_{\text{lim}} < f_{\text{F-D}}$, we *also cannot reach* $f_{\text{F-D}}$. Accordingly, there may be cases where it is entropically beneficial to undershoot f_{lim} at some energy *so that occupation rates at lower energies can overshoot it*.

The investigation of whether this is the case is actually made quite simple by the fact that the entropy is a concave functional [23]. This means that for any given occupation function $g(\varepsilon)$ that satisfies the known statistics, there exists a sequence of small perturbations which transform $g(\varepsilon)$ to the maximum entropy occupation function, where each perturbation increases the total entropy and maintains the established constraints.[4] Accordingly, it is sufficient to determine conditions under which no energy and upper limit conserving redistribution of particles can be made that increases entropy.

We can express an arbitrary, small particle and energy preserving transformation to the maximum entropy occupation function as the *sum* of transfers of c particles (where $0 < c \leq 1$) between any two single particle states at ε_i and ε_f, that are energy balanced by *any* small redistribution of the occupation function in the domain $\mathfrak{e}_{\text{F-D}}$ that changes its energy by $-c(\varepsilon_f - \varepsilon_i)$. The reason we can leave the latter redistribution unspecified is that *all* such transformations result in the *same* change in entropy equal to $-c(\varepsilon_f - \varepsilon_i)\beta$. Since the entropy is extensive, the total entropy change associated with associated with the arbitrary, small transformation is the sum of the entropy change associated with each transformation. Accordingly, the entropy change associated with an arbitrary, small particle and energy preserving transformation can only be positive if the entropy change associated with one of these *specified* small particle and energy preserving transformations.

Suppose we express the *entire* occupation function ansatz in a Fermi-Dirac form:

$$f(\varepsilon) = \frac{1}{1 + e^{\,\alpha_\varepsilon^{\text{eff}}(\varepsilon) + \beta\varepsilon}}. \tag{G.12}$$

In this form, $\alpha_\varepsilon^{\text{eff}}(\varepsilon)$ is now the constant α_ε over the domain $\mathfrak{e}_{\text{F-D}}$, and a function of energy *bounded below by* α_ε on the domain $\mathfrak{e}_{\text{lim}}$. We can now write down the rate of change of entropy due to the two-state, energy balanced tranformation described, with respect to the transformation scale c:

[4] This statement is true because the subspace of occupation functions that satisfy the known statistics is (path) connected.

$$\frac{\partial S}{\partial c} = \overbrace{-\Big(\alpha_\varepsilon^{\text{eff}}(\varepsilon_i) + \beta\varepsilon_i\Big)}^{\text{remove } c \text{ particles from state at } \varepsilon_i} + \overbrace{\Big(\alpha_\varepsilon^{\text{eff}}(\varepsilon_f) + \beta\varepsilon_f\Big)}^{\text{add } c \text{ particles from state at } \varepsilon_f} + \overbrace{-\Big(\varepsilon_f - \varepsilon_i\Big)\beta}^{\text{remove energy change from } \mathfrak{e}_{\text{F-D}}}$$

$$= \alpha_\varepsilon^{\text{eff}}(\varepsilon_f) - \alpha_\varepsilon^{\text{eff}}(\varepsilon_i). \tag{G.13}$$

We now make the following observations:

- *If* $\varepsilon_i \in \mathfrak{e}_{\text{F-D}}$ and $\varepsilon_f \in \mathfrak{e}_{\text{F-D}}$, *then* there is no entropy change since $\alpha_\varepsilon^{\text{eff}}(\varepsilon_f) = \alpha_\varepsilon^{\text{eff}}(\varepsilon_i) = \alpha_\varepsilon$.

- *If* $\varepsilon_i \in \mathfrak{e}_{\text{F-D}}$ and $\varepsilon_f \in \mathfrak{e}_{\text{lim}}$, *then* the transformation will break the upper limit for the number of particles above $\varepsilon_f - \delta\varepsilon$.

- *If* $\varepsilon_i \in \mathfrak{e}_{\text{lim}}$ and $\varepsilon_f \in \mathfrak{e}_{\text{F-D}}$, *then* the entropy change is less than or equal to zero since $\alpha_\varepsilon^{\text{eff}}(\varepsilon_f) = \alpha_\varepsilon \leq \alpha_\varepsilon^{\text{eff}}(\varepsilon_i)$.

- *If* $\varepsilon_i \in \mathfrak{e}_{\text{lim}}$ and $\varepsilon_f \in \mathfrak{e}_{\text{lim}}$, *and* $\varepsilon_f > \varepsilon_i$, *then* the transformation will break the upper limit for the number of particles above $\varepsilon_f - \delta\varepsilon$.

- *If* $\varepsilon_i \in \mathfrak{e}_{\text{lim}}$ and $\varepsilon_f \in \mathfrak{e}_{\text{lim}}$, *and* $\varepsilon_f \leq \varepsilon_i$, *then* the entropy change *is positive if and only if* $\alpha_\varepsilon(\varepsilon_f) > \alpha_\varepsilon(\varepsilon_i)$.

Thus given the concavity of the entropy functional, we now have proved that, *if* the occupation function for a given particle density, energy density, and upper limit on tail density is of the form given in eq. (G.11), *and* $\alpha_\varepsilon^{\text{eff}}(\varepsilon)$ as derived from eq. (G.12) is a monotonically increasing function, *then* the occupation function is the maximum entropy occupation function.

As a quick aside, it may help the reader to consider this thought experiment which is yields some intuition into the proof we have outlined. Suppose every energy is associated with seperate reservoir of particles in thermal equilibrium with all other reservoirs at some temperature $T = \frac{1}{k\beta}$. Suppose each reservoir has its own chemical potential, equal to $\mu_\varepsilon^{\text{eff}}(\varepsilon) = -kT\alpha_\varepsilon^{\text{eff}}(\varepsilon)$. And suppose all the states are connected by pipes allow-

ing particle transfer, however the pipes connected to energies in $\mathfrak{e}_{\text{lim}}$ contain one-way valves that only allow net particle transfer to lower energy. If we leave such a set up to reach equilibrium, there will be particle transfer until such time that the chemical potential in all the states of $\mathfrak{e}_{\text{F-D}}$ is the same, and is greater of equal to the chemical potential in all the states of $\mathfrak{e}_{\text{lim}}$. In turn, throughout $\mathfrak{e}_{\text{lim}}$, there will be particle transfer until the chemical potential is a montonically decreasing function of energy, at which point particle transfer will stop. Such an end state is thus an equilibrium, or maximum entropy, state.

Returning to our main point, we note the fact that $\alpha_\varepsilon^{\text{eff}}(\varepsilon)$ is a monotonically increasing function implies that there exists a single crossover energy ε^*— possibly equal to zero or infinity— such that $\mathfrak{e}_{\text{F-D}} = [0, \varepsilon^*)$ and $\mathfrak{e}_{\text{lim}} = [\varepsilon^*, \infty)$. Accordingly, we can rewrite eq. (G.14) as follows:

$$f(\varepsilon) = \begin{cases} \frac{1}{1+e^{\alpha_\varepsilon + \beta\varepsilon}} & \varepsilon < \varepsilon^*, \\ f_{\text{lim}}(\varepsilon) & \varepsilon \geq \varepsilon^*. \end{cases} \tag{G.14}$$

In this thesis, we are interested in the case where the distribution is a CHEMICALLY CONSTRAINED QUASI-EQUILIBRIUM distribution, which simply means that $f_{\text{lim}}(\varepsilon)$ above ε^* is a *lattice temperature equilibrium distribution with a chemical potential* μ_ε^{max}. So finally, we note as a short corollary that for this case, so long as $\beta \leq \frac{1}{kT_L}$, $\alpha_\varepsilon^{\text{eff}}(\varepsilon)$ will be a monotonically increasing function of energy and the form given in eq. (G.14) will be a maximum entropy occupation function. Accordingly, the Chemically Constrained Quasi-Equilibrium f_{CCE} occupation function is defined as follows:

$$f_{\text{CCE}}(\varepsilon) = \begin{cases} \frac{1}{1+e^{\alpha_\varepsilon + \beta\varepsilon}} & \varepsilon < \varepsilon^*, \\ \frac{1}{1+e^{\frac{\varepsilon - \mu_\varepsilon^{\max}}{kT_L}}} & \varepsilon \geq \varepsilon^*. \end{cases} \tag{G.15}$$

In the non-degenerate limit, this simplifies to the following:

$$f_{\text{CCE}}^{\text{non-degenerate}}(\varepsilon) = \begin{cases} e^{-(\alpha_\varepsilon + \beta\varepsilon)} & \varepsilon < \varepsilon^*, \\ e^{-\frac{\varepsilon - \mu_\varepsilon^{\max}}{kT_L}} & \varepsilon \geq \varepsilon^*. \end{cases} \tag{G.16}$$

In both the degenerate and non-degenerate case, the Chemically Constrained Quasi-

Equilibrium occupation function must be continuous at the KNEE ENERGY ε^*, implying that the following relation holds:

$$\alpha_\varepsilon = \frac{\varepsilon^* - \mu_\varepsilon^{\max}}{kT_L} - \beta\varepsilon^*. \tag{G.17}$$

Glossary

a	Lattice constant
α_ε	Slotboom alpha
α_H	Thermodynamic alpha
$\hat{a}_{\mathbf{q}\eta}$	Annihilation operator for phonon at $\mathbf{q}\eta$
$\hat{a}^+_{\mathbf{q}\eta}$	Creation operator for phonon at $\mathbf{q}\eta$
a_{sph}	Lattice radius for a sphere the volume of unit cell
β	Thermodynamic beta
β_S	Inverse screening length
$\int_{\text{BZ}}$	Integral over Brillouin Zone
D	Density of states
$\mathcal{D}$	Diffusion parameter
$\mathcal{D}^\varepsilon$	Diffusion parameter at single kinetic energy
$\Delta_{\eta,\nu}$	Isotropic coupling constant (acoustic style)
$(\Delta q)_{\eta,\nu}$	Isotropic coupling constant (optical style)
$\left\langle \Delta\varepsilon_{\text{pho}} \right\rangle^\eta$	Average energy exchanged with phonons in band η

e	Fundamental charge
$\mathbf{e}$	Strain tensor
$\mathbf{E}$	Electric field vector
ϵ	Permittivity
ε^*	Knee Energy
ε_{cut}	Cutoff Energy
$\varepsilon_i^{\text{thr}}$	Impact ionization threshold energy
ε_{gap}	Band gap
ε_C	Conduction band potential energy
ε	Kinetic Energy
$\varepsilon_{\mathbf{k}\nu}$	Kinetic Energy of carrier at $\mathbf{k}\nu$
$\mathbf{F}$	External force
f	Distribution function
f_A	Antisymmetric distribution function
f_ε	Kinetic energy dependent distribution function
$f_{\text{equilibrium}}$	Thermal equilibrium distribution
$f_{\text{F-D}}$	Fermi–Dirac distribution
f_H	Total energy dependent distribution function
$f_{\text{M-B}}$	Maxwell–Boltzmann distribution
f_S	Symmetric distribution function
$f_{S/es}$	Symmetric perturbation to energy dependent distribution function
g	State multiplicity/degeneracy
$\mathbf{G}$	Reciprocal lattice vector
Γ	Number of states per unit volume of $\mathbf{k}$-space
$\mathbf{G}^*$	Reciprocal lattice vector which makes vector sum finish in first Brillouin zone
H	Total energy
$\hat{H}$	Hamiltonian operator
$\hat{H}^{\mathbf{r}}$	Hamiltonian as a function of carrier position
$\hat{H}^{\mathbf{q}}$	Fourier component of Hamiltonian as a function of carrier position
$\hat{H}_{\text{car–par}}$	Carrier–partner interaction Hamiltonian
$\hat{H}_{\text{pseudo}}$	Pseudo-Hamiltonian

$\mathscr{I}^{\mathbf{G}}$	Overlap integral
$\mathscr{I}^{\text{mod}}$	Modified overlap integral
$\mathbf{j}$	Carrier flux density
$\mathbf{j}^{\varepsilon}$	Carrier flux density per unit energy
$\mathbf{k}$	Crystal momentum/wavevector of carrier Bloch state
$\mathbf{K}$	Total initial carrier momentum minus total final carrier momentum
μ	Band index of secondary carrier Bloch state
μ_{ε}	Slotboom chemical potential (max Slotboom chemical potential of stunted electron population in Results II)
μ_H	Thermodynamic chemical potential (max thermodynamic chemical potential of stunted electron population in Results II)
n	Electron density
ν	Band index of carrier Bloch state
$n_{\text{B-E}}$	Bose-Einstein distribution function
N^i_{dop}	Density of i^{th} dopant field]
$n_{\mathbf{q}\eta}$	Numder of phonons occupying $\mathbf{q}\eta$ mode
ω	Frequency (radians)
Ω	Volume of unit cell
$\omega_{\mathbf{q}\eta}$	Frequency of $\mathbf{q}\eta$ phonons
$\omega^{\text{max}}_{\eta}$	Maximum frequency of phonons in η band
p	Hole density
$\mathbf{p}$	Crystal momentum/wavevector of secondary carrier Bloch state
ϕ	Electric potential
$\varphi^{\text{pseudo}}_{\mathbf{k}\nu}$	Pseudoeigenstate at $\mathbf{k}\nu$
$\psi_{\mathbf{k}\nu}$	Bloch wavefunction at $\mathbf{k}\nu$
$\mathbf{q}$	Crystal momentum of phonon mode
$\mathbf{r}$	Carrier position
$\mathbf{R}$	Lattice vector
ρ	Density (charge density normally, mass density if in reference to phonon scattering rate)
R_i	Rate constant for i^{th} impact ionization threshold
ρ^i	Rho functional equivalent to integral weighted by ε^i over all $\mathbf{k}\nu$ states between two given energies

S	Scattering operator with all partners
$\mathbf{S}$	Energy flux density
s_{par}	Scattering operator with particular partner, ignoring energy conservation delta function
$\mathfrak{s}_{\text{par}}$	Scattering partner state
S_{par}	Scattering operator with particular partner
σ^{ε}	Sigma functional equivalent to integral over constant energy surface
t	Time
T	Temperature
T_L	Lattice temperature
τ_{ii}	Impact ionization time
τ_{pho}^{η}	Average scattering time of phonons in band η
$\tau_{\text{relax}}^{\text{equilibrium}}$	Thermal equilibrium relaxation time
τ_{relax}	Elastic equilibrium relaxation time
τ_{scat}	Scattering time
τ^{v}	Velocity relaxation time
$u_{\mathbf{k}\nu}$	Periodic part of Bloch state at $\mathbf{k}\nu$
$\hat{\mathbf{u}}_{\mathbf{q}\eta}$	Displacement vector field operator associated with phonon state $\mathbf{q}\eta$
$\mathbf{v}$	Velocity of carrier
V	Volume
$\hat{V}_{\text{pseudo}}$	Pseudopotential operator
$\tilde{V}_{\text{pseudo}}$	Fourier components of pseudopotential
w	Energy density
X	Generic variable
$\boldsymbol{\xi}_{\eta}$	Phonon polarization vector
$\boldsymbol{\Xi}$	Deformation potential
Z_{dop}^{i}	Number of dopant protons relative to lattice ions for i^{th} dopant field

Bibliography

[1] K. Gobok. (2017, Feb.) Synopsys corporate overview for investors. [Online]. Available: https://www.synopsys.com/content/dam/synopsys/company/investor-relations/corporate-overview-investor-feb2017.pdf

[2] J.-C. Barbe, A. Benvenuti, A. Burenkov, M. Ciappa, W. Demmerle, A. Erdmann, W. Grabinski, T. Grasser, V. Huard, B. Huizing, H. Jaouen, A. Juge, A. De Keersgieter, G. Klimeck, W. Lerch, J. Lorenz, W. Molzer, V. Moroz, C. Mouli, J. Niess, Y.-S. Pang, H. Park, P. Pfäffli, P. Pichler, Y.-M. Sheu, and I. C. Yang, "International technology roadmap for semiconductors: Modeling and simulation," *ITRS Reports*, pp. 1–48, Apr. 2013.

[3] M. Pourfath, V. Sverdlov, and S. Selberherr, "Transport modeling for nanoscale semiconductor devices," in *Tenth IEEE International Conference on Solid-State and Integrated Circuit Technology*. IEEE, 2010, pp. 1737–1740.

[4] D. L. Scharfetter and H. K. Gummel, "Large-signal analysis of a silicon read diode oscillator," *IEEE Trans Electron Devices*, 1969.

[5] L. Reggiani and M. Asche, *Hot-electron transport in semiconductors*. Springer, 1985.

[6] M. V. Fischetti, S. E. Laux, and E. Crabbé, "Understanding hot-electron transport in silicon devices: Is there a shortcut?" *J Appl Phys*, vol. 78, no. 2, pp. 1058–1087, 1995.

[7] T. Grasser, C. Jungemann, H. Kosina, B. Meinerzhagen, and S. Selberherr, "Advanced transport models for sub-micrometer devices," *Proc SISPAD*, pp. 1–8, 2004.

[8] R. Stratton, "Diffusion of hot and cold electrons in semiconductor barriers," *Phys Rev*, vol. 126, no. 6, pp. 2002–&, 1962.

[9] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, "A review of hydrodynamic and energy-transport models for semiconductor device simulation," *Proc IEEE*, vol. 91, no. 2, pp. 251–274, 2003.

[10] K. Bløtekjær, "Transport equations for electrons in two-valley semiconductors," *IEEE Trans Electron Devices*, vol. ED17, no. 1, pp. 38–&, 1970.

[11] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, "Using six moments of Boltzmann's transport equation for device simulation," *J Appl Phys*, vol. 90, no. 5, pp. 2389–2396, 2001.

[12] V. Sverdlov, E. Ungersboeck, H. Kosina, and S. Selberherr, "Current transport models for nanoscale semiconductor devices," *Mater Sci Eng R*, vol. 58, no. 6, pp. 228–270, 2008.

[13] M. V. Fischetti, S. E. Laux, P. M. Solomon, and A. Kumar, "Thirty years of Monte Carlo simulations of electronic transport in semiconductors: Their relevance to science and mainstream VLSI technology," *J Comput Electron*, vol. 3, no. 3, pp. 287–293, 2004.

[14] M. Vasicek, J. Cervenka, D. Esseni, P. Palestri, and T. Grasser, "Applicability of macroscopic transport models to decananometer MOSFETs," *IEEE Trans Electron Devices*, vol. 59, no. 3, pp. 639–646, 2012.

[15] T. Grasser, R. Kosik, C. Jungemann, H. Kosina, and S. Selberherr, "Nonparabolic macroscopic transport models for device simulation based on bulk Monte Carlo data," *J Appl Phys*, vol. 97, no. 9, p. 093710, 2005.

[16] M. V. Fischetti and S. E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Phys Rev B*, vol. 38, no. 14, p. 9721, 1988.

[17] D. Vasileska and S. M. Goodnick, *Computational electronics*. Morgan & Claypool, 2005, vol. 1.

[18] H. Kosina, M. Nedjalkov, and S. Selberherr, "Theory of the Monte Carlo method for semiconductor device simulation," *IEEE Trans Electron Devices*, vol. 47, no. 10, pp. 1898–1908, 2000.

[19] T. Ma, V. Moroz, R. Borges, and L. Smith, "TCAD: Present state and future challenges," *IEDM Tech Digest*, 2010.

[20] S. E. Laux and M. V. Fischetti, "The DAMOCLES Monte Carlo device simulation program," *IBM J Res Dev*, pp. 87–92, 1991.

[21] P. Dmitruk, A. Saúl, and L. G. Reyna, "High electric field approximation to charge transport in semiconductor devices," *App Math Lett*, 1992.

[22] E. T. Jaynes, "Gibbs vs Boltzmann entropies," *Am J Phys*, vol. 33, p. 391, 1965.

[23] A. Wehrl, "General properties of entropy," *Rev Mod Phys*, vol. 50, no. 2, p. 221, 1978.

[24] E. T. Jaynes, "Where do we stand on maximum entropy," in *MIT Max Entropy Formalism Conf*, Massachusetts Institute of Technology. Springer, 1978, pp. 341–350.

[25] J. Zak, "Dynamics of electrons in solids in external fields," *Phys Rev*, vol. 168, no. 3, p. 686, 1968.

[26] N. W. Ashcroft and N. D. Mermin, *Solid state physics*. Saunders College Publishing, 1976.

[27] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys Rev*, vol. 40, no. 5, p. 749, 1932.

[28] J. E. Moyal, "Quantum mechanics as a statistical theory," *Math Proc Camb Philos Soc*, vol. 45, no. 01, pp. 99–124, 1949.

[29] C. Gérard, "Resonance theory for periodic Schrödinger operators," *Bul Soc Math France*, vol. 118, no. 1, pp. 27–54, 1990.

[30] P. A. Markowich, N. J. Mauser, and F. Poupaud, "A Wigner-function approach to (semi) classical limits: Electrons in a periodic potential," *J Math Phys*, vol. 35, p. 1066, 1994.

[31] N. Bohr, *Atomic theory and the description of nature*. Cambridge University Press, 1934, vol. 1.

[32] C. Jacoboni and P. Bordone, "The Wigner-function approach to non-equilibrium electron transport," *Rep Prog Phys*, vol. 67, no. 7, pp. 1033–1071, 2004.

[33] D. J. Bohm and D. Pines, "A collective description of electron interactions: III. Coulomb interactions in a degenerate electron gas," *Phys Rev*, vol. 92, no. 3, pp. 609–625, 1953.

[34] W. H. Zurek, "Decoherence and the transition from quantum to classical— Revisited," *Los Alamos Sci*, vol. 27, pp. 86–109, 2002.

[35] ——, "Decoherence, einselection, and the quantum origins of the classical," *Rev Mod Phys*, vol. 75, no. 3, pp. 1–61, May 2003.

[36] R. C. Tolman, *The principles of statistical mechanics*. Clarendon Press, 1938.

[37] J. C. Phillips and L. Kleinman, "New method for calculating wave functions in crystals and molecules," *Phys Rev*, vol. 116, no. 2, pp. 287–294, Oct. 1959.

[38] M. H. Cohen and V. Heine, "Cancellation of kinetic and potential energy in atoms, molecules, and solids," *Phys Rev*, vol. 122, no. 6, p. 1821, 1961.

[39] J. R. Chelikowsky and M. L. Cohen, "Electronic structure of silicon," *Phys Rev B*, vol. 10, no. 12, p. 5095, 1974.

[40] L. Onsager, "Reciprocal relations in irreversible processes. I." *Phys Rev*, vol. 37, no. 4, p. 405, 1931.

[41] ——, "Reciprocal relations in irreversible processes. II." *Phys Rev*, vol. 38, no. 12, p. 2265, 1931.

[42] B. K. Ridley, "Lucky-drift mechanism for impact ionisation in semiconductors," *J Phys C*, vol. 16, p. 3373, 1983.

[43] C. Jacoboni and L. Reggiani, "The Monte-Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev Mod Phys*, vol. 55, no. 3, pp. 645–705, 1983.

[44] J. M. Ziman, *Electrons and Phonons*. Oxford University Press, 1960.

[45] J. R. Meyer and F. J. Bartoli, "Phase-shift calculation of ionized impurity scattering in semiconductors," *Phys Rev B*, vol. 23, no. 10, pp. 5413–5427, May 1981.

[46] M. V. Fischetti, D. J. Frank, and S. E. Laux, "Monte Carlo analysis of semiconductor devices: The DAMOCLES program," *IBM J Res Dev*, vol. 34, no. 4, pp. 466–494, Jul. 1990.

[47] B. K. Ridley, "Reconciliation of the Conwell-Weisskopf and Brooks-Herring formulae for charged-impurity scattering in semiconductors: Third-body interference," *J Phys C*, vol. 10, no. 10, p. 1589, 1977.

[48] S. M. Goodnick and P. Lugli, "Effect of electron-electron scattering on nonequilibrium transport in quantum-well systems," *Phys Rev B*, vol. 37, no. 5, pp. 2578–2588,

Feb. 1988.

[49] E. A. Cartier, M. V. Fischetti, E. A. Eklund, and F. R. McFeely, "Impact ionization in silicon," *App Phys Lett*, vol. 62, no. 25, pp. 3339–3341, 1993.

[50] E. O. Kane, "Electron scattering by pair production in silicon," *Phys Rev*, vol. 159, no. 3, p. 624, 1967.

[51] L. V. Keldysh, "Concerning the theory of impact ionization in semiconductors," *Sov Phys JETP*, vol. 21, no. 6, pp. 1135–1144, 1965.

[52] C. Kittel, *Quantum Theory of Solids*, 2nd ed. John Wiley & Sons, 1987.

[53] K. Hess, *Monte Carlo device simulation: Full band and beyond*. Springer, 1991, vol. 144.

[54] G. U. Jensen, B. Lund, T. A. Fjeldly, and M. S. Shur, "Monte-Carlo simulation of semiconductor devices," *Comput Phys Commun*, vol. 67, no. 1, pp. 1–61, 1991.

[55] J. M. Higman, K. Hess, C. G. Hwang, and R. W. Dutton, "Coupled Monte Carlo-drift diffusion analysis of hot-electron effects in MOSFETs," *IEEE Trans Electron Devices*, vol. 36, no. 5, pp. 930–937, 1989.

[56] M. V. Fischetti and S. E. Laux, "Monte Carlo simulation of electron transport in Si: The first 20 years," *IBM Res Rep*, pp. 813–820, 1996.

[57] C. Herring and E. Vogt, "Transport and deformation-potential theory for many-valley semiconductors with anisotropic scattering," *Phys Rev*, vol. 101, no. 3, p. 944, 1956.

[58] S. M. Hong and C. Jungemann, "A fully coupled scheme for a Boltzmann-Poisson equation solver based on a spherical harmonics expansion," *J Comput Electron*, vol. 8, no. 3-4, pp. 225–241, 2009.

[59] C. Jungemann, A. T. Pham, B. Meinerzhagen, C. Ringhofer, and M. Bollhoefer, "Stable discretization of the Boltzmann equation based on spherical harmonics, box integration, and a maximum entropy dissipation principle," *J Appl Phys*, vol. 100, no. 2, 2006.

[60] M. C. Vecchi and M. Rudan, "Modeling electron and hole transport with full-band structure effects by means of the spherical-harmonics expansion of the BTE," *IEEE Trans Electron Devices*, vol. 45, no. 1, pp. 230–238, 1998.

[61] M. C. Vecchi and L. G. Reyna, "Generalized energy transport models for semiconductor device simulation," *Solid State Electron*, 1994.

[62] N. Ben Abdallah and P. Degond, "On a hierarchy of macroscopic models for semiconductors," *J Math Phys*, vol. 37, no. 7, pp. 3306–3333, 1996.

[63] P. Degond, "An infinite system of diffusion equations arising in transport theory: the coupled spherical harmonics expansion model," *Math Models Meth Appl Sci*, vol. 11, no. 05, pp. 903–932, 2001.

[64] E. Bringuier, "Statistical mechanics of high-field transport in semiconductors," *Phys Rev B*, vol. 52, no. 11, pp. 8092–8105, 1995.

[65] ——, "Fokker-Planck approach to nonlocal high-field transport," *Phys Rev B*, vol. 56, no. 9, p. 5328, 1997.

[66] V. I. Kolobov, "Fokker–Planck modeling of electron kinetics in plasmas and semiconductors," *Comput Mat Sci*, vol. 28, no. 2, pp. 302–320, Oct. 2003.

[67] B. Meinerzhagen, "Consistent gate and substrate current modeling based on energy transport and the lucky electron concept," *IEDM Tech Digest*, 1988.

[68] P. G. Scrobohaci and T.-W. Tang, "Modeling of the hot-electron subpopulation and its application to impact ionization in submicron silicon devices— Part II: Numerical solutions," *IEEE Trans Electron Devices*, vol. 41, no. 7, pp. 1206–1212, Jul. 1994.

[69] J. G. Ahn, C.-S. Yao, Y.-J. Park, H.-S. Min, and R. W. Dutton, "Impact ionization modeling using simulation of high-energy tail distributions," *IEEE Electron Device Lett*, vol. 15, no. 9, pp. 348–350, Sep. 1994.

[70] K. I. Sonoda, S. T. Dunham, M. Yamaji, K. Taniguchi, and C. Hamaguchi, "Impact ionization model using average energy and average square energy of distribution function," *Jpn J Appl Phys*, vol. 35, no. 2B, pp. 818–825, Feb. 1996.

[71] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, "Characterization of the hot electron distribution function using six moments," *J Appl Phys*, vol. 91, no. 6, pp. 3869–3879, 2002.

[72] K. I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. T. Dunham, "Moment expansion approach to calculate impact ionization rate in submicron silicon devices," *J Appl Phys*, vol. 80, no. 9, pp. 5444–5448, 1996.

[73] I. Bork, C. Jungemann, B. Meinerzhagen, and W. L. Engl, "Influence of heat flux on the accuracy of hydrodynamic models for ultra-short Si MOSFETs," *Proc NUPAD*, pp. 63–66, 1994.

[74] R. Thoma, A. Emunds, B. Meinerzhagen, H. J. Peifer, and W. L. Engl, "Hydrodynamic equations for semiconductors with nonparabolic band-structure," *IEEE Trans Electron Devices*, vol. 38, no. 6, pp. 1343–1353, 1991.

[75] T.-W. Tang, S. Ramaswamy, and J. Nam, "An improved hydrodynamic transport model for silicon," *IEEE Trans Electron Devices*, vol. 40, no. 8, pp. 1469–1477, 1993.

[76] S.-C. Lee and T.-W. Tang, "Transport coefficients for a silicon hydrodynamic model extracted from inhomogeneous Monte-Carlo calculations," *Solid State Electron*, vol. 35, no. 4, pp. 561–569, 1992.

[77] D. Chen, E. C. Kan, U. Ravaioli, C.-W. Shu, and R. W. Dutton, "An improved energy-transport model including nonparabolicity and non-Maxwellian distribution effects," *IEEE Electron Device Lett*, vol. 13, no. 1, pp. 26–28, 1992.

[78] M. Trovato and L. Reggiani, "Maximum entropy principle for hydrodynamic transport in semiconductor devices," *J Appl Phys*, vol. 85, no. 8, pp. 4050–4065, 1999.

[79] A. M. Anile, V. Romano, and G. Russo, "Extended hydrodynamical model of carrier transport in semiconductors," *Siam J Appl Math*, vol. 61, no. 1, pp. 74–101, 2000.

[80] E. T. Jaynes, "Information theory and statistical mechanics," *Phys Rev*, vol. 106, pp. 620–630, 1957.

[81] R. Landauer, "Statistical physics of machinery: Forgotten middle-ground," *Physica A*, vol. 194, no. 1, pp. 551–562, 1993.

[82] T. J. Bordelon, V. M. J. Agostinelli, X. Wang, C. M. Maziar, and A. F. Tasch, "Relaxation time approximation and mixing of hot and cold electron populations," *Electron Lett*, vol. 28, no. 12, Jun. 1992.

[83] D. Ventura, A. Gnudi, G. Baccarani, and F. Odeh, "Multidimensional spherical harmonics expansion of Boltzmann equation for transport in semiconductors," *App Math Lett*, 1992.

[84] A. Gnudi, D. Ventura, G. Baccarani, and F. Odeh, "Two-dimensional MOSFET simulation by means of a multidimensional spherical harmonics expansion of the Boltzmann transport equation," *Solid State Electron*, 1993.

[85] P. Degond and C. Schmeiser, "Macroscopic models for semiconductor heterostructures," *J Math Phys*, vol. 39, no. 9, pp. 4634–4663, 1998.

[86] H. Struchtrup, "Extended moment method for electrons in semiconductors," *Physica A*, vol. 275, no. 1-2, pp. 229–255, 2000.

[87] P. Degond, "Mathematical modelling of microelectronics semiconductor devices," *AMS/IP Stud Adv Math*, vol. 15, pp. 77–110, 2000.

[88] G. Gilat and L. J. Raubenheimer, "Accurate numerical method for calculating

frequency-distribution functions in solids," *Phys Rev*, vol. 144, no. 2, p. 390, 1966.

[89] H. J. Monkhorst and J. D. Pack, "Special points for Brillouin-zone integrations," *Phys Rev B*, vol. 13, no. 12, pp. 5188–5192, 1976.

[90] E. Sangiorgi, P. Palestri, D. Esseni, C. Fiegna, and L. Selmi, "The Monte Carlo approach to transport modeling in deca-nanometer MOSFETs," *Solid State Electron*, vol. 52, no. 9, pp. 1414–1423, Sep. 2008.

[91] E. Pop, "Energy dissipation and transport in nanoscale devices," *Nano Res*, vol. 3, no. 3, pp. 147–169, 2010.

[92] H. Kroemer, "On the derivation of $\hbar \mathrm{dk/dt} = \mathrm{F}$, the k-space form of Newton's law for Bloch waves," *Am J Phys*, vol. 54, no. 2, pp. 177–178, Feb. 1986.

[93] J. M. Zhang and Y. Liu, "Fermi's golden rule: Its derivation and breakdown by an ideal model," *Eur J Phys*, vol. 37, no. 6, p. 065406, Oct. 2016.