ASC Report No. 10/1010

Matrix Compression for Spherical Harmonics Expansions of the Boltzmann Transport Equation for Semiconductors

Karl Rupp, Ansgar Jüngel, Karl-Tibor Grasser

Institute for Analysis and Scientific Computing
Vienna University of Technology — TU Wien
www.asc.tuwien.ac.at ISBN 978-3-902627-03-2

Most recent ASC Reports

9/2010 *Markus Aurada, Samuel Ferraz-Leite, Dirk Praetorius*
Estimator Reduction and Convergence of Adaptive BEM

8/2010 *Robert Hammerling, Othmar Koch, Christa Simon, Ewa B. Weinmüller*
Numerical Solution of singular Eigenvalue Problems for ODEs with a Focus on Problems Posed on Semi-Infinite Intervals

7/2010 *Robert Hammerling, Othmar Koch, Christa Simon, Ewa B. Weinmüller*
Numerical Solution of Singular ODE Eigenvalue Problems in Electronic Structure Computations

6/2010 *Markus Aurada, Michael Feischl, Dirk Praetorius*
Convergence of Some Adaptive FEM-BEM Coupling

5/2010 *Marcus Page, Dirk Praetorius*
Convergence of Adaptive FEM for some Elliptic Obstacle Problem

4/2010 *Ansgar Jüngel, Josipa-Pina Milišić*
A Simplified Quantum Energy-Trans port Model for Semiconductors

3/2010 *Georg Kitzhofer, Othmar Koch, Gernot Pulverer, Christa Simon, Ewa Weinmüller*
The New MATLAB Code bvpsuite for the Solution of Singular Implicit BVPs

2/2010 *Maike Löhndorf, Jens Markus Melenk*
Wavenumber-explicit hp-BEM for High Frequency Scattering

1/2010 *Jens Markus Melenk*
Mapping Properties of Combined Field Helmholtz Boundary Integral Operators

49/2009 *Markus Aurada, Jens Markus Melenk, Dirk Praetorius*
Mixed Conforming Elements for the Large-Body Limit in Micromagnetics

Institute for Analysis and Scientific Computing
Vienna University of Technology
Wiedner Hauptstraße 8–10
1040 Wien, Austria

E-Mail: admin@asc.tuwien.ac.at
WWW: http://www.asc.tuwien.ac.at
FAX: +43-1-58801-10196

ISBN 978-3-902627-03-2

ASC TU WIEN

Matrix Compression for Spherical Harmonics Expansions of the Boltzmann Transport Equation for Semiconductors

K. Rupp[a], A. Jüngel[b], T. Grasser[c]

[a] *Christian Doppler Laboratory for Reliability Issues in Microelectronics at the Institute for Microelectronics, TU Wien Gußhausstraße 27–29/E360, A-1040 Wien, Austria*
[b] *Institute for Analysis and Scientific Computing, TU Wien Wiedner Hauptstraße 8-10/E101, A-1040 Wien, Austria*
[c] *Institute for Microelectronics, TU Wien Gußhausstraße 27–29/E360, A-1040 Wien, Austria*

Abstract

We investigate the numerical approximation of the semiconductor Boltzmann transport equation using an expansion of the distribution function in spherical harmonics. A complexity analysis shows that traditional implementations of higher order spherical harmonics expansions suffer from huge memory requirements, especially for two and three dimensional devices. To overcome these complexity limitations, a compressed matrix storage scheme using Kronecker products is proposed, which reduces the memory requirements for the storage of the system matrix significantly. Furthermore, the total memory requirements are asymptotically dominated only by the memory required for the unknowns. We discuss the increased importance of the selection of an appropriate linear solver and show that execution times for matrix-vector multiplications using the compressed matrix scheme are even smaller than those for an uncompressed system matrix. Numerical results demonstrate the applicability of our method and confirm our theoretical results.

Keywords: Boltzmann equation, Spherical Harmonics, Kronecker product

Email addresses: rupp@iue.tuwien.ac.at (K. Rupp), juengel@asc.tuwien.ac.at (A. Jüngel), grasser@iue.tuwien.ac.at (T. Grasser)

1. Introduction

While in the early years of the semiconductor industry macroscopic models such as the drift-diffusion model or the hydrodynamic model have been sufficient for device simulation, accurate simulations of modern nanoscale devices require the use of more precise models. As long as quantum mechanical effects in transport direction are not dominant, the microscopic electron transport may be described by the Boltzmann Transport Equation (BTE), which may be considered to be the most appropriate semi-classical description of electrons in a semiconductor.

A direct solution of the BTE has been pursued for several decades and many ingenious techniques have been developed for this purpose. However, direct solution approaches are limited by the high dimensionality of the problem: Three spatial dimensions and three momentum dimensions lead to a six-dimensional problem already for stationary simulations, thus only coarse grids can be used for direct solutions [1, 2]. Therefore, the most commonly used technique is the non-deterministic Monte Carlo method, primarily because it is very flexible and allows one to incorporate modeling details such as complicated band structures and scattering processes. The main disadvantage of the Monte Carlo method is its computational cost, especially when attempting to reduce the statistical noise in the low density tails of the distribution function [3, 4].

As an alternative to the stochastic Monte Carlo method and high-dimensional direct approaches, the deterministic spherical harmonics expansion method of first order was introduced in the early 1990s for one-dimensional devices [5, 6]. Later, the method has been extended to arbitrary expansion order [3, 7] and two spatial dimensions [8, 9, 10, 11]. Furthermore, numerous contributions from the physics point of view [12, 13, 14, 15, 16] and some results from the mathematics point of view [17, 18, 19, 20, 21, 22] are available. However, there are only a few contributions on improvements of the treatment of the discrete system of equations [23, 24].

The major challenge for SHE is the huge – but much smaller than other direct methods – memory need reported already for two-dimensional devices [25]. The reason is that the model contains an additional energy variable leading to an increased set of space-energy grid points $(\boldsymbol{x}, \varepsilon)$ and spherical expansion coefficients. In particular, for three-dimensional devices, this requires the discretization in a four-dimensional $(\boldsymbol{x}, \varepsilon)$-space with a tuple of unknowns associated with each grid point, which is out of reach even for

modern computers. In current implementations, most of the required memory is used for the storage of the global system matrix of size $M \times M$. In this paper we propose a method to reduce the memory required by the system matrix such that most of the memory is actually consumed by the M unknowns of the system. On a machine with 8 Gigabytes of memory, this allows us to store 10^9 unknowns, which is by two orders of magnitude higher than the largest SHE simulations reported so far [25]. This method paves the way for real spatial three-dimensional simulations.

This work is organized as follows: We briefly review the derivation of the SHE equations in Sec. 2. In Sec. 3 we show that the unknown expansion coefficients are only weakly coupled, which leads to a very sparse system matrix for the discretized equations. The decoupling of spherical harmonics expansion coefficient interactions from the underlying discretization is used in the main section of this work (Sec. 4) to derive a matrix compression scheme based on sums of Kronecker products, which reduces the memory requirements for the system matrix considerably. Sec. 5 shows how non-spherical bands can be incorporated into the matrix compression scheme, while Sec. 6 deals with the inclusion of stabilization schemes. The selection of appropriate linear solvers is discussed in Sec. 7. Numerical results are given in Sec. 8, confirming our theoretical results. Finally, we conclude in Sec. 9.

2. SHE of the BTE

We briefly sketch the equations resulting from a SHE of the BTE, following the derivation given in more detail by Jungemann et. al. [26]. Here and in the following, function arguments are suppressed whenever appropriate to increase the readability of the equations. The electron distribution is described by a distribution function $f(\boldsymbol{x}, \boldsymbol{k}, t)$, where $\boldsymbol{x} \in \mathbb{R}^3$ is the position in real space, $\boldsymbol{k} \in \mathbb{R}^3$ is the wave vector and $t > 0$ is the time. The distribution function is assumed to fulfill the BTE

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f + \frac{1}{\hbar} \boldsymbol{F} \cdot \nabla_{\boldsymbol{k}} f = Q\{f\} \ ,$$

where $\boldsymbol{v} = \nabla_{\boldsymbol{k}} \varepsilon / \hbar$ is the group velocity induced by the band energy $\varepsilon(\boldsymbol{k})$ (relative to its minimum) and $\boldsymbol{F} = -\nabla_{\boldsymbol{x}}(\mathrm{q}\psi + \varepsilon_{\mathrm{b}})$ is the effective force acting on a particle with charge $\mathrm{q} = \pm e$ (where e is the modulus of the electron charge and the positive sign refers to holes and the negative one to electrons)

induced by the quasi-static potential ψ and the band edge ε_b. The scattering operator Q is assumed to be linear and given by

$$Q\{f\} = \frac{\Omega_\mathrm{s}}{(2\pi)^3} \int s(\boldsymbol{x}, \boldsymbol{k}', \boldsymbol{k}) f(\boldsymbol{x}, \boldsymbol{k}', t) - s(\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{k}') f(\boldsymbol{x}, \boldsymbol{k}, t) \, \mathrm{d}\boldsymbol{k} \ ,$$

where Ω_s denotes a sample volume. According to Fermi's Golden Rule, the scattering terms are assumed to be of the form

$$s(\boldsymbol{x}, \boldsymbol{k}', \boldsymbol{k}) = \frac{1}{\Omega_\mathrm{s}} \sum_\eta c_\eta(\boldsymbol{x}, \boldsymbol{k}', \boldsymbol{k}) \delta\Big(\varepsilon(\boldsymbol{k}) - \varepsilon(\boldsymbol{k}') \pm \hbar\omega_\eta\Big) \ ,$$

where we have assumed for simplicity that the energy transfer $\hbar\omega_\eta$ for each scattering process η does not depend on the initial and final wave vector.

For reasons of numerical stability it is advantageous to define the generalized distribution function [26]

$$g(\boldsymbol{x}, \varepsilon, \theta, \varphi, t) = 2Z(\varepsilon, \theta, \varphi) f(\boldsymbol{x}, \boldsymbol{k}(\varepsilon, \theta, \varphi), t) \ ,$$

where the generalized density of states Z for one spin direction is given by

$$Z(\varepsilon, \theta, \varphi) = \frac{|\boldsymbol{k}|^2}{(2\pi)^3} \frac{\partial |\boldsymbol{k}|}{\partial \varepsilon} \ .$$

In the following it is assumed that the mapping $\varepsilon \mapsto \boldsymbol{k}$ is a bijection, otherwise a spherical harmonics expansion can not be carried out on equi-energy surfaces.

We expand the generalized distribution function into orthonormal and real valued spherical harmonics $Y_{l,m}(\theta, \varphi)$, and truncate after $(L+1)^2$ terms:

$$g(\boldsymbol{x}, \varepsilon, \theta, \varphi, t) \approx \sum_{l=0}^{L} \sum_{m=-l}^{l} g_{l,m}(\boldsymbol{x}, \varepsilon, t) Y_{l,m}(\theta, \varphi) \ . \tag{1}$$

The expansion coefficients are obtained from the generalized distribution function by the projections

$$\begin{aligned} g_{l,m}(\boldsymbol{x}, \varepsilon, t) &= \int Y_{l,m}(\theta, \varphi) g(\boldsymbol{x}, \varepsilon, \theta, \varphi, t) \, \mathrm{d}\Omega \\ &= 2 \int Y_{l,m}(\theta, \varphi) Z(\varepsilon, \theta, \varphi) f(\boldsymbol{x}, \boldsymbol{k}(\varepsilon, \theta, \varphi), t) \, \mathrm{d}\Omega \ , \end{aligned}$$

where the integration is carried out over the unit sphere, Ω is the solid angle and $\mathrm{d}\Omega = \sin\theta\,\mathrm{d}\theta\mathrm{d}\varphi$. Equations for the coefficients $g_{l,m}$ are directly obtained from a projection of the BTE, resulting in

$$\frac{\partial g_{l,m}}{\partial t} + \frac{\partial \boldsymbol{F}\cdot\boldsymbol{j}_{l,m}}{\partial \varepsilon} + \nabla_{\boldsymbol{x}}\cdot\boldsymbol{j}_{l,m} - \boldsymbol{F}\cdot\boldsymbol{\Gamma}_{l,m} = Q_{l,m}\{g\}\ , \tag{2}$$

where the generalized current density

$$\boldsymbol{j}_{l,m}(\boldsymbol{x},\varepsilon,t) = \int \boldsymbol{v} g Y_{l,m}\,\mathrm{d}\Omega \tag{3}$$

and the angular force coupling term

$$\boldsymbol{\Gamma}_{l,m}(\boldsymbol{x},\varepsilon,t) = \int \frac{1}{\hbar|\boldsymbol{k}|}\Big(\frac{\partial Y_{l,m}}{\partial\theta}\boldsymbol{e}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_{l,m}}{\partial\varphi}\boldsymbol{e}_\varphi\Big) g\,\mathrm{d}\Omega \tag{4}$$

have been introduced, and $\boldsymbol{e}_\theta$ and $\boldsymbol{e}_\varphi$ denote the angular unit vectors. The projection of the scattering operator $Q_{l,m}\{g\}$ is detailed below. We substitute (1) into (3) and (4) and then substitute these into (2). Using Einstein's summation convention, we obtain the system of partial differential equations

$$\frac{\partial g_{l,m}}{\partial t} + \frac{\partial \boldsymbol{F}\cdot\boldsymbol{v}_{l,m}^{l',m'} g_{l',m'}}{\partial\varepsilon} + \boldsymbol{v}_{l,m}^{l',m'}\cdot\nabla_{\boldsymbol{x}} g_{l',m'} - \boldsymbol{F}\cdot\boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'} = Q_{l,m}\{g\} \tag{5}$$

for all $l = 0,\ldots,L$, $m = -l,\ldots,l$, where

$$\boldsymbol{v}_{l,m}^{l',m'}(\varepsilon) = \int \boldsymbol{v} Y_{l,m} Y_{l',m'}\,\mathrm{d}\Omega\ , \tag{6}$$

$$\boldsymbol{\Gamma}_{l,m}^{l',m'}(\varepsilon) = \int \frac{1}{\hbar|\boldsymbol{k}|}\Big(\frac{\partial Y_{l,m}}{\partial\theta}\boldsymbol{e}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_{l,m}}{\partial\varphi}\boldsymbol{e}_\varphi\Big) Y_{l',m'}\,\mathrm{d}\Omega\ . \tag{7}$$

Prior to projection of the scattering operator, we split $Q\{f\} = Q^{\mathrm{in}}\{f\} - Q^{\mathrm{out}}\{f\}$, where

$$Q^{\mathrm{in}}\{f\} = \frac{\Omega_{\mathrm{s}}}{(2\pi)^3}\int s(\boldsymbol{x},\boldsymbol{k}',\boldsymbol{k}) f(\boldsymbol{x},\boldsymbol{k}',t)\,\mathrm{d}\boldsymbol{k}'\ ,$$

$$Q^{\mathrm{out}}\{f\} = \frac{\Omega_{\mathrm{s}}}{(2\pi)^3}\int s(\boldsymbol{x},\boldsymbol{k},\boldsymbol{k}') f(\boldsymbol{x},\boldsymbol{k},t)\,\mathrm{d}\boldsymbol{k}\ .$$

Under the assumption of velocity randomizing scattering rates [4], a spherical harmonics projection leads to [26]

$$
\begin{aligned}
Q_{l,m}^{\mathrm{in}}(\boldsymbol{x},\varepsilon,t) &= \sum_{\eta} s_{l,m;\eta}^{l',m';\mathrm{in}}\, g_{l',m'}(\boldsymbol{x},\varepsilon \mp \hbar\omega_\eta,t) \ , \\
s_{l,m;\eta}^{l',m';\mathrm{in}}(\boldsymbol{x},\varepsilon) &= \frac{1}{Y_{0,0}} Z_{l,m}(\varepsilon) c_\eta(\boldsymbol{x},\varepsilon \pm \hbar\omega_\eta,\varepsilon)\delta_{0,l'}\delta_{0,m'} \ ,
\end{aligned} \tag{8}
$$

and

$$
\begin{aligned}
Q_{l,m}^{\mathrm{out}}(\boldsymbol{x},\varepsilon,t) &= s_{l,m}^{l',m';\mathrm{out}}\, g_{l',m'}(\boldsymbol{x},\varepsilon,t) \ , \\
s_{l,m}^{l',m';\mathrm{out}}(\boldsymbol{x},\varepsilon) &= \frac{1}{Y_{0,0}} \sum_{\eta} Z_{0,0}(\varepsilon \mp \hbar\omega_\eta) c_\eta(\boldsymbol{x},\varepsilon,\varepsilon \pm \hbar\omega_\eta)\delta_{l,l'}\delta_{m,m'} \ ,
\end{aligned} \tag{9}
$$

where δ denotes the Kronecker delta, the upper and lower signs refer to scattering to higher and lower energies respectively, and

$$
Z_{l,m} = \int_{\Omega} Z(\varepsilon,\theta,\varphi) Y_{l,m}\, \mathrm{d}\Omega \ . \tag{10}
$$

Substitution of the projected scattering terms into (5) yields the full system of partial differential equations

$$
\begin{aligned}
&\frac{\partial g_{l,m}}{\partial t} + \frac{\partial \boldsymbol{F}\cdot \boldsymbol{v}_{l,m}^{l',m'} g_{l',m'}}{\partial \varepsilon} + \boldsymbol{v}_{l,m}^{l',m'} \cdot \nabla_{\boldsymbol{x}} g_{l',m'} - \boldsymbol{F}\cdot \boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'} = \\
&\qquad = \sum_{\eta} s_{l,m;\eta}^{l',m';\mathrm{in}} g_{l',m'}(\boldsymbol{x},\varepsilon \mp \hbar\omega_\eta,t) - s_{l,m}^{l',m';\mathrm{out}} g_{l',m'}(\boldsymbol{x},\varepsilon,t)
\end{aligned} \tag{11}
$$

for all $l = 0,\ldots,L$ and $m = -l,\ldots,l$.

In the case of several energy bands, a BTE has to be written for each band and scattering rates between these subbands have to be given. In the following we assume a single energy band only. This allows us to keep the expressions simpler, but it does not imply that our approach is limited to a single energy band only.

3. Sparse Coupling for Spherical Bands

The representation (11) obscures the physical interpretation of the individual terms, but it exposes the full coupling structure. If all coupling coefficients $\boldsymbol{v}_{l,m}^{l',m'}$, $\boldsymbol{\Gamma}_{l,m}^{l',m'}$, $s_{l,m;\eta}^{l',m';\mathrm{in}}$ and $s_{l,m}^{l',m';\mathrm{out}}$ were multiples of the Kronecker delta

$\delta_{l,l'}\delta_{m,m'}$, all equations would be decoupled and could be solved individually. Conversely, nonzero coupling coefficients for all quadruples (l, m, l', m') indicate a tight coupling, which usually complicates the solution process. This is in analogy to systems of linear equations: If the system matrix is diagonal, the solution is found immediately, but if the matrix is dense, typically a lot of computational effort is required to solve the system.

According to (8) and (9), the scattering coefficients $s_{l,m;\eta}^{l',m';\mathrm{in}}$ and $s_{l,m}^{l',m';\mathrm{out}}$ vanish except for the case that $l' = m' = 0$ or $l = l'$, $m = m'$, respectively. This leads to a very weak coupling: The first term couples all differential equations with $g_{0,0}$, while the second term does not couple any equations at all. Moreover, under the assumption of spherical bands, the generalized density of states is spherically symmetric, hence $Z_{l,m} \equiv 0$ for $(l, m) \neq (0, 0)$. Consequently, in this case, the scattering terms do not couple any unknowns. The remainder of this section is thus devoted to the investigation of the couplings induced by $\boldsymbol{v}_{l,m}^{l',m'}$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ (see (6) and (7)).

For general band structures, the symmetry of the underlying processes leads to the following result.

Theorem 1 (Jungemann et. al.)**.** *For a spherical harmonics expansion up to order $L = 2I + 1$ with $I \in \mathbb{N}$, there holds for all $i, i' \in \{0, \ldots, I\}$, $m \in \{-i, \ldots, i\}$ and $m' \in \{-i', \ldots, i'\}$*

$$\boldsymbol{v}_{2i,m}^{2i',m'} = \boldsymbol{v}_{2i+1,m}^{2i'+1,m'} = \boldsymbol{0}, \quad \boldsymbol{\Gamma}_{2i,m}^{2i',m'} = \boldsymbol{\Gamma}_{2i+1,m}^{2i'+1,m'} = \boldsymbol{0} \ .$$

The essence of this theorem is that all nonzero coupling coefficients possess different parities in the leading indices. This minor structural information about the coupling was already used for a preprocessing step for the solution of the discretized equations in [26].

Under the assumption of spherical energy bands, i.e. $\varepsilon(\boldsymbol{k}) = \tilde{\varepsilon}(|\boldsymbol{k}|)$, the velocity $\boldsymbol{v}$, the modulus of the wave vector $|\boldsymbol{k}|$ and the generalized density of states only depend on the energy ε, but not on the angles θ, φ. Consequently, we rewrite

$$\boldsymbol{v}_{l,m}^{l',m'}(\varepsilon) = v(\varepsilon) \int Y_{l,m} \boldsymbol{e}_\varepsilon Y_{l',m'} \, \mathrm{d}\Omega =: v(\varepsilon) \boldsymbol{a}_{l,m}^{l',m'} \ , \tag{12}$$

$$\boldsymbol{\Gamma}_{l,m}^{l',m'}(\varepsilon) = \frac{1}{\hbar|\boldsymbol{k}|} \int \Big(\frac{\partial Y_{l,m}}{\partial \theta} \boldsymbol{e}_\theta + \frac{1}{\sin\theta} \frac{\partial Y_{l,m}}{\partial \varphi} \boldsymbol{e}_\varphi \Big) Y_{l',m'} \, \mathrm{d}\Omega =: \frac{1}{\hbar|\boldsymbol{k}|} \boldsymbol{b}_{l,m}^{l',m'} \ . \tag{13}$$

The coupling between index pairs (l, m) and (l', m') is determined by the integral terms $\boldsymbol{a}_{l,m}^{l',m'}$ and $\boldsymbol{b}_{l,m}^{l',m'}$ only. It turns out that the coupling is rather weak:

Theorem 2. *For spherical energy bands, the following holds true for indices* $l, l' \in \{0, \dots, L\}$, $m \in \{-l, \dots, l\}$ *and* $m' \in \{-l', \dots, l'\}$*:*

1. *If* $\boldsymbol{v}_{l,m}^{l',m'}$ *is nonzero, then* $l \in \{l' \pm 1\}$ *and* $m \in \{\pm|m'| \pm 1, m'\}$.
2. *If* $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ *is nonzero, then* $l \in \{l' \pm 1\}$ *and* $m \in \{\pm|m'| \pm 1, m'\}$.

The proof is given in Appendix B; it makes use of recurrence relations and orthogonalities of trigonometric functions and associated Legendre functions. Theorem 2 is very important for large order expansions: The total number of unknown expansion coefficients is $(L+1)^2$, but according to (8), (9) and (11), each $g_{l,m}$ is directly coupled with at most ten other coefficients. The weak coupling stated in Theorem 2 has already been observed for less general situations in earlier publications [3, 9].

4. Matrix Compression

In this section we investigate the discretization of the projected SHE system (11) for spherical energy bands. Substitution of (12) and (13) into (11) yields

$$\begin{aligned} \frac{\partial g_{l,m}}{\partial t} + \boldsymbol{a}_{l,m}^{l',m'} \cdot \left[\boldsymbol{F} \frac{\partial v g_{l',m'}}{\partial \varepsilon} + v \nabla_{\boldsymbol{x}} g_{l',m'} \right] - \boldsymbol{b}_{l,m}^{l',m'} \cdot \boldsymbol{F} \frac{g_{l',m'}}{\hbar |\boldsymbol{k}|} \\ = \sum_{\eta} s_{l,m;\eta}^{l',m';\mathrm{in}} g_{l',m'}(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t) - s_{l,m}^{l',m';\mathrm{out}} g_{l',m'} \ . \end{aligned} \tag{14}$$

Let us consider a spatial discretization for $g_{l,m}$ in the $(\boldsymbol{x}, \varepsilon)$ plane using a finite element or finite volume scheme: We select a space of trial functions U with basis $(\varphi_i)_{i=1}^N$ and a space of test functions V with basis $(\chi_j)_{j=1}^N$, making the usual assumption of equal dimensionality of the two spaces. The particular choice of these spaces depends on the particular finite element or finite volume scheme, but it does not affect the next steps. Moreover, a compression scheme for finite difference methods is obtained analogously by taking suitable limits in the choice of basis functions in the distributional sense.

A weak form of (14) is derived as usual by multiplication with a test function and integration over the whole domain. We make the ansatz

$$g_{l,m} = \sum_{i=1}^{N} \alpha_{i;l,m}(t) \varphi_i(\boldsymbol{x}, \varepsilon) \ ,$$

so that we have to solve for the $N \times (L+1)^2$ unknowns $\alpha_{i;l,m}(t)$. For the numbering of the unknowns, we introduce the mapping

$$\begin{aligned} &\pi_L : \mathbb{N} \times \{0, \ldots, L\} \times \{-L, \ldots, L\} \to \mathbb{N} \ , \\ &(i, l, m) \mapsto i(L+1)^2 + l^2 + l + m \ , \end{aligned} \tag{15}$$

which is a bijection for $-l \leq m \leq l$. Also other mappings of the form $(i, l, m) \mapsto i(L+1)^2 + \kappa(l, m)$ for a bijection κ from admissible values for l, m into the set $\{0, \ldots, (L+1)^2 - 1\}$ can be used. In the following, we set $\kappa(l, m) = l^2 + l + m$.

Similar to finite element and finite volume methods, we define the matrix valued bilinear mapping $\boldsymbol{w} : U \times V \to \mathbb{R}^{(L+1)^2 \times (L+1)^2}$ by

$$\begin{aligned} \Big(\boldsymbol{w}(\varphi, \chi)\Big)_{\kappa(l,m),\kappa(l',m')} = \int \Bigg[&\frac{\partial \varphi}{\partial t} \delta_{l,l'} \delta_{m,m'} + \sum_{l',m'} \boldsymbol{a}_{l,m}^{l',m'} \cdot \Big(\boldsymbol{F} \frac{\partial v \varphi}{\partial \varepsilon} + v \nabla_{\boldsymbol{x}} \varphi \Big), \\ &- \sum_{l',m'} \boldsymbol{b}_{l,m}^{l',m'} \cdot \boldsymbol{F} \frac{\varphi}{\hbar |\boldsymbol{k}|} + \sum_{l',m'} s_{l,m}^{l',m';\mathrm{out}} \varphi \\ &- \sum_{l',m';\eta} s_{l,m;\eta}^{l',m';\mathrm{in}} \varphi(\boldsymbol{x}, \varepsilon \mp \hbar \omega_\eta, t) \Bigg] \chi \, \mathrm{d}\boldsymbol{x} \mathrm{d}\varepsilon \ , \end{aligned} \tag{16}$$

where the integration is carried out over the simulation domain. Depending on the actual discretization method, the integral terms may be rearranged using integration by parts, but this does not affect the following arguments. In the above definition of the bilinear mapping, the time derivative may be discretized by a backward Euler scheme or omitted when considering steady states.

With the numbering (15), the system matrix for the discrete system is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{w}(\varphi_1, \chi_1) & \boldsymbol{w}(\varphi_2, \chi_1) & \ldots & \boldsymbol{w}(\varphi_N, \chi_1) \\ \boldsymbol{w}(\varphi_1, \chi_2) & \boldsymbol{w}(\varphi_2, \chi_2) & \ldots & \boldsymbol{w}(\varphi_N, \chi_2) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}(\varphi_1, \chi_N) & \boldsymbol{w}(\varphi_2, \chi_N) & \ldots & \boldsymbol{w}(\varphi_N, \chi_N) \end{pmatrix} , \tag{17}$$

which is the common matrix structure for Galerkin methods such as the finite element method. Moreover, the sparsity of $\boldsymbol{S}$ becomes now apparent: If the intersection of the support of φ_i and χ_j is empty (taking into account shifts by $\pm\hbar\omega_\eta$ along the energy axis coming from the scattering operator),

the full block $\boldsymbol{w}(\varphi_i, \chi_j)$ vanishes, see (16). Note that, in general, $\boldsymbol{w}(\varphi_i, \chi_j) \neq \boldsymbol{w}(\varphi_j, \chi_i)$ and therefore $\boldsymbol{S}$ is not symmetric, which must be taken into account for the selection of a proper linear solver.

For the complexity analysis, we introduce the following notation.

Definition 1. *Given a triangulation $\mathcal{T}$ and trial and test spaces U, V with basis $(\varphi)_{i=1}^N$ and $(\chi)_{j=1}^N$, respectively, we define the* sparsity indicator

$$C_{\text{sparse}} := \max_{\chi \in \{\chi_1, \dots, \chi_N\}} \left| \left\{ \varphi \in \{\varphi_1, \dots, \varphi_N\} \mid \exists (\boldsymbol{x}, \varepsilon) \in G \ : \right. \right. \\ \left. \left. \varphi\chi \neq 0 \text{ or } \exists \eta \ : \ \varphi(\boldsymbol{x}, \varepsilon \pm \hbar\omega_\eta)\chi \neq 0 \right\} \right| ,$$

where the notation $|A|$ denotes the number of elements of the set A and G is the simulation domain in the $(\boldsymbol{x}, \varepsilon)$-space.

From the definition of C_{sparse} we directly see that there are at most C_{sparse} blocks in each row of the block structure (17) of the system matrix $\boldsymbol{S}$. In the following we assume that the triangulations are sufficiently regular such that C_{sparse} does not increase when the mesh is refined. With Landau's notation, we assume that $C_{\text{sparse}} = \mathcal{O}(1)$. This allows us to show the following statement about memory requirements.

Theorem 3. *Assume spherical energy bands, a spherical harmonics expansion up to degree L and a discretization of the $(\boldsymbol{x}, \varepsilon)$-domain using N degrees of freedom. Then it holds:*

1. *A straightforward assembly of the matrix $\boldsymbol{S}$, defined in* (17)*, needs a storage of $C_{\text{sparse}} N (L+1)^4$ entries.*
2. *There exists an assembly of $\boldsymbol{S}$ needing a storage of $11 C_{\text{sparse}} N \ (L+1)^2$ entries only.*

Proof. The matrix $\boldsymbol{S}$ is of size $N(L+1)^2 \times N(L+1)^2$. In each of the N rows of the block structure (17) there are at most C_{sparse} blocks. Each block is of dimension $(L+1)^2 \times (L+1)^2$, hence there are at most $C_{\text{sparse}} N (L+1)^4$ nonzero entries in $\boldsymbol{S}$, which proves the first statement.

Since each block in the block structure is sparse due to Thm. 2, (8) and (9), each block carries at most $11(L+1)^2$ nonzero entries, thus there are in fact at most $11 C_{\text{sparse}} N (L+1)^2$ nonzero entries in $\boldsymbol{S}$. □

It has been observed in [26] that the expansion order L should be at least nine in order to obtain good agreement with Monte Carlo simulations. Taking $L = 9$ for demonstration purposes, we see that a straightforward assembly leads to $10{,}000 C_{\text{sparse}} N$ entries, whereas the number of nonzero entries is at most $1{,}100 C_{\text{sparse}} N$ entries, thus more than 90 percent of the memory is wasted in a straightforward assembly.

Even though a careful assembly reduces the required memory by an order of magnitude, total memory requirements of at most $11 C_{\text{sparse}} N (L+1)^2$ entries are still very large. Compared to a linear finite element or finite volume scheme for the Poisson equation, the coupling between the expansion coefficients in the SHE equations requires an additional factor of $11(L+1)^2$ to the storage need. Since $L \geq 9$, we have to take into account an additional factor of at least 1100. This leads to huge memory requirements for two-dimensional devices and makes the simulation of three-dimensional devices using the SHE model impossible so far. In the following, we derive a matrix compression scheme that requires much less memory.

Writing $\boldsymbol{a}_{l,m}^{l',m'}$, $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ and $\boldsymbol{F}(\boldsymbol{x})$ in components,

$$\boldsymbol{a}_{l,m}^{l',m'} = \begin{pmatrix} (\boldsymbol{a}_{l,m}^{l',m'})_1 \\ (\boldsymbol{a}_{l,m}^{l',m'})_2 \\ (\boldsymbol{a}_{l,m}^{l',m'})_3 \end{pmatrix}, \quad \boldsymbol{b}_{l,m,l',m'} = \begin{pmatrix} (\boldsymbol{b}_{l,m}^{l',m'})_1 \\ (\boldsymbol{b}_{l,m}^{l',m'})_2 \\ (\boldsymbol{b}_{l,m}^{l',m'})_3 \end{pmatrix}, \quad \boldsymbol{F}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{F}_1(\boldsymbol{x}) \\ \boldsymbol{F}_2(\boldsymbol{x}) \\ \boldsymbol{F}_3(\boldsymbol{x}) \end{pmatrix},$$

a rearrangement of (16) leads to the following nine integrals

$$\begin{aligned} \Big(\boldsymbol{w}(\varphi,\chi)\Big)_{\kappa(l',m'),\kappa(l,m)} = \quad & \delta_{l,l'}\delta_{m,m'} \int \frac{\partial \varphi}{\partial t} \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \\ & + \sum_{l',m'} \int s_{l,m}^{l',m';\text{out}} \varphi \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \\ & - \sum_{l',m',\eta} \int s_{l,m;\eta}^{l',m';\text{in}} \varphi(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t) \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \\ & + \sum_{p=1}^{3} \sum_{l',m'} (\boldsymbol{a}_{l,m}^{l',m'})_p \int \Big(\boldsymbol{F}_p \frac{\partial v\varphi}{\partial \varepsilon} + v \frac{\partial \varphi}{\partial (\boldsymbol{x})_p} \Big) \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \\ & - \sum_{p=1}^{3} \sum_{l',m'} (\boldsymbol{b}_{l,m}^{l',m'})_p \int \boldsymbol{F}_p \frac{\varphi}{\hbar|\boldsymbol{k}|} \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \,. \end{aligned} \tag{18}$$

The crucial observation is that after substitution of (8) and (9) into (18), all summands are products in which one factor only depends on l, m, l' and m',

and the other factor involves the integrals and depends only on the indices of φ_j and χ_i. In particular, the full system matrix (17) can be written as

$$\boldsymbol{S} = \sum_{i=1}^{9} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i \ , \tag{19}$$

where $\otimes$ denotes the Kronecker product (cf. Appendix A for the definition). The spatial discretization matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_9$ are given by

$$(\boldsymbol{Q}_1)_{i,j} = \int \frac{\partial \varphi_j}{\partial t} \chi_i \, \mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \ , \tag{20}$$

$$(\boldsymbol{Q}_2)_{i,j} = \frac{1}{Y_{0,0}} \int Z_{0,0}(\varepsilon \mp \hbar\omega_\eta) c_\eta(\varepsilon, \varepsilon \pm \hbar\omega_\eta) \varphi_j \chi_i \, \mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \ , \tag{21}$$

$$(\boldsymbol{Q}_3)_{i,j} = -\frac{1}{Y_{0,0}} \int Z_{0,0} c_\eta(\varepsilon \pm \hbar\omega_\eta, \varepsilon) \varphi_j(\boldsymbol{x}, \varepsilon \pm \hbar\omega_\eta, t) \chi_i \, \mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \ , \tag{22}$$

$$(\boldsymbol{Q}_p)_{i,j} = \int \Big[\boldsymbol{F}_{p-1} \frac{\partial v \varphi_j}{\partial \varepsilon} + v \frac{\partial \varphi_j}{\partial x_{p-1}} \Big] \chi_i \, \mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \quad p = 4, 5, 6 \ , \tag{23}$$

$$(\boldsymbol{Q}_p)_{i,j} = -\int \boldsymbol{F}_{p-4} \frac{\varphi_j}{\hbar|\boldsymbol{k}|} \chi_i \, \mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon \ , \quad p = 7, 8, 9 \ , \tag{24}$$

and the coupling matrices $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_9$ by

$$\big(\boldsymbol{R}_1\big)_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'} \delta_{m,m'} \ , \tag{25}$$

$$\big(\boldsymbol{R}_2\big)_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'} \delta_{m,m'} \ , \tag{26}$$

$$\big(\boldsymbol{R}_3\big)_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'} \delta_{m,m'} \delta_{l,0} \delta_{m,0} \ , \tag{27}$$

$$\big(\boldsymbol{R}_p\big)_{\kappa(l,m),\kappa(l',m')} = (\boldsymbol{a}_{l,m}^{l',m'})_{p-3} \ , \quad p = 4, 5, 6 \ , \tag{28}$$

$$\big(\boldsymbol{R}_p\big)_{\kappa(l,m),\kappa(l',m')} = (\boldsymbol{b}_{l,m}^{l',m'})_{p-6} \ , \quad p = 7, 8, 9 \ . \tag{29}$$

Hence, we can represent the full system matrix $\boldsymbol{S}$, which has up to $11 C_{\mathrm{sparse}} N (L+1)^2$ nonzero entries, by nine matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_9$ (with at most $C_{\mathrm{sparse}} N$ entries each) and nine matrices $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_9$ (with at most $4(L+1)^2$ entries each due to the fact that for given (l, m), each component of $\boldsymbol{a}_{l,m}^{l',m'}$ and $\boldsymbol{a}_{l,m}^{l',m'}$ couples with at most four other pairs (l', m')). Since the matrices $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ do not need to be stored at all, we can store $\boldsymbol{S}$ in a compressed form using $24(L+1)^2 + 9 C_{\mathrm{sparse}} N$ entries only. As $(L+1)^2$ is for two- and three-dimensional devices typically much smaller than the degree of freedom

N, the total memory requirements for $\boldsymbol{S}$ can be reduced down to the order $9C_{\text{sparse}}N = \mathcal{O}(N)$. This leads to the situation that the number of unknowns $N(L+1)^2$ is the only limitation with respect to memory for large order expansions. Even in the case of very high order expansions such as $L = 19$ we can still use 312,500 grid nodes in $(\boldsymbol{x}, \varepsilon)$-space in order to fit all unknowns into one gigabyte of memory in double precision.

5. Non-spherical bands

The matrix compression described in the previous section relies on the factorizations (12) and (13) of the coupling terms $\boldsymbol{v}_{l,m}^{l',m'}(\varepsilon)$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}(\varepsilon)$, whose factors depend on the energy or on the indices l, m, l' and m'. In the case of non-spherical bands, the velocity and the modulus of the wave vector depend on the energy *and* on the angles.

In order to decouple the radial (energy) contributions from the angular ones, we perform a spherical projection up to order L' of the coupled terms in the integrands by approximating

$$\boldsymbol{v}(\varepsilon, \theta, \varphi) \approx \sum_{l''=0}^{L'} \sum_{m''=-l''}^{l''} \boldsymbol{v}^{l'',m''}(\varepsilon) Y_{l'',m''}(\theta, \varphi) \ , \tag{30}$$

$$\frac{1}{\hbar|\boldsymbol{k}(\varepsilon, \theta, \varphi)|} \approx \sum_{l''=0}^{L'} \sum_{m''=-l''}^{l''} \Gamma^{l'',m''}(\varepsilon) Y_{l'',m''}(\theta, \varphi) \ , \tag{31}$$

where the expansion coefficients are given by

$$\boldsymbol{v}^{l'',m''}(\varepsilon) = \int \boldsymbol{v}(\varepsilon, \theta, \varphi) Y_{l'',m''}(\theta, \varphi) \, \mathrm{d}\Omega \ ,$$

$$\Gamma^{l'',m''}(\varepsilon) = \int \frac{1}{\hbar|\boldsymbol{k}(\varepsilon, \theta, \varphi)|} Y_{l'',m''}(\theta, \varphi) \, \mathrm{d}\Omega \ .$$

For simplicity, the expansion order L' is the same for both $\boldsymbol{v}_{l,m}^{l',m'}$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}$. It depends on the complexity of the band structure, but values of $L' = 1$ or $L' = 2$ should usually be sufficient to obtain a good approximation of the non-spherical bands of interest.

Substitution of the expansions (30) and (31) into (6) and (7) yields

$$\begin{aligned}
\boldsymbol{v}_{l,m}^{l',m'} &= \boldsymbol{v}^{l'',m''}(\varepsilon) \int Y_{l,m} Y_{l',m'} Y_{l'',m''} \,\mathrm{d}\Omega =: \boldsymbol{v}^{l'',m''}(\varepsilon) a_{l,m;l'',m''}^{l',m'} \ , \\
\boldsymbol{\Gamma}_{l,m}^{l',m'} &= \Gamma^{l'',m''}(\varepsilon) \int \Big(\frac{\partial Y_{l,m}}{\partial \theta} \boldsymbol{e}_\theta + \frac{1}{\sin\theta} \frac{\partial Y_{l,m}}{\partial \varphi} \boldsymbol{e}_\varphi \Big) Y_{l',m'} Y_{l'',m''} \,\mathrm{d}\Omega \\
&=: \Gamma^{l'',m''}(\varepsilon) \boldsymbol{b}_{l,m;l'',m''}^{l',m'} \ ,
\end{aligned}$$

so that we have to deal with a sum of $(L'+1)^2$ decoupled terms in contrast to the case of spherical bands, where the sum degenerates to a single term. Repeating the steps from the previous section, the system matrix $\boldsymbol{S}$ can be written similar to (19) in the form

$$\boldsymbol{S} = \sum_{i=1}^{3+6(L'+1)^2} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i \ . \tag{32}$$

Some of the coupling matrices $\boldsymbol{R}_4, \ldots, \boldsymbol{R}_{3+6(L'+1)^2}$ are sparse: If $l'' = m'' = 0$, $Y_{l'',m''}$ is a constant and the sparsity is assured by Thm. 2. For coupling matrices involving $a_{l,m;l'',m''}^{l',m'}$, with row indices $\kappa(l,m)$ and column indices $\kappa(l',m')$ for each pair (l'',m''), the entries are directly obtained from the Wigner 3jm-symbols, cf. Appendix Appendix C. The sparsity of the coupling matrices, arising from $\boldsymbol{b}_{l,m;l'',m''}^{l',m'}$ in the same way as for $a_{l,m;l'',m''}^{l',m'}$, is not clear at present, but we presume that the structure is similar. Since the total memory required for the coupling matrices induced by $\boldsymbol{b}_{l,m;l'',m''}^{l',m'}$ is still negligible even if they are dense, we assume for simplicity dense spherical harmonics coupling matrices, so $(L+1)^4$ memory is required for each. With this, the system matrix can be stored using at most $[3+6(L'+1)^2][(L+1)^4 + C_{\text{sparse}}N] = \mathcal{O}(L'^2(L^4+N))$ matrix entries. In typical applications, $L' \ll L$ and $L^4 \ll N$, so the total memory requirement is still dominated by the storage of the NL^2 unknowns.

6. Stabilization Schemes

Due to the strong gradients in the distribution function and the large numerical range of values, spurious oscillations in the numerical approximation show up if no stabilization scheme is applied [3, 26]. For very small devices, a combination of staggered grids, the maximum entropy dissipation scheme (MEDS) [20] and the H-transform [8] was reported by Hong et. al. [25] to

yield stable numerical results. In the following we extend our matrix compression scheme such that it can be used with these stabilization schemes.

For staggered grids, unknowns associated with spherical harmonics of even order are associated with different basis than unknowns associated with odd order spherical harmonics. Consequently, for the even order unknowns we select a space of trial functions U^{even} with basis $(\varphi_i^{\text{even}})_{i=1}^N$ and a space of test functions V^{even} with basis $(\chi_j^{\text{even}})_{j=1}^N$. Similarly, a space of trial functions U^{odd} with basis $(\varphi_i^{\text{odd}})_{i=1}^N$ and a space of test functions V^{odd} with basis $(\chi_j^{\text{odd}})_{j=1}^N$ is chosen for the odd order harmonics. The total trial space is $U = U^{\text{even}} \cup U^{\text{odd}}$ and the test space $V = V^{\text{even}} \cup V^{\text{odd}}$.

Moreover, we first enumerate the even order unknowns and test functions and then the odd order unknowns and test functions. Unknowns associated with the same trial function carry are enumerated consecutivly similar to (15). Repeating the steps in Sec. 4, the full system matrix $\boldsymbol{S}$ can be written in the block-structure

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}^{\text{ee}} & \boldsymbol{S}^{\text{eo}} \\ \boldsymbol{S}^{\text{oe}} & \boldsymbol{S}^{\text{oo}} \end{pmatrix} = \sum_{i=1}^{p} \begin{pmatrix} \boldsymbol{Q}_i^{\text{ee}} \otimes \boldsymbol{R}_i^{\text{ee}} & \boldsymbol{Q}_i^{\text{eo}} \otimes \boldsymbol{R}_i^{\text{eo}} \\ \boldsymbol{Q}_i^{\text{oe}} \otimes \boldsymbol{R}_i^{\text{oe}} & \boldsymbol{Q}_i^{\text{oo}} \otimes \boldsymbol{R}_i^{\text{oo}} \end{pmatrix} . \tag{33}$$

The even-to-even coupling matrix $\boldsymbol{S}^{\text{ee}}$ and the odd-to-odd coupling matrix $\boldsymbol{S}^{\text{oo}}$ are square matrices and determined according to Thm. 1 or Thm. 2 only by the projected time derivative $\partial g_{l,m}/\partial t$ and the projected scattering operator $Q_{l,m}\{g\}$. The even-to-odd coupling matrix $\boldsymbol{S}^{\text{eo}}$ is non-square and determined by the free-streaming operator with sparsity pattern given by Thm. 2. The odd-to-even coupling matrix $\boldsymbol{S}^{\text{oe}}$ is also non-square and determined by the free-streaming operator and for non-spherical bands also by the scattering operator $Q_{l,m}\{g\}$, cf. (8).

The spatial matrices $\boldsymbol{Q}_i^{\text{ee}}$, $\boldsymbol{Q}_i^{\text{eo}}$, $\boldsymbol{Q}_i^{\text{eo}}$ and $\boldsymbol{Q}_i^{\text{oo}}$ in (33) are obtained by evaluating the underlying bilinear mapping for trial functions from U^{even} and U^{odd} and test functions from V^{even} and V^{odd} respectively. Similarly, the spherical coupling matrices $\boldsymbol{R}_i^{\text{ee}}$, $\boldsymbol{R}_i^{\text{eo}}$, $\boldsymbol{R}_i^{\text{eo}}$ and $\boldsymbol{R}_i^{\text{oo}}$ are obtained by taking only the rows and columns of $\boldsymbol{R}_i$ that correspond to even or odd harmonics respectively.

Since the coupling structure of the scattering operator is explicitly given in (8) and (9), the structure of $\boldsymbol{S}^{\text{ee}}$ and $\boldsymbol{S}^{\text{oo}}$ is as follows:

Theorem 4. *For spherical harmonics expansions in steady state, the following statements for staggered grids hold true:*

1. *The matrix* $\boldsymbol{S}^{\mathrm{oo}}$ *is diagonal.*
2. *For spherical energy bands without considering inelastic scattering,* $\boldsymbol{S}^{\mathrm{ee}}$ *is also diagonal.*

This structural information is very important for the construction of solution schemes in the next section.

To employ the H-transform, variables are changed from $(\boldsymbol{x}, \varepsilon)$ to $(\tilde{\boldsymbol{x}}, H)$ by the transformation

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} \ , \qquad\qquad H = \varepsilon + \mathrm{q}\psi(\boldsymbol{x}) \ ,$$

where ψ denotes the electrostatic potential and q is the charge of the carriers (negative for electrons, positive for holes). Since this transformation effects only the $(\boldsymbol{x}, \varepsilon)$-space, the decouplings (12) and (13) are unchanged and the proposed matrix compression scheme can be applied. Clearly, the expressions (20) to (24) for the spatial matrices $\boldsymbol{Q}_i$ have to be adapted due to the application of the H-transform, but can be derived in analogy to the derivation in Sec. 4.

Similarly, an application of MEDS modifies the odd order equations only and does not interfere with the decoupling given by (12) and (13). Thus, the entries in $\boldsymbol{Q}_i^{\mathrm{oe}}$ and $\boldsymbol{Q}_i^{\mathrm{oo}}$ as in (33) are modified, but the matrix compression scheme can be applied without additional difficulties.

7. Solution of the Linear System

In the previous sections we have introduced a matrix compression scheme. However, such a scheme is of use only if the resulting scheme can be solved without recovering the full matrix again. Such a reconstruction is, in principle, necessary if direct solvers such as the Gauss algorithm are used, because the matrix structure is altered in a way that destroys the block structure. For many popular iterative solvers from the family of Krylov methods, it is usually sufficient to provide matrix-vector multiplications. Consequently, we first discuss methods to compute the matrix-vector product $\boldsymbol{S}\boldsymbol{x}$ for a given vector $\boldsymbol{x}$ in the case that the system matrix $\boldsymbol{S}$ is given in the compressed form

$$\boldsymbol{S} = \sum_{i=1}^{p} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i \ .$$

The number of summands p and the entries of $\boldsymbol{Q}_i$ and $\boldsymbol{R}_i$ depend on the underlying band structure and discretization schemes as discussed in the previous sections.

It is well known that a row-by-row reconstruction of the compressed matrix $\boldsymbol{S}$ is not efficient. Therefore, we decompose the vector $\boldsymbol{x}$ into N blocks of size $(L+1)^2$ by

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix} = \sum_{j=1}^{(L+1)^2} \boldsymbol{e}_j \otimes \boldsymbol{x}_j \ , \tag{34}$$

where $\boldsymbol{e}_j$ is the j-th column vector of the identity matrix. The matrix-vector product can now be written as

$$\boldsymbol{S}\boldsymbol{x} = \Big[\sum_{i=1}^{p} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i\Big]\Big[\sum_{j=1}^{(L+1)^2} \boldsymbol{e}_j \otimes \boldsymbol{x}_j\Big] = \sum_{i=1}^{p} \sum_{j=1}^{(L+1)^2} (\boldsymbol{Q}_i \boldsymbol{e}_j) \otimes (\boldsymbol{R}_i \boldsymbol{x}_j) \ .$$

The product $\boldsymbol{Q}_i \boldsymbol{e}_j$ is simply the j-th column of $\boldsymbol{Q}_i$. The computation of $\boldsymbol{R}_i \boldsymbol{x}_j$ requires $\mathcal{O}(C_{\text{sparse}} N)$ additions and multiplications. Building the Kronecker products and adding nonzero entries to the resulting vector requires $4N$ operations for each index pair (i, j). Thus, $\mathcal{O}((4 + C_{\text{sparse}}) p N L^2) \approx \mathcal{O}(C_{\text{sparse}} p N L^2)$ additions and multiplications are needed in total, since in typical situations $C_{\text{sparse}} \gg 4$.

For spherical energy bands ($p = 9$), the matrix-vector multiplication requires slightly less computational effort than the uncompressed case, where the scalar prefactor is 11. Thus, the proposed matrix compression reduces both the computational effort and memory requirements. Non-spherical bands lead to larger values of p as discussed in Sec. 5, thus leading to a higher computational effort for the matrix-vector multiplications compared to the uncompressed case. Nevertheless, the additional computational effort is increased only moderately, while the memory requirements are significantly reduced.

Due to the coupling structure, recent publications report the elimination of odd order unknowns in a preprocessing step [25, 26]. Moreover, it has been shown that for first order expansion the system matrix after elimination of odd order unknowns is an M-matrix [25]. Moreover, numerical experiments indicate a considerable improvement in the convergence of iterative solvers.

For a matrix structure as given by (33), a direct elimination of odd order unknows would destroy the representation of the system matrix $\boldsymbol{S}$ as a sum of Kronecker products. Writing the system as

$$\boldsymbol{S}\boldsymbol{g} = \begin{pmatrix} \boldsymbol{S}^{\mathrm{ee}} & \boldsymbol{S}^{\mathrm{eo}} \\ \boldsymbol{S}^{\mathrm{oe}} & \boldsymbol{S}^{\mathrm{oo}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}^{\mathrm{e}} \\ \boldsymbol{g}^{\mathrm{o}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}^{\mathrm{e}} \\ \boldsymbol{r}^{\mathrm{o}} \end{pmatrix} \tag{35}$$

with the vector of unknowns $\boldsymbol{g}$ split into $\boldsymbol{g}^{\mathrm{e}}$ and $\boldsymbol{g}^{\mathrm{e}}$ as unknowns associated with even and odd order harmonics respectively and analogously for the right hand side vector $\boldsymbol{r}$, the elimination of odd order unknowns is carried out using the Schur complement:

$$(\boldsymbol{S}^{\mathrm{ee}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{S}^{\mathrm{oe}})\boldsymbol{g}^{\mathrm{e}} = \boldsymbol{r}^{\mathrm{e}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{r}^{\mathrm{o}} \ . \tag{36}$$

Since $\boldsymbol{S}^{\mathrm{oo}}$ is according to Thm. 4 a diagonal matrix, the inverse is directly available. The other matrix-vector products are carried out as discussed in the beginning of this section.

In contrast to matrix-vector multiplication with the full system matrix $\boldsymbol{S}$, where the proposed matrix compression scheme requires approximately the same computational effort, a matrix-vector multiplication with the condensed matrix $(\boldsymbol{S}^{\mathrm{ee}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{S}^{\mathrm{oe}})$ is more expensive than a matrix-vector multiplication with a fully set up condensed matrix. To estimate the additional effort, we assume that the number of even spherical harmonics is equal to the number of odd spherical harmonics and is given by $(L+1)^2/2$, which is a good approximation for $L \geq 5$. Since $\boldsymbol{S}^{\mathrm{ee}}$ is diagonal or at least close to diagonal, the most computational effort is needed for the computation of $\boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{S}^{\mathrm{oe}}\boldsymbol{g}^{\mathrm{e}}$. Neglecting the cost of inverting the diagonal matrix $\boldsymbol{S}^{\mathrm{oo}}$, the operation boils down to the computation of two matrix-vector products. Summing up, a runtime penalty for matrix vector multiplication of a factor slightly above two is expected.

The total memory needed for the SHE equations is essentially given by the memory required for the unknowns, which adds another perspective on the selection of the iterative solver. From (16) we see that the system matrix $\boldsymbol{S}$ is not symmetric, since $\boldsymbol{\Gamma}_{l,m}^{l',m'} \neq \boldsymbol{\Gamma}_{l',m'}^{l,m}$. Moreover, numerical experiments indicate that the matrix $\boldsymbol{S}$ is indefinite, thus many popular solvers cannot be used. A popular solver for indefinite problems is GMRES [27, 28]. It is typically restarted after, say, s steps, denoted by GMRES(s). This method was used in recent publications on SHE simulations [25, 26]. For a system with N' unknowns, the memory required during the solution process is $\mathcal{O}(sN')$.

In typical applications, in which the system matrix is uncompressed, this additional memory is approximately the amount of memory needed for the storage of the system matrix; thus, it is not a major concern. However, using the proposed matrix compression scheme, the memory needed for the unknowns is dominant, so the additional memory for GMRES(s) directly pushes the overall memory requirements to $\mathcal{O}(sNL^2)$. The number of steps s is typically chosen between 20 and 30 as smaller values have been reported to lead to slower convergence rates. Hence, we conclude that GMRES(s) might be too expensive for SHE simulations. Instead, iterative solvers with smaller memory consumption such as BiCGStab [29] should be used.

8. Numerical Results

In the preceding sections we have derived asymptotic memory requirements for large expansion orders L and high numbers of spatial degrees of freedom N with $L^2 \ll N$. In this section we report the CPU times observed from our in-house SHE simulator running on a single core of a machine with a Core 2 Quad 9550 CPU.

All simulations were carried out for a stationary two-dimensional device on a regular staggered grid with $5 \times 50 \times 50$ nodes in $(\boldsymbol{x}, H)$-space for various expansion orders. We assumed spherical energy bands and applied the H-transform and MEDS for stabilization. A fixed potential distribution was applied to the device to obtain comparable results. For self-consistency with the Poisson equation using a Newton scheme, similar results can in principle be obtained by application of the matrix compression scheme to the Jacobian.

First we compared memory requirements for the storage of the system matrix. We extracted the total number of entries stored in the matrix, multiplied by three to account for row and column indices and assumed 8 bytes per entry. In this way, the influence of different sparse matrix storage schemes is eliminated. The results in Tab. 1 and Fig. 1 clearly demonstrate the asymptotic advantage of our approach: Already at an expansion order of $L = 5$, memory savings of a factor of 18 are observed. At $L = 13$, this factor grows to 145. In particular, the memory requirement for the matrix compression scheme shows only a weak dependence on L, which is due to the additional memory needed for the coupling matrices $\boldsymbol{R}_i$ in (25)-(29). The memory required at $L = 1$ is essentially determined by the degrees of freedom in the $(\boldsymbol{x}, H)$-space. With increasing expansion order L, the additional memory requirements for the compressed scheme grow quadratically with L (because

L	$\boldsymbol{S}$	$\sum \boldsymbol{Q}_i \otimes \boldsymbol{R}_i$	Unknowns
1	3.7 MB	4.7 MB	0.2 MB
3	28.4 MB	4.7 MB	1.4 MB
5	83.1 MB	4.7 MB	3.5 MB
7	168 MB	4.8 MB	6.6 MB
9	263 MB	4.8 MB	10.7 MB
11	470 MB	4.8 MB	15.7 MB
13	709 MB	4.9 MB	21.6 MB

Table 1: Memory requirements for the uncompressed and the compressed system matrix compared to the memory needed for the unknowns for different expansion orders L on a grid in the three-dimensional $(\boldsymbol{x}, H)$-space with $5 \times 50 \times 50$ nodes.

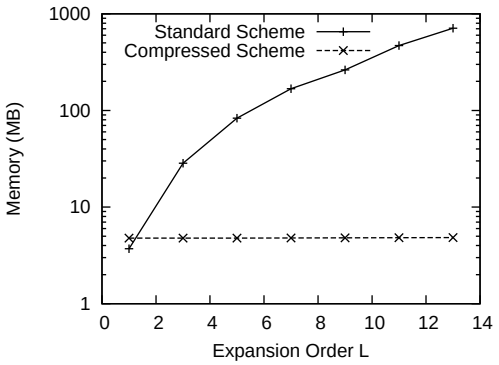

Figure 1: Memory used for the uncompressed and the compressed system matrix for different expansion orders L on a three-dimensional $(\boldsymbol{x}, H)$-grid with 12.500 nodes.

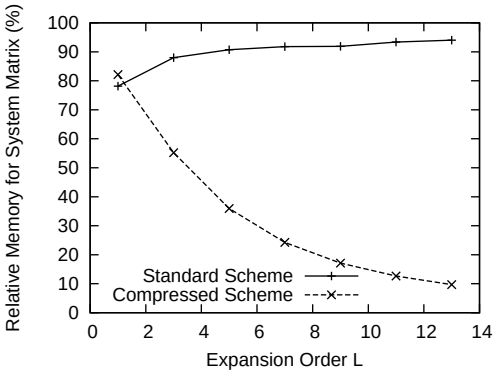

Figure 2: Memory used for the system matrix in relation to the total amount of memory used (i.e. system matrix, unknowns and right hand side).

	Full system		Condensed	
L	$\boldsymbol{S}$	compr.	$\boldsymbol{S}_{\text{cond}}$	compr.
1	3.9	7.4	0.2	9.2
3	28.4	19.3	4.0	17.9
5	73.9	53.2	15.7	48.9
7	134.8	98.3	36.5	92.2
9	228.1	160.7	68.2	149.8

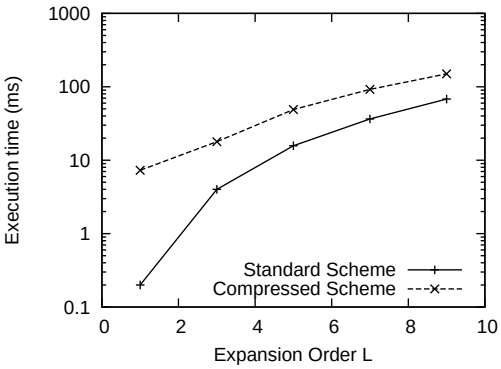

Figure 3: Comparison of execution times (milliseconds) for matrix-vector multiplication at different expansion orders L for the fully set up system matrix and the proposed compressed matrix scheme. Both the full system of linear equations and the condensed system with odd order unknowns eliminated in a preprocessing step are compared.

there are $(L+1)^2$ spherical harmonics of degree smaller or equal to L), but even at $L = 13$ the additional memory compared to $L = 1$ is less than one megabyte. Moreover, the memory used for the unknowns dominates even for moderate values of L, cf. Fig. 2.

In order to quantify the impact of the matrix compression on the runtime performance of iterative solvers, a comparison of execution times for the matrix-vector multiplications was carried out, cf. Fig. 3. We compared execution times for the full system matrix and the condensed system matrix, where unknowns associated with odd order spherical harmonics had been eliminated. For the lowest expansion order $L = 1$, matrix compression does not pay off, the execution times are by a factor of two larger. This is due to the additional structural overhead of the compressed scheme at expansion order $L = 1$, where no compression effect occurs. However, for larger values of L, the matrix compression scheme leads to faster matrix-vector multiplications with the full system of linear equations as predicted in Sec. 7. The predicted asymptotic performance gain of a factor slightly above one can readily be seen.

Comparing execution times for the condensed system, where odd order unknowns have been eliminated in a preprocessing step, the runtime penalty for matrix-vector multiplication is a factor of 15 at $L = 1$, but in this case there is no compression effect anyway. At $L = 5$, the runtime penalty is only a factor of three and drops to slightly above two at $L = 13$. Better caching possibilities and less limitations due to memory bandwidth appear to be the cause for the smaller relative differences in execution times at higher expansion orders.

As discussed in Sec. 7, GMRES leads to higher memory requirements than

L	GMRES(50)	GMRES(30)	GMRES(10)	BiCGStab	Unknowns
1	10.2 MB	6.2 MB	2.2 MB	1.2 MB	0.2 MB
3	71.4 MB	43.4 MB	15.4 MB	8.4 MB	1.4 MB
5	178.5 MB	108.5 MB	38.5 MB	21.0 MB	3.5 MB
7	336.6 MB	204.7 MB	72.6 MB	39.6 MB	6.6 MB
9	545.7 MB	331.7 MB	117.7 MB	64.2 MB	10.7 MB
11	800.7 MB	486.7 MB	172.7 MB	93.5 MB	15.7 MB
13	1101.6 MB	669.6 MB	237.6 MB	129.6 MB	21.6 MB

Table 2: Additional memory requirements of the linear solvers GMRES(s) with different values of s and BiCGStab compared to the memory needed for the unknowns.

many other Krylov methods such as BiCGStab. A comparison of additional memory required by GMRES(50), GMRES(30), GMRES(10) and BiCGStab is shown in Tab. 2 and Fig. 4. For GMRES(s), our implementation used $s + 1$ auxiliary vectors of the same length as the vector of unknowns, while BiCGStab uses six auxiliary vectors of that size. It can clearly be seen that the memory required by GMRES(50) is by one order of magnitude larger than the memory needed for the compressed system (i.e. second and third column in Tab. 1) and BiCGStab. On the other hand, without system matrix compression, the additional memory needed by GMRES(50) is comparable to the memory needed for the system matrix and is thus less of a concern.

9. Conclusions

We have investigated the coupling structure of the SHE equations and shown the weak coupling of the expansion coefficients. This guarantees that the total memory requirements for the storage of the system matrix, obtained from a discretization with N degrees of freedom in $(\boldsymbol{x}, \varepsilon)$-space and SHE order L, is of order $\mathcal{O}(NL^2)$ in contrast to $\mathcal{O}(NL^4)$ that would be required for a dense coupling. Since $L \geq 9$ have been reported to be needed for sufficiently accurate results, the memory savings are significant compared to straightforward implementations.

The matrix compression scheme presented in this work further reduces the memory requirements for the system matrix from order $\mathcal{O}(NL^2)$ to $\mathcal{O}(N+L^2)$ at only slightly increased runtime efficiency of matrix-vector multiplications.

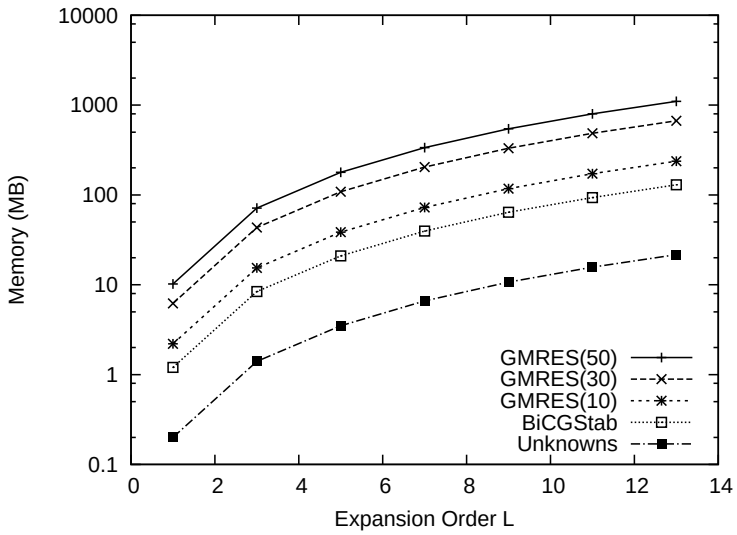

Figure 4: Additional memory requirements of the linear solvers GMRES(s) with different values of s and BiCGStab compared to the memory needed for the unknowns.

While the huge memory requirements for the storage of the full system matrix prohibited the simulation of three-dimensional devices so far, our proposed scheme paves the way for such simulations even for sufficiently large expansion order L. Assuming a $50 \times 50 \times 50 \times 50$ grid in $(\boldsymbol{x}, H)$-space for the simulation of a three-dimensional device, approximately 400 MB of memory is required at lowest expansion order $L = 1$ for the storage of the unknowns only. This amount is proportional to $(L + 1)^2$, hence with expansion order $L = 9$, roughly 10 GB of memory is needed for the storage of the unknowns only. Using the proposed matrix compression scheme and BiCGStab as linear solver, this would result in a total memory footprint of around 60 GB, which is available already on workstations today. Without matrix compression scheme, the memory needed for the system matrix would then be approximately 1 TB, which is certainly out of reach on mainstream computers. Execution times certainly increase with the number of unknowns, but since the proposed matrix compression scheme is also attractive for parallelization, execution times are expected to be reasonably small.

Furthermore, we have shown that the memory requirements of the chosen linear solver affects the total memory needs for SHE simulations using the proposed matrix compression scheme much more than in many other circumstances. A comparison between GMRES and BiCGStab shows that a

careless choice can increase the total memory consumption by up to an order of magnitude.

Appendix A. The Kronecker Product

For matrices $\boldsymbol{Q} = (Q_{i,j})_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{R} \in \mathbb{R}^{p\times q}$, the Kronecker product is defined as the block matrix

$$\boldsymbol{Q}\otimes\boldsymbol{R} = \begin{pmatrix} Q_{1,1}\boldsymbol{R} & Q_{1,2}\boldsymbol{R} & \dots & Q_{1,m-1}\boldsymbol{R} & Q_{1,m}\boldsymbol{R} \\ Q_{2,1}\boldsymbol{R} & Q_{2,2}\boldsymbol{R} & \dots & Q_{2,m-1}\boldsymbol{R} & Q_{2,m}\boldsymbol{R} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Q_{n-1,1}\boldsymbol{R} & Q_{n-1,2}\boldsymbol{R} & \dots & Q_{n-1,m-1}\boldsymbol{R} & Q_{n-1,m}\boldsymbol{R} \\ Q_{n,1}\boldsymbol{R} & Q_{n,2}\boldsymbol{R} & \dots & Q_{n,m-1}\boldsymbol{R} & Q_{n,m}\boldsymbol{R} \end{pmatrix} \in \mathbb{R}^{np\times mq} .$$

The Kronecker product is bilinear and associative, but not commutative. Moreover, if the matrices $\boldsymbol{Q}$, $\boldsymbol{R}$, $\boldsymbol{S}$ and $\boldsymbol{T}$ are such that the products $\boldsymbol{QS}$ and $\boldsymbol{RT}$ can be formed, there holds

$$(\boldsymbol{Q}\otimes\boldsymbol{R})(\boldsymbol{S}\otimes\boldsymbol{T}) = (\boldsymbol{QS})\otimes(\boldsymbol{RT}) .$$

Appendix B. Sparsity of Coupling Coefficients

To prove the sparsity of $\boldsymbol{v}_{l,m}^{l',m'}$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ as stated in Thm. 2, it is sufficient to prove the sparsity for the integrals $\boldsymbol{a}_{l,m}^{l',m'}$ and $\boldsymbol{b}_{l,m}^{l',m'}$ as defined in (12) and (13). We give a proof for the first components $(\boldsymbol{a}_{l,m}^{l',m'})_1$ and $(\boldsymbol{b}_{l,m}^{l',m'})_1$ only, the proof for the second and third components follows the same arguments and is thus omitted. We note that the spherical harmonics are given by

$$Y_{l,m}(\theta,\varphi) = N_{l,m}P_l^{|m|}(\cos\theta)\times\begin{cases}\cos(m\varphi), & m>0 ,\\ 1, & m=0 ,\\ \sin(m\varphi), & m<0 ,\end{cases}$$

where $N_{l,m}$ denotes a suitable normalization constant and $P_l^{|m|}$ is an associated Legendre function. Substitution of the definition of spherical harmonics and splitting the integral leads to

$$\begin{aligned}(\boldsymbol{a}_{l,m}^{l',m'})_1 &= N_{l,m}N_{l',m'}\int_0^\pi P_l^{|m|}(\cos\theta)\sin^2\theta P_{l'}^{|m'|}(\cos\theta)\,\mathrm{d}\theta \\ &\quad\times\int_0^{2\pi}\cos(\varphi)\times\begin{Bmatrix}\cos(m\varphi), & m>0\\ 1, & m=0\\ \sin(m\varphi), & m<0\end{Bmatrix}\times\begin{Bmatrix}\cos(m'\varphi), & m'>0\\ 1, & m'=0\\ \sin(m'\varphi), & m'<0\end{Bmatrix}\mathrm{d}\varphi .\end{aligned}$$

The orthogonality of the trigonometric functions $\cos(m\varphi)$ and $\cos(m'\varphi)$ shows that $(\boldsymbol{a}_{l,m}^{l',m'})_1$ vanishes if $m' \neq m \pm 1$. Thus, it is sufficient to consider the case $m' = m \pm 1$. We distinguish two cases:

First, let $|m'| = |m| - 1$. Then we write

$$\begin{aligned}(\boldsymbol{a}_{l,m}^{l',m'})_1 &= A_{l,m,l',m'} \int_0^\pi P_l^{|m|}(\cos\theta) \sin^2\theta P_{l'}^{|m|-1}(\cos\theta)\,\mathrm{d}\theta \\ &= A_{l,m,l',m'} \int_{-1}^1 P_l^{|m|}(\mu)(1-\mu^2)^{1/2} P_{l'}^{|m|-1}(\mu)\,\mathrm{d}\mu,\end{aligned}$$

with a constant $A_{l,m,l',m'}$ depending on l, m, l' and m'. The recurrence relation

$$(l+m)P_{l-1}^m(\mu) = (1-\mu^2)^{1/2} P_l^{m+1}(\mu) + (l-m)\mu P_l^m(\mu) \tag{B.1}$$

for associated Legendre functions yields

$$\begin{aligned}(\boldsymbol{a}_{l,m}^{l',m'})_1 = A_{l,m,l',m'} \int_{-1}^1 \Big[&(l+|m|-1)P_{l-1}^{|m|-1}(\mu) \\ &- (l-|m|+1)\mu P_l^{|m|-1}(\mu)\Big] P_{l'}^{|m|-1}\,\mathrm{d}\mu\ .\end{aligned}$$

Using the recurrence relation

$$(l-m+2)P_{l+2}^m(\mu) = (2l+3)\mu P_{l+1}^m(\mu) - (l+m+1)P_l^m(\mu) \tag{B.2}$$

for the second term, we obtain

$$\begin{aligned}(\boldsymbol{a}_{l,m}^{l',m'})_1 = A_{l,m,l',m'} \int_{-1}^1 \Big[&(l+|m|-1)P_{l-1}^{|m|-1}(\mu) - \frac{(l-|m|+1)}{2l+1} \\ &\times \Big((l-|m|+2)P_{l+1}^{|m|-1}(\mu) \\ &+ (l+|m|-1)P_{l-1}^{|m|-1}(\mu)\Big)\Big] P_{l'}^{|m|-1}\,\mathrm{d}\mu \\ = A_{l,m,l',m'} \Big[&(l+|m|-1)\delta_{l-1,l'} \\ &- \frac{(l-|m|+1)}{2l+1}\Big((l-|m|+2)\delta_{l+1,l'} + (l+|m|-1)\delta_{l-1,l'}\Big)\Big]\ .\end{aligned}$$

Therefore, in view of the orthogonality of the associated Legendre functions, we have $(\boldsymbol{a}_{l,m}^{l',m'})_1 = 0$ for $l' \neq l \pm 1$.

Next, we consider the case $|m'| = |m| + 1$. Then

$$\begin{aligned}(\boldsymbol{a}_{l,m}^{l',m'})_1 &= B_{l,m,l',m'} \int_0^\pi P_l^{|m|}(\cos\theta) \sin^2\theta P_{l'}^{|m|+1}(\cos\theta)\,\mathrm{d}\theta \\ &= B_{l,m,l',m'} \int_{-1}^1 P_l^{|m|}(\mu)(1-\mu^2)^{1/2} P_{l'}^{|m|+1}(\mu)\,\mathrm{d}\mu,\end{aligned}$$

with a constant $B_{l,m,l',m'}$ depending on l, m, l' and m'. Arguing similarly as above, we conclude that $l' = l \pm 1$ is required for nonzero $(\boldsymbol{a}_{l,m}^{l',m'})_1$.

For the term $(\boldsymbol{a}_{l,m}^{l',m'})_2$ one finds that nonzero values are obtained only if $l' \in \{l-1,\ l+1\}$ and $m' \in \{-m-1,\ -m+1\}$. The coefficient $(\boldsymbol{a}_{l,m}^{l',m'})_3$ vanishes except for $l' \in \{l-1,\ l+1\}$ and $m' = m$.

The sparsity of $\boldsymbol{b}_{l,m}^{l',m'}$ with respect to the indices m and m' is proved in the same way as for $\boldsymbol{a}_{l,m}^{l',m'}$. However, proving sparsity with respect to the indices l and l' is more cumbersome because of the derivatives in the integrands.

First, let $|m'| = |m| - 1$. We have

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 &= C_{l,m,l',m'} \int_0^\pi \bigg[\frac{\mathrm{d}P_l^{|m|}(\cos\theta)}{\mathrm{d}\theta}\cos\theta\sin\theta \\ &\qquad + |m| P_l^{|m|}(\cos\theta)\bigg] P_{l'}^{|m|-1}(\cos\theta)\,\mathrm{d}\theta \\ &= C_{l,m,l',m'} \int_{-1}^1 \bigg[-\frac{\mathrm{d}P_l^{|m|}(\mu)}{\mathrm{d}\mu}\mu(1-\mu^2)^{1/2} \\ &\qquad + |m| P_l^{|m|}(\mu)(1-\mu^2)^{-1/2}\bigg] P_{l'}^{|m|-1}(\mu)\,\mathrm{d}\mu\end{aligned}$$

with some constant $C_{l,m,l',m'}$. Using the recursion formula

$$(1-\mu^2)\frac{\mathrm{d}P_l^m(\mu)}{\mathrm{d}\mu} = (l+m)P_{l-1}^m(\mu) - l\mu P_l^m(\mu)$$

to resolve the derivative yields

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 = C_{l,m,l',m'} \int_{-1}^1 \bigg[&l\mu^2 P_l^{|m|}(\mu) - (l+|m|)\mu P_{l-1}^{|m|}(\mu) \\ &+ |m| P_l^{|m|}(\mu)\bigg] P_{l'}^{|m|-1}(\mu)(1-\mu^2)^{-1/2}\,\mathrm{d}\mu.\end{aligned}$$

To use the orthogonality of associated Legendre functions, the term $(1-\mu^2)^{-1/2}$ has to be eliminated and the upper index of associated Legendre functions has to be equal. To this end, we employ the relation

$$\mu P_l^m(\mu) = (l-m+1)(1-\mu^2)^{1/2} P_l^{m-1}(\mu) + P_{l-1}^m(\mu)$$

to obtain

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 = C_{l,m,l',m'} \int_{-1}^{1} \Big[& l(l-|m|+1)\mu P_l^{|m|-1}(\mu)(1-\mu^2)^{1/2} \\ & - |m|\mu P_{l-1}^{|m|}(\mu) + |m| P_l^{|m|}(\mu) \Big] P_{l'}^{|m|-1}(\mu)(1-\mu^2)^{-1/2}\,\mathrm{d}\mu\ .\end{aligned}$$

Applying the recursion (B.2) to the first term and

$$P_{l+1}^m(\mu) = \mu P_l^m(\mu) + (l+m)(1-\mu^2)^{1/2} P_l^{m-1}(\mu)$$

to the remaining terms, we find that

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 &= C_{l,m,l',m'} \int_{-1}^{1} \Big[\frac{l(l-|m|+1)^2}{2l+1} P_{l+1}^{|m|-1}(\mu) \\ &\quad + \frac{l(l-|m|+1)(l+|m|)}{2l+1} P_{l-1}^{|m|-1}(\mu) \\ &\quad + |m|(l+|m|-1) P_{l-1}^{|m|-1}(\mu) \Big] P_{l'}^{|m|-1}(\mu)\,\mathrm{d}\mu \\ &= C_{l,m,l',m'} \Big[\frac{l(l-|m|+1)^2}{2l+1} \delta_{l+1,l'} \\ &\quad + \frac{l(l-|m|+1)(l+|m|) + (2l+1)|m|(l+|m|-1)}{2l+1} \delta_{l-1,l'} \Big]\ .\end{aligned}$$

Thus, $l = l' \pm 1$ is required for nonvanishing $(\boldsymbol{b}_{l,m}^{l',m'})_1$.

Next, let $|m'| = |m| + 1$. Starting from

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 &= D_{l,m,l',m'} \int_0^{\pi} \Big[\frac{\mathrm{d}P_l^{|m|}(\cos\theta)}{\mathrm{d}\theta} \cos\theta \sin\theta \\ &\quad - |m| P_l^{|m|}(\cos\theta) \Big] P_{l'}^{|m|+1}(\cos\theta)\,\mathrm{d}\theta \\ &= D_{l,m,l',m'} \int_{-1}^{1} \Big[-\frac{\mathrm{d}P_l^{|m|}(\mu)}{\mathrm{d}\mu} \mu (1-\mu^2)^{1/2} \\ &\quad - |m| P_l^{|m|}(\mu)(1-\mu^2)^{-1/2} \Big] P_{l'}^{|m|+1}(\mu)\,\mathrm{d}\mu\end{aligned}$$

for some constant $D_{l,m,l',m'}$, we arrive similarly as above at

$$(\boldsymbol{b}_{l,m}^{l',m'})_1 = D_{l,m,l',m'} \int_{-1}^{1} \Big[l\mu^2 P_l^{|m|}(\mu) - (l+|m|)\mu P_{l-1}^{|m|}(\mu) \\ - |m| P_l^{|m|}(\mu) \Big] P_{l'}^{|m|+1}(\mu)(1-\mu^2)^{-1/2} \,\mathrm{d}\mu \ .$$

With the recurrence relation

$$(l+m+1)\mu P_l^m(\mu) = (l-m+1)P_{l+1}^m(\mu) + (1-\mu^2)^{1/2} P_l^{m+1}(\mu)$$

applied to the first and the second term we find that

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 = D_{l,m,l',m'} \int_{-1}^{1} \Big[& \frac{l}{l+|m|+1}(1-\mu^2)^{1/2}\mu P_l^{|m|+1}(\mu) \\ & - (1-\mu^2)^{1/2} P_{l-1}^{|m|+1}(\mu) + l\frac{l-|m|+1}{l+|m|+1}\mu P_{l+1}^{|m|}(\mu) \\ & - l P_l^{|m|}(\mu) \Big] P_{l'}^{|m|+1}(\mu)(1-\mu^2)^{-1/2} \,\mathrm{d}\mu \ .\end{aligned}$$

The recurrence relations (B.2) applied to the first term and (B.1) applied to the last two terms yields

$$\begin{aligned}(\boldsymbol{b}_{l,m}^{l',m'})_1 &= D_{l,m,l',m'} \int_{-1}^{1} \Big[\frac{l}{l+|m|+1}\frac{l-|m|+1}{2l+1} P_{l+1}^{|m|+1}(\mu) \\ &\quad + \frac{l}{l+|m|+1}\frac{l+|m|}{2l+1} P_{l-1}^{|m|+1}(\mu) - P_{l-1}^{|m|+1}(\mu) \\ &\quad + \frac{l}{l+|m|+1} P_{l+1}^{|m|+1}(\mu) \Big] P_{l'}^{|m|+1}(\mu) \,\mathrm{d}\mu \\ &= D_{l,m,l',m'} \Big[\frac{l}{l+|m|+1} \left(\frac{l-|m|+1}{2l+1} + 1 \right) \delta_{l+1,l'} \\ &\quad + \left(\frac{l}{l+|m|+1}\frac{l+|m|}{2l+1} - 1 \right) \delta_{l-1,l'} \Big] \ .\end{aligned}$$

Summarizing, $l' \in \{l-1, l+1\}$ and $m' \in \{m+1, m-1\}$ is required for nonzero $(\boldsymbol{b}_{l,m}^{l',m'})_1$. The coefficient $(\boldsymbol{b}_{l,m}^{l',m'})_2$ requires $l' \in \{l-1, l+1\}$ and $m' \in \{-m+1, -m-1\}$ in order to have nonzero values, while $(\boldsymbol{b}_{l,m}^{l',m'})_3 \neq 0$ requires $l' \in \{l-1,\ l+1\}$ and $m' = m$. Hence, the sparsity structure of $\boldsymbol{b}_{l,m}^{l',m'}$ is the same as that of $\boldsymbol{a}_{l,m}^{l',m'}$.

Appendix C. Wigner 3jm Symbols

The symbol

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \tag{C.1}$$

with parameters being either integers or half-integers is called a *Wigner 3jm symbol* arising in coupled angular momenta in two quantum systems. It is zero unless all of the following selection rules apply:

1. $m_1 \in \{-|j_1|, \dots |j_1|\}$, $m_2 \in \{-|j_2|, \dots |j_2|\}$ and $m_3 \in \{-|j_3|, \dots |j_3|\}$,
2. $m_1 + m_2 + m_3 = 0$,
3. $|j_1 - j_2| \leq j_3 \leq j_1 + j_2$.

The connection with spherical harmonics is the following:

$$\int_\Omega Y_{l_1,m_1} Y_{l_2,m_2} Y_{l_3,m_3} \, \mathrm{d}\Omega = \sqrt{\frac{(2l_1+1)(2l_2+1)(2l_3+1)}{4\pi}} \times \begin{pmatrix} l_1 & l_2 & l_3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} ,$$

where the left hand side is often termed *Slater integral*.

References

[1] J. A. Carillo, I. M. Gamba, A. Majorana, C. W. Shu, 2D Semiconductor Device Simulations by WENO-Boltzmann Schemes: Efficiency, Boundary Conditions and Comparison to Monte Carlo Methods, Journal of Computational Physics 214 (2006) 55–80.

[2] C. Ertler, F. Schürrer, Simulation of Silicon Semiconductor Devices by means of a Direct Boltzmann-Poisson Solver, COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering 25 (4) (2006) 979–994.

[3] K. Rahmat, J. White, D. A. Antoniadis, Simulation of Semiconductor Devices Using a Galerkin/Spherical Harmonic Expansion Approach to Solving the Coupled Poisson-Boltzmann System, IEEE Trans. Computer-Aided Design Integr. Circuits Sys. 15 (10) (1996) 1181–1195.

[4] C. Jungemann, B. Meinerzhagen, Hierarchical Device Simulation, Computational Microelectronics, Springer-Verlag, 2003.

[5] N. Goldsman, L. Hendrickson, J. Frey, A Physics-Based Analytical/Numerical Solution to the Boltzmann Transport Equation for the Use in Device Simulation, Solid-State Electr. 34 (1991) 389–396.

[6] A. Gnudi, D. Ventura, G. Baccarani, One-Dimensional Simulation of a Bipolar Transistor by means of Spherical Harmonics Expansion of the Boltzmann Transport Equation, in: W. Fichtner, D. Aemmer (Eds.), Proc. of SISDEP, Vol. 4, 1991, pp. 205–213.

[7] K. A. Hennacy, Y. J. Wu, N. Goldsman, I. D. Mayergoyz, Deterministic MOSFET Simulation Using a Generalized Spherical Harmonic Expansion of the Boltzmann Equation, Solid-State Electr. 38 (8) (1995) 1485–1495.

[8] A. Gnudi, D. Ventura, G. Baccarani, F. Odeh, Two-Dimensional MOSFET Simulation by Means of a Multidimensional Spherical Harmonics Expansion of the Boltzmann Transport Equation, Solid-State Electr. 36 (4) (1993) 575–581.

[9] K. A. Hennacy, N. Goldsman, I. D. Mayergoyz, 2-Dimensional Solution to the Boltzmann Transport Equation to Arbitrarily High-Order Accuracy, in: Proceedings of IWCE, 1993, pp. 118–122.

[10] D. Ventura, A. Gnudi, G. Baccarani, F. Odeh, Multidimensional Spherical Harmonics Expansion of Boltzmann Equation for Transport in Semiconductors, Appl. Math. Lett. 5 (3) (1992) 85–90.

[11] S. M. Hong, C. Jungemann, M. Bollhofer, A Deterministic Boltzmann Equation Solver for Two-Dimensional Semiconductor Devices, in: Simulation of Semiconductor Processes and Devices (SISPAD) 2008, 2008, pp. 293–296.

[12] H. Lin, N. Goldsman, I. D. Mayergoyz, Deterministic BJT Modeling by Self-Consistent Solution to the Boltzmann, Poisson and Hole-Continuity Equations, in: Proceedings of IWCE, 1993, pp. 55–59.

[13] H. Lin, N. Goldsman, I. D. Mayergoyz, Improved Self-Consistent Device Modeling by Direct Solution to Boltzmann and Poisson Equations, in: Proceedings of IWCE, 1992, pp. 143–146.

[14] M. C. Vecchi, D. Ventura, A. Gnudi, G. Baccarani, Incorporating Full Band-Structure Effects in the Spherical Harmonics Expansion of the Boltzmann Transport Equation, in: Numerical Modeling of Processes and Devices for Integrated Circuits (NUPAD) 1994, 1994, pp. 55–58.

[15] M. C. Vecchi, 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 45 (1998) 230–238.

[16] S. M. Hong, C. Jungemann, Deterministic Simulation of SiGe HBTs Based on the Boltzmann Equation, in: Proceedings ESSDERC, 2008, pp. 170–173.

[17] C. Ringhofer, Dissipative Discretization Methods for Approximations to the Boltzmann Equation, Math. Models Meth. Appl. Sci. 11 (2001) 133–149.

[18] C. Ringhofer, A Mixed Spectral-Difference Method for the Steady State Boltzmann-Poisson System, SIAM J. Numer. Anal. 41 (1) (2003) 64–89.

[19] C. Ringhofer, C. Schmeiser, A. Zwirchmayr, Moment Methods for the Semiconductor Boltzmann Equation on Bounded Position Domains, SIAM J. Numer. Anal. 39 (3) (2001) 1078–1095.

[20] C. Ringhofer, Numerical Methods for the Semiconductor Boltzmann Equation Based on Spherical Harmonics Expansions and Entropy Discretizations, Transp. Theory Stat. Phys. 31 (2002) 431–452.

[21] C. Ringhofer, Space-Time Discretization of Series Expansion Methods for the Boltzmann Transport Equation, SIAM J. Numer. Anal. 38 (2) (2000) 442–465.

[22] O. Hansen, A. Jüngel, Analysis of a Spherical Harmonics Expansion Model of Plasma Physics, Math. Models Meth. Appl. Sci. 14 (2004) 759–774.

[23] A. T. Pham, C. Jungemann, B. Meinerzhagen, A Convergence Enhancement Method for Deterministic Multisubband Device Simulations of Double Gate PMOSFET, in: Proceedings of SISPAD, 2009, pp. 115–118.

[24] M. C. Vecchi, J. Mohring, M. Rudan, An efficient Solution Scheme for the Spherical-Harmonics Expansion of the Boltzmann Transport Equation [MOS Transistors], IEEE Trans. Computer-Aided Design Integr. Circuits Sys. 16 (4) (1997) 353–361.

[25] S. M. Hong, C. Jungemann, A Fully Coupled Scheme for a Boltzmann-Poisson Equation Solver Based on a Spherical Harmonics Expansion, J. Comput. Electron. 8 (2009) 225–241.

[26] C. Jungemann, A. T. Pham, B. Meinerzhagen, C. Ringhofer, M. Bollhöfer, Stable Discretization of the Boltzmann Equation based on Spherical Harmonics, Box Integration, and a Maximum Entropy Dissipation Principle, J. Appl. Phys. 100 (2) (2006) 024502.

[27] Y. Saad, M. H. Schultz, GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems, SIAM J. Sci. Stat. Comput. 7 (3) (1986) 856–869.

[28] H. F. Walker, L. Zhou, A Simpler GMRES, Numer. Linear Algebra Appl. 1 (6) (1994) 571–581.

[29] H. A. van der Vorst, Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Non-Symmetric Linear Systems, SIAM Journal on Scientific and Statistical Computing 12 (1992) 631–644.