Journal of Computational Physics 229 (2010) 8750–8765

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Matrix compression for spherical harmonics expansions of the Boltzmann transport equation for semiconductors

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

[a] Christian Doppler Laboratory for Reliability Issues in Microelectronics, 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

ARTICLE INFO

Article history:
Received 27 April 2010
Received in revised form 31 July 2010
Accepted 10 August 2010
Available online 14 August 2010

Keywords:
Boltzmann equation
Spherical harmonics
Kronecker product

ABSTRACT

We investigate numerical solution schemes for the semiconductor Boltzmann transport equation using an expansion of the distribution function in spherical harmonics. A complexity analysis shows that traditional implementations using higher-order 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 asymptotic memory requirements for the storage of the system matrix significantly. The total memory requirements are then dominated by the memory required for the unknowns. Numerical results demonstrate the applicability of our method and confirm our theoretical results.

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 [1–3]. 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 [4,5].

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 [6,7]. Later, the method was extended to arbitrary expansion order [4,8] and two spatial dimensions [9–12]. Furthermore, numerous contributions from the physics point of view [13–17] and some results from the mathematics point of view [18–23] are available. However, there are only a few contributions on improvements of the treatment of the discrete system of equations [24,25].

* Corresponding author at: Institute for Microelectronics, TU Wien, Gußhausstraße 27–29/E360, A-1040 Wien, Austria.
E-mail addresses: rupp@iue.tuwien.ac.at (K. Rupp), juengel@asc.tuwien.ac.at (A. Jüngel), grasser@iue.tuwien.ac.at (T. Grasser).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.08.008

The major challenge for SHE are the still huge memory requirements reported already for two-dimensional devices [26]. The reason is that the model contains up to three spatial variables and 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 device simulations, 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. 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 unknowns of the system. On a machine with 8 Gb 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 [26]. Therefore, the method proposed in this work paves the way for three-dimensional device simulations using a SHE approach.

This work is organized as follows: We briefly review the derivation of the SHE equations in Section 2. In Section 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 (Section 4) to derive a matrix compression scheme using sums of Kronecker products, which reduces the memory requirements for the system matrix considerably. Section 5 shows how non-spherical bands can be incorporated into the matrix compression scheme, while Section 6 deals with the inclusion of stabilization schemes and Section 7 discusses the handling of boundary conditions. The selection of appropriate linear solvers is discussed in Section 8. Numerical results are given in Section 9, confirming our theoretical results. Finally, we conclude in Section 10.

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. [27]. 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, $\hbar\boldsymbol{k} \in \mathbb{R}^3$ is the momentum vector (with modified Plack constant $\hbar$) and $t > 0$ denotes 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

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

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 energy distribution function [27]

$$g(\boldsymbol{x}, \varepsilon, \theta, \varphi, t) = 2Z(\varepsilon, \theta, \varphi) f(\boldsymbol{x}, \boldsymbol{k}(\varepsilon, \theta, \varphi), t), \tag{1}$$

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:

$$g(\boldsymbol{x}, \varepsilon, \theta, \varphi, t) \approx \sum_{l=0}^{L}\sum_{m=-l}^{l} g_{l,m}(\boldsymbol{x}, \varepsilon, t) Y_{l,m}(\theta, \varphi). \tag{2}$$

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

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

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

$$\frac{\partial g_{l,m}}{\partial t} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{j}_{l,m} + \frac{\partial \boldsymbol{F} \cdot \boldsymbol{j}_{l,m}}{\partial \varepsilon} - \boldsymbol{F} \cdot \boldsymbol{\Gamma}_{l,m} = Q_{l,m}\{g\}, \tag{3}$$

where the generalized current density

$$\boldsymbol{j}_{l,m}(\boldsymbol{x}, \varepsilon, t) = \int \boldsymbol{v} g Y_{l,m}\,\mathrm{d}\Omega \tag{4}$$

and the angular force coupling term

$$\boldsymbol{\Gamma}_{l,m}(\boldsymbol{x}, \varepsilon, t) = \int \frac{1}{\hbar|\boldsymbol{k}|}\left(\frac{\partial Y_{l,m}}{\partial \theta}\boldsymbol{e}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_{l,m}}{\partial \varphi}\boldsymbol{e}_\varphi\right) g\,\mathrm{d}\Omega \tag{5}$$

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 (2) into (4) and (5) and then substitute these into (3). Using Einstein's summation convention, we obtain the system of partial differential equations

$$\frac{\partial g_{l,m}}{\partial t} + \boldsymbol{v}_{l,m}^{l',m'} \cdot \nabla_{\boldsymbol{x}} g_{l',m'} + \frac{\partial \boldsymbol{F} \cdot \boldsymbol{v}_{l,m}^{l',m'} g_{l',m'}}{\partial \varepsilon} - \boldsymbol{F} \cdot \boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'} = Q_{l,m}\{g\} \tag{6}$$

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

$$\boldsymbol{v}_{l,m}^{l',m'}(\varepsilon) = \int \boldsymbol{v} Y_{l,m} Y_{l',m'}\,\mathrm{d}\Omega, \tag{7}$$

$$\boldsymbol{\Gamma}_{l,m}^{l',m'}(\varepsilon) = \int \frac{1}{\hbar|\boldsymbol{k}|}\left(\frac{\partial Y_{l,m}}{\partial \theta}\boldsymbol{e}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_{l,m}}{\partial \varphi}\boldsymbol{e}_\varphi\right) Y_{l',m'}\,\mathrm{d}\Omega. \tag{8}$$

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 [5], a spherical harmonics projection leads to [27]

$$Q_{l,m}^{\mathrm{in}}(\boldsymbol{x}, \varepsilon, t) = \sum_\eta s_{l,m;\eta}^{l',m';\mathrm{in}} g_{l',m'}(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t),$$

$$s_{l,m;\eta}^{l',m';\mathrm{in}}(\boldsymbol{x}, \varepsilon) = \frac{1}{Y_{0,0}} Z_{l,m}(\varepsilon) c_\eta(\boldsymbol{x}, \varepsilon \pm \hbar\omega_\eta, \varepsilon)\delta_{0,l'}\delta_{0,m'} \tag{9}$$

and

$$Q_{l,m}^{\mathrm{out}}(\boldsymbol{x}, \varepsilon, t) = s_{l,m}^{l',m';\mathrm{out}} g_{l',m'}(\boldsymbol{x}, \varepsilon, t),$$

$$s_{l,m}^{l',m';\mathrm{out}}(\boldsymbol{x}, \varepsilon) = \frac{1}{Y_{0,0}}\sum_\eta Z_{0,0}(\varepsilon \mp \hbar\omega_\eta) c_\eta(\boldsymbol{x}, \varepsilon, \varepsilon \pm \hbar\omega_\eta)\delta_{l,l'}\delta_{m,m'}, \tag{10}$$

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

$$Z_{l,m} = \int_\Omega Z(\varepsilon, \theta, \varphi) Y_{l,m}\,\mathrm{d}\Omega. \tag{11}$$

Substitution of the projected scattering terms into (6) yields the full system of partial differential equations

$$\frac{\partial g_{l,m}}{\partial t} + \boldsymbol{v}_{l,m}^{l',m'} \cdot \nabla_{\boldsymbol{x}} g_{l',m'} + \frac{\partial \boldsymbol{F} \cdot \boldsymbol{v}_{l,m}^{l',m'} g_{l',m'}}{\partial \varepsilon} - \boldsymbol{F} \cdot \boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'} = \sum_\eta s_{l,m;\eta}^{l',m';\mathrm{in}} g_{l',m'}(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t) - s_{l,m}^{l',m';\mathrm{out}} g_{l',m'}(\boldsymbol{x}, \varepsilon, t) \tag{12}$$

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 energy bands

The representation (12) 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 (9) and (10), 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 energy bands, the generalized density of states is spherically symmetric, hence $Z_{l,m} \equiv 0$ for $(l,m) \neq (0,0)$. Consequently, the scattering terms do not couple any unknowns in this case. 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 (7) and (8)).

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, $i' \in \{0,\dots,I\}$, $m \in \{-i,\dots,i\}$ and $m' \in \{-i',\dots,i'\}$*

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

The essence of this theorem is that all nonzero coupling coefficients possess different parities in the leading indices. This small structural information about the coupling was already used for a preprocessing step for the solution of the discretized equations in [27].

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

$$\boldsymbol{v}_{l,m}^{l',m'}(\varepsilon) = v(\varepsilon)\int Y_{l,m}\boldsymbol{e}_\varepsilon Y_{l',m'}\,\mathrm{d}\Omega =: v(\varepsilon)\boldsymbol{a}_{l,m}^{l',m'}, \tag{13}$$

$$\boldsymbol{\Gamma}_{l,m}^{l',m'}(\varepsilon) = \frac{1}{\hbar|\boldsymbol{k}|}\int\left(\frac{\partial Y_{l,m}}{\partial\theta}\boldsymbol{e}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_{l,m}}{\partial\varphi}\boldsymbol{e}_\varphi\right)Y_{l',m'}\,\mathrm{d}\Omega =: \frac{1}{\hbar|\boldsymbol{k}|}\boldsymbol{b}_{l,m}^{l',m'}. \tag{14}$$

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 9, 10 and 12, 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 [4,10].

4. System matrix compression for spherical energy bands

In this section we investigate the discretization of the projected SHE system (12) for spherical energy bands.

Substitution of (13) and (14) into (12) yields

$$\frac{\partial g_{l,m}}{\partial t} + \boldsymbol{a}_{l,m}^{l',m'}\cdot\left[v\nabla_{\boldsymbol{x}}g_{l',m'} + \boldsymbol{F}\frac{\partial v g_{l',m'}}{\partial\varepsilon}\right] - \boldsymbol{b}_{l,m}^{l',m'}\cdot\boldsymbol{F}\frac{g_{l',m'}}{\hbar|\boldsymbol{k}|} = \sum_\eta s_{l,m;\eta}^{l',m';\mathrm{in}}g_{l',m'}(\boldsymbol{x},\varepsilon\mp\hbar\omega_\eta,t) - s_{l,m}^{l',m';\mathrm{out}}g_{l',m'}. \tag{15}$$

Let us consider a spatial discretization for $g_{l,m}$ in the $(\boldsymbol{x},\varepsilon)$-space 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 (15) 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

$$\begin{aligned} &\pi_L : \mathbb{N} \times \{0,\dots,L\} \times \{-L,\dots,L\} \to \mathbb{N}, \\ &(i,l,m) \mapsto i(L+1)^2 + l^2 + l + m, \end{aligned} \tag{16}$$

which is a bijection for $-l \leqslant m \leqslant l$. More general 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 $\{0,\dots,(L+1)^2-1\}$ such as $\kappa(l,m) = l^2 + l + m$ can also be used.

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

$$\begin{aligned} (\boldsymbol{w}(\varphi,\chi))_{\kappa(l,m),\kappa(l',m')} = \int \Bigg[&\frac{\partial\varphi}{\partial t}\delta_{l,l'}\delta_{m,m'} + \sum_{l',m'} \boldsymbol{a}_{l,m}^{l',m'} \cdot \left(v\nabla_{\boldsymbol{x}}\varphi + \boldsymbol{F}\frac{\partial v\varphi}{\partial\varepsilon} \right) - \sum_{l',m'} \boldsymbol{b}_{l,m}^{l',m'} \cdot \boldsymbol{F}\frac{\varphi}{\hbar|\boldsymbol{k}|} + \sum_{l',m'} s_{l,m}^{l',m';\mathrm{out}}\varphi \\ &- \sum_{l',m';\eta} s_{l,m;\eta}^{l',m';\mathrm{in}}\varphi(\boldsymbol{x},\varepsilon \mp \hbar\omega_\eta,t) \Bigg] \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \end{aligned} \tag{17}$$

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 (16), the system matrix for the discrete system is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{w}(\varphi_1,\chi_1) & \boldsymbol{w}(\varphi_2,\chi_1) & \dots & \boldsymbol{w}(\varphi_N,\chi_1) \\ \boldsymbol{w}(\varphi_1,\chi_2) & \boldsymbol{w}(\varphi_2,\chi_2) & \dots & \boldsymbol{w}(\varphi_N,\chi_2) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}(\varphi_1,\chi_N) & \boldsymbol{w}(\varphi_2,\chi_N) & \dots & \boldsymbol{w}(\varphi_N,\chi_N) \end{pmatrix}, \tag{18}$$

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 there is no common support of φ_i and χ_j (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 (17). 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 a 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*

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

where the notation $|A|$ denotes the number of elements of the set A and G is the simulation domain in the (x,ε)-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 (18) 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 $C_{\mathrm{sparse}} = \mathcal{O}(1)$. 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 (x,ε)-domain using N degrees of freedom. Then it holds:*

1. *A straightforward assembly of the matrix S, defined in (18), needs a storage of $C_{sparse}N(L+1)^4$ entries.*
2. *There exists an assembly of S needing a storage of $11C_{sparse}N(L+1)^2$ entries only.*

Proof. The matrix S is of size $N(L+1)^2 \times N(L+1)^2$. In each of the N rows of the block structure (18) there are at most C_{sparse} blocks. Each block is of dimension $(L+1)^2 \times (L+1)^2$, hence there are at most $C_{\mathrm{sparse}}N(L+1)^4$ nonzero entries in S, which proves the first statement. Since each block in the block structure is sparse due to Theorem 2, 9 and 10, each block carries at most $11(L+1)^2$ nonzero entries, thus there are in fact at most $11C_{\mathrm{sparse}}N(L+1)^2$ nonzero entries in S. □

Jungemann et al. [27] observed that the expansion order L has to be at least three to five in order to obtain good results for e.g. electron velocities in a silicon bipolar transistor. In [26] expansions up to order nine have been compared for a n^+nn^+ diode including magnetic fields. For such a high-order simulation with $L = 9$, 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% of the memory is wasted in a straightforward assembly.

Even though a careful assembly reduces the required memory at $L = 9$ by an order of magnitude, total memory requirements of $11 C_{\text{sparse}} N (L+1)^2$ entries are still considerable. Compared to a finite element or finite volume scheme for the Poisson equation with memory requirements $C_{\text{sparse}} N$, the coupling between the expansion coefficients in the SHE equations requires memory of an additional factor $11(L+1)^2$. For a high-order expansion such as $L = 9$, this additional factor is 1100. This leads to huge memory requirements for two-dimensional devices and has rendered the simulation of three-dimensional devices using a higher-order 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,

$$\boldsymbol{a}_{l,m}^{l',m'} = \begin{pmatrix} (\boldsymbol{a}_{l,m}^{l',m'})_1 \\ (\boldsymbol{a}_{l,m}^{l',m'})_2 \\ (\boldsymbol{a}_{l,m}^{l',m'})_3 \end{pmatrix}, \quad \boldsymbol{b}_{l,m,l',m'} = \begin{pmatrix} (\boldsymbol{b}_{l,m}^{l',m'})_1 \\ (\boldsymbol{b}_{l,m}^{l',m'})_2 \\ (\boldsymbol{b}_{l,m}^{l',m'})_3 \end{pmatrix}, \quad \boldsymbol{F}(\boldsymbol{x}) = \begin{pmatrix} F_1(\boldsymbol{x}) \\ F_2(\boldsymbol{x}) \\ F_3(\boldsymbol{x}) \end{pmatrix},$$

a rearrangement of (17) leads to the following nine integrals

$$\begin{aligned}(\boldsymbol{w}(\varphi,\chi))_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'}\delta_{m,m'} \int \frac{\partial \varphi}{\partial t}\chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon + \sum_{l',m'} \int s_{l,m}^{l',m';\text{out}} \varphi \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon - \sum_{l',m',\eta} \int s_{l,m;\eta}^{l',m';\text{in}} \varphi(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t)\chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon + \sum_{p=1}^{3} \\ \times \sum_{l',m'} (\boldsymbol{a}_{l,m}^{l',m'})_p \int \left(v \frac{\partial \varphi}{\partial (\boldsymbol{x})_p} + F_p \frac{\partial v \varphi}{\partial \varepsilon} \right) \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon - \sum_{p=1}^{3} \sum_{l',m'} (\boldsymbol{b}_{l,m}^{l',m'})_p \int F_p \frac{\varphi}{\hbar|\boldsymbol{k}|} \chi \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon. \end{aligned} \tag{19}$$

The crucial observation is that after substitution of the scattering terms (9) and (10) into (19), 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 (18) can be written as

$$\boldsymbol{S} = \sum_{i=1}^{9} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i, \tag{20}$$

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

$$(\boldsymbol{Q}_1)_{i,j} = \int \frac{\partial \varphi_j}{\partial t} \chi_i \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \tag{21}$$

$$(\boldsymbol{Q}_2)_{i,j} = \frac{1}{Y_{0,0}} \int Z_{0,0}(\varepsilon \mp \hbar\omega_\eta) c_\eta(\varepsilon, \varepsilon \pm \hbar\omega_\eta) \varphi_j \chi_i \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \tag{22}$$

$$(\boldsymbol{Q}_3)_{i,j} = -\frac{1}{Y_{0,0}} \int Z_{0,0} c_\eta(\varepsilon \pm \hbar\omega_\eta, \varepsilon) \varphi_j(\boldsymbol{x}, \varepsilon \pm \hbar\omega_\eta, t) \chi_i \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \tag{23}$$

$$(\boldsymbol{Q}_p)_{i,j} = \int \left[v \frac{\partial \varphi_j}{\partial x_{p-1}} + F_{p-1} \frac{\partial v \varphi_j}{\partial \varepsilon} \right] \chi_i \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \quad p = 4,5,6, \tag{24}$$

$$(\boldsymbol{Q}_p)_{i,j} = -\int F_{p-4} \frac{\varphi_j}{\hbar|\boldsymbol{k}|} \chi_i \,\mathrm{d}\boldsymbol{x}\mathrm{d}\varepsilon, \quad p = 7,8,9, \tag{25}$$

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

$$(\boldsymbol{R}_1)_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'}\delta_{m,m'}, \tag{26}$$

$$(\boldsymbol{R}_2)_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'}\delta_{m,m'}, \tag{27}$$

$$(\boldsymbol{R}_3)_{\kappa(l,m),\kappa(l',m')} = \delta_{l,l'}\delta_{m,m'}\delta_{l',0}\delta_{m',0}, \tag{28}$$

$$(\boldsymbol{R}_p)_{\kappa(l,m),\kappa(l',m')} = (\boldsymbol{a}_{l,m}^{l',m'})_{p-3}, \quad p = 4,5,6, \tag{29}$$

$$(\boldsymbol{R}_p)_{\kappa(l,m),\kappa(l',m')} = (\boldsymbol{b}_{l,m}^{l',m'})_{p-6}, \quad p = 7,8,9. \tag{30}$$

Hence, we can represent the full system matrix $\boldsymbol{S}$, which has up to $11 C_{\text{sparse}} N (L+1)^2$ nonzero entries, by nine matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_9$ (with at most $C_{\text{sparse}} N$ 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 $\boldsymbol{b}_{l,m}^{l',m'}$ 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 $24(L+1)^2 + 9 C_{\text{sparse}} N$ 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 high-order expansions. Even in the case of very high-order expansions such as $L = 19$, still 312,500 grid nodes can be used in $(\boldsymbol{x}, \varepsilon)$-space in order to fit all unknowns into one gigabyte of memory in double precision.

5. Non-spherical energy bands

The matrix compression described in the previous section relies on the factorizations (13) and (14) 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'. Moreover, it was used that the expansion coefficients $Z_{l,m}$ in (11) vanish for nonzero l or m using spherical energy bands. However, in the case of non-spherical energy bands, the velocity and the modulus of the wave vector as well as the generalized density of states 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 coupling terms by approximating

$$\boldsymbol{v}(\varepsilon,\theta,\varphi) \approx \sum_{l''=0}^{L'} \sum_{m''=-l''}^{l''} \boldsymbol{v}^{l'',m''}(\varepsilon) Y_{l'',m''}(\theta,\varphi), \tag{31}$$

$$\frac{1}{\hbar|\boldsymbol{k}(\varepsilon,\theta,\varphi)|} \approx \sum_{l''=0}^{L'} \sum_{m''=-l''}^{l''} \Gamma^{l'',m''}(\varepsilon) Y_{l'',m''}(\theta,\varphi), \tag{32}$$

where the expansion coefficients are given for $\varepsilon > 0$ 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; values of $L' = 4$ have been reported to yield a good approximation of the non-spherical energy bands of interest [28].

The expansion order of the generalized density of states Z, which is also assumed to be equal to L' for simplicity, is implicitly coupled to the expansion order L of the distribution function by (9). Thus, even if Z is expanded up to order $L' > L$, only expansion terms up to order L can be considered. For this reason we assume in the following that $L \geqslant L'$.

Substitution of the expansions (31) and (32) into (7) and (8) yields

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

$$\boldsymbol{\Gamma}_{l,m}^{l',m'} = \Gamma^{l'',m''}(\varepsilon) \int \left(\frac{\partial Y_{l,m}}{\partial\theta}\boldsymbol{e}_\theta + \frac{1}{\sin\theta}\frac{\partial Y_{l,m}}{\partial\varphi}\boldsymbol{e}_\varphi \right) Y_{l',m'} Y_{l'',m''}\,\mathrm{d}\Omega =: \Gamma^{l'',m''}(\varepsilon) \boldsymbol{b}_{l,m;l'',m''}^{l',m'},$$

so that we obtain in both cases a sum of $(L'+1)^2$ decoupled terms. This is in contrast to the case of spherical energy 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 (20) in the form

$$\boldsymbol{S} = \sum_{i=1}^{2+(L+1)^2+6(L'+1)^2} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i. \tag{33}$$

The coupling matrices $\boldsymbol{R}_3, \ldots, \boldsymbol{R}_{2+(L+1)^2}$ arising from the in-scattering term consist of a single entry only. 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 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

$$\left[2 + (L+1)^2 + 6(L'+1)^2\right]\left[(L+1)^4 + C_{\text{sparse}}N\right] = \mathcal{O}((L'^2 + L^2)(L^4 + N)) \tag{34}$$

matrix entries. With $L^4 \ll N$ and $L \approx L' \geqslant 3$ in typical applications, the total memory requirements for the system matrix are roughly $7(L+1)^2 C_{\text{sparse}}N$, which is at first sight similar to the memory requirements for the uncompressed system matrix for spherical energy bands, cf. Theorem 3. However, due to symmetries present in non-spherical energy bands, several coupling coefficients vanish [28]. As a consequence, the estimate $7(L'+1)^2 C_{\text{sparse}}N$ considerably overestimates the true memory requirements, and the proposed scheme still results in significant savings compared to setting up the full system matrix.

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 [27,4]. For very small devices, a combination of staggered grids, the maximum entropy dissipation scheme (MEDS) [21] and the H-transform [9] was reported by Hong et al. [26] 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 N^{even} even order unknowns we select a space of trial functions U^{even} with basis $(\varphi_i^{\text{even}})_{i=1}^{N^{\text{even}}}$ and a space of test functions V^{even} with basis $(\chi_j^{\text{even}})_{j=1}^{N^{\text{even}}}$. Similarly, a space of trial functions U^{odd} with basis $(\varphi_i^{\text{odd}})_{i=1}^{N^{\text{odd}}}$ and a space of test functions V^{odd} with basis $(\chi_j^{\text{odd}})_{j=1}^{N^{\text{odd}}}$ is chosen for the N^{odd} odd order unknowns. The total trial space is $U = U^{\text{even}} \oplus U^{\text{odd}}$ and the test space $V = V^{\text{even}} \oplus V^{\text{odd}}$.

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 are enumerated consecutively similar to (16). Repeating the steps in Section 4, the full system matrix $\boldsymbol{S}$ can be written in the block-structure

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}^{\text{ee}} & \boldsymbol{S}^{\text{eo}} \\ \boldsymbol{S}^{\text{oe}} & \boldsymbol{S}^{\text{oo}} \end{pmatrix} = \sum_{i=1}^{p} \begin{pmatrix} \boldsymbol{Q}_i^{\text{ee}} \otimes \boldsymbol{R}_i^{\text{ee}} & \boldsymbol{Q}_i^{\text{eo}} \otimes \boldsymbol{R}_i^{\text{eo}} \\ \boldsymbol{Q}_i^{\text{oe}} \otimes \boldsymbol{R}_i^{\text{oe}} & \boldsymbol{Q}_i^{\text{oo}} \otimes \boldsymbol{R}_i^{\text{oo}} \end{pmatrix}. \tag{35}$$

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 Theorem 1 or Theorem 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 Theorem 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. (9).

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 (35) are obtained by evaluating the underlying bilinear mapping for a basis of trial functions from U^{even} and U^{odd} and a basis of 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 (9) and (10), 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 $\boldsymbol{S}^{oo}$ is diagonal.*
2. *For spherical energy bands without considering inelastic scattering, $\boldsymbol{S}^{ee}$ 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

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

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 (13) and (14) are unchanged and the proposed matrix compression scheme can be applied. Clearly, the expressions (21)–(25) 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 Section 4.

Similarly, an application of MEDS modifies the odd order equations only and does not interfere with the decoupling given by (13) and (14). Thus, the entries in $\boldsymbol{Q}_i^{\text{oe}}$ and $\boldsymbol{Q}_i^{\text{oo}}$ as in (35) are modified, but the matrix compression scheme can be applied without additional difficulties.

7. Boundary conditions

So far we have considered the discretization of the resulting system of partial differential equations in the interior of the simulation domain. At the boundary, suitable conditions need to be imposed and incorporated into the proposed compression scheme.

At all non-contact boundaries, homogeneous Neumann boundary conditions are imposed [4,27,26], which can be directly included in the proposed compression scheme, because no additional boundary terms appear on non-contact boundaries if the weak formulation (17) is integrated by parts. Using a box discretization scheme, this means that box contributions outside the simulation domain are simply ignored.

At the contact boundaries, two different types of boundary conditions are typically imposed. The first type are Dirichlet boundary conditions [4], where the distribution function is set to a Maxwell distribution. Hence, the generalized energy distribution function $g_{0,0}$ is set according to (1) and (2), while $g_{l,m}$ is set to zero at the contact for $(l,m) \neq (0,0)$. This it either

enforced by replacing the corresponding matrix row with unity in the diagonal and zeros elsewhere, and setting the appropriate value at the load vector, or by directly absorbing the known values to the load vector. For the proposed compression scheme, the second way is of advantage, because in that case boundary conditions do not alter the matrix structure.

The second type of contact boundary conditions is a Robin-type generation/recombination process [27]

$$\gamma_{l,m}(\varepsilon) = -\frac{g_{l,m}(\varepsilon) - Z_{l,m}(\varepsilon) f^{\mathrm{eq}}(\varepsilon)}{\tau_0},$$

where f^{eq} denotes the equilibrium Maxwell distribution, or, similar in structure, a surface generation process of the form [26]

$$I^{\mathrm{s}} = \left[f^{\mathrm{eq}}(\boldsymbol{k}')\theta(\boldsymbol{v}\cdot\boldsymbol{n}) + f(\boldsymbol{k}')\theta(-\boldsymbol{v}\cdot\boldsymbol{n})\right]\boldsymbol{v}\cdot\boldsymbol{n},$$

where θ denotes the step function and $\boldsymbol{n}$ is the unit surface normal vector pointing into the device. This type of boundary conditions leads to additional entries in the system matrix due to the additional surface boundary integrals, hence the compressed matrix scheme has to account for them. We propose to write the system matrix in the form

$$\boldsymbol{S} = \boldsymbol{S}^{\mathrm{inner}} + \boldsymbol{S}^{\mathrm{contact}}, \tag{36}$$

where $\boldsymbol{S}^{\mathrm{inner}}$ contains the discretized equations for all interior points as given by (20), (33) or (35), and $\boldsymbol{S}^{\mathrm{contact}}$ consists of the discretized contact boundary terms. Since the number of contact boundary points is much smaller than the total number of grid points N, the sparse matrix $\boldsymbol{S}^{\mathrm{contact}}$ can be set up without compression scheme and the additional memory requirements are negligible.

The additional matrix $\boldsymbol{S}^{\mathrm{contact}}$ in (36) can be easily included in the following derivations, especially since it can be written as a sum of Kronecker products again, as one can easily verify. Therefore, Dirichlet boundary conditions are considered in the following for reasons of clarity.

8. Solution of the linear system

The matrix compression scheme introduced in the previous sections 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

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix} = \sum_{j=1}^{N} \boldsymbol{e}_j \otimes \boldsymbol{x}_j, \tag{37}$$

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

$$\boldsymbol{S}\boldsymbol{x} = \left[\sum_{i=1}^{p} \boldsymbol{Q}_i \otimes \boldsymbol{R}_i\right]\left[\sum_{j=1}^{N} \boldsymbol{e}_j \otimes \boldsymbol{x}_j\right] = \sum_{i=1}^{p}\sum_{j=1}^{N} (\boldsymbol{Q}_i\boldsymbol{e}_j) \otimes (\boldsymbol{R}_i\boldsymbol{x}_j).$$

The product $\boldsymbol{Q}_i\boldsymbol{e}_j$ is simply the jth column of $\boldsymbol{Q}_i$ with, say, C_{sparse} entries on average. The computation of $\boldsymbol{R}_i\boldsymbol{x}_j$ requires roughly $4(L+1)^2$ additions and multiplications. Building the Kronecker product of the intermediate vectors $\boldsymbol{Q}_i\boldsymbol{e}_j$ and $\boldsymbol{R}_i\boldsymbol{x}_j$ and adding nonzero entries to the resulting vector requires $C_{\mathrm{sparse}}(L+1)^2$ operations for each index pair (i,j). Thus, $\mathcal{O}((4 + C_{\mathrm{sparse}})pNL^2) \approx \mathcal{O}(C_{\mathrm{sparse}}pNL^2)$ additions and multiplications are needed in total, since in typical situations $C_{\mathrm{sparse}} \gg 4$.

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 Section 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 [27,26]. Moreover, it has been shown that for a first-order expansion the system matrix after elimination of odd order un-

knowns is an M-matrix [26]. Furthermore, numerical experiments indicate a considerable improvement in the convergence of iterative solvers.

For a matrix structure as given by (35), a direct elimination of odd order unknowns would destroy the representation of the system matrix $\boldsymbol{S}$ as a sum of Kronecker products. Writing the system as

$$\boldsymbol{S}\boldsymbol{g} = \begin{pmatrix} \boldsymbol{S}^{ee} & \boldsymbol{S}^{eo} \\ \boldsymbol{S}^{oe} & \boldsymbol{S}^{oo} \end{pmatrix} \begin{pmatrix} \boldsymbol{g}^{e} \\ \boldsymbol{g}^{o} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}^{e} \\ \boldsymbol{r}^{o} \end{pmatrix} \tag{38}$$

with the vector of unknowns $\boldsymbol{g}$ split into $\boldsymbol{g}^{e}$ and $\boldsymbol{g}^{e}$ 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:

$$(\boldsymbol{S}^{ee} - \boldsymbol{S}^{eo}(\boldsymbol{S}^{oo})^{-1}\boldsymbol{S}^{oe})\boldsymbol{g}^{e} = \boldsymbol{r}^{e} - \boldsymbol{S}^{eo}(\boldsymbol{S}^{oo})^{-1}\boldsymbol{r}^{o}. \tag{39}$$

Since $\boldsymbol{S}^{oo}$ is according to Theorem 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 a 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 $(\boldsymbol{S}^{ee} - \boldsymbol{S}^{eo}(\boldsymbol{S}^{oo})^{-1}\boldsymbol{S}^{oe})$ 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 $L \geqslant 5$. Since $\boldsymbol{S}^{ee}$ is diagonal or at least close to diagonal, the most computational effort is needed for the computation of $\boldsymbol{S}^{eo}(\boldsymbol{S}^{oo})^{-1}\,\boldsymbol{S}^{oe}\boldsymbol{g}^{e}$. Neglecting the cost of inverting the diagonal matrix $\boldsymbol{S}^{oo}$, 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 (17) 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 [29,30]. It is typically restarted after, say, s steps, denoted by GMRES (s). This method was used in recent publications on SHE simulations [27,26]. 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 from $\mathcal{O}(NL^2)$ to $\mathcal{O}(sNL^2)$. The number of steps s is typically chosen between 20 and 30 as smaller values may lead to smaller convergence rates or the solver may even fail to converge within a reasonable number of iterations. Hence, we conclude that GMRES (s) might be too expensive for SHE simulations. Instead, iterative solvers with smaller memory consumption such as BiCGStab [31] should be used.

9. 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.

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.

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

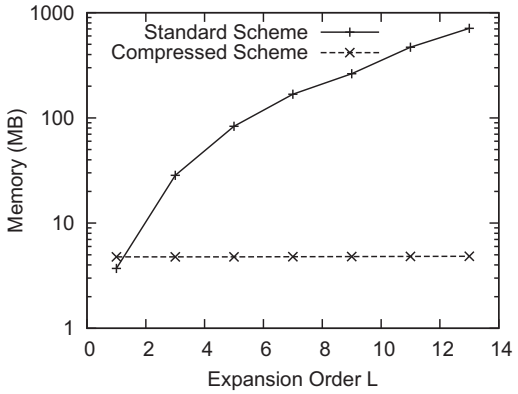

Fig. 1. Memory used for the uncompressed and the compressed system matrix for different expansion orders L on a three-dimensional $(\boldsymbol{x},H)$-grid with 12.500 nodes.

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 Table 1 and Fig. 1 clearly demonstrate the asymptotic advantage of our approach: While no savings are observed at $L = 1$, memory savings of a factor of 18 are obtained already at an expansion order of $L = 5$. 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 and is determined only by the additional memory needed for the coupling matrices $\boldsymbol{R}_i$ in (26)–(30). With increasing expansion order L, the additional memory requirements for the compressed scheme grow quadratically with L (because 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. Consequently, 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, execution times for the matrix–vector multiplications are compared in Fig. 3. Execution times for the full system matrix and the condensed system matrix, where unknowns associated with odd order spherical harmonics have been eliminated, are depicted. 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 Section 8. 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 $L = 13$.

As discussed in Section 8, GMRES leads to higher memory requirements than 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 Table 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 Table

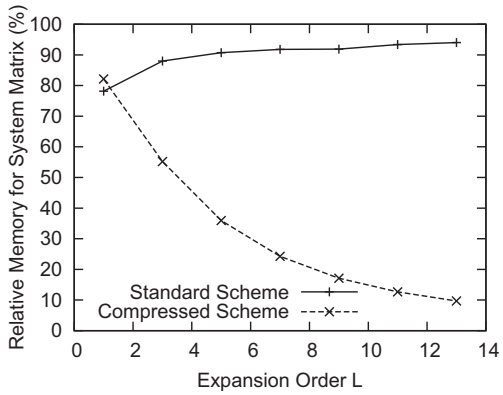

Fig. 2. Memory used for the system matrix in relation to the total amount of memory used (i.e. system matrix, unknowns and right hand side vector).

	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

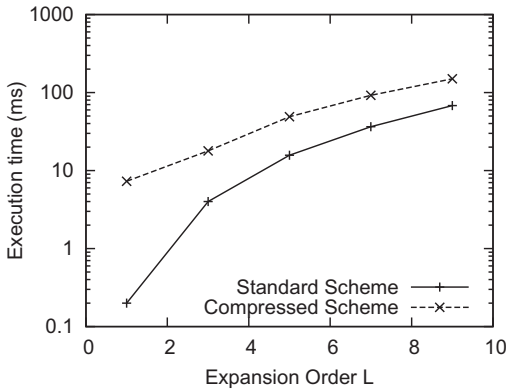

Fig. 3. Comparison of execution times (milliseconds) for matrix–vector multiplication at different expansion orders L for the fully set up system matrix and the proposed compressed matrix scheme. Both the full system of linear equations and the condensed system with odd order unknowns eliminated in a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

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.

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

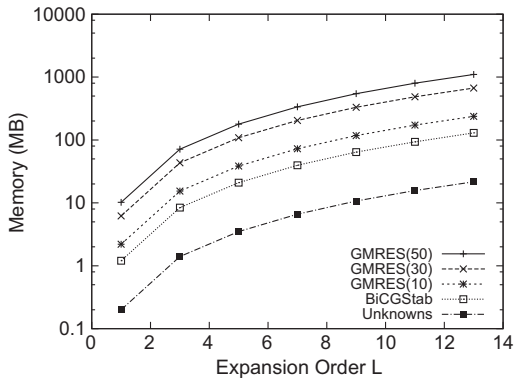

Fig. 4. Additional memory requirements of the linear solvers GMRES (s) with different values of s and BiCGStab compared to the memory needed for the unknowns.

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.

The convergence of iterative solvers strongly depends on the condition number of the system matrix, which can be substantially improved by preconditioners. Recent publications on SHE [27,17] used a black-box incomplete LU factorization preconditioner with threshold (ILUT), which requires approximately the same amount of memory than the full system matrix and is thus not attractive for the proposed approach. On the other hand, simpler preconditioners such as a diagonal or a Jacobi preconditioner do not lead to convergence even for devices under moderate bias. Thus, the construction of good preconditioners taking the structural information in the proposed matrix compression scheme into account is recognized but postponed to future work.

10. Conclusions

We investigated the coupling structure of the SHE equations and showed 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 $L \geqslant 5$ is often used in order to obtain 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 execution times for matrix–vector multiplications. 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 already available on high-end 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.

The proposed compressed matrix scheme is attractive for parallel computing architectures. While most sparse matrix–vector multiplications are memory-bandwidth limited on modern computers, the proposed matrix compression scheme is much more limited by the computational speed. Therefore, we expect that matrix–vector multiplications can considerably benefit from parallel architectures such as multi-core CPUs and GPUs. This would result in higher speed compared to the full system matrix setting due to the lower requirements on memory bandwidth.

Furthermore, we showed that the memory requirements of the chosen linear solver affects the total memory footprint for SHE simulations using the proposed matrix compression scheme much more than in cases where the full matrix is set up. A comparison between GMRES and BiCGStab shows that a careless choice can increase the total memory consumption by up to an order of magnitude.

The proposed scheme can be extended to Jacobian matrices arising from couplings of the BTE with, e.g., Poisson's equation or from nonlinear iterations due to the inclusion of the Pauli principle. Unfortunately, this further complicates the design of suitable preconditioners for iterative solvers, which is the subject of future work.

Acknowledgements

The authors thank Prof. C. Jungemann for numerous discussions and for hosting K. Rupp at the Bundeswehr University, Neubiberg. A. Jüngel acknowledges partial support from the Austrian Science Fund (FWF), grants P20214 and WK Differential Equations; the German Science Foundation (DFG), grant JU 359/7; the Austria-France, Austria-Spain, and Austria-Croatia Projects of the Austrian Exchange Service (ÖAD). Support by the Graduate School PDEtech at the Vienna University of Technology is gratefully acknowledged.

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 Theorem 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 (13) and (14).

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. 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. The recurrence relations used in the following may slightly differ from those given by other authors due to the deliberate choice of an additional phase factor $(-1)^m$ in the definition of $Y_{l,m}$.

Substitution of the definition of spherical harmonics and splitting the integral leads to

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

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

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

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

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

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

Using the recurrence relation

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

for the second term, we obtain

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

Therefore, in view of the orthogonality of 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

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

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 non-zero $(\boldsymbol{a}_{l,m}^{l',m'})_1$.

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

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

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

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

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

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

on the first term to obtain

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

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

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

to the remaining terms, we find that

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

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

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

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

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

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

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

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

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

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

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 between two quantum systems. It is zero unless all of the following selection rules apply:

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

The connection with spherical harmonics is the following:

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

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] M. Galler, Multigroup Equations for the Description of the Particle Transport in Semiconductors, fifth ed., Series on Advances in Mathematics for Applied Sciences, World Scientific, 2005.
[4] 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 Transactions on Computer-Aided Design Integrated Circuits and Systems 15 (10) (1996) 1181–1195.
[5] C. Jungemann, B. Meinerzhagen, Hierarchical Device Simulation, Computational Microelectronics, Springer-Verlag, 2003.
[6] 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 Electronics 34 (1991) 389–396.
[7] 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.), Proceedings of SISDEP, vol. 4, 1991, pp. 205–213.
[8] 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 Electronics 38 (8) (1995) 1485–1495.
[9] 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 Electronics 36 (4) (1993) 575–581.
[10] 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.
[11] D. Ventura, A. Gnudi, G. Baccarani, F. Odeh, Multidimensional spherical harmonics expansion of Boltzmann equation for transport in semiconductors, Applied Mathematics Letters 5 (3) (1992) 85–90.
[12] S.M. Hong, C. Jungemann, M. Bollhofer, A deterministic Boltzmann equation solver for two-dimensional semiconductor devices, in: Proceedings of SISPAD, 2008, pp. 293–296.
[13] 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.
[14] 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.
[15] 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.
[16] 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 Transactions on Electron Devices 45 (1998) 230–238.
[17] S.M. Hong, C. Jungemann, Deterministic simulation of SiGe HBTs based on the Boltzmann equation, in: Proceedings of ESSDERC, 2008, pp. 170–173.
[18] C. Ringhofer, Dissipative discretization methods for approximations to the Boltzmann equation, Mathematical Models and Methods in Applied Sciences 11 (2001) 133–149.
[19] C. Ringhofer, A mixed spectral-difference method for the steady state Boltzmann–Poisson system, SIAM Journal on Numerical Analysis 41 (1) (2003) 64–89.
[20] C. Ringhofer, C. Schmeiser, A. Zwirchmayr, Moment methods for the semiconductor Boltzmann equation on bounded position domains, SIAM Journal on Numerical Analysis 39 (3) (2001) 1078–1095.
[21] C. Ringhofer, Numerical methods for the semiconductor Boltzmann equation based on spherical harmonics expansions and entropy discretizations, Transport Theory and Statistical Physics 31 (2002) 431–452.
[22] C. Ringhofer, Space–time discretization of series expansion methods for the Boltzmann transport equation, SIAM Journal on Numerical Analysis 38 (2) (2000) 442–465.
[23] O. Hansen, A. Jüngel, Analysis of a spherical harmonics expansion model of plasma physics, Mathematical Models and Methods in Applied Sciences 14 (2004) 759–774.
[24] 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.
[25] M.C. Vecchi, J. Mohring, M. Rudan, An efficient solution scheme for the spherical-harmonics expansion of the Boltzmann transport equation [MOS transistors], IEEE Transactions on Computer-Aided Design Integrated Circuits and Systems 16 (4) (1997) 353–361.
[26] S.M. Hong, C. Jungemann, A fully coupled scheme for a Boltzmann–Poisson equation solver based on a spherical harmonics expansion, Journal of Computational Electronics 8 (2009) 225–241.
[27] 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, Journal of Applied Physics 100 (2) (2006) 024502.
[28] A.T. Pham, C. Jungemann, B. Meinerzhagen, 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.
[29] Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM Journal on Scientific and Statistical Computing 7 (3) (1986) 856–869.
[30] H.F. Walker, L. Zhou, A simpler GMRES, Numerical Linear Algebra with Applications 1 (6) (1994) 571–581.
[31] 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.