ECE539 - Advanced Theory of Semiconductors and Semiconductor Devices

Numerical Methods and Simulation / Umberto Ravaioli

Review of Conventional Semiconductor Device Models Based on Partial Differential Equations

1 Boltzmann equation

The standard semi–classical transport theory is based on the Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_r f + \frac{e\mathbf{E}}{\hbar} \cdot \nabla_k f = \sum_{\mathbf{k}'} \quad [S(\mathbf{k}', \mathbf{k}) f(\mathbf{r}, \mathbf{k}', t)[1 - f(\mathbf{r}, \mathbf{k}, t)] - [S(\mathbf{k}, \mathbf{k}') f(\mathbf{r}, \mathbf{k}, t)[1 - f(\mathbf{r}, \mathbf{k}', t)] \tag{1}$$

where $\mathbf{r}$ is the position, $\mathbf{k}$ is the momentum, $f(\mathbf{k}, t)$ is the distribution function (for instance, a Fermi–Dirac in equilibrium), $\mathbf{v}$ is the group velocity, $\mathbf{E}$ is the electric field, $S(\mathbf{k}, \mathbf{k}')$ is the transition probability between the momentum states $\mathbf{k}$ and $\mathbf{k}'$, and $[1 - f(\mathbf{k}', t)]$ is the probability of non–occupation for a momentum state $\mathbf{k}'$. The summation on the right hand side is the collison term, which accounts for all the scattering events. The terms on the left hand side indicate, respectively, the dependence of the distribution function on time, space (explicitly related to velocity), and momentum (explicitly related to electric field).

The Boltzmann equation is valid under assumptions of semi–classical transport: effective mass approximation (which incorporates the quantum effects due to periodicity of the crystal); Born approximation for the collisions, in the limit of small perturbation for the electron–phonon interaction and instantaneous collisions; no memory effects, i.e. no dependence on initial condition terms. The phonons are usually treated as in equilibrium, although the condition of non–equilibrium phonons may be included through an additional equation.

Analytical solutions of the Boltzmann equation are possible only under very restrictive assumptions. Direct numerical methods for device simulation have been limited by the complexity of the equation, which in the complete 3–D time–dependent form requires seven independent variables for time, space and momentum. In recent times, more powerful computational platforms have spurred a renewed interest in numerical solutions based on the spheroidal harmonics expansion of the distribution function. To–date, most semiconductor applications have been based on stochastic solution methods (Monte Carlo), which involve the simulation of particle trajectories rather than the direct solution of partial differential equations.

The vast majority of device simulations are normally based on the numerical solution of approximate models which are related to the Boltzmann equation, coupled to Poisson's equation for self–consistency.

In the simplest approach, the collision term on the right hand side of (1) is substituted with a *phenomenological* term

$$\frac{f_{eq} - f(\mathbf{r}, \mathbf{k}, t)}{\tau} \tag{2}$$

where f_{eq} indicates the (local) equilibrium distribution function, and τ is a microscopic relaxation time. It is very useful to express the distribution function in terms of velocity, rather than momentum, since it will be easier to calculate electrical currents. In equilibrium we may use the Maxwell–Boltzmann distribution function

$$f_{eq}(\mathbf{r},\mathbf{v}) = n(\mathbf{r}) \left(\frac{2\pi k_B T_o}{m^*}\right)^{-3/2} \exp\left(-\frac{m^*|\mathbf{v}|^2}{2k_B T_o}\right) \tag{3}$$

where $n(\mathbf{r})$ is the carrier density, T_o is the lattice temperature, k_B is the Boltzmann constant, and m^* is the effective mass. The use of (3) for semiconductors is justified in equilibrium as long as degeneracy is not present. The carrier density $n(\mathbf{r})$ is directly related to the distribution function according to

$$n(\mathbf{r}) = \int d\mathbf{v} f(\mathbf{r},\mathbf{v}) \tag{4}$$

which is of general applicability. The significance of the momentum relaxation time can be understood if the electric field is switched off instantaneously and a space–independent distribution is considered. The resulting Boltzmann equation is then

$$\frac{\partial f}{\partial t} = \frac{f_{eq} - f}{\tau} \tag{5}$$

which shows that the relaxation time is a characteristic decay constant for the return to the equilibrium state.

2 Drift–diffusion model

The popular drift-diffusion current equations can be easily derived directly from the Boltzmann equation. Let's consider a steady state situation and for simplicity a 1–D geometry. With the use of a relaxation time approximation as in (2) the Boltzmann equation becomes

$$\frac{eE}{m^*}\frac{\partial f}{\partial v} + v\frac{\partial f}{\partial x} = \frac{f_{eq} - f(v,x)}{\tau} \tag{6}$$

Here, we have used the relation $m^* v = \hbar k$, which is valid for a parabolic energy band. Note that the charge e has to be taken with the proper sign of the particle (positive for holes and negative for electrons). A general definition of current density is given by

$$J(x) = e\int v f(v,x) dv \tag{7}$$

where the integral on the right hand side represents the first moment of the distribution function. This definition of current can be related to (6) after multiplying both sides by v and integrating over v. From the left hand side we get

$$\frac{1}{\tau}\left[\int v f_{eq} dv - \int v f(v,x) dv\right] = -\frac{J(x)}{e\tau} \tag{8}$$

The equilibrium distribution function is symmetric in v, and its integral is zero. Therefore, we have from (6)

$$J(x) = -e\frac{e\tau}{m^*} E\int v\frac{\partial f}{\partial v} dv - e\tau\frac{d}{dx}\int v^2 f(v,x) dv \tag{9}$$

Integrating by parts we have

$$\int v\frac{\partial f}{\partial v} dv = \underbrace{[v f(v,x)]_{-\infty}^{\infty}}_{0} - \int f(v,x) dv = -n(x) \tag{10}$$

and we can write

$$\int v^2 f(v,x)dv = n(x)\left\langle v^2 \right\rangle \tag{11}$$

where $\langle v^2 \rangle$ is the average of the square of the velocity defined as

$$\left\langle v^2 \right\rangle = \frac{1}{n}\int v^2 f(v,x)dv \tag{12}$$

For a purely 1–D treatment, the $-\frac{3}{2}$ exponent in (3) may be replaced with $-\frac{1}{2}$, while the appropriate thermal kinetic energy becomes $\frac{k_B T}{2}$ instead of $\frac{3k_B T}{2}$.

The drift–diffusion equations are derived introducing the mobility $\mu = \frac{e\tau}{m^*}$ and replacing $\langle v^2 \rangle$ with its average equilibrium value $\frac{k_B T}{m^*}$, therefore neglecting thermal effects. The diffusion coefficient $D = \frac{\mu k_B T_o}{e}$ (Einstein's relation) is also introduced, and the resulting drift–diffusion current is

$$J_n = qn(x)\mu_n E(x) + qD_n \frac{dn}{dx} \tag{13}$$

$$J_p = qp(x)\mu_p E(x) - qD_p \frac{dp}{dx} \tag{14}$$

where q is used to indicate the absolute value of the electronic charge. Although no direct assumptions on the non–equilibrium distribution function $f(v,x)$ was made in the derivation of (13) and (14), in effect, the choice of equilibrium (thermal) velocity means that the drift–diffusion equations are only valid for very small perturbations of the equilibrium state (low fields). The validity of the drift–diffusion equations is *empirically* extended by introducing field–dependent mobility $\mu(E)$ and diffusion coefficient $D(E)$, obtained from empirical models or detailed calculations.

3 Physical Limitations on Numerical Drift–Diffusion Schemes

The complete drift–diffusion model is based on the following set of equations

1. Current equations

$$\mathbf{J}_n = qn\mu_n \mathbf{E} + qD_n \nabla n \tag{15}$$

$$\mathbf{J}_p = qp\mu_p \mathbf{E} - qD_p \nabla p \tag{16}$$

2. Continuity equations

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla\cdot\mathbf{J}_n + U_n \tag{17}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla\cdot\mathbf{J}_p + U_p \tag{18}$$

3. Poisson's equation

$$\nabla^2 V = \frac{q(n - p + N_A^- - N_D^+)}{\epsilon} \tag{19}$$

where U_n and U_p are the net generation–recombination rates. The continuity equations (17) and (18) are the conservation laws for the carriers. A numerical scheme which solves the continuity equations should

1. Conserve the total number of particles inside the device simulated.

2. Respect local positivity of carrier density. Negative density is unphysical.

3. Respect monotonicity of the solution (i.e. it should not introduce spurious space oscillations).

Conservative schemes are usually achieved by subdivision of the computational domain into patches (boxes) surrounding the mesh points. The currents are then defined on the boundaries of these elements, thus enforcing conservation (the current exiting one element side is exactly equal to the current entering the neighboring element through the side in common). For example, on a uniform 2–D grid with mesh size Δ, the electron continuity equation could be discretized in an explicit form as follows

$$\begin{aligned}\frac{n(i,j;k+1)-n(i,j;k)}{\Delta t} &= \frac{J^x(i+\frac{1}{2},j;k)-J^x(i-\frac{1}{2},j;k))}{q\Delta} \\ &+\frac{J^y(i,j+\frac{1}{2};k)-J^y(i,j-\frac{1}{2};k)}{q\Delta}+U(i,j;k) \end{aligned} \tag{20}$$

This simple approach has certainly practical limitations, but is sufficient to exemplify the idea behind the conservative scheme. With the present convention for positive and negative components, it is easy to see that in the absence of generation-recombination terms, the only contributions to the overall device current arise from the contacts. Remember that, since electrons have negative charge, the particle flux is opposite to the current flux. The actual determination of the current densities appearing in (20) will be discussed later.

When the equations are discretized, using finite differences for instance, there are limitations on the choice of mesh size and time step:

1. The mesh size Δx is limited by the Debye length.

2. The time step is limited by the dielectric relaxation time.

A mesh size must be smaller than the Debye length where one has to resolve charge variations in space. A simple example is the carrier redistribution at an interface between two regions with different doping levels. Carriers diffuse into the lower doped region creating excess carrier distribution which at equilibrium decays in space down to the bulk concentration with approximately exponential behavior. The space decay constant is the Debye length

$$L_D = \sqrt{\frac{\epsilon k_B T}{q^2 N}} \tag{21}$$

where N is the doping. In $GaAs$ and Si, at room temperature the Debye length is approximately 400 Å when $N \approx 10^{16}\ cm^{-3}$ and decreases to about only 50 Å when $N \approx 10^{18}\ cm^{-3}$.

The dielectric relaxation time is the characteristic time for charge fluctuations to decay under the influence of the field that they produce. The dielectric relaxation time may be estimated using

$$t_{dr} = \frac{\epsilon}{qN\mu} \tag{22}$$

To see the importance of respecting the limitations related to the dielectric relaxation time, imagine to have a space fluctuation of carrier concentration which is going to relax to equilibrium with a law

$$\frac{\partial \Delta n}{\partial t} = -\frac{\Delta n(t=0)}{t_{dr}} \tag{23}$$

A finite difference discretization of this equation gives at the first time step

$$\Delta n(\Delta t) = \Delta n(0) - \frac{\Delta t \Delta n(0)}{t_{dr}} \tag{24}$$

Clearly, if $\Delta t > t_{dr}$, the value of $\Delta n(\Delta t)$ is negative, which means that the actual concentration is evaluated to be less than the equilibrium value, and it is easy to see that the solution at higher time steps will decay oscillating between positive and negative values of Δn. An excessively large Δt may lead, therefore, to nonphysical results. In the case of high mobility the dielectric relaxation time can be very small. For instance, a sample of $GaAs$ with a mobility of $6,000 \frac{cm^2}{V-s}$ and doping $10^{18}\ cm^{-3}$ has approximately $t_{dr} \approx 10^{-15}\ s$, and in a simulation the time step Δt should be chosen to be considerably smaller.

4 Steady State Solution of Bipolar Semiconductor Equations

The general semiconductor equations may be rewritten as

$$\nabla \cdot (\epsilon \nabla V) = q(n - p + N_B) \tag{25}$$

$$\nabla \cdot \mathbf{J}_n = qU(n,p) + q\,\frac{\partial n}{\partial t} \tag{26}$$

$$\nabla \cdot \mathbf{J}_p = -qU(n,p) + q\,\frac{\partial p}{\partial t} \tag{27}$$

$$\mathbf{J}_n = q\mu_n \left(-n\nabla V + \frac{k_B T}{q}\,\nabla n \right) \tag{28}$$

$$\mathbf{J}_p = q\mu_p \left(-p\nabla V - \frac{k_B T}{q}\,\nabla p \right) \tag{29}$$

with $N_B = N_A - N_D$. Note that the above equations are valid in the limit of small deviations from equilibrium, since the Einstein relations have been used for the diffusion coefficient, normally valid for low fields or large devices. The generation–recombination term U will be in general a a function of the local electron and hole concentrations, according to possible different physical mechanisms, to be examined later in more detail. We will consider from now on steady state, with the time dependent derivatives set to zero.

The semiconductor equations constitute a coupled nonlinear set. It is not possible, in general, to obtain a solution directly in one step, but a nonlinear iteration method is required. The two more popular methods for solving the discretized equations are the Gummel's iteration method and the Newton's method. It is very difficult to determine an optimum strategy for the solution, since this will depend on a number of details related to the particular device under study.

There are in general three possible choices of variables.

1. Natural variable formulation (V, n, p)
2. Quasi–Fermi level formulation (V, ϕ_n, ϕ_p), where the quasi–Fermi levels derive from the following definition of carrier concentration out of equilibrium (non–degenerate case)

$$n = n_i \exp \frac{q(V - \phi_n)}{k_B T} \tag{30}$$

$$p = n_i \exp \frac{q(\phi_p - V)}{k_B T} \tag{31}$$

3. Slotboom formulation (V, Φ_n, Φ_p) where the Slotboom variables are defined as

$$\Phi_n = n_i \exp\left(\frac{-q\phi_n}{k_B T}\right) \tag{32}$$

$$\Phi_p = n_i \exp\left(\frac{q\phi_p}{k_B T}\right) \tag{33}$$

   The Slotboom variables are therefore related to the carrier definitions (30) and (31), and the extension to degenerate conditions is cumbersome.

Normally, there is a preference for the quasi–Fermi level formulation in steady state simulation, and for the natural variables n and p in transient simulation.

4.1 Normalization and Scaling

For the sake of clarity, all formulae have been presented without the use of simplifications or normalization. It is however common practice to perform the actual calculation using normalized units to make the algorithms more efficient, and in cases to avoid numerical overflow and underflow. It is advisable to input the data in M.K.S. or practical units (the use of centimeters is for instance very common in semiconductor practice, instead of meters) and then provide a conversion block before and after the computation blocks to normalize and denormalize the variables. It is advisable to use consistently scaling, rather than set certain constants to arbitrary values. In computational physics, it is common practice to use $\hbar = 1$, for instance. This simplifies a lot the derivation of formulae, but if this is not done carefully, when denormalization is applied to recover quantitative results in M.K.S. or other appropriate units, it may be very difficult to verify which terms are multiplied by $\hbar$, $\hbar^2$, $\hbar^3$, etc. Problems may arise when formulae are taken from the published literature, since c.g.s. and M.K.S. units are practically interchangeable, unless the dielectric constant is involved. In c.g.s. units, the vacuum permittivity is $\epsilon_o = 1$ or $1/4\pi$ but it is about 8.85×10^{-12} [F/m] in M.K.S. units. A simple change of centimeters into meters and grams into kilograms, will leave many orders of magnitude unaccounted for when the dielectric permittivity is involved. According to our experience, the majority of mistakes in computational applications prepared by beginners are caused by improper scaling or choice of units. Great care and systematic dimensional analysis of the formulae is suggested, at least until a good feeling is developed for the orders of magnitudes of the variables involved in a computation.

The most common scaling factors for normalization of semiconductor equations are listed in Table I.

4.2 Gummel's Iteration Method

Gummel's method solves the equations with a *decoupled* procedure. If we choose the quasi–Fermi level formulation, we solve first a *nonlinear* Poisson's equation. The potential obtained is substituted into the continuity equations, which are now linear, and are solved directly to conclude the

TABLE I - Scaling Factors

Space	Intrinsic Debye length ($N = n_i$)	$L = \sqrt{\frac{\epsilon k_B T}{q^2 N}}$
	Extrinsic Debye length ($N = N_{max}$)	
Potential	Thermal voltage	$V^* = \frac{k_B T}{q}$
Carrier concentration	Intrinsic concentration	$N = n_i$
	Maximum doping concentration	$N = N_{max}$
Diffusion coefficient	Practical unit	$D = 1\frac{cm^2}{s}$
	Maximum diffusion coeff.	$D = D_{max}$
Mobility		$M = \frac{D}{V^*}$
Gen–Recomb		$R = \frac{DN}{L^2}$
Time		$T = \frac{L^2}{D}$

iteration. The result in terms of quasi–Fermi levels is then substituted back into Poisson's equation until convergence is reached. In order to check for convergence, one can calculate the *residuals* obtained by positioning all the terms to the left hand side of the equations and substituting the variables with the iteration values. For exact solution values, the residuals should be zero. Convergence is assumed when the residuals are smaller than a set tolerance. The rate of convergence of the Gummel method is faster when there is little coupling between the different equations.

The computational cost of one Gummel iteration is one matrix solution for each carrier type plus one iterative solution for the linearization of Poisson's equation. Note that in conditions of equilibrium (zero bias) only the solution of Poisson's equation is necessary, since the equilibrium Fermi level is constant and coincides with both quasi–Fermi levels.

We give some examples of the quasi–linearization of Poisson equation, as necessary when Gummel's method is implemented. Let's consider the 1–D case in equilibrium first. As said earlier, one has to solve only Poisson's equation, since exact expressions for the carrier concentrations are known. In the non–degenerate case, (30) and (31) are substituted into Poisson's equation to give

$$\frac{d^2V}{dx^2} = \frac{q}{\epsilon}\left[n_i \exp(-q\phi_n)\exp\left(\frac{qV}{k_BT}\right) - n_i \exp(q\phi_p)\exp\left(-\frac{qV}{k_BT}\right) + N_A - N_D\right] \tag{34}$$

In equilibrium it is $\phi_n = \phi_p = 0$ (Fermi level taken as reference for the potential energy). Furthermore, the equation may be scaled by using the (minimum) extrinsic Debye length for the space coordinate x, and the thermal voltage k_BT/q for the potential V. Indicating with $\bar{V}$ and $\bar{x}$ for the normalized potential and space coordinates

$$\frac{d^2\bar{V}}{d\bar{x}^2} = \frac{n_i}{N}\left[\exp(\bar{V}) - \exp(-\bar{V}) + \frac{N_A - N_D}{n_i}\right] \tag{35}$$

The equilibrium Poisson's equation (35) can be solved with the following quasi–linearization procedure

1. Set initial guess for the potential $\bar{V}$.
2. Write the potential at the next iteration as $\bar{V}_{new} = \bar{V} + \delta V$, and write Poisson's equation for $\bar{V}_{new}$ with the above substitution

$$\frac{d^2\bar{V}}{d\bar{x}^2} + \frac{d^2\delta V}{d\bar{x}^2} = \frac{n_i}{N}\left[\exp(\bar{V})\exp(\delta V) - \exp(-\bar{V})\exp(-\delta V) + \frac{N_A - N_D}{n_i}\right] \tag{36}$$

3. Use the linearization $\exp(\pm\delta V) \approx 1 \pm \delta V$ and discretize (36). This equation has a tridiagonal matrix and is readily solved for $\delta V(i)$.

$$\begin{aligned} &\delta V(i-1) - \left[2 + \frac{n_i}{N}\,\Delta^2 x[\exp(\bar{V}(i)) - \exp(-\bar{V}(i))]\right]\delta V(i) + \delta V(i+1) = \\ &-\bar{V}(i-1) + 2\bar{V}(i) - \bar{V}(i+1) + \frac{n_1}{N}\,\Delta^2 x\left[\exp(\bar{V}(i)) - \exp(-\bar{V}(i)) + \frac{N_A - N_D}{n_i}\right] \end{aligned} \tag{37}$$

4. Check for convergence. The residual of (35) is calculated and convergence is achieved if the norm of the residual is smaller than a preset tolerance. If convergence is not achieved, return to step 2. In practice one might simply check the norm of the error

$$||\delta V||_2 \leq Tol \quad or \quad ||\delta V||_\infty \leq Tol \tag{38}$$

Note that for the solution of Poisson's equation the boundary conditions are referenced to the equilibrium Fermi level. One may use the separation between the Fermi level and the intrinsic Fermi level at the contacts for the boundary conditions.

After the solution in equilibrium is obtained, the applied voltage is increased in steps $\Delta V \leq k_B T/q$. Now the scaled nonlinear Poisson equation becomes

$$\frac{d^2V}{dx^2} = \frac{n_i}{N}\left[\exp(-\phi_n)\exp(V) - \exp(\phi_p)\exp(-V) + \frac{N_A - N_D}{n_i}\right] \tag{39}$$

where the quasi–Fermi levels are also normalized. The continuity equations, as long as the Einstein's relations are valid, may be written as

$$\begin{aligned} J_n &= -q\mu_n n\,\frac{\partial V}{\partial x} + q\mu_n \frac{k_B T}{q}\,\frac{\partial}{\partial x}\left[n_i \exp\left(\frac{q(V-\phi_n)}{k_B T}\right)\right] \\ &= -q\mu_n n\,\frac{\partial V}{\partial x} + q\mu_n \frac{k_B T}{q}\,n\,\frac{q}{k_B T}\left[\frac{\partial V}{\partial x} - \frac{\partial \phi_n}{\partial x}\right] \\ &= -q\mu_n n\,\frac{\partial \phi_n}{\partial x} \\ &= -q\mu_n n_i \exp\left[\frac{q(V-\phi_n)}{k_B T}\right]\frac{\partial \phi_n}{\partial x} \\ &= -q\mu_n n_i \exp\left(\frac{qV}{k_B T}\right)\frac{-k_B T}{q}\,\frac{\partial}{\partial x}\,\exp\left(\frac{-q\phi_n}{k_B T}\right) \end{aligned} \tag{40}$$

which can be compacted, including quasi–Fermi level normalization as

$$J_n = a_n(x)\,\frac{\partial}{\partial x}\,\exp(-\phi_n) \tag{41}$$

A similar formula is obtained for the holes

$$J_p = a_p(x)\,\frac{\partial}{\partial x}\,\exp(\phi_p) \tag{42}$$

and the continuity equations are

$$\frac{\partial}{\partial x}\left[a_n(x)\,\frac{\partial}{\partial x}\,\exp(-\phi_n)\right] = qU(x) \tag{43}$$

$$\frac{\partial}{\partial x}\left[a_p(x)\,\frac{\partial}{\partial x}\,\exp(\phi_p)\right] = qU(x) \tag{44}$$

The continuity equations may be discretized with a straightforward finite difference approach (here for simplicity with uniform mesh)

$$\frac{\frac{a_n(i+\frac{1}{2})[\Phi_n(i+1)-\Phi_n(i)]}{\Delta x} - \frac{a_n(i-\frac{1}{2})[\Phi_n(i)-\Phi_n(i-1)]}{\Delta x}}{\Delta x} = U \tag{45}$$

where the Slotboom variables have been used for simplicity of notation. Note that the inner derivative has been discretized with centered differences around the points $(i \pm \frac{1}{2})$ of the interleaved mesh. Variables on the interleaved mesh must be determined very carefully, using consistent interpolation schemes for potential and carrier density, as discussed later. The continuity equations give the tridiagonal system

$$a_n(i-\frac{1}{2})\Phi_n(i-1) - [a_n(i+\frac{1}{2}) + a_n(i-\frac{1}{2})]\Phi_n(i) + a_n(i+\frac{1}{2})\Phi_n(i+1) = \Delta^2 x U(i) \tag{46}$$

$$a_p(i-\frac{1}{2})\Phi_p(i-1) - [a_p(i+\frac{1}{2}) + a_p(i-\frac{1}{2})]\Phi_p(i) + a_p(i+\frac{1}{2})\Phi_p(i+1) = -\Delta^2 x U(i) \tag{47}$$

The decoupled iteration solves now Poisson's equation (39), initially with a guess for the quasi–Fermi levels. The voltage distribution obtained for the previous voltage considered is normally a good initial guess for the potential. Since the quasi–Fermi levels are inputs for Poisson's equation, the quasi–linearization procedure for equilibrium can be used again. The potential is then used to update the $a_n(i)$ and $a_p(i)$, and (46) and (47) are solved to provide new quasi–Fermi level values for Poisson's equation, and the process is repeated until convergence is reached. The generation–recombination term depends on the electron and hole concentrations, therefore it has to be updated at each iteration. It is possible to update the generation–recombination term also intermediately, using the result of (46) for the electron concentration.

The examples given here to illustrate the Gummel's approach are limited to the nondegenerate case. If field dependent mobility and diffusion coefficients are introduced, minimal changes should be necessary, as long as it is still justified to use of Einstein's relations. Extension to nonuniform mesh is left as an exercise for the reader.

In the 2–D case, the quasi–linearized Poisson's equation becomes

$$\begin{aligned} -\left(4 + h^2\,\frac{n_i}{N}\,[\Phi_n(i,j)\exp(V(i,j)) + \Phi_p(i,j)\exp(-V(i,j))]\right)\delta V(i,j) \\ +[\delta V(i-1,j) + \delta V(i+1,j) + \delta V(i,j-1) + \delta V(i,j+1)] = \\ 4V(i,j) - V(i-1,j) - V(i+1,j) - V(i,j-1) - V(i,j+1) + \\ h^2\,\frac{n_i}{N}\left[\Phi_n(i,j)\exp(V(i,j)) + \Phi_p(i,j)\exp(-V(i,j)) + \frac{N_A+N_B}{n_i}\right] \end{aligned} \tag{48}$$

The normalized mesh size is $h = \Delta x = \Delta y$. As before, the thermal voltage $k_B T/q$ has been used to normalize the potential V and the quasi–Fermi levels ϕ_n and ϕ_p included in the Slotboom variables $\Phi_{n,p} = \exp(\pm\phi_{n,p})$.

The continuity equations with the form $\nabla\cdot(a(x,y)\nabla\Phi) = \pm U(x,y)$ are discretized as

$$
\begin{aligned}
&-\left[a(i+\frac{1}{2},j)+a(i-\frac{1}{2},j)+a(i,j+\frac{1}{2})+a(i,j-\frac{1}{2})\right]\Phi(i,j)+a(i+\frac{1}{2},j)\Phi(i+1,j)\\
&+a(i-\frac{1}{2},j)\Phi(i-1,j)+a(i,j+\frac{1}{2})\Phi(i,j+1)+a(i,j-\frac{1}{2})\Phi(i,j-1) = \pm h^2 U(i,j)
\end{aligned}
\tag{49}
$$

4.3 Newton's Method

Newton's method is a *coupled* procedure which solves the equations simultaneously, through a generalization of the Newton–Raphson method for determining the roots of an equation. We rewrite (25)–(27) in the residual form

$$
W_V(V,n,p) = 0 \quad W_n(V,n,p) = 0 \quad W_V(V,n,p) = 0 \tag{50}
$$

Starting from an initial guess V_o, n_o, and p_o, the corrections V, Δn, and Δp are calculated from the Jacobian system

$$
\begin{vmatrix} \frac{\partial W_V}{\partial V} & \frac{\partial W_V}{\partial n} & \frac{\partial W_V}{\partial p} \\ \frac{\partial W_n}{\partial V} & \frac{\partial W_n}{\partial n} & \frac{\partial W_n}{\partial n} \\ \frac{\partial W_p}{\partial V} & \frac{\partial W_p}{\partial n} & \frac{\partial W_p}{\partial p} \end{vmatrix} \begin{vmatrix} \Delta V \\ \Delta n \\ \Delta p \end{vmatrix} = - \begin{vmatrix} W_V \\ W_n \\ W_p \end{vmatrix} \tag{51}
$$

which is obtained by Taylor expansion. The solutions are then updated according to the scheme

$$
\begin{aligned}
V(k+1) &= V(k) + \Delta V(k) \\
n(k+1) &= n(k) + \Delta n(k) \\
p(k+1) &= p(k) + \Delta p(k)
\end{aligned}
\tag{52}
$$

where k indicates the iteration number. In practice, a relaxation approach is also applied to avoid excessive variations of the solutions at each iteration step.

The system (51) has 3 equations for each mesh point on the grid. This indicates the main disadvantage of a full Newton iteration, related to the computational cost of matrix inversion (one may estimate that a $3N \times 3N$ matrix takes typically 20 times longer to invert than an analogous $N \times N$ matrix!). On the other hand convergence is usually fast for the Newton method, provided that the initial condition is reasonably close to the solution, and is in the neighborhood where the solution is unique. There are several viable approaches to alleviate the computational requirements of the Newton's method. In the Newton–Richardson approach, the Jacobian matrix in (51) is updated only when the norm of the error does not decrease according to a preset criterion.

In general, the Jacobian matrix is not symmetric positive definite, and fairly expensive solvers are necessary. Iterative schemes have been proposed to solve each step of Newton's method by reformulating (51) as

$$
\begin{vmatrix} \frac{\partial W_V}{\partial V} & 0 & 0 \\ \frac{\partial W_n}{\partial V} & \frac{\partial W_n}{\partial n} & 0 \\ \frac{\partial W_p}{\partial V} & \frac{\partial W_p}{\partial n} & \frac{\partial W_p}{\partial p} \end{vmatrix} \underbrace{\begin{vmatrix} \Delta V \\ \Delta n \\ \Delta p \end{vmatrix}}_{k+1} = - \begin{vmatrix} W_V \\ W_n \\ W_p \end{vmatrix} - \begin{vmatrix} 0 & \frac{\partial W_V}{\partial n} & \frac{\partial W_V}{\partial p} \\ 0 & 0 & \frac{\partial W_n}{\partial p} \\ 0 & 0 & 0 \end{vmatrix} \underbrace{\begin{vmatrix} \Delta V \\ \Delta n \\ \Delta p \end{vmatrix}}_{k} \tag{53}
$$

Since the matrix on the left hand side is lower triangular, one may solve decoupling into three systems of equations solved in sequence. First, one solves the block of equations (again, one for each grid point)

$$\frac{\partial W_V}{\partial V}(\Delta V)_{k+1} = -W_V - \frac{\partial W_V}{\partial n}(\Delta n)_k - \frac{\partial W_V}{\partial p}(\Delta p)_k \tag{54}$$

and the result is used in the next block of equations

$$\frac{\partial W_n}{\partial n}(\Delta n)_{k+1} = -W_n - \frac{\partial W_n}{\partial V}(\Delta V)_{k+1} - \frac{\partial W_n}{\partial p}(\Delta p)_k \tag{55}$$

Similarly, for the third block

$$\frac{\partial W_p}{\partial p}(\Delta p)_{k+1} = -W_p - \frac{\partial W_p}{\partial V}(\Delta V)_{k+1} - \frac{\partial W_p}{\partial n}(\Delta n)_{k+1} \tag{56}$$

The procedure achieves a decoupling of the equations as in a block Gauss-Seidel iteration, and can be intended as a generalization of the Gummel method. A block–SOR method is obtained if the left hand sides are premultiplied by a relaxation parameter. This iteration procedure has better performance if the actual variables are (V, ϕ_n, ϕ_p).

In general, Gummel's method is preferred at low bias because of its faster convergence and low cost per iteration. At medium and high bias the Newton's method becomes more convenient, since the convergence rate of Gummel's method becomes worse as the coupling between equations becomes stronger at hogher bias. But since Gummel's method has a fast initial error reduction, it is often convenient to couple the two procedures, using Newton's method after several Gummel's iterations. Remember that it is very important for the Newton's iteration to start as close as possible to the true solution. Close to convergence, the residual in Newton's iteration should decrease quadratically from one iteration to the other.

5 Generation and Recombination

The Shockley–Reed–Hall model is very often used for the generation–recombination term due to trap levels

$$U_{SRH} = \frac{np - n_i^2}{\tau_p \left[n + n_i \exp\left(\frac{q(E_t - E_i)}{k_B T}\right)\right] + \tau_n \left[p + n_i \exp\left(\frac{q(E_i - E_t)}{k_B T}\right)\right]} \tag{57}$$

where E_t is the trap energy level involved and τ_n and τ_p are the electron and hole lifetimes. Surface rates may be included with a similar formula, in which the lifetimes are substituted by $\frac{1}{s_{n,p}}$ where $s_{n,p}$ is the surface recombination velocity.

The Auger recombination may be accounted for by using the formula

$$U_{Aug} = C_n[pn^2 - nn_i^2] + C_p[np^2 - pn_i^2] \tag{58}$$

where C_n and C_p are appropriate constants. The Auger effect is for instance very relevant in the modeling of highly doped emitter regions in bipolar transistors.

The generation process due to impact ionization can be included using the field–dependent rate

$$U_I = \frac{a_n^\infty \exp\left(-\frac{E_n^{crit}}{E}\right)^{\beta n} |\mathbf{J}_n| + a_p^\infty \exp\left(-\frac{E_p^{crit}}{E}\right)^{\beta p} |\mathbf{J}_p|}{q} \tag{59}$$

6 Time-dependent simulation

The time–dependent form of the drift–diffusion equations can be used both for steady–state and transient calculations. Steady–state analysis is accomplished by starting from an initial guess, and letting the numerical system evolve until a stationary solution is reached, within set tolerance limits. This approach is seldom used in practice, since now robust steady–state simulators are widely available. It is nonetheless an appealing technique for beginners since a relatively small effort is necessary for simple applications and elementary discretization approaches. If an explicit scheme is selected, no matrix solutions are necessary, but it is normally the case that stability is possible only for extremely small timesteps.

The simulation of transients requires the knowledge of a physically meaningful initial condition, which can be obtained from a steady-state calculation. The same time–dependent numerical approaches used for steady-state simulation are suitable, but there must be more care for the boundary conditions, because of the presence of displacement current during transients. In a transient simulation to determine the steady–state, the displacement current can be neglected because it goes to zero when a stationary condition is reached. Therefore, it is sufficient to impose on the contacts the appropriate potential values provided by the bias network. In a true transient regime, however, the presence of displacement currents manifests itself as a potential variation at the contacts, superimposed to the bias, which depends on the external circuit in communication with the contacts. Neglect of the displacement current in a transient is equivalent to the application of bias voltages using *ideal* voltage generators, with zero internal impedance. In such a situation, the potential variations due to displacement current drop across a short circuit, and are therefore cancelled. In this arrangement, one will observe the shortest possible switching time attainable with the structure considered, but in practice an external load and parasitics will be present, and the switching times will be normally longer. A simulation neglecting displacement current effects may be useful to assess the ultimate speed limits of a device structure.

When a realistic situation is considered, it is necessary to include a displacement term in the current equations. It is particularly simple to deal with a 1–D situation. Consider a 1–D device with length W and a cross–sectional area A. The total current flowing in the device is

$$I_D(t) = I_n(x,t) + \epsilon A \frac{\partial E(x,t)}{\partial t} \tag{60}$$

The displacement term makes the total current constant at each possition x. This property can be exploited to perform an integration along the device

$$I_D(t) = \frac{1}{W}\int_0^W I_n(x,t)dx + \frac{\epsilon A}{W}\frac{\partial V^*}{\partial t} \tag{61}$$

where $V^*(t)$ is the total voltage drop across the structure, with the ground reference voltage applied at $x = W$. The term $\frac{\epsilon A}{W}$ is called *cold capacitance*. The 1–D device, therefore, can be studied as the parallel of a current generator and of the cold capacitance which is in parallel with the (linear) load circuit. At every time step, V^* has to be updated, since it depends on the charge stored by the capacitors.

To illustrate the procedure, consider a simple Gunn diode in parallel with an RLC resonant load containing the bias source. Calling C_o the parallel of cold and load capacitance, it is

$$I(t) = C_o \frac{dV^*(t)}{dt} + I_o(t) \tag{62}$$

where $I_o(t)$ is the particle current given by the first term on the right hand side of (61), calculated at the given time step with drift–diffusion (or any other suitable scheme). It is also

$$I(t) = -\frac{V^*(t)}{R} - \int \frac{V^*(t) - V_b}{L}\, dt \tag{63}$$

Upon time differencing this last equation, with the use of finite differences we obtain

$$V^*(t+\Delta t) = V^*(t) + [I(t) - I_o(t)]\,\frac{\Delta t}{C_o} \tag{64}$$

$$I(t+\Delta t) = I(t) - \frac{V^*(t+\Delta t) - V^*(t)}{R} - [V^*(t) - V_b]\,\frac{\Delta t}{L} \tag{65}$$

This set of difference equations allows one to update the boundary conditions for Poisson's equation at every time step to fully include displacement current.

A robust approach for transient simulation should be based on the same numerical apparatus established for purely steady-state models. It is usually preferred to use fully implicit schemes, which require a matrix solution at each iteration, because the choice of the timestep is more likely to be limited by the physical time constants of the problem rather than by stability of the numerical scheme. In order to estimate the timestep limits, let's assume a typical electron velocity $v = 10^7 cm/s$ and a spatial mesh $\Delta x = 0.01\mu m$. The C.F.L. condition necessary to resolve correctly a purely drift process on this mesh requires $\Delta t \leq \Delta x/v = 10^{-15} s$. As calculated earlier, this value is not too far from typical values of the dielectric relaxation time in practical semiconductor structures.

When dealing with unipolar devices, as often used in many microwave applications, it is possible to formulate very simple time–dependent drift–diffusion models, which can be solved with straightforward finite difference techniques and are suitable for small student projects. If we can neglect the generation–recombination effects, the 1–D unipolar drift–diffusion model is reduced to the following system of equations

$$\frac{\partial n}{\partial t} = -\frac{d}{dx}[n v_d(E)] + \frac{d}{dx}\left[D(E)\frac{d}{dx}n\right] \tag{66}$$

$$\frac{d^2V}{dx^2} = \frac{q(n - N_D)}{\epsilon} \tag{67}$$

where $v_d(E) = -\mu_n(E)E$ is the drift velocity. There are two physical processes involved: drift (advection) expressed by the first term on the right hand side of (66), and diffusion described by the second term. The continuity equation (66) is an admixture of competing hyperbolic and parabolic behavior whose relative importance depends on the local electric field strength.

The system (66) and (67) can be used for both transient or steady state conditions if the simulation is run $\partial n/\partial t = 0$. A basic simple algorithm could consists of the following steps

1. Guess the carrier distribution $n(x)$.
2. Solve Poisson's equation to obtain the field distribution.
3. Compute one iteration of the discretized continuity equation with time step Δt. $v(E)$ and $D(E)$ are updated according to the local field value.
4. Check for convergence. If convergence is obtained, stop. Otherwise, go back to step (2) updating the charge distribution.

This is an *uncoupled* procedure, since (66) and (67) are not solved simultaneously. Usually, explicit methods are used for computational speed. The time step must respect the limitations due to the C.F.L. condition (related to the advective component) and to the dielectric relaxation time. A simple discretization scheme could employ an explicit finite difference approach

$$\begin{aligned} n(i;k+1) &= n(i;k) + \frac{\Delta t}{\Delta x}\{[v_d(i-1;k)n(i-1;k) - v_d(i;k)n(i;k)] \\ &\quad + \frac{1}{\Delta x} D(i;k)[n(i-1;k) - 2n(i;k) + n(i+1;k)]\}; \qquad v_d < 0 \end{aligned} \tag{68}$$

$$\begin{aligned} n(i;k+1) &= n(i;k) + \frac{\Delta t}{\Delta x}\{[v_d(i;k)n(i;k) - v_d(i+1;k)n(i+1;k)] \\ &\quad + \frac{1}{\Delta x} D(i;k)[n(i-1;k) - 2n(i;k) + n(i+1;k)]\}; \qquad v_d > 0 \end{aligned} \tag{69}$$

where we have introduced upwinding for the drift term and we have assumed that the diffusion coefficient is slowly varying in space. There are of course many other possible explicit and implicit discretizations. Such simple finite difference approaches are in general a compromise which cannot provide at one time an optimal treatment of both advective and diffusive components. Because of spatially varying drift velocity, spurious diffusion and dispersion are present. This could be mitigated by using a nonuniform grid discretization, where the mesh size is locally adapted to achieve $v_d = \Delta x/\Delta t$ everywhere, which would involve interpolation to the new grid-points. The discretization for a diffusive process is better behaved with a fully implicit scheme (if the Crank-Nicholson approach is used, one needs to make sure that spurious oscillations in the solution do not develop). On the other hand, the fully implicit algorithm for advection is not conservative. From these conflicting requirements, it emerges that it would be beneficial to split the drift and diffusion processes, and apply an optimal solution procedure to each. There are 1–D situations where this is known to be nearly exact. In well known experiments, a small concentration of excess carriers is generated in a semiconductor sample with a uniform electric field, and the motion of the centroid of the carrier envelope can be studied independently of the diffusive spread of the spatial distribution around the centroid itself. For an initial Gaussian distribution in space, a simple analytical solution shows that drift and diffusion can be treated as a sequential process, each using the total duration of the observation as simulation time. In analogy with this, the 1–D continuity equation can be solved in two steps, for instance

$$n^*(j,i+1) = n(j,i) + v_d(j)[n(j-1,i) - n(j,i))\,\frac{\Delta t}{\Delta x}; \qquad v_d < 0 \tag{70}$$

$$\begin{aligned} n(j,i+1) &= n^*(j,i+1) + D(j)[n(j-1,i+1) - 2n(j,i-1) \\ &\quad + n(j+1,i+1)]\frac{\Delta t}{\Delta^2 x} \end{aligned} \tag{71}$$

where again a simple explicit upwinding scheme is used for the drift, while a fully implicit scheme is used for the diffusion.

7 Scharfetter-Gummel approximation

The discretization of the continuity equations in conservation form, requires the determination of the currents on the mid–points of mesh lines connecting neighboring grid nodes. Since the solutions are accessible only on the grid nodes, interpolation schemes are needed to determine the currents.

For consistency with Poisson's equation, it is common to assume that the potential varies linearly between two neighboring nodes. This is equivalent to assume a constant field along the mesh lines, and the field at the mid–point is obtained by centered finite differences of the potential values. In order to evaluate the current, it is also necessary to estimate the carrier density at the mid–points. The simplest approximation which comes to mind is to also assume a linear variation of the carrier density, by taking the arithmetic average between to neighboring nodes. This simple approach is only acceptable for very small potential variation between the nodes, and indeed is exact only if the field between two nodes is zero, which implies the same exact carrier density on the two points.

In order to illustrate this, let's consider a 1–D mesh where we want to discretize the electron current

$$J_n \;=\; q\mu_n n\left(-\frac{d\psi}{dx}\right) \;+\; qD_n\frac{dn}{dx} \tag{72}$$

Here, the field is explicitly expressed by the derivative of the potential. The discretization on the mid–point of the mesh line between nodes x_i and x_{i+1} is given by

$$J_{i+\frac{1}{2}} \;=\; -\,q\mu_n n_{i+\frac{1}{2}}\;\frac{\psi_{i+1}-\psi_i}{\Delta x} \;+\; qD_n\;\frac{n_{i+1}-n_i}{\Delta x} \tag{73}$$

In the simple approach indicated above, the carrier density $n_{i+\frac{1}{2}}$ is expressed as

$$n_{i+\frac{1}{2}} \;\approx\; \frac{n_{i+1}+n_i}{2} \tag{74}$$

In (73), the assumed linearity of the potential between meshes, is implied by the use of the centered finite differences to express the field on the mid–point. We can now rewrite (73) including the approximation in (74) as

$$\begin{aligned} J_{i+\frac{1}{2}} \;&=\; n_{i+1}\left[-q\frac{\mu_n}{2}\;\frac{\psi_{i+1}-\psi_i}{\Delta x} \;+\; q\;\frac{D_n}{\Delta x}\right] \\ &-\; n_i\left[\underbrace{q\frac{\mu_n}{2}\;\frac{\psi_{i+1}-\psi_i}{\Delta x}}_{a} \;+\; \underbrace{q\;\frac{D_n}{\Delta x}}_{b}\right] \end{aligned} \tag{75}$$

If we assume a condition where $J_n = 0$ (equilibrium) and $a >> b$ (negligible diffusion) it is easy to see that positivity of the carrier density is not guaranteed, since the solution oscillates as $n_{i+1} \approx -n_i$. Also, it can be shown that for stability we need to have $\psi_{i+1} - \psi i \;>\; 2k_BT/q$, which requires very small mesh spacing to be verified.

The approach by Scharfetter and Gummel (1969) has provided an optimal solution to this problem, although the mathematical properties of the proposed scheme have been fully recognized much later. We consider again a linear potential variation between neighboring mesh points, which is consistent with the use of finite differences to express the field. We express the current in the interval $[x_i; x_{i+1}]$ as a truncated expansion about the value at the mid–point $x_{i+\frac{1}{2}}$

$$J_n(x) \;=\; J_n(x_{i+\frac{1}{2}}) \;+\; (x - x_{i+\frac{1}{2}})\frac{\partial}{\partial x}J_n(x) \tag{76}$$

From (76) we obtain a first order differential equation for J_n which can be solved to provide $n(x)$ in the mesh interval, using as boundary conditions the values of carrier density n_i and n_{i+1}. We obtain

$$n(x) \;=\; [1-g(x,\psi)]\, n_i \;+\; g(x,\psi) n_{i+1}; \qquad x \in [x_i; x_{i+1}] \tag{77}$$

where $g(x,\psi)$ is the growth function

$$g(x,\psi) \;=\; \left[1 \;-\; \exp\left(\frac{\psi_{i+1}-\psi_i}{k_B T/q}\frac{x-x_i}{\Delta x}\right)\right] / \left[1 \;-\; \exp\left(\frac{\psi_{i+1}-\psi_i}{k_B T/q}\right)\right] \tag{78}$$

The result in (77) can be used to evaluate $n(x_{i+\frac{1}{2}})$ for the discretization of the current in (73). It is easy to see that only when $\psi(i+i)-\psi(i)=0$ we have

$$n_{1+\frac{1}{2}} \;=\; (1-\frac{1}{2})n_i + \frac{1}{2}n_{i+1} \;=\; \frac{n_i \;+\; n_{i+1}}{2} \tag{79}$$

The continuity equation can be easily discretized on rectangular uniform and nonuniform meshes using the above results for the currents, because the mesh lines are aligned exactly.

For a more in depth reading on advanced aspects of drift–diffusion simulation, we recommend to consult the book by S. Selberherr, cited below.

Some classic references:

H.K. Gummel, "A self–consistent iterative scheme for one–dimensional steady state transistor calculation," *IEEE Transactions on Electron Devices*, vol. ED-11, pp.455–465, 1964.

A. DeMari, "An accurate numerical steady state one–dimensional solution of the p–n junction," *Solid–state Electronics*, vol. 11, pp. 33–59, 1968.

A. DeMari, "An accurate numerical one–dimensional solution of the p–n junction under arbitrary transient conditions," *Solid–state Electronics*, vol. 11, pp. 1021–2053, 1968.

S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer–Verlag, 1984.

D. L. Scharfetter and D. L. Gummel, "Large signal analysis of a Silicon Read diode oscillator," *IEEE Transaction on Electron Devices*, vol. ED-16, pp.64–77, 1969.

J. W. Slotboom, "Iterative scheme for 1 and 2–dimensional d.c. transistor simulation," *Electronics Letters*, vol. 5, pp. 677–678, 1969.

J.W. Slotboom, "Computer–aided two–dimensional analysis of bipolar transistors," *IEEE Transactions on Electron Devices*, vol. ED-20, pp. 669–679, 1973.