NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Positivity Preserving Discretization of Time Dependent Semiconductor Model Equations

by

Markus Brunk and Anne Kværnø

PREPRINT
NUMERICS NO. 6/2009

NORWEGIAN UNIVERSITY OF
SCIENCE AND TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
`http://www.math.ntnu.no/preprint/numerics/2009/N6-2009.pdf`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

POSITIVITY PRESERVING DISCRETIZATION OF TIME DEPENDENT SEMICONDUCTOR MODEL EQUATIONS [*]

MARKUS BRUNK[†] AND ANNE KVÆRNØ[‡]

Abstract. Positivity preserving space discretization of the semiconductor drift-diffusion equations is considered. The drift-diffusion equations are spatially discretized by mixed hybrid finite elements. It is shown that this leads to a positive ODE or DAE system with index of at most one. For time discretization a second order splitting technique based on a combination of explicit exponential integration and implicit one-step methods is proposed. The technique allows for positivity preservation with larger step sizes than corresponding one-step methods of higher order. A algorithm is presented coupling the proposed splitting technique with the Gummel iteration scheme for semiconductor equations allowing for efficient positivity preserving device simulation. Numerical results for a one-dimensional *pn*-diode are given, showing that the proposed scheme allows for runtime acceleration.

Key words. Drift-diffusion equations, semiconductor model, positivity preservation, mixed finite elements, splitting method

1. Introduction.

In the field of semiconductor modeling we can distinguish two classes of classical models. The kinetic models like the semiconductor Boltzmann equation, and the fluid-dynamical models like the drift-diffusion (DD) or the energy-transport (ET) model. The semiconductor Boltzmann equations allows for accurate simulation results, however the numerical methods, like the Monte-Carlo method, to solve the equation are very time consuming and numerically expensive. Good accuracy to cheaper numerical costs is obtained by the solution of macroscopic fluid-dynamical models derived from the Boltzmann equation [1], like the drift-diffusion model. For sub-micron devices the energy-transport model might give better results, as it additionally allows for the consideration of thermal effects. As in semiconductor application the trend goes to devices driven by higher frequencies, transient model equations are indispensable for accurate simulation.

The drift-diffusion and the energy-transport model consist of continuity equations for the electron and hole densities where for the ET-model an additional continuity equation for the energy density occurs. After suitable scaling the occurring continuity equations can be written as

$$\partial_t g - \mathrm{div} J + \sigma g = f \qquad\qquad J = \nabla g - g\nabla V \qquad \text{in } \Omega \qquad (1.1)$$

$$g = g_D \quad \text{on } \Gamma_D \qquad J\cdot\nu = 0 \qquad \text{on } \Gamma_N, \qquad (1.2)$$

where $\Omega \subset \mathbb{R}^d$ is the domain occupied by the device, g denotes a particle density and J the corresponding current density. It is reasonable to keep the physical properties of the continuous solution also for the discrete solution. In this paper we will apply spatial and time discretization to (1.1) that allows to keep the continuity of the current density as well as the positivity of the particle density.

For the stationary version of (1.1) and the case of vanishing zero-order term $\sigma = 0$ a scheme based on the lowest order Raviart-Thomas elements has been proposed in [17]. In [4] and [5] another discretization scheme has been discussed for the cases $f = 0$

[*] The first author was supported by the ERCIM “Alain Bensoussan” fellowship programme.

[†] Department C, Institute of Applied Mathematics and Numerical Analysis, University Wuppertal, 42097 Wuppertal, Germany.

[‡] Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway

and $f \geq 0$, respectively. It has been shown that in case of weakly acute triangulation, the matrix associated with this scheme is an M-matrix. This guarantees a discrete maximum principle and thus the non-negativity of the discrete particle density for non-negative boundary conditions.

In [14] Marini and Pietra presented a mixed finite element scheme that allows to keep the M-matrix property also for $\sigma \geq 0$ in the stationary equations. Moreover, the presented finite element scheme guarantees the continuity of the current density across interelement boundaries.

In [6] the Marini-Pietra elements are used for spatial discretization of the transient energy-transport equations which partially fit into the framework of (1.1). There, the equations have been discretized in time first and for positivity preservation solution dependent time step size restrictions occur for methods of order larger than one.

In this paper we discretize the transient equation (1.1) by use of the method of lines. For spatial discretization the Marini-Pietra elements are employed in one or two space dimensions. Different approximations will be made leading to different dynamical systems. Positivity of the different systems will be proven and as the system might be of differential algebraic type, an index statement is given.

For time discretization with one-step methods of order larger than one, severe time step size restriction occurs. We will present a splitting technique based on a combination of explicit exponential integration and implicit one-step methods that allows to increase the time step size and still keeps positivity. We will show, that the presented splitting technique is efficient for semiconductor application when coupled to the iterative Gummel-algorithm [11].

The paper is arranged as follows. In Section 2 we shortly present the semiconductor drift-diffusion model. We fit the model equations into the framework of (1.1). In Section 3 the spatial discretization is described. Three different approximative dynamical systems are derived under different approximations. The positivity of the systems is proven and it is shown that the index of the system is at most one.

In Section 4 we present the splitting technique for time discretization of order two. The coupled Gummel-splitting algorithm is presented. In Section 5 we present numerical examples for the simulation of a pn-diode. We show the good approximative behavior of our dynamical systems. Moreover the efficiency of the coupled Gummel-splitting algorithm is clarified. Finally we conclude in Section 6.

2. Semiconductor device modeling. A standard model in semiconductor device modeling is the well-known drift-diffusion model. It consists of continuity equations for the electron and hole densities n and p, respectively, and is stated in the bounded domain Ω in scaled formulation

$$\partial_t n - \mathrm{div} J_n = -R(n,p), \qquad J_n = (\nabla(\mu_n n) - \mu_n n \nabla V) \tag{2.1}$$

$$\partial_t p + \mathrm{div} J_p = -R(n,p), \qquad J_p = -(\nabla(\mu_p p) + \mu_p p \nabla V), \tag{2.2}$$

coupled self-consistently to the Poisson equation for the electric potential V,

$$\lambda^2 \Delta V = n - p - C(x). \tag{2.3}$$

For scaling see [6], for instance. We notice that the more enhanced energy-transport model is also based on the drift-diffusion equations (2.1)–(2.2).

Here, the function $C(x)$ models fixed charged background ions in the semiconductor crystal (doping profile). The physical parameters are the (scaled) electron and hole mobilities μ_n and μ_p and the Debye length λ, given by

$$\lambda^2 = \frac{\varepsilon_s U_T}{q C_0 L},$$

where ε_s is the permittivity constant, $U_T = k_B T_L/q$ with the Boltzmann constant k_B, the lattice temperature T_L and the elementary charge q denotes the thermal voltage, C_0 is the maximal doping value, and L is the device diameter. The function

$$R(n,p) = \frac{np - n_i^2}{\tau_p(n+n_i) + \tau_n(p+n_i)} \tag{2.4}$$

models recombination-generation processes with the (scaled) intrinsic density n_i and the material-depending electron and hole lifetimes τ_n and τ_p, respectively.

For the model equations (2.1)-(2.2) we impose appropriate initial and boundary conditions. The initial conditions are given by $n(\cdot,0) = n_I, \quad p(\cdot,0) = p_I, \quad$ in Ω. The device boundary is assumed to split into two parts, the union of Ohmic contacts Γ_D and the union of insulating boundary segments Γ_N, where $\partial\Omega = \Gamma_D \cup \Gamma_N$. On the insulating parts, it is assumed that the normal components of the current densities and of the electric field vanish,

$$J_n \cdot \nu = J_p \cdot \nu = \nabla V \cdot \nu = 0 \qquad \text{on } \Gamma_N, \ t > 0. \tag{2.5}$$

At the contacts, the electric potential, and the particle densities are assumed to be known. The electric potential equals the sum of the applied voltage U and the so-called built-in potential V_{bi},

$$V = U + V_{\text{bi}} \quad \text{on } \Gamma_D, \ t > 0, \qquad V_{\text{bi}}(x) = \operatorname{arsinh}\Big(\frac{C(x)}{2n_i}\Big). \tag{2.6}$$

The boundary conditions for the particle densities are derived under the assumptions of charge neutrality, $n - p - C(x) = 0$, and thermal equilibrium, $np = n_i^2$. Solving these equations for n and p gives

$$n = \frac{1}{2}\Big(C + \sqrt{C^2 + 4n_i^2}\Big), \quad p = \frac{1}{2}\Big(-C + \sqrt{C^2 + 4n_i^2}\Big) \quad \text{on } \Gamma_D. \tag{2.7}$$

We will apply an iterative algorithm to solve the coupled system (see Algorithm 2). We semi-linearize the continuity equations by approximating the values of n and p in the denominator of (2.4) by values from the former iteration step. Using the variables $g_n = \mu_n n$ and $g_p = \mu_p p$ we can write (2.1)-(2.2) as

$$\begin{aligned} \mu_n^{-1}\partial_t g_n - \operatorname{div} J_n + \sigma_n g_n &= f, & J_n &= \nabla g_n - g_n \nabla V, \\ \mu_p^{-1}\partial_t g_p + \operatorname{div} J_p + \sigma_p g_p &= f, & J_p &= -(\nabla g_p + g_p \nabla V), \end{aligned}$$

with

$$\begin{aligned} \sigma_n = \widetilde{r}\mu_n^{-1}\mu_p^{-1} g_p, \qquad \sigma_p = \widetilde{r}\mu_p^{-1}\mu_n^{-1} g_n, \qquad f = \widetilde{r} n_i^2, \\ \widetilde{r} = (\tau_p(\mu_n^{-1}\widetilde{g}_n + n_i) + \tau_n(\mu_p^{-1}\widetilde{g}_p + n_i))^{-1} \end{aligned}$$

to fit the semiconductor model equations into the framework given by (1.1) – (1.2). There, $\widetilde{g}_n$ and $\widetilde{g}_p$ denote the values of g_n, g_p in the former iteration step. We note that the hole equations fit into the given framework with slightly different signs. However, the discretization technique described in the following can be done for the hole equations analogously.

3. Positivity preserving spatial discretization. In the following we describe the spatial discretization of (1.1)–(1.2). Let $\Omega \subset \mathbb{R}^d$ with $d = 1, 2$ be a polygonal domain and let $\mathcal{T}_h$ be a regular family of decompositions of Ω into triangles K such that there is no element across the interface of the Dirichlet boundary Γ_D and the Neumann boundary Γ_N. Let $\mathcal{E}_h$ be the set of edges e of $\mathcal{T}_h$ and $\mathcal{E}_{h,D}$ the set of edges belonging to Γ_D. The idea of the discretization of the stationary version of (1.1)–(1.2) described in [14] is to symmetrize the equations by introducing the Slotboom-variable and to discretize the symmetric form by use of mixed finite elements. This enforces the introduction of an independent current variable. A suitable discrete change of variables then allows to return to the natural variable. We will use this approach for the transient problem. Instead of a positivity preserving linear algebraic equation we derive a positivity conserving linear DAE or ODE, respectively, for the semi-discretized system.

With the notation $v \geq 0$ for a vector $v \in \mathbb{R}^n$ with non-negative components we state the following:

Definition 3.1. *Consider an ODE-system in $\mathbb{R}^n$ for $t \geq 0$*

$$y'(t) = f(t, y(t)). \tag{3.1}$$

The system will be called <u>*positive*</u> *if*

$$y(0) \geq 0 \qquad \Rightarrow \qquad y(t) \geq 0 \quad \textit{for all } t > 0.$$

Theorem 3.2 (See [12]). *Suppose that $f(t, v)$ is continuous and satisfies a Lipschitz condition with respect to v. Then the system* (3.1) *is positive iff for any vector $v \in \mathbb{R}^n$ and all $i = 1, \ldots, n$ and $t \geq 0$,*

$$v \geq 0, \quad v_i = 0 \qquad \Rightarrow \qquad f_i(t, v) \geq 0.$$

Corollary 3.3. *The system*

$$y'(t) = \mathcal{A}(t)y(t) + f(t)$$

is positive for $f(t) \geq 0$ with $t \geq 0$, if $a_{ij} \geq 0$ for all $j \neq i$.

3.1. Approximation and test spaces. We assume the potential V to be piecewise linear. Thus, ∇V is constant on each element K and we introduce the Slotboom-variable $y = e^{-V}g$. Thus equations (1.1) turn into

$$\partial_t(e^V y) - \text{div}J + \sigma e^V y = f, \qquad e^{-V}J = \nabla y \qquad \text{in } \Omega. \tag{3.2}$$

To define the approximation and test spaces for the mixed finite element approach developed in [15] we introduce a set of polynomial vectors to approximate the current density on the element K:

$$\Sigma(K) = \text{span}(\tau_1, \tau_2, \tau_3) \qquad \text{with} \qquad \tau_1 = (1, 0), \quad \tau_2 = (0, 1), \quad \tau_3 = (\omega_1, \omega_2).$$

Let e_1 be the edge of the triangle K connecting the vertices with the smallest values of the potential V. (We note that for the corresponding discretization of the hole equation we denote by e_1 the edge connecting the vertices with the largest values of

V.) We number the remaining edges counter-clockwise. We choose $\tau_3 = (\omega_1, \omega_2)$ with $\omega_1, \omega_2 \in \Pi_2(K)$ fulfilling the following conditions:

$$\tau_3 \cdot \nu_2|_{e_2} = \tau_3 \cdot \nu_3|_{e_3} = 0, \qquad \tau_3 \cdot \nu_1|_{e_1} = 1, \tag{3.3}$$

$$\int_K \tau_3 \cdot \mathrm{curl}(\lambda_1, \lambda_2, \lambda_3)\ dx\ dy = 0, \tag{3.4}$$

$$\int_K \omega_1\ dx\ dy = \int_K \omega_2\ dx\ dy = 0, \tag{3.5}$$

where λ_i denotes the $i-$th barycentric coordinate. It can be easily checked that $\dim(\mathrm{div}\Sigma(K)) = 1$.

With that we define the finite dimensional spaces

$$V_h = \{\tau \in [L^2(\Omega)]^2 : \tau|_K \in \Sigma(K) \forall K \in \mathcal{T}_h\} \tag{3.6}$$

$$W_h = \{\phi \in L^2(\Omega) : \phi|_K \in \Pi_0(K) \forall K \in \mathcal{T}_h\}, \tag{3.7}$$

$$\Lambda_{h,\xi} = \{q \in L^2(\mathcal{E}_h) : q|_K \in \Pi_0(e) \quad \forall e \in \mathcal{E}_h; \int_e (q - \xi)\ ds = 0 \quad \forall e \in \mathcal{E}_{h,D}\}, \tag{3.8}$$

with any function $\xi \in L^2(\Gamma_D)$. The spaces $\Pi_0(K)$ and $\Pi_0(e)$ denote the sets of constant functions in K and on e, respectively, whereas the space $\Pi_2(K)$ denotes the set of degree two polynomials on K. After going back to the natural variable g, the mixed-hybrid formulation of the transient problem (1.1)–(1.2) reads:

Find $J^h \in V_h, \overline{g}^h \in W_h, g^h \in \Lambda_{h,g_D}$ such that

$$\sum_{K \in \mathcal{T}_h} \left(\int_K Q_K J^h \tau\ dx + \int_K S_K \overline{g}^h \mathrm{div}\tau\ dx - \sum_{e_i(K)} \int_{e_i} S_{e_i} g^h \tau \cdot \nu\ ds \right) = 0, \tag{3.9}$$

$$\sum_{K \in \mathcal{T}_h} \left(\int_K \partial_t \overline{g}^h \phi\ dx - \int_K \mathrm{div} J^h \phi\ dx + \int_K \sigma \overline{g}^h \phi\ dx \right) = \sum_{K \in \mathcal{T}_h} \int_K f \phi\ dx, \tag{3.10}$$

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} q J^h \cdot \nu\ ds = 0 \tag{3.11}$$

for all $\tau \in V_h, \phi \in W_h$ and $q \in \Lambda_{h,0}$. The functions Q and S denote piecewise constant functions defined on each triangle K by

$$Q_K = \frac{1}{|K|} \int_K e^{-V}\ dx, \qquad S_K = \max_i S_{e_i(K)} = S_{e_1(K)}, \qquad S_{e_i(K)} := \frac{1}{|e_i|} \int_{e_i} e^{-V}\ ds,$$

where the piecewise constant function S is defined via the average on the edges in order to approximate large gradients of the potential with the correct order of magnitude (see [15]).

We notice, that the edge e_1 connects the vertices with the smallest values of V. Moreover, ν denotes the outer normal unit vector on ∂K and $e_i(K)$ denotes the set of edges of the triangle K. $J^h \in V_h$ denotes the approximation to the current density J, $\overline{g}^h \in W_h$ denotes the piecewise constant approximation of g and $g^h \in \Lambda_{h,g_D}$ is the approximation on the edges of the triangulation. Equations (3.9)–(3.10) are the weak formulation of the equations in (3.2), (3.11) imposes a continuity requirement for the normal component of the current density at the interelement boundaries.

We remark that the continuity of the current variable is not demanded in the ansatz space (3.3) directly. It is stated as a constraint in the weak formulation, such that the variable g^h can be considered as corresponding Lagrange multiplier.

In the stationary case, the variables J^h and $\overline{g}^h$ can be eliminated by static condensation, thus that we end up with a system on the Lagrange multipliers g^h only, where the resulting stiffness matrix is an M-matrix. In the following we adopt this approach to the transient system in order to derive a positivity preserving ODE or DAE, respectively. For this we state the weak formulation in the matrix-vector notation:

$$\begin{pmatrix} 0 \\ E \\ 0 \end{pmatrix} \partial_t \overline{g}^h + \begin{pmatrix} A & \widetilde{B}^\top & -\widetilde{C}^\top \\ -B & D & 0 \\ C & 0 & 0 \end{pmatrix} \begin{pmatrix} J^h \\ \overline{g}^h \\ g^h \end{pmatrix} = \begin{pmatrix} 0 \\ F \\ 0 \end{pmatrix}. \tag{3.12}$$

For the number of elements n_K and the number of internal edges n_I the matrices $A \in \mathbb{R}^{2n_K \times 2n_K}, B \in \mathbb{R}^{n_K \times 2n_K}, C \in \mathbb{R}^{n_I \times 2n_K}, D \in \mathbb{R}^{n_K \times n_K}$ and $E \in \mathbb{R}^{n_K \times n_K}$ are given by the corresponding elementary matrices denoted by the superscript K:

$$A^K_{jk} = Q_K \int_K \tau_j \tau_k \, dx, \qquad B^K_{jk} = \int_K \phi_j \mathrm{div} \tau_k \, dx, \qquad C^K_{jk} = \int_{\partial K} q_j \tau_k \cdot \nu \, ds,$$
$$D^K_{jk} = \int_K \sigma \phi_j \phi_k \, dx, \qquad E^K_{jk} = \int_K \phi_j \phi_k \, dx,$$

where τ_k, ϕ_k and q_k are the canonical basis functions of the corresponding spaces in (3.6)–(3.8). Moreover it holds

$$\widetilde{B}^K = B^K S_K, \qquad \widetilde{C}^K = \mathrm{diag}(S_{e_1}, S_{e_2}, S_{e_3}) C^K.$$

Considering the terms containing the time derivative as right hand side, we can accomplish the static condensation procedure applied in [15]. The complete elimination of the variable $\overline{g}^h$ enforces further approximations on the time derivative $\partial_t \overline{g}^h$.

3.2. DAE for g^h. In the following we omit the superscript K denoting the element matrices. Thus, in the following the notation B is used for the contribution of one element to the overall matrix B as well as for the overall matrix itself. From the context it will be clear, which matrix is meant.

The first weak equation then leads to

$$J^h = A^{-1} \widetilde{C}^\top g^h - A^{-1} \widetilde{B}^\top \overline{g}^h \tag{3.13}$$

Inserting (3.13) into the second and third weak equations leads to

$$E \partial_t \overline{g}^h + (BA^{-1} \widetilde{B}^\top + D) \overline{g}^h - BA^{-1} \widetilde{C}^\top g^h = F, \tag{3.14}$$

$$CA^{-1} \widetilde{C}^\top g^h = CA^{-1} \widetilde{B}^\top \overline{g}^h. \tag{3.15}$$

The matrix A has diagonal structure an can be easily inverted. The same holds for the matrix $BA^{-1}\widetilde{B}^\top + D$. The static condensation leads to equation

$$\overline{\mathcal{A}} \partial_t \overline{g}^h + \mathcal{M} g^h = G, \tag{3.16}$$

with

$$\begin{aligned} \overline{\mathcal{A}} &= CA^{-1} \widetilde{B}^\top (BA^{-1} \widetilde{B}^\top + D)^{-1} E, \\ \mathcal{M} &= CA^{-1} \widetilde{C}^\top - CA^{-1} \widetilde{B}^\top (BA^{-1} \widetilde{B}^\top + D)^{-1} BA^{-1} \widetilde{C}^\top, \\ G &= CA^{-1} \widetilde{B}^\top (BA^{-1} \widetilde{B}^\top + D)^{-1} F. \end{aligned}$$

Here we face the problem, of both approximations g^h and $\overline{g}^h$ occurring. At this step we make use of the fact, that the Lagrange multipliers g^h have good approximation property on the value of the density variable g at the element boundaries. In stationary case we get the relation

$$\overline{g}^h = [BA^{-1}\widetilde{B}^\top + D]^{-1}(F + BA^{-1}\widetilde{C}^\top g^h), \tag{3.17}$$

what for $\sigma = 0$ and $f = 0$ in (1.1) reduces to a simple upwind scheme, see [15]. For the exact shape of the matrices see below. Using this, we approximate the time derivative term $\partial_t \overline{g}^h_K \approx \partial_t g^h_{e_1}$ by use of an upwind scheme. Thus we end up with a system of equations operating on the Lagrange multipliers g^h, only:

$$\mathcal{A}\partial_t g^h + \mathcal{M} g^h = G. \tag{3.18}$$

3.3. Shape of the matrices. In order to analyze the DAE, respectively ODE, (3.18) with respect to index and positivity, we have to consider the occurring matrices: We introduce the following notation

$$n^i = v^i |e_i| \qquad i = 1, 2, 3, \tag{3.19}$$

$$\gamma = \int_K (\omega_1^2, \omega_2^2)\, dx, \tag{3.20}$$

with v^i being the outer unit normal vector on edge e_i. Then the element matrices are

$$A^K = Q_K \begin{pmatrix} |K| & 0 & 0 \\ 0 & |K| & 0 \\ 0 & 0 & \gamma \end{pmatrix} \qquad C^K = \begin{pmatrix} n_1^1 & n_2^1 & |e_1| \\ n_1^2 & n_2^2 & 0 \\ n_1^3 & n_2^3 & 0 \end{pmatrix} \tag{3.21}$$

$$B^K = \begin{pmatrix} 0 & 0 & |e_1| \end{pmatrix} \qquad D^K = \sigma |K| \qquad E^K = |K|. \tag{3.22}$$

This leads to the (elementary) matrices where the superscript K is omitted

$$[BA^{-1}\widetilde{B}^\top + D]^{-1} = \frac{Q_K \gamma}{S_K |e_1^2| + \sigma |K| Q_K \gamma},$$

$$(CA^{-1}\widetilde{C}^\top)_{ij} = \begin{cases} \frac{S_K}{Q_K}\left(\frac{n^1 \cdot n^1}{|K|} + \frac{|e_1|^2}{\gamma}\right) & \text{if } i = j = 1, \\ \frac{S_{e_j}}{Q_K}\frac{n^i \cdot n^j}{|K|} & \text{else} \end{cases} \qquad CA^{-1}\widetilde{B}^\top = \begin{pmatrix} \frac{|e_1^2| S_K}{\gamma Q_K} \\ 0 \\ 0 \end{pmatrix},$$

Thus, the final elementary matrices $\mathcal{M}^K = (m_{ij})$ and $\mathcal{A}^K = (a_{ij})$ in (3.18) are given as:

$$m_{ij} = \begin{cases} \frac{S_K}{Q_K}\frac{n^1 \cdot n^1}{|K|} + \sigma\beta(\sigma) & \text{if } i = j = 1, \\ \frac{S_{e_i}}{Q_K}\frac{n^i \cdot n^j}{|K|} & \text{else} \end{cases} \qquad a_{ij} = \begin{cases} \beta(\sigma) & \text{if } i = j = 1, \\ 0 & \text{else} \end{cases}$$

with

$$\beta(\sigma) = |e_1|^2 |K| \left(|e_1|^2 + \sigma |K| \gamma \frac{Q_K}{S_K} \right)^{-1}.$$

In [15] it is shown, that the elementary matrix $\mathcal{M}^K$ is an M-matrix for triangulation of weakly acute type. This property holds for the assembled matrix $\mathcal{M}$. We notice, that the elementary mass-matrix $\mathcal{A}^K$ is a diagonal but singular matrix. The assembled matrix $\mathcal{A}$ then is still a diagonal matrix that is not necessarily singular.

Remark 3.4. *In one-dimensional case, the elementary mass-matrix $\mathcal{A}^K$ is a singular diagonal matrix. The complete mass-matrix $\mathcal{A}$ is singular if and only if the potential V has a local maximum in Ω that is not on Γ.*

Knowing the shape of the corresponding matrices, we can make statements about the index and positivity of (3.18).

Theorem 3.5. *For given potential V, equation* (3.18) *describes a DAE with index of at most 1.*

Proof. If the assembled mass-matrix $\mathcal{A}$ turns out to be regular, the system is an ODE and the index is zero.

If, however, the mass-matrix is singular, we notice, that it has diagonal structure. Renumbering the edges allows us to write the DAE (3.18) in the shape

$$\mathcal{A}_1 \partial_t g_1^h = -\mathcal{M}_{11} g_1^h - \mathcal{M}_{12} g_2^h + f_1(t), \tag{3.23}$$

$$0 = -\mathcal{M}_{21} g_1^h - \mathcal{M}_{22} g_2^h + f_2(t), \tag{3.24}$$

with a regular diagonal matrix $\mathcal{A}_1$. Independent of the numbering of elements and edges, the assembled stiffness-matrix $\mathcal{M}$ is an M-matrix, what holds for the submatrix $\mathcal{M}_{22}$ as well. Thus, $\mathcal{M}_{22}$ is invertible and the DAE has index 1. □

Theorem 3.6. *Using the described upwind approximation for the time derivative, i.e. $\partial_t \overline{g}_K^h \approx \partial_t g_{e_1}^h$ the solution of the DAE* (3.18) *is non-negative.*

Proof. If $\mathcal{A}$ is non-singular, the statement follows from the M-matrix property of $\mathcal{M}$ and the non-negativity of the diagonal of $\mathcal{A}$ with corollary 3.3. Otherwise according to (3.23), the DAE can be written in semi-explicit shape. As $\mathcal{M}_{22}$ is a M-matrix it can be inverted and the inverse has only positive entries. Moreover $\mathcal{A}_1$ can be inverted and the system can be written as a descriptor system:

$$\begin{aligned} \partial_t g_1^h &= \mathcal{H}_1 g_1^h + \widetilde{f}_1(t), \\ g_2 &= \mathcal{H}_2 g_1^h + \widetilde{f}_2(t), \end{aligned}$$

with

$$\mathcal{H}_1 = \mathcal{A}_1^{-1}(-\mathcal{M}_{11} + \mathcal{M}_{12}\mathcal{M}_{22}^{-1}\mathcal{M}_{21}), \qquad \mathcal{H}_2 = -\mathcal{M}_{22}^{-1}\mathcal{M}_{21}, \tag{3.25}$$

$$\widetilde{f}_1(t) = \mathcal{A}_1^{-1}\left(f_1(t) - \mathcal{M}_{12}\mathcal{M}_{22} f_2(t)\right), \qquad \widetilde{f}_2(t) = \mathcal{M}_{22}^{-1} f_2(t). \tag{3.26}$$

As $\mathcal{M}_{11}$ and $\mathcal{M}_{22}$ are both M-matrices and $\mathcal{M}_{12}, \mathcal{M}_{21}$ contain only non-positive entries (namely off-diagonal entries of the matrix $\mathcal{M}$), the matrix $\mathcal{H}_1$ is a $-$Z-matrix (the off-diagonal elements are non-negative). Moreover the matrix $\mathcal{H}_2$ and the source terms $\widetilde{f}_1, \widetilde{f}_2$ are non-negative and thus according to [9, 22] the solution of the descriptor system is non-negative. □

We notice, that the matrix $\mathcal{A}_1$ in (3.23) has to be an inverse positive matrix in order to ensure positivity for the descriptor system above. However, use of the trapezoidal rule in 1D or use of all three edge-approximations $g_{e_i}^h$ to approximate the integral $\int_K \partial_t \overline{g}^h \, ds$ in two dimensions, leads - with corresponding numbering of the edges - to a tridiagonal mass matrix $\mathcal{A}_1$ with non-negative elements only. It is easy to verify, that for positive off-diagonal entries, this matrix is not positive inverse, see [10]. Thus, for another choice than the upwind scheme, the positivity of the DAE can't be ensured.

3.4. ODE for g^h. Theorem 3.5 states, that for given potential V the index of the system is at most one. However, for semiconductor application, the continuity equation is coupled to the Poisson equation modeling the potential distribution in the device. If the potential distribution changes, even the rank of the leading mass matrix $\mathcal{A}$ might change and we face the problem of rank deficiency of the coupled system described in Section 2.

A simple approximation of the mass matrix $\mathcal{A}$ allows to get rid of the rank deficiency problem. For each edge, that does not contribute to the mass-matrix, i.e. the corresponding entry on the diagonal is zero, we choose one of the elements, the edge belongs to and approximate the elementary mass matrix by

$$\mathcal{A}^K = \beta(\sigma)\begin{pmatrix} 1-\epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \text{or} \qquad \mathcal{A}^K = \beta(\sigma)\begin{pmatrix} 1-\epsilon & 0 & 0 \\ 0 & \frac{\epsilon}{2} & 0 \\ 0 & 0 & \frac{\epsilon}{2} \end{pmatrix}.$$

The complete mass matrix then will be a regular diagonal matrix with positive values. Thus, according to Theorem 3.2 the resulting system will be a positive ODE for the lagrange multipliers g^h. In the numerical experiments below, we have chosen a value of $\varepsilon = \frac{1}{4}$. The experience shows, that compared to the number of edges, the number of edges not contributing to the mass-matrix is comparatively small.

Moreover, we remark that the singular mass-matrix results from the upwind approximation for $\partial_t \overline{g}$. The introduction of the parameter $\varepsilon > 0$ counterworks this approximation as it somehow also incorporates the the influence of the other nodes on the corresponding interval in the one-dimensional case (and analogous in two dimensions). In experiments we even observed better accuracy after introduction of the parameter $\varepsilon = \frac{1}{4}$.

3.5. Generalization to Petrov-Galerkin approach. We recall, that to ensure positivity of the space discretized system (3.18), the off-diagonal elements of $\mathcal{M}$ have to be non-positive and the diagonal elements of $\mathcal{A}$ have to be non-negative. The Marini-Pietra mixed finite elements allow us to keep positivity with the restriction that the final equation is of differential-algebraic type and suffers from rank deficiency when coupled to the Poisson equation.

We notice, that the conditions (3.5) are chosen such that the matrix A in (3.12) has block diagonal structure, what makes it easily invertible. If we keep this condition in order to keep the numerical effort low, no other choice of τ_3 allows to derive a positive ODE for the semi-discretized system. Not even the generalization to a Petrov-Galerkin approach using different ansatz and test space would lead to a semi-discretized system fulfilling these conditions.

If we namely assume the ansatz functions given as in (3.6)–(3.8) with general values for the ansatz function τ_3 and the test function $\widetilde{\tau}_3$

$$\int_{e_1} \tau_3 \cdot \nu \, ds = \eta_1, \quad \int_{e_2} \tau_3 \cdot \nu \, ds = \eta_2, \quad \int_{e_3} \tau_3 \cdot \nu \, ds = \eta_3, \quad \int_K \text{div}\tau_3 \, dx = \delta$$
$$\int_{e_1} \widetilde{\tau}_3 \cdot \nu \, ds = \widetilde{\eta}_1, \quad \int_{e_2} \widetilde{\tau}_3 \cdot \nu \, ds = \widetilde{\eta}_2, \quad \int_{e_3} \widetilde{\tau}_3 \cdot \nu \, ds = \widetilde{\eta}_3, \quad \int_K \text{div}\widetilde{\tau}_3 \, dx = \widetilde{\delta}$$

a simple computation shows the shape of the resulting mass-matrix $\overline{\mathcal{A}} = (\overline{a}_{ij})$ and stiffness matrix $\mathcal{M} = (m_{ij})$ in (3.16) is

$$m_{ij} = \frac{S_{e_i}}{Q_K}\frac{n^i \cdot n^j}{|K|} + \sigma|K|\widetilde{\beta}(\sigma)\widetilde{\eta}_i\eta_j \qquad \overline{a}_{i1} = \widetilde{\delta}|K|\widetilde{\beta}(\sigma)\eta_i. \qquad \text{for } i,j = 1,2,3,$$

with

$$\widetilde{\beta}(\sigma) = \left(\delta\widetilde{\delta} + \sigma|K|\gamma\frac{Q_K}{S_K}\right)^{-1}.$$

We assume $\widetilde{\beta} > 0$ as it is for $\widetilde{\delta} = \delta$. For $\widetilde{\delta} > 0$, in order to get a positive contribution into each line of the mass matrix, it has to hold $\eta_1, \eta_2, \eta_3 > 0$. In order to keep positivity of the final ODE, we have to ensure that the off-diagonal elements of $\mathcal{M}$ are non-positive. Thus it has to hold $\widetilde{\eta}_1, \widetilde{\eta}_2, \widetilde{\eta}_2 \leq 0$. This leads to contradiction as $\widetilde{\eta}_1 + \widetilde{\eta}_2 + \widetilde{\eta}_3 = \widetilde{\delta} > 0$. A similar arguments leads to contradiction for $\widetilde{\delta} < 0$. If we allow $\widetilde{\beta}(\sigma)$ to be smaller than zero (what can't be ensured for any choice $\sigma > 0$), again similar argumentation as above leads to contradiction.

3.6. ODE for $\overline{g}^h$. In section 3.2 we used the weak formulation of the problem to derive a DAE or ODE for the Lagrange multipliers using several simplifying assumptions. On the other hand, using (3.13)–(3.15) we can derive an ODE for the approximation $\overline{g}^h$. Under the assumption of homogenous Dirichlet boundary conditions, i. e. $g_D = 0$ this leads to :

$$E\partial_t\overline{g}^h + \overline{\mathcal{M}}\overline{g}^h = F, \tag{3.27}$$

$$\overline{\mathcal{M}} = BA^{-1}\widetilde{B}^\top + D - BA^{-1}\widetilde{C}^\top[CA^{-1}\widetilde{C}^\top]^{-1}CA^{-1}\widetilde{B}^\top. \tag{3.28}$$

From section 3.3 we see, that the matrix $BA^{-1}\widetilde{B}^\top + D$ is a positive 1×1 matrix per element. Thus the corresponding overall matrix is a diagonal matrix with positive entries. Moreover, the (elementary)matrices $BA^{-1}\widetilde{C}^\top$, $CA^{-1}\widetilde{B}^\top$ contain only non-negative values. Finally, for triangulation of weakly acute type the (elementary) matrix $CA^{-1}\widetilde{C}^\top$ is a diagonally dominant M-matrix. Thus, the off-diagonal elements of $\overline{\mathcal{M}}$ are non-positive. As E is a regular diagonal matrix with positive values on the diagonal, for non-negative initial conditions, ODE (3.27) is positive.

However, in contrast to the procedure in section 3.2 the computation is not that straightforward, as the matrix to invert, $CA^{-1}\widetilde{C}^\top$, does not have diagonal structure. Thus the computation can not be accomplished elementwise. In Section 5.2 we will see that this is the bottleneck and makes it preferable to use the approximation from the previous section for simulation on fine grids.

3.6.1. Incorporation of boundary conditions. In case of inhomogeneous Dirichlet boundary conditions, the positivity of the system can still be ensured, as long as the boundary conditions are non-negative.

We notice that the elementary matrices $CA^{-1}\widetilde{C}^\top$ for boundary elements reduce to 1×3 or 2×3 matrices, depending on wether the boundary element has two or one boundary edges, respectively. The entries corresponding to the internal edges enter the assembled matrix, whereas the 'off-diagonal' elements enter the right hand side. Thus equation (3.15) changes with boundary conditions into:

$$CA^{-1}\widetilde{C}^\top g^h = CA^{-1}\widetilde{B}^\top\overline{g}^h + N_{D,1}g_D, \tag{3.29}$$

where g_D denotes the vector containing the values of g on the boundary edges and $N_{D,1}$ is a $n_I \times n_D$ matrix, with n_I the number of internal edges and n_D the number of boundary edges. The entries of $N_{D,1} = (n^1)_{ij}$ are

$$n^1_{ij} = \begin{cases} -\frac{S_{e(\Gamma_{D,i})}}{Q_K} \frac{n^{I_i} \cdot n^{\Gamma_j}}{|K|}, & \text{if } \Gamma_j \text{ and } e_{I_i} \text{belong to element } K, \\ 0 & \text{else,} \end{cases}$$

where e_{I_i} denotes the i-th internal edge of the triangulation. We notice that the matrix $BA^{-1}\widetilde{C}^\top$ in (3.14) causes, that only the values of g^h in the minimum edges, i.e. the edge connecting the vertices with the smallest values of V, contribute to the equation. As g^h only denotes the internal edges, the case that the minimum edge at the boundary elements are the boundary edges is not incorporated.

Considering this case we end up with

$$E\partial_t \overline{g} + \overline{\mathcal{M}}\overline{g} = \overline{F}, \tag{3.30}$$

$$\overline{F} = F + BA^{-1}\widetilde{C}^\top [CA^{-1}\widetilde{C}^\top]^{-1} N_{D,1} g_D + N_{D,2} g_D \tag{3.31}$$

with $N_{D,2} = (n^2)_{ij}$ having the shape

$$n^2_{ij} = \begin{cases} \frac{|e_{\Gamma_j}|^2}{Q_K \gamma}, & \text{if } \Gamma_j \text{ and } e_{I_i} \text{belong to element } K \text{and } \Gamma_j \text{ is the min. edge,} \\ 0 & \text{else.} \end{cases}$$

The matrices $N_{D,1}, N_{D,2}$ and $BA^{-1}\widetilde{C}^\top$ are all non-negative and $CA^{-1}\widetilde{C}^\top$ is an M-matrix. Thus non-negative Dirichlet boundary values lead to non-negative contribution to the right hand side and ODE (3.30) is positive.

4. Positivity preserving time discretization. So far we described to ways to use mixed finite elements for space discretization of the drift-diffusion equations in order to achieve a positive system. In both cases, the ODE for g^h or $\overline{g}^h$, the system can be written as a linear inhomogeneous system

$$y'(t) = \mathcal{A}(t)y(t) + f(t) + b(t), \tag{4.1}$$

where $f(t)$ denotes the time dependent source term and $b(t)$ contains the contribution of the boundary conditions. We assume, that $b(t) \geq 0$ and $f(t) \geq 0$.

We have shown that the system is positive. However this is not necessarily true for the numerical solution.

For linear systems there exists a complete theory developed by Bolley & Crouzeix [2] concerning the positivity preservation of one-step methods. They showed that for linear systems of type

$$y'(t) = \mathcal{A}(t)y(t)$$

where the matrix $\mathcal{A}$ satisfies

$$a_{ij} \geq 0 \text{ for } i \neq j \qquad \text{and} \qquad a_{ii} \geq -\alpha \tag{4.2}$$

with problem parameter α and furthermore $\mathcal{A}$ has no eigenvalues on the positive real axis, the applied one-step method is positivity preserving if the condition

$$\tau\alpha \leq \gamma_R$$

is fulfilled. There, τ denotes the applied time step size and γ_R stands for the threshold factor of the method. It is known, that there does not exist any one-step method of order $p > 1$ with $\gamma_R = \infty$. Thus, for any higher order method time step size restriction occurs. For all known second order one-step methods it holds $\gamma_R = O(1)$. In fact, for the drift-diffusion equations solved for simulation of the semiconductor device below, the (unscaled) time step size restriction is in the order of 10^{-15} s.

By use of exponential integrators positivity preserving second order methods can be constructed. This enforces the knowledge or computation of the exponential $\exp(\mathcal{A})$ in each time or iteration step, respectively, what is numerically expensive.

In the following we will describe a method based on a splitting method and combining explicit exponential integration with implicit one-step methods. The resulting method is of order two. It is not unrestricted positivity preserving but it allows to increase the restriction to the time step size to the underlying one-step method by a factor up to 10. The scheme is given in Algorithm 1.

Algorithm 1: Splitting scheme

1. At time t_n split the linear system (4.1) for $t \in [t_n, t_{n+1}]$ as follows

$$y'(t) = (\mathcal{A}_0 + \mathcal{A}_1(t))\, y(t) + f(t) + b_0 + b_1(t), \qquad y(t_n) \approx y_n$$

where

$$\begin{aligned} \mathcal{A}_0 &= \omega \mathcal{A}(t_n) & b_0 &= \omega b(t_n), \\ \mathcal{A}_1(t) &= \mathcal{A}(t) - \mathcal{A}_0 & b_1(t) &= b(t) - b_0. \end{aligned}$$

2. Compute $y_{n+1} \approx y(t_n + \tau)$ with $f_1(t, y) = \mathcal{A}_1(t) y(t) + f(t) + b_1(t)$ by

$$y_{n+1} = \mathcal{RE}\left(\frac{\tau}{2}, \mathcal{A}_0, b_0\right) \circ \mathcal{RK}\left(\tau, f_1(t, y)\right) \circ \mathcal{RE}\left(\frac{\tau}{2}, \mathcal{A}_0, b_0\right) \circ y_n.$$

There, $\mathcal{RK}(\tau, f_1(t)) \circ y_0$ denotes the solution of the problem

$$y'(t) = f_1(t, y) \qquad \text{with} \qquad y(0) = y_0$$

after one time step with time step size τ obtained by a one-step method of order two.

$\mathcal{RE}\left(\frac{\tau}{2}, \mathcal{A}_0, b_0\right) \circ y_0$ denotes the exact solution of the problem

$$y'(t) = \mathcal{A}_0 y(t) + b_0 \qquad \text{with} \qquad y(0) = y_0$$

solved by use of the exponential matrix

$$y\left(\frac{\tau}{2}\right) = \exp(\mathcal{A}_0) y_0 + [\exp(\mathcal{A}_0) - I]\left[\mathcal{A}_0^{-1} b_0\right].$$

If the applied one-step method is of at least order two, the described method is of order two. The explicit exponential solution namely is exact and the error of the applied splitting method is known to be of order two. The method $\mathcal{RE}$ is positivity preserving. As the matrix $\mathcal{A}$ has been split, the time step size restriction for positivity preservation of the method $\mathcal{RK}$ now is approximately given by $(1 - \omega)\alpha\tau \leq \gamma_R$. Thus for $\omega = 0.5$ the time step size can be increased by the factor 2 compared to the application of the one-step method only.

At that point the method does not seem to have any use, as the computation of the exponential is necessary and that does not bring any advantage compared to exponential integrators. However, for the semiconductor application described in Section 2 it will turn out to be efficient. After semi-discretization according to Section 3.4 of the drift-diffusion equations for electrons and holes are in the shape

$$g_n'(t) = \mathcal{A}_n(\cdot)g_n(t) + f_n(\cdot) + b_n(\cdot), \qquad g_n(t_0) = g_{n,0}, \tag{4.3}$$
$$g_p'(t) = \mathcal{A}_p(\cdot)g_p(t) + f_p(\cdot) + b_p(\cdot), \qquad g_p(t_0) = g_{p,0}, \tag{4.4}$$

where the matrices, source terms and terms incorporating the boundary conditions depend on time, potential and the densities themselves, namely t, V, g_n, g_p.

In semiconductor application, efficient iterative algorithms for solution of the coupled system of Poisson and drift-diffusion equations for electrons and holes exist, e.g. Gummel-iteration. We observe, that the leading matrices $\mathcal{A}_n, \mathcal{A}_p$ occurring in (4.3)–(4.4) don't change significantly in between iterations. That allows us to keep the matrix $\mathcal{A}_0$ and thus its exponential constant over several iteration steps and update the matrix $\mathcal{A}_1$ only. For (4.3)–(4.4) the combined Gummel-splitting-algorithm is given in Algorithm 2.

Computation of the exponential of the matrix. Compared to application of the one-step method only, we are able to increase the step size for positivity preservation by the factor $(1-\omega)^{-1}$. The necessary extra numerical effort is governed by the computation of the exponential of the matrix, what especially for large matrices enforces high effort. Thus an efficient algorithm for the computation of the matrix exponential is indispensable for this scheme. For matrices $\mathcal{A}$, where $\|\mathcal{A}\|$ is large the computation using Padé-Approximation or Taylor series is instable as it suffers from severe roundoff error difficulties. Therefore we make use of the fundamental property

$$\exp(\mathcal{A}) = \exp(\mathcal{A}/m)^m.$$

We use m to be a power of two, such that $\|\mathcal{A}/m\| \leq 1$. Then $e^{\frac{\mathcal{A}}{m}}$ can be satisfactorily computed by the Padé-approximation $R_{33}(\mathcal{A}/m)$ and $e^{\mathcal{A}}$ is computed by repeated squaring, see [16]. To increase efficiency we apply Strassen's algorithm [21] for the occurring matrix multiplication for large matrices. The algorithm is implemented recursively. For matrices with dimension smaller than 256, we then use the standard matrix multiplication in MATLAB. This combination proved to be the fastest. For larger matrices, the MATLAB multiplication is significantly slower than Strassen's algorithm. Moreover, due to it's recursive structure, the algorithm turned out to be slower than the MATLAB-solver for smaller matrices.

5. Numerical Results. In the following we apply the different discretization schemes from Section 3 and compare the performance of the suggested Algorithm 2 to those of the corresponding underlying standard one-step method. As one-step method we employ the second order Rosenbrock method used in the MATLAB-solver ode23s. For the ODE $g' = f(g)$ it is given by

$$\begin{aligned}
(I - a\tau J)k_1 &= f(g_i), \\
(I - a\tau J)k_2 &= f(g_i + \frac{\tau}{2}k_1) - a\tau J k_1, \\
g_{i+1} &= g_i + \tau k_2,
\end{aligned}$$

with $a = (2+\sqrt{2})^{-1}$, compare [19]. There, J denotes the Jacobian, I the identity matrix, τ the step size and g_i, g_{i+1} the approximation on $g(t_i), g(t_{i+1})$, respectively.

```
Algorithm 2: Coupling of Gummel-Iteration and Splitting method
for timesteps $\ell = 1 : N_T$ do
    1. Let $g_n^\star = g_n^{(\ell-1)}, g_p^\star = g_p^{(\ell-1)}, V = V^{(\ell-1)}$.
    2. repeat
        (a) Solve electron equation
            i. Determine $\mathcal{A}_n(t^{(\ell-1)}, V, g_n^\star, g_p^\star), b_n(t^{(\ell-1)}, V, g_n^\star, g_p^\star)$
            ii. if first iteration step in each 2nd time step $\ell$
                split $\mathcal{A}_{n,0} = \omega \mathcal{A}_n$ and $b_{n,0} = \omega b_n$
            iii. with $\mathcal{A}_{n,1}(\cdot) = \mathcal{A}_n(\cdot) - \mathcal{A}_{n,0}$ and $b_{n,1}(\cdot) = b_n(\cdot) - b_{n,0}$
                find $g_n$ at time $t_\ell$ using Algorithm 1.2 such that
                $\begin{cases} g_n'(t) = (\mathcal{A}_{n,0} + \mathcal{A}_{n,1}(\cdot))g_n(t) + f_n(\cdot) + b_{n,0} + b_{n,1}(\cdot), \\ g_n(t_{\ell-1}) = g_n^{(\ell-1)} \end{cases}$
        (b) Solve hole equation
            i. Determine $\mathcal{A}_p(V, g_n^\star, g_p^\star), f_p(V, g_n^\star, g_p^\star), b_p(V, g_n^\star, g_p^\star)$
            ii. if first iteration step in each 2nd time step $\ell$
                split $\mathcal{A}_{p,0} = \omega \mathcal{A}_p$ and $b_{p,0} = \omega b_p$
            iii. with $\mathcal{A}_{p,1}(\cdot) = \mathcal{A}_p(\cdot) - \mathcal{A}_{p,0}$ and $b_{p,1}(\cdot) = b_p(\cdot) - b_{p,0}$
                find $g_p$ at time $t_\ell$ using Algorithm 1.2 such that
                $\begin{cases} g_p'(t) = (\mathcal{A}_{p,0} + \mathcal{A}_{p,1}(\cdot))g_p(t) + f_p(\cdot) + b_{p,0} + b_{p,1}(\cdot), \\ g_p(t_{\ell-1}) = g_p^{(\ell-1)} \end{cases}$
        (c) Set $V_1 = V + \delta V$ with $n = \mu_n^{-1} g_n, p = \mu_p^{-1} g_p$ and
            $\lambda^2 \Delta(\delta V) - (p + n)\delta V = -\lambda^2 \Delta V + n - p - C,$
        (d) Set $g_n^\star := g_n, g_p^\star := g_p$ and $V := V_1$
    until $\|\delta V\|_2 < tol$ ;
    3. Set $g_n^{(\ell)} := g_n, g_p^{(\ell)} := g_p$ and $V^{(\ell)} = V$.
end
```

The threshhold factor of this method can be proved to be $\gamma_R = (1-2a)^{-1} \approx 2.41$. For the splitting method we apply the splitting factor $\omega = 0.9$ and update the exponential at each second time step.

We employ a one-dimensional simulation of a 400 nm *pn*-diode consisting of a 200 nm *p*-doped region with a doping concentration C_0 and a 200 nm *n*-doped region with a doping concentration $-C_0$. The physical parameters of the device are listed in Table 5.1. The device is modeled by the drift-diffusion equations as given in Section 2 and discretized in 1D by use of the Marini-Pietra mixed finite element approach presented in Section 3.

As initial conditions we assume the device to be in thermal equilibrium, see [6, 18]. The diode is backward biased once with 2 V and once with 0.25 V. We choose these two values in order to consider the case of strong backward bias (2 V) and the case when the potential distribution has a local maximum what causes the leading mass-matrix to be singular (0.25 V). A rough estimation assuming $n, p \approx C_0$ and assuming a linear potential distribution leads to a (unscaled) step size restriction of $\tau \leq 10^{-15}$ s for a one-step method with threshold factor $\gamma_R \approx 1$. In numerical examples, larger

Parameter	Physical meaning	Numerical value
L	length of device	$4 \cdot 10^{-7}\,\mathrm{m}$
q	elementary charge	$1.6 \cdot 10^{-19}\,\mathrm{As}$
ε_s	permittivity constant	$10^{-12}\,\mathrm{As/Vcm}$
U_T	thermal voltage at $T_L = 300K$	$0.026\,\mathrm{V}$
μ_n/μ_p	low-field carrier mobilities	$1500/450\,\mathrm{cm}^2/\mathrm{Vs}$
τ_n/τ_p	carrier lifetimes	$10^{-6}/10^{-5}\,\mathrm{s}$
n_i	intrinsic density	$10^{16}\,\mathrm{m}^{-3}$
C_0	maximum doping concentration	$10^{22}\,\mathrm{m}^{-3}$

Table 5.1
Physical parameters for a silicon pn-junction diode.

step sizes were applicable but we will see that the step size restriction still will be severe.

We restrict our numerical examples to the backward biased diode, as especially in these cases strong depletion zones occur that can lead to negative values for the densities, if standard discretization is applied.

We simulate the device and compute the nodal approximation to the electron density distribution g_n^h after 1 ps. This is before the stationary state is reached and thus allows for a comparison of the simulated transient behavior. In the following we distinguish between the different discretization approaches and numerical methods, respectively:

- ODE($\overline{g}^h$): Simulation using the ODE for $\overline{g}^h$ in (3.30). As numerical method the given Rosenbrock scheme is applied and finally g^h is computed according to (3.17).
- R-ODE(g^h):Using the ODE for g^h derived by upwind approximation and modification of the mass-matrix. As numerical method the given Rosenbrock method is is applied.
- S-ODE(g^h): Using the ODE for g^h as in R-ODE. As numerical method the coupled Gummel-splitting algorithm is applied with $\omega = 0.9$ and the given Rosenbrock scheme as underlying one-step method.

5.1. Accuracy. In order to compare the different approaches in terms of accuracy we compare them to a numerical reference solution. The reference solution is computed using ODE($\overline{g}^h$) on a very fine grid with time step size $\tau = 10^{-15}$ s and spatial step size $h = 1/500$.

Firstly, in Figure 5.1 we compare the values of the nodal and piecewise constant approximation of the electron density in the diode for the different applied bias. The values have been obtained by use of ODE($\overline{g}^h$). We observe that in both cases (high and low bias) the upwind approximation is appropriate, ever for low bias, where the local extremum of the potential occurs.

In Figure 5.2 we depict the relative L^2 deviation of the different approaches from the reference solution. For all approaches we applied a time step size of $\tau = 10^{-14}$ s and we depict the deviation for different spatial step sizes. We see that even for the relatively coarse grid of 32 discretization nodes the accuracy of the different approaches is comparable. This justifies our approximations to derive the ODE for g^h.

In Figure 5.3 we compare the relative error of the different approximations on a grid with 64 discretization nodes. The resulting ODEs thereby have been solved with

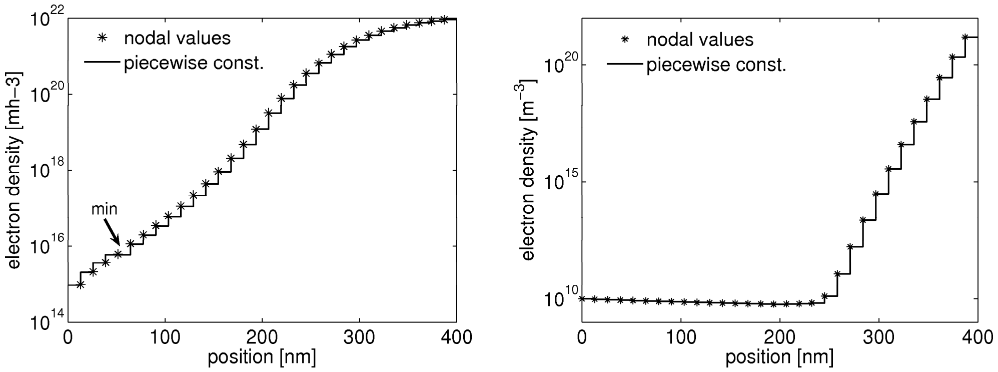

FIG. 5.1. *Comparison of piecewise constant and nodal approximation for the electron density. after 1 ps for backward bias of 0.25 V (left) and 2 V (right)*

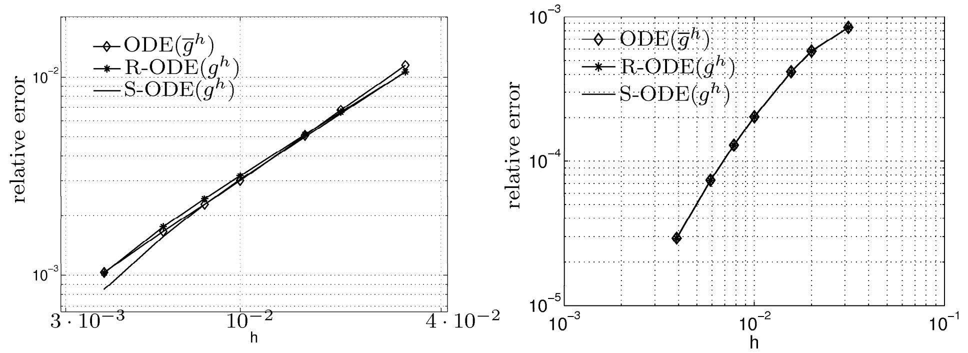

FIG. 5.2. *Relative deviation of the different approaches from the reference solution after 1 ps for backward bias of 0.25 V (left) and 2 V (right).*

different time step sizes. Again we observe that the results are all comparable and our approximations done in Section 3 do not introduce larger errors. Moreover, we see that even for rather big time step sizes in the range of $\tau = 10^{-13}$ s the time discretization error is smaller than the space discretization error in the described example. This shows that with respect to accuracy smaller time steps are not needed.

5.2. Runtime acceleration by splitting method. Finally we compare the different approaches with respect to runtime and positivity preservation. We recall that the bottleneck of ODE($\overline{g}^h$) is the inversion of the matrix $CA^{-1}\widetilde{C}^\top$ (see Section 3.6). The bottleneck of S-ODE(g^h) is the computation of the exponential of the matrix. On the other hand for R-ODE(g^h) and again ODE($\overline{g}^h$) we face the problem of severe step size restriction for positivity preservation.

In Table 5.2 we depict the runtime needed for the different approaches for the performance of one iteration step for different fine grids. We notice that the implemented matrix multiplication using Strassen's algorithm is most efficient for matrices with a dimension given by a power of two, what explains the chosen number of discretization nodes. As expected R-ODE(g^h) performs fastest. As for finer grids the computation of the matrix inverse and the exponential of the matrix need significantly more effort, the other two approaches fall behind for finer grids. However, we notice

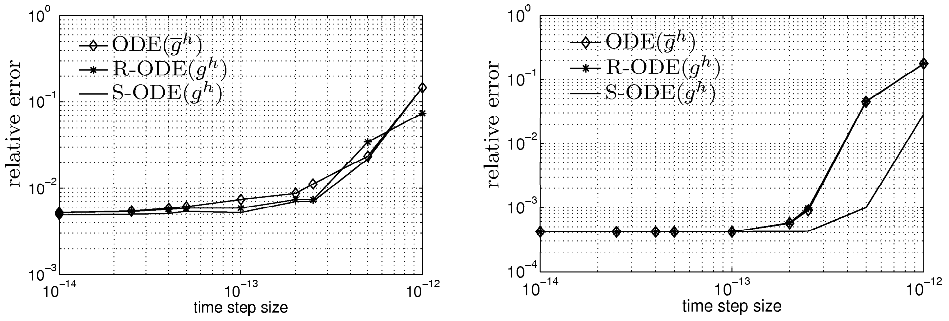

FIG. 5.3. *Relative deviation of the different approaches from the reference solution after 1 ps for bias of 0.25 V (left) and 2 V (right).*

# nodes	64	128	256	512	1024
ODE($\overline{g}^h$)	$5.1 \cdot 10^{-3}$	$1.4 \cdot 10^{-2}$	0.12	1.24	6.44
R-ODE(g^h)	$1.46 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$4.2 \cdot 10^{-3}$	$7.2 \cdot 10^{-3}$
S-ODE(g^h)	$3.9 \cdot 10^{-3}$ $(8.3 \cdot 10^{-4})$	$1.3 \cdot 10^{-2}$ $(1.0 \cdot 10^{-3})$	$9.1 \cdot 10^{-2}$ $(1.7 \cdot 10^{-3})$	1.01 $(3.1 \cdot 10^{-3})$	10.25 $(7.1 \cdot 10^{-3})$

TABLE 5.2

Runtime for iteration step with different methods for different number of discretization intervals: The number in parentheses denote the runtime without consideration of the computation of the exponential.

that the computation time for one iteration step does not reflect the entire performance of the algorithm S-ODE(g^h). In the coupled Gummel-splitting algorithm, the exponential only has to be computed once per time step (or even less) instead of each iteration step. On the other hand the numerical effort for the approximation ODE($\overline{g}^h$) increases significantly with finer grids. As the matrix has to be inverted in each iteration step the slightly better accuracy this approach might have does not pay off for the significantly higher effort.

In Tables 5.3–5.4 we compare the runtime of the different approaches for the simulation of the *pn*-diode for 1 ps backward biased with 2 V. Moreover we list the minimum nodal density value g^h computed by the different approaches. Obviously the effort of the ODE($\overline{g}^h$)-approach is very high and out of question for large matrices as they easily occur for two-dimensional geometries.

Comparing the Rosenbrock scheme and the splitting algorithm we see that for small matrices the extra numerical effort for the computation of the exponential is almost negligible. Only if we update the exponential once per iteration step, the numerical costs would be significantly higher. On the other hand, the splitting algorithm allows us to increase the time step size for positivity preservation. For the relatively coarse grid of 64 nodes we observe that increasing the step size by the factor five or ten and applying the splitting algorithm allows us to speed up our computation almost by the factor of three. As we already saw in Figure 5.3 the smaller step size for the Rosenbrock scheme does not even lead to better accuracy.

For finer grids and larger matrices, the costs for the exponential of the matrix increase, as we see from Table 5.4. But even for a discretization with 256 nodes we see that with a ten times larger time step size the splitting algorithm keeps positivity and is approximately 20% faster than the applied Rosenbrock scheme. For larger

time step	$5 \cdot 10^{-13}$	$2 \cdot 10^{-13}$	10^{-13}	$5 \cdot 10^{-14}$	$2 \cdot 10^{-14}$
min value					
ODE($\overline{g}^h$)	$-7.57 \cdot 10^{20}$	$-4.08 \cdot 10^{20}$	$-1.94 \cdot 10^{19}$	$-7.58 \cdot 10^{14}$	$5.93 \cdot 10^{9}$
R-ODE(g^h)	$-7.6 \cdot 10^{20}$	$-4.23 \cdot 10^{20}$	$-2.24 \cdot 10^{19}$	$-1.91 \cdot 10^{15}$	$6.03 \cdot 10^{9}$
S-ODE(g^h)	$5.94 \cdot 10^{9}$	$6.01 \cdot 10^{9}$	$6.03 \cdot 10^{9}$	$6.03 \cdot 10^{9}$	$6.03 \cdot 10^{9}$
run time [s]					
ODE($\overline{g}^h$)	0.39	0.51	0.49	1.25	2.08
R-ODE(g^h)	0.06	0.22	0.11	0.39	0.56
S-ODE(g^h)	0.08	0.18	0.33	0.45	1.07

TABLE 5.3

Runtime and minimal value for the nodal approximation of electron density g by use of the different approximations for 64 discretization nodes different time step sizes. Backward bias: 2 V.

time step	$5 \cdot 10^{-13}$	$2 \cdot 10^{-13}$	10^{-13}	$5 \cdot 10^{-14}$	$2 \cdot 10^{-14}$
min value					
ODE($\overline{g}^h$)	$-7.75 \cdot 10^{20}$	$-4.40 \cdot 10^{20}$	$-2.63 \cdot 10^{19}$	$-4.53 \cdot 10^{15}$	$5.94 \cdot 10^{9}$
R-ODE(g^h)	$-7.75 \cdot 10^{20}$	$-4.42 \cdot 10^{20}$	$-2.66 \cdot 10^{19}$	$-4.89 \cdot 10^{15}$	$6.03 \cdot 10^{9}$
S-ODE(g^h)	$5.94 \cdot 10^{9}$	$6.01 \cdot 10^{9}$	$6.03 \cdot 10^{9}$	$6.03 \cdot 10^{9}$	$6.03 \cdot 10^{9}$
run time [s]					
ODE($\overline{g}^h$)	2.53	5.39	9.77	17.92	39.21
R-ODE(g^h)	0.06	0.09	0.18	0.36	0.88
S-ODE(g^h)	0.49	0.71	1.28	2.18	4.70

TABLE 5.4

Runtime and minimal value for the nodal approximation of electron density g by use of the different approximations for 256 discretization nodes different time step sizes. Backward bias: 2 V.

matrices, however, the one-step method with small step size will be faster than the proposed splitting algorithm.

Lastly we notice the following. In the numerical examples we observed that the Rosenbrock method keeps positivity for small time steps and for time step sizes larger than 10^{-12} s. This is due to the fact that the final matrices $\mathcal{A}_n, \mathcal{A}_p$ in (4.3)–(4.4) not only fulfill the conditions in (4.2) but are M-matrices. This corresponds to positivity preservation in the stationary case. For time steps larger than 10^{-12} s the M-matrix property 'dominates' the system and leads to positivity preservation. Thus, by use of the splitting method with a splitting factor of $\omega = 0.9$ we were not only able to increase the allowed step sizes but we also were able to close this 'gap'. This holds only for the numerical experiments and could not been proved yet. However, in the numerical experiments step sizes larger than $3 \cdot 10^{-13}$ s led to very inaccurate results.

6. Conclusion. In this paper we discretize the transient drift-diffusion equations occurring in semiconductor simulation such that the positivity of the particle densities is kept. Therefor we apply the mixed finite element scheme presented by Marini and Pietra in [14] for spatial discretization of the transient drift-diffusion equations. We present several approximative methods leading to a positive ODE or DAE, respectively, for the semi-discretized model equations. We have shown, that for known potential distribution the index of the system is at most one. The coupled system of Poisson and drift-diffusion equations suffers from rank deficiency. The index computation of the coupled system is subject of current work.

For preservation of positivity after time discretization a splitting technique is proposed as a combination of explicit exponential integration and implicit one-step methods. In combination with the Gummel-iteration the suggested splitting technique allows to increase the step size for positivity preservation. Thus, for one-dimensional positivity preserving semiconductor device simulation and even for two-dimensional simulation on coarse grids, the suggested algorithm performs faster than the underlying one-step method with the smaller time step size. The extension of the suggested algorithm to the application of the more enhanced energy-transport model is postponed to future work.

REFERENCES

[1] [illegible] *Math. Phys.* 37 (1996), 3308-3333.

[2] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d´évolution paraboliques. *RAIRO Anal. Numer.* 12, 237–245, 1978.

[3] S. C. Brenner and L. R. Scott. The mathematical theory of finite element methods, Springer, 2002.

[4] F. Brezzi, L. Marini and P.Pietra. Two-dimensional exponential fitting and applications to drift-diffusion models. *SIAM J. Num. Anal.*, 26, 1342–1355, 1989.

[5] F. Brezzi, L. Marini and P.Pietra. Numerical simulation of semiconductor devices. Comp. Meth. Appl. Mech. Engrg., 75, 493–514, 1989.

[6] M. Brunk and A. Jüngel, *Numerical coupling of electric circuit equations and energy-transport models for semiconductors*, SIAM J. Sci. Comput., 30 , 873–894, 2008.

[7] P. Degond, A. Jüngel, and P. Pietra. Numerical discretization of energy-transport models for semiconductors with non-parabolic band structure. *SIAM J. Sci. Comp.* 22 (2000), 986-1007.

[8] A. Ern and J. Guermond. Theory and practice of finite elements. Appl. Math. Sci. 159, Springer, 2004.

[9] L. Farina and S. Rinaldi. Positive Linear Systems: Theory and its Applications. John Wiley and Sons Inc., New York, 2000.

[10] T. Fujimoto and R. R. Ranade. Two characterizations of inverse-positive matrices: The Hawkins-Simon condition and the Le Chatelier-Braun principle. *Electron. J. Linear Algebra* 11 (2004), 59–65.

[11] H. K. Gummel. A self-consistent iterative scheme for one-dimensional steady-state transistor calculations. *IEEE Trans. El. Dev.*, 11, 455–465, 1964.

[12] W. Hundsdorfer and J. Verwer. Numerical solution of time-dependent advection-diffusion-reaction equations. *Series in Comp. Math.* 33, Springer, Berlin, Heidelberg, 2007.

[13] S. Holst, A. Jüngel, and P. Pietra. A mixed finite-element discretization of the energy-transport equations for semiconductors. *SIAM J. Sci. Comp.* 24 (2003), 2058-2075.

[14] L. D. Marini and P. Pietra. An abstract theory for mixed approximations of second order elliptic equations. *Mat. Aplic. Comp.* 8 (1989), 219-239.

[15] L. D. Marini and P. Pietra. New mixed finite element schemes for current continuity equations. *COMPEL* 9 (1990), 257-268.

[16] C. Moler and C. V. Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five-years later. *SIAM Rev.* 45 (2003), no. 1, 3–49.

[17] P. Raviart and J. Thomas. A mixed finite element method for second order elliptic equations. In: *Mathematical Aspects of the Finite Element Method*, Proc. Conf. Rome 1975. Lecture Notes in Mathematics 606 (1977), 292-315.

[18] S. Selberherr. Analysis and simulation of semiconductor devices. Springer, Wien, New York, 1984.

[19] L. F. Shampine and M. W. Reichelt. The MATLAN ODE Suite. *SIAM J. Sci. Comput.* **18**, 1-22, 1997.

[20] S. M. Sze. Physical semiconductor devices. John Wiley and Sons Inc., New York, 1981.

[21] V. Strassen. Gaussian elimination is not optimal. *Numer. Math.* 13 (1969), 354-356.

[22] E. Virnik. *Analysis of positive descriptor systems.* PhD-thesis, Berlin, Technical University, 2008.