

Numerical Solution of the Electron Transport Equation in the Upper Atmosphere

M. Woods^{a,*}, W. Sailor^a, M. Holmes^b

^a*Department of Phenomenology and Sensor Science, Sandia National Laboratories, Albuquerque, NM 87185*

^b*Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180*

Abstract

A new approach for solving the electron transport equation in the upper atmosphere is derived. The problem is a very stiff boundary value problem, and to obtain an accurate numerical solution, matrix factorizations are used to decouple the fast and slow modes. A stable finite difference method is applied to each mode. This solver is applied to a simplified problem for which an exact solution exists using various versions of the boundary conditions that might arise in a natural auroral display. The numerical and exact solutions are found to agree with each other to at least two significant digits.

Keywords: integro-differential equation, stiff boundary value problem, upwind difference method, electron transport, upper atmosphere

1. Introduction

The principle cause of the aurora is charged particles from the sun interacting with the earth's upper atmosphere. During a solar storm, the sun releases charged particles called the solar wind, which consists mostly of electrons and protons. Some of these particles reach the earth where they accelerate along the magnetic field lines which converge at the poles. Once the particles reach sufficiently low altitudes they scatter off atmospheric atoms and molecules. This scattering causes the atoms and molecules to enter excited states, and they return to their ground states via collisional quenching and fluorescence. Most of the auroral light is due to electron impact excitations, which is why we focus on electron transport. A more comprehensive explanation of how this occurs can be found in Rees [1].

The electron transport problem can be modeled as an integro-differential equation. What is challenging about the problem is the unusual way the boundary conditions are split between the upper and lower limits of the atmosphere, and the pronounced boundary-layer structure of the solution due to the stiffness of the equation. Both Monte Carlo and deterministic methods have been used, but only the latter are more relevant to our study. The most well-known approach is due to Stamnes et al. [2]. This method

*Corresponding author

Email address: mcwood@sandia.gov (M. Woods)

Preprint submitted to Journal of Computational Physics

July 3, 2017

subdivides the atmosphere into layers, and in each layer the problem is replaced by one with constant coefficients. The source term is approximated by an exponential, and the system of equations is solved exactly within each layer. These layers are patched together by requiring the solution to be continuous. This method has been used in Stamnes [3], Lummerzheim et al. [4], Min et al. [5], Lummerzheim and Lilensten [6], and most recently in Lanchester and Gustavsson [7].

What has been missed in the above and other numerical methods is an accurate accounting of the lower atmosphere. To explain, at the upper boundary a downward electron distribution is specified. Similarly, at some low altitude the upward electron distribution is set to zero. The top of the upper atmosphere is simply chosen to be an altitude where the density is relatively small and scattering effects are negligible. The bottom of the atmosphere is more troublesome. Theoretically, the ground (an altitude of zero) could be chosen because there are no free electrons at ground level. However, the electron transport equation has the property of becoming exponentially stiff at lower altitudes. This region can be avoided by prescribing a lower boundary that is far from ground level, but this yields a model which is not physically meaningful. On the other hand, if a more realistic lower boundary is used then standard numerical methods, such as collocation at Gaussian or Lobatto points, will produce negative and oscillating solutions due to the exponential stiffness.

Regardless of what low altitude is chosen, the solution must smoothly approach zero as the altitude decreases. The stiffness creates an interior layer somewhere between 100 and 200 km. For illustration purposes, we show one such example layer in Figures 1 and 2. Each line represents electron intensity at 2 eV for some pitch angle. The downward moving electron streams are shown with the solid lines and the upward moving streams with the dashed lines. The very rapid drop in intensity as the electrons approach the lower altitudes causes most solvers, even those designed for stiff problems, to overshoot or oscillate. The resulting negative intensities are physically meaningless, and the problem is unstable in such cases.

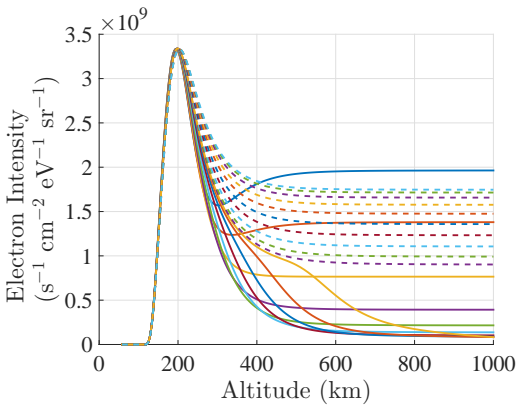


Figure 1: An example electron intensity solution at 2 eV. The interior layer region is at altitudes between 100 and 200 km.

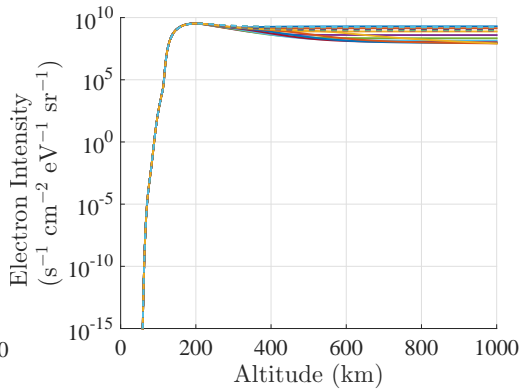


Figure 2: The same intensity on a logarithmic scale. The intensity approaches zero as altitude decreases, but does not overshoot or oscillate.

For the most part, studies of the electron transport problem deemphasize the bound-

ary conditions. Although they may be briefly mentioned, there is little discussion or explanation of the numerical complications they introduce. Some do not mention the boundary conditions at all (see Stamnes [3] for example) and some do not give enough detail to understand what exactly they used as a boundary condition (see Porter et al. [8] for example). In particular, almost all articles neglect to mention the consequences of how to handle the lower boundary. Only Mantas [9] discusses the difficulty of choosing the lower boundary and that negative intensities lead to instabilities and meaningless solutions. In a later article however, this same author uses an arbitrary reflecting lower boundary condition (see Mantas and Bowhill [10]). He forces 60% of all electrons reaching an altitude of 120 km back upward. The claim is made that although this is not realistic, the condition does not adversely affect the solution.

The computational algorithm developed here to solve the electron transport equation uses a generalized Legendre polynomial approximation for the elastic scattering integral followed by an eigenvalue decomposition to delineate the relevant scales needed to resolve the solution. To begin, in Section 2, the electron transport equation is described. In Section 3 the generalized Legendre polynomial approximation is used to reduce the electron transport equation to the form that is solved numerically. Section 4 provides the methodology for solving a stiff boundary value problem by adapting the work of Kreiss et al. [11]. Finally, in Section 5, our numerical solution is compared to a case where an exact solution is possible, showing the effectiveness of the numerical method.

2. Mathematical Model

In order to quantify electron transport, we use the electron intensity $I(\mathbf{r}, \mathbf{v}, t)$ where $\mathbf{r} = (x, y, z)$ is electron position in Cartesian coordinates, $\mathbf{v} = (v, \theta, \phi)$ is electron velocity in spherical coordinates, and t is time. Electron intensity is related to the electron distribution function, and its units are $\text{s}^{-1} \text{cm}^{-2} \text{eV}^{-1} \text{sr}^{-1}$ (see Rees [1]). An equation for electron intensity can be derived from the continuity equation (see Woods [12] for the derivation). To obtain this equation, a few assumptions are made.

First, we consider the steady state problem. This is reasonable because the time it takes for an electron to penetrate the atmosphere is small compared to the time it takes for the aurora to change. The second assumption is that the atmosphere can be modeled as horizontally stratified. This is reasonable because the atmospheric density gradient in the local vertical direction is much larger than in any local horizontal direction. Finally, we assume a uniform magnetic field. For the altitudes we are considering, the curvature of the earth's magnetic field has a negligible effect, so we consider it to be uniform. A consequence of this final assumption is that the electron intensity is invariant under rotations about the magnetic field lines. This is because an electron travels in a helical path in a uniform magnetic field. With these assumptions, the electron intensity becomes a function of three variables: altitude z (which we take in the same direction as the magnetic field for simplicity), kinetic energy E , and the cosine of the pitch angle $\mu = \cos \theta$ (the angle between the electron velocity vector and the magnetic field is the pitch angle – downward moving electrons have a negative pitch angle).

If we assume the only processes are in-scattering and out-scattering, then the electron

transport equation is given by

$$\begin{aligned}
\mu \frac{\partial I(z, E, \mu)}{\partial z} = & \sum_{\substack{\text{species} \\ \xi}} n_{\xi}(z) \left(-\sigma_{\xi}^{\text{tot}}(E) I(z, E, \mu) \right. \\
& + \sigma_{\xi}^{\text{el}}(E) \int_{-1}^1 P_{\xi}(E', \mu, \mu') I(z, E, \mu') d\mu' + \sum_{\substack{\text{channels} \\ \eta}} \sigma_{\xi}^{\eta}(E + T_{\xi}^{\eta}) I(z, E + T_{\xi}^{\eta}, \mu) \\
& + \int_{E+T_{\xi}^{\text{ion}}}^{\infty} \sigma_{\xi}^{\text{ion}}(E') R_{\xi}^{\text{pri}}(E, E') I(z, E', \mu) dE' \\
& \left. + \frac{1}{2} \int_{E+T_{\xi}^{\text{ion}}}^{\infty} \int_{-1}^1 \sigma_{\xi}^{\text{ion}}(E') R_{\xi}^{\text{sec}}(E, E') I(z, E', \mu') d\mu' dE' \right) \quad (1)
\end{aligned}$$

where the species under consideration are N_2 , O_2 , O , N , Ar , NO , H , and He and the channels are the most significant excited states (the states whose cross sections are the largest). Of the species listed, the most important are N_2 , O_2 , and O . The others are included because they can significantly contribute to the production of auroral light. The other terms are defined in Table 1.

Table 1: The functions of (1) are defined.

Function	Description	Units
$n_{\xi}(z)$	atmospheric number density of species ξ	cm^{-3}
$\sigma_{\xi}^{\text{el}}(E)$	cross section for elastic scattering of species ξ	cm^2
$\sigma_{\xi}^{\eta}(E)$	cross section for excitation scattering of species ξ to state η	cm^2
$\sigma_{\xi}^{\text{ion}}(E)$	cross section for single ionization scattering of species ξ	cm^2
$\sigma_{\xi}^{\text{tot}}(E)$	sum of all cross sections for species ξ	cm^2
$P_{\xi}(E, \mu, \mu')$	probability density function for an incident electron at energy E and direction μ' to elastically scatter to direction μ for species ξ	—
$R_{\xi}^{\text{pri}}(E, E')$	probability density function for an incident electron at energy E' to create a primary electron at energy E via ionization for species ξ	eV^{-1}
$R_{\xi}^{\text{sec}}(E, E')$	probability density function for an incident electron at energy E' to create a secondary electron at energy E via ionization for species ξ	eV^{-1}
T_{ξ}^{η}	excitation energy threshold for species ξ and state η	eV
T_{ξ}^{ion}	single ionization energy threshold for species ξ	eV

Equation (1) describes how the electron intensity changes with respect to altitude. For every species, there is a chance for elastic scattering (second term on the right-hand-side), excitation scattering (third term on the right-hand-side), and single ionization scattering (final two terms on the right-hand-side). We model these events as the production of a new electron (or two new electrons in the case of single ionization). Although it is physically the same electron (except for the secondary electron produced in the ionization), it has a new pitch angle if the scattering was elastic or a new kinetic energy if the scattering was inelastic. The first term on the right-hand-side serves to remove the electron with the pre-scattering pitch angle and kinetic energy.

The elastic scattering probability density functions (also called the phase functions) are given by (see Rees [1])

$$P_{\xi}(E, \mu, \mu') = \frac{2\varepsilon_{\xi}(1 + \varepsilon_{\xi})(1 + 2\varepsilon_{\xi} - \mu\mu')}{[(1 + 2\varepsilon_{\xi} - \mu\mu')^2 - (1 - \mu^2)(1 - \mu'^2)]^{3/2}} \quad (2)$$

where $\varepsilon_{\xi} = \varepsilon_{\xi}(E)$ is an energy dependent parameter that will be discussed in more detail along with (2) in Section 3.

The ionization probability density functions are given by (see again Rees [1])

$$R_{\xi}^{\text{pri}}(E, E') = \frac{1}{N_{\xi}(E')} \frac{1}{E' - E} \exp \left[-\frac{E' - E}{31.5} - 339 \exp \left(-\frac{E' - E}{2.49} \right) \right] \\ \times \log \left(\frac{\sqrt{E'} + \sqrt{E}}{\sqrt{E'} - \sqrt{E}} \right) \quad (3)$$

and

$$R_{\xi}^{\text{sec}}(E, E') = R_{\xi}^{\text{pri}}(E' - E - T_{\xi}^{\text{ion}}, E'), \quad (4)$$

where the normalization functions $N_{\xi}(E')$ are determined from

$$\int_0^{E' - T_{\xi}^{\text{ion}}} R_{\xi}^{\text{pri}}(E, E') dE = 1. \quad (5)$$

There exist sophisticated models of the atmosphere, and obtaining $n_{\xi}(z)$ for all species listed above except NO can be found in Hedin [13]. The number densities for NO can be found in Sharma et al. [14]. In addition, numerous scattering experiments and theoretical calculations have been performed that give the necessary cross sections. For this study, we used Laher and Gilmore [15], Itikawa and Ichimura [16], Salvat et al. [17], Pancheshnyi et al. [18], Morgan [19], Itikawa [20], Phelps [21], Biagi [22], Alves [23], Itikawa [24], Ionin et al. [25], Kim et al. [26], Hayashi [27], and Cartwright et al. [28].

Finally, the ranges for the electron kinetic energy E and pitch angle cosines μ are

$$0 < E < \infty, \quad -1 \leq \mu \leq 1, \quad (6)$$

and the domain of equation (1) is

$$z_{\text{bot}} \leq z \leq z_{\text{top}}. \quad (7)$$

The boundary conditions are given by

$$I(z_{\text{top}}, E, \mu) = I_{\text{top}}(E, \mu) \text{ for } \mu < 0, \quad I(z_{\text{bot}}, E, \mu) = 0 \text{ for } \mu > 0. \quad (8)$$

These conditions give the intensity entering at each end of the domain. Both endpoints must be chosen. The top of the atmosphere z_{top} can simply be chosen to be any altitude where the atmosphere is sufficiently thin so that scattering is negligible. We shall choose $z_{\text{top}} = 1000$ km. The bottom of the atmosphere z_{bot} is more difficult. This is an altitude where there is no upward intensity. We could simply choose the ground ($z_{\text{bot}} = 0$), but this is not a good choice due to the extreme stiffness of the problem at low altitudes. This stiffness causes rapid oscillations and negative intensities (which are meaningless)

in the numerical solution for most solvers unless a very large number of points are placed in the mesh. We will soon see that the number of points becomes so large that it is impractical. On the other hand, if z_{bot} is chosen too high (say 100 km), then we will not find a true solution because we will be ignoring the physics of how the intensity decays to zero. The approach we will take is to choose an altitude where the electron intensity is clearly zero. Figures 1 and 2 show that $z_{\text{bot}} = 50$ km is adequate.

3. The δ - M Method

In order to discretize (1), the integrals should be approximated by appropriate quadrature sums. Many quadrature techniques approximate the integrand by a low-order polynomial or a piecewise polynomial. The problem here is that the phase function for elastic scattering $P_\xi(E, \mu, \mu')$ is not well approximated by a low-order polynomial due to the fact that it contains sharp peaks at $\mu = \mu'$. The formula for the phase function was given by (2), and the problem is that ε_ξ is an energy-dependent parameter that approaches zero as E becomes large. The result is that for large energies, $P_\xi(E, \mu, \mu')$ is sharply peaked at $\mu = \mu'$ and an inordinate number of points is required to approximate it with polynomials. There is an idea called the δ - M method, originally formulated by Wiscombe [29] for radiative transfer, that overcomes this difficulty. The formulas used here are slightly different than the original formulation, but the concepts are the same.

If we expand the phase function in an infinite series

$$P_\xi(E, \mu, \mu') = \sum_{m=0}^{\infty} \frac{2m+1}{2} \chi_{\xi,m}(E) P_m(\mu) P_m(\mu') \quad (9)$$

where $P_m(\cdot)$ is the m th degree Legendre polynomial and the $\chi_{\xi,m}$ are the phase function moments, then it can be shown that (see Woods [12])

$$\chi_0 = 1, \quad (10)$$

$$\chi_1 = 1 + 2\varepsilon + 2\varepsilon(1 + \varepsilon) \log\left(\frac{\varepsilon}{1 + \varepsilon}\right), \quad (11)$$

and the remaining moments are found through recursion

$$\chi_m = \frac{1}{m-1} [(2m-1)(1+2\varepsilon)\chi_{m-1} - m\chi_{m-2}], \quad m = 2, 3, \dots \quad (12)$$

As mentioned earlier, the phase function is not well approximated by a low-order polynomial due to the sharp peak at $\mu = \mu'$. The idea behind the δ - M method is to approximate $P_\xi(E, \mu, \mu')$ by a Dirac delta function and M Legendre polynomials. This way, the delta function can capture the sharp peak and the polynomials can capture the rest. Let our approximation be

$$P_{\xi,M}^*(E, \mu, \mu') = f_\xi(E)\delta(\mu - \mu') + [1 - f_\xi(E)] \sum_{m=0}^{M-1} \frac{2m+1}{2} \chi_{\xi,m}^*(E) P_m(\mu) P_m(\mu') \quad (13)$$

where $f_\xi(E)$ is the “fraction” that represents the sharp peak and $\chi_{\xi,m}^*(E)$ are the modified moments. From this, we can find

$$\chi_{\xi,m}^*(E) = \frac{\chi_{\xi,m}(E) - f_\xi(E)}{1 - f_\xi(E)}, \quad m = 0, 1, \dots, M-1, \quad (14)$$

$$f_\xi(E) = \chi_{\xi,M}(E). \quad (15)$$

Putting everything together, the electron transport equation is now given by

$$\begin{aligned} \mu \frac{\partial I(z, E, \mu)}{\partial z} = & \sum_{\substack{\text{species} \\ \xi}} n_\xi(z) \left(- [\sigma_\xi^{\text{tot}}(E) - f_\xi(E) \sigma_\xi^{\text{el}}(E)] I(z, E, \mu) \right. \\ & + [1 - f_\xi(E)] \sigma_\xi^{\text{el}}(E) \sum_{m=0}^{M-1} \frac{2m+1}{2} \chi_{\xi,m}^*(E) P_m(\mu) \int_{-1}^1 P_m(\mu') I(z, E, \mu') d\mu' \\ & + \sum_{\substack{\text{channels} \\ \eta}} \sigma_\xi^\eta(E + T_\xi^\eta) I(z, E + T_\xi^\eta, \mu) + \int_{E+T_\xi^{\text{ion}}}^{\infty} \sigma_\xi^{\text{ion}}(E') R_\xi^{\text{pri}}(E, E') I(z, E', \mu) dE' \\ & \left. + \frac{1}{2} \int_{E+T_\xi^{\text{ion}}}^{\infty} \int_{-1}^1 \sigma_\xi^{\text{ion}}(E') R_\xi^{\text{sec}}(E, E') I(z, E', \mu') d\mu' dE' \right). \end{aligned} \quad (16)$$

The power of the δ - M method is that M does not need to be large to accurately approximate the elastic scattering integral. This means that a small number of points can be used, which greatly reduces the amount of computation.

4. Numerical Solution

Equations (16) and (8) together form a two-point linear boundary value problem (BVP). A common method for solving this type of problem is either Gaussian or Lobatto collocation (see Ascher et al. [30]). The idea behind these methods is to choose a set of points in the domain (called the collocation points) and satisfy the differential equation at those points using piecewise polynomials of a certain degree. The difference between Gaussian and Lobatto collocation is the choice of collocation points. These methods are common because they are excellent for non-stiff and moderately stiff problems and software is freely available.

However, upon inspecting (16), we see that the atmospheric number density multiplies the entire right-hand-side, meaning that the eigenvalues of the system will be approximately proportional to the dominant density at every altitude. From Hedin [13], we know that at an altitude of 0 km, the density is on the order of 10^{19} cm^{-3} and at 1000 km, the density is on the order of 10^5 cm^{-3} . Since these densities span many orders of magnitude, so do the eigenvalues of the system. Therefore at the lower altitudes, equation (16) is very stiff.

The largest eigenvalues for some representative cases are shown in Table 2. We see that the eigenvalues indeed span many orders of magnitude. Using a low order BVP solver, it is not unreasonable to require that the product of the local step size and largest local eigenvalue be less than 2. It is well known that keeping this product small helps keep

the error small (see Ascher and Petzold [31]). Table 2 shows that this is not a problem for the high altitudes. Step sizes on the order of kilometers are possible. However, at the low altitudes this requires step sizes on the order of centimeters for 100000 eV and microns for 10 eV. Clearly, if we wish to accurately solve the problem without using hundreds of thousands of points, something else must be done.

Table 2: The largest eigenvalues for some representative cases.

Altitude (km)	Energy (eV)	Eigenvalue (cm^{-1})
50	10	1.639×10^3
50	100000	1.513×10^{-1}
1000	10	7.960×10^{-9}
1000	100000	1.194×10^{-13}

The problem under consideration qualifies as being a very stiff, or an exponentially stiff BVP. We make a distinction between a stiff and a very stiff BVP. For the latter, the relevant length scales change dramatically in the boundary layer with the result that standard stiff methods are not capable of accurately finding the solution. The fact that the electron transport equation qualifies as being very stiff is evident in Figure 2. This is likely the reason why all previous numerical attempts at solving this problem have made spurious assumptions such as 60% of all electrons reaching 120 km are reflected back upward. Numerical methods for very stiff BVPs are not widely used or well-known. The purpose of this section is to describe an upwind numerical method that avoids these spurious assumptions.

4.1. Overview of the Upwind Method

To explain what is done in the upwind method, consider the problem

$$\frac{d\mathbf{y}(x)}{dx} = \mathbf{A}(x)\mathbf{y}(x) + \mathbf{f}(x), \quad a \leq x \leq b \quad (17)$$

with boundary conditions

$$\mathbf{B}_a\mathbf{y}(a) + \mathbf{B}_b\mathbf{y}(b) = \mathbf{c} \quad (18)$$

where $\mathbf{y}(x), \mathbf{f}(x), \mathbf{c} \in \mathbb{R}^M$ and $\mathbf{A}(x), \mathbf{B}_a, \mathbf{B}_b \in \mathbb{R}^{M \times M}$. In Woods [12], it is shown how (16) can be reduced to this form. It turns out that the eigenvalues for the electron transport problem are real, so we will assume real eigenvalues throughout this section. The method we will use is originally from Kreiss et al. [11]. These authors were principally focused on proving error estimates and did not adequately explain its implementation. Consequently, we will here be focused on implementation.

Throughout this section, we will assume some mesh $a = x_0 < x_1 < x_2 < \dots < x_{N-1} < x_N = b$ with $\Delta x_n = x_n - x_{n-1}$ for $n = 1, 2, \dots, N$ and

$$\Delta x = \max_{1 \leq n \leq N} \Delta x_n. \quad (19)$$

In addition, all norms used denote the infinity norm. The notation

$$\|\mathbf{z}(x)\|_{(\alpha, \beta)} = \sup_{\alpha < x < \beta} \|\mathbf{z}(x)\|_{\infty} \quad (20)$$

will be used throughout for the norm of a vector function $\mathbf{z}(x)$ on an interval (α, β) .

As stated in Kreiss et al. [11], a function $\mathbf{y}(x)$ is resolved on an interval (α, β) if

$$\left\| \frac{d^\nu \mathbf{y}(x)}{dx^\nu} \right\|_{(\alpha, \beta)} \leq K (\|\mathbf{y}(x)\|_{(\alpha, \beta)} + 1) \quad (21)$$

for $\nu = 0, 1, \dots, p$ where $K\Delta x \ll 1$. The degree of smoothness p can vary with the BVP. The importance of using a mesh that resolves the solution follows from the error analysis for any finite difference method. For example, the local truncation error for the trapezoidal rule is bounded by $\|\tau_n\| \leq \frac{1}{12} \Delta x^3 \|\mathbf{y}'''(x)\|_{(x_{n-1}, x_n)}$. For non-stiff BVPs, Δx can be made small enough so that this error is small. However, for a stiff BVP the derivatives of the solution can be very large so that the error is large unless if Δx is made prohibitively small. For this reason, a solution is only resolved if a number p of its derivatives are bounded by a constant K that is not too large (i.e. $K\Delta x \ll 1$). Further, if $\mathbf{y}(x)$ is resolved on (α, β) and (β, γ) , then it is resolved on (α, γ) . This means that we only need to worry about resolving $\mathbf{y}(x)$ in the neighborhood of every point $x \in (a, b)$.

The obvious problem with (21) is that in order to know if the mesh resolves the solution, it appears we need to already have the solution. It turns out that we can find a mesh that resolves the solution using only information about the BVP coefficients $\mathbf{A}(x)$ and $\mathbf{f}(x)$. This means that an adequate mesh can be found before obtaining the solution.

Definition 1. Suppose a matrix function $\mathbf{D}(x) \in \mathbb{R}^{M \times M}$ can be partitioned into the form

$$\mathbf{D}(x) = \begin{bmatrix} \mathbf{D}^{11}(x) & \mathbf{D}^{12}(x) & \mathbf{D}^{13}(x) \\ \mathbf{D}^{21}(x) & \mathbf{D}^{22}(x) & \mathbf{D}^{23}(x) \\ \mathbf{D}^{31}(x) & \mathbf{D}^{32}(x) & \mathbf{D}^{33}(x) \end{bmatrix} \quad (22)$$

where $\mathbf{D}^{ij}(x) \in \mathbb{R}^{m_i \times m_j}$ for $i, j = 1, 2, 3$ and $m_1 + m_2 + m_3 = M$. $\mathbf{D}(x)$ is *essentially diagonally dominant* on (α, β) if $\mathbf{D}^{11}(x)$ and $\mathbf{D}^{33}(x)$ are strictly diagonally dominant,

$$\left\| [\tilde{\mathbf{D}}^{ii}(x)]^{-1} \mathbf{D}^{ij}(x) \right\|_{(\alpha, \beta)} \leq K_0 \quad (23)$$

for $i = 1, 3$ and $j = 1, 2, 3$, and

$$\|\mathbf{D}^{2j}(x)\|_{(\alpha, \beta)} \leq K_0 \quad (24)$$

for $j = 1, 2, 3$. Here $\tilde{\mathbf{D}}^{ii}(x)$ is the diagonal matrix containing the diagonal elements of $\mathbf{D}^{ii}(x)$ and $K_0\Delta x \ll 1$.

Note that in this definition, there is no requirement that $\|\mathbf{D}^{ij}(x)\|_{(\alpha, \beta)}$ for $i = 1, 3$ and $j = 1, 2, 3$ is small. As we will shortly see, the first and last row blocks will correspond to the stiff portion of the BVP. Also note that there is nothing special about submatrices $\mathbf{D}^{11}(x)$ and $\mathbf{D}^{33}(x)$. The definition is written with those blocks being strictly diagonally dominant because the algorithm we will use to solve the BVP will put the system in this form. Now we are in a position to determine if a given mesh resolves the solution to our BVP.

Theorem 1 (Kreiss et al. [11]). *Consider the BVP*

$$\frac{d\mathbf{w}(x)}{dx} = \mathbf{D}(x)\mathbf{w}(x) + \mathbf{g}(x), \quad a \leq x \leq b \quad (25)$$

where $\mathbf{D}(x)$ is partitioned as in (22). Let $\mathbf{w}(x)$ and $\mathbf{g}(x)$ be similarly partitioned so that $\mathbf{w}^i(x), \mathbf{g}^i(x) \in \mathbb{R}^{m_i}$ for $i = 1, 2, 3$. If $\mathbf{D}(x)$ is essentially diagonally dominant on (α, β) and there are constants K_1 and K_2 such that

$$\left\| [\tilde{\mathbf{D}}^{ii}(x)]^{-1} \frac{d^\nu \mathbf{D}^{ij}(x)}{dx^\nu} \right\|_{(\alpha, \beta)} \leq K_1 \quad \text{and} \quad \left\| [\tilde{\mathbf{D}}^{ii}(x)]^{-1} \frac{d^\nu \mathbf{g}^i(x)}{dx^\nu} \right\|_{(\alpha, \beta)} \leq K_1 \quad (26)$$

for $i = 1, 3, j = 1, 2, 3$, and $\nu = 0, 1, \dots, p$;

$$\left\| \frac{d^\nu \mathbf{D}^{2j}(x)}{dx^\nu} \right\|_{(\alpha, \beta)} \leq K_1 \quad \text{and} \quad \left\| \frac{d^\nu \mathbf{g}^2(x)}{dx^\nu} \right\|_{(\alpha, \beta)} \leq K_1 \quad (27)$$

for $j = 1, 2, 3$ and $\nu = 0, 1, \dots, p$; and

$$\left\| \tilde{\mathbf{D}}^{11}(a) \right\|_{(\alpha, \beta)} \leq K_2 \quad \text{and} \quad \left\| \tilde{\mathbf{D}}^{33}(b) \right\|_{(\alpha, \beta)} \leq K_2, \quad (28)$$

then $\mathbf{w}(x)$ is resolved on (α, β) . Here again, $\tilde{\mathbf{D}}^{ii}(x)$ is the diagonal matrix containing the diagonal elements of $\mathbf{D}^{ii}(x)$, $K_1 \Delta x \ll 1$, and $K_2 \Delta x \ll 1$.

Theorem 1 requires the matrix of the BVP to be essentially diagonally dominant. This will not in general be the case. Thus, we must find a way to transform (17) to this form. Suppose we are able to find an invertible matrix function $\mathbf{V}(x) \in \mathbb{R}^{M \times M}$ such that

$$\mathbf{V}^{-1}(x) \mathbf{A}(x) \mathbf{V}(x) = \mathbf{\Lambda}(x) = \begin{bmatrix} \mathbf{\Lambda}^{11}(x) & & \\ & \mathbf{\Lambda}^{22}(x) & \\ & & \mathbf{\Lambda}^{33}(x) \end{bmatrix} \quad (29)$$

where $\mathbf{\Lambda}^{ii}(x) \in \mathbb{R}^{m_i \times m_i}$ and $m_1 + m_2 + m_3 = M$. Here, the eigenvalues of $\mathbf{\Lambda}^{11}(x)$ are large and negative, the eigenvalues of $\mathbf{\Lambda}^{33}(x)$ are large and positive, and the eigenvalues (both positive and negative) of $\mathbf{\Lambda}^{22}(x)$ are small. Using (29), Equation (17) becomes

$$\frac{d\mathbf{y}(x)}{dx} = \mathbf{V}(x) \mathbf{\Lambda}(x) \mathbf{V}^{-1}(x) \mathbf{y}(x) + \mathbf{f}(x). \quad (30)$$

Let $\mathbf{w}(x) = \mathbf{V}^{-1}(x) \mathbf{y}(x)$. Then

$$\frac{d\mathbf{w}(x)}{dx} = \mathbf{D}(x) \mathbf{w}(x) + \mathbf{g}(x) \quad (31)$$

where $\mathbf{D}(x) = \mathbf{\Lambda}(x) - \mathbf{V}^{-1}(x) \mathbf{V}'(x)$ (the prime denotes a derivative with respect to x) and $\mathbf{g}(x) = \mathbf{V}^{-1}(x) \mathbf{f}(x)$. With Theorem 1 in mind, suppose that $\mathbf{\Lambda}^{11}(x)$ and $\mathbf{\Lambda}^{33}(x)$ are strictly diagonally dominant. Now if $\left\| \mathbf{V}^{-1}(x) \mathbf{V}'(x) \right\|_{(\alpha, \beta)}$ is small, then $\mathbf{D}(x)$ is essentially diagonally dominant on (α, β) . Hence, the application of Theorem 1 relies on our ability to find an appropriate matrix function $\mathbf{V}(x)$.

Before we do this however, (29) and (31) are suggestive of a finite difference method. If an initial value problem (IVP) or a terminal value problem (TVP) have large eigenvalues, it is known that methods such as backward Euler work very well. If either an IVP or TVP has only small eigenvalues, then the trapezoidal rule is more accurate. With this in mind, let us use backward Euler for the $\mathbf{\Lambda}^{11}(x)$ block integrating from a to b , the

trapezoidal rule for the $\Lambda^{22}(x)$ block (integration direction does not matter since it is a symmetric method), and backward Euler for the $\Lambda^{33}(x)$ block integrating from b to a . This finite difference method gives us

$$\begin{aligned}
& - \begin{bmatrix} \mathbf{I}_{m_1} & & \\ \frac{\Delta x_n}{2} \mathbf{D}_{n-1}^{21} & \mathbf{I}_{m_2} + \frac{\Delta x_n}{2} \mathbf{D}_{n-1}^{22} & \frac{\Delta x_n}{2} \mathbf{D}_{n-1}^{23} \\ \Delta x_n \mathbf{D}_{n-1}^{31} & \Delta x_n \mathbf{D}_{n-1}^{32} & \mathbf{I}_{m_3} + \Delta x_n \mathbf{D}_{n-1}^{33} \end{bmatrix} \mathbf{V}_{n-1}^{-1} \mathbf{u}_{n-1} \\
& + \begin{bmatrix} \mathbf{I}_{m_1} - \Delta x_n \mathbf{D}_n^{11} & -\Delta x_n \mathbf{D}_n^{12} & -\Delta x_n \mathbf{D}_n^{13} \\ -\frac{\Delta x_n}{2} \mathbf{D}_n^{21} & \mathbf{I}_{m_2} - \frac{\Delta x_n}{2} \mathbf{D}_n^{22} & -\frac{\Delta x_n}{2} \mathbf{D}_n^{23} \\ & & \mathbf{I}_{m_3} \end{bmatrix} \mathbf{V}_n^{-1} \mathbf{u}_n = \Delta x_n \begin{bmatrix} \mathbf{g}_n^1 \\ \frac{1}{2}(\mathbf{g}_{n-1}^2 + \mathbf{g}_n^2) \\ \mathbf{g}_{n-1}^3 \end{bmatrix}
\end{aligned} \tag{32}$$

for $n = 1, 2, \dots, N$ where \mathbf{I}_m is the $m \times m$ identity matrix and $\mathbf{u}_n \approx \mathbf{y}(x_n)$. Also, we use the shorthand $\mathbf{V}_n^{-1} = \mathbf{V}^{-1}(x_n)$ and similarly for $\mathbf{D}^{ij}(x_n)$ and $\mathbf{g}^i(x_n)$. This notation will be used throughout this section.

To numerically solve (17), the finite difference equations (32) and the boundary conditions (18) are assembled into a system of equations of size $M(N + 1)$. This is a very sparse system. To solve it, we use an incomplete LU factorization as a preconditioner and a stabilized biconjugate gradient method.

Finally, for stability reasons, we consider an eigenvalue small if the product of itself and the local mesh spacing is less than 2. Otherwise, we consider the eigenvalue to be large.

4.2. The Schur Method

We now turn our attention to finding an appropriate $\mathbf{V}(x)$. From (29), we see that we need a similarity transform. One way to find $\mathbf{V}(x)$ is to find a series of similarity transforms and construct $\mathbf{V}(x)$ from those. At any mesh point x_n we can find the Schur decomposition of $\mathbf{A}(x_n)$. This is given by

$$\tilde{\mathbf{V}}_n^{-1} \mathbf{A}_n \tilde{\mathbf{V}}_n = \tilde{\Lambda}_n = \begin{bmatrix} \tilde{\Lambda}^{11} & \tilde{\Lambda}^{12} & \tilde{\Lambda}^{13} \\ & \tilde{\Lambda}^{22} & \tilde{\Lambda}^{23} \\ & & \tilde{\Lambda}^{33} \end{bmatrix} \tag{33}$$

where $\tilde{\mathbf{V}}_n$ is an orthogonal matrix and $\tilde{\Lambda}_n$ is an upper triangular matrix. This implies that $\tilde{\Lambda}^{11}$, $\tilde{\Lambda}^{22}$, and $\tilde{\Lambda}^{33}$ are upper triangular, so the eigenvalues of \mathbf{A}_n are on the diagonal of $\tilde{\Lambda}_n$. Further, the Schur decomposition can be done in such a way that the large negative eigenvalues are on the diagonal of $\tilde{\Lambda}^{11}$, the large positive eigenvalues are on the diagonal of $\tilde{\Lambda}^{33}$, and the small eigenvalues (both positive and negative) are on the diagonal of $\tilde{\Lambda}^{22}$.

Next, we zero out the remaining off-diagonal blocks. Let

$$\hat{\mathbf{V}}_n = \begin{bmatrix} \mathbf{I}_{m_1} & \mathbf{S}_1 & \mathbf{S}_3 \\ & \mathbf{I}_{m_2} & \mathbf{S}_2 \\ & & \mathbf{I}_{m_3} \end{bmatrix} \iff \hat{\mathbf{V}}_n^{-1} = \begin{bmatrix} \mathbf{I}_{m_1} & -\mathbf{S}_1 & \mathbf{S}_1 \mathbf{S}_2 - \mathbf{S}_3 \\ & \mathbf{I}_{m_2} & -\mathbf{S}_2 \\ & & \mathbf{I}_{m_3} \end{bmatrix} \tag{34}$$

so that we obtain

$$\hat{\mathbf{V}}_n^{-1} \tilde{\Lambda}_n \hat{\mathbf{V}}_n = \hat{\Lambda}_n = \begin{bmatrix} \hat{\Lambda}_n^{11} & & \\ & \hat{\Lambda}_n^{22} & \\ & & \hat{\Lambda}_n^{33} \end{bmatrix}. \tag{35}$$

This will occur if \mathbf{S}_1 , \mathbf{S}_2 , and \mathbf{S}_3 solve the Sylvester equations

$$\tilde{\mathbf{\Lambda}}^{11}\mathbf{S}_1 - \mathbf{S}_1\tilde{\mathbf{\Lambda}}^{22} = -\tilde{\mathbf{\Lambda}}^{12}, \quad (36)$$

$$\tilde{\mathbf{\Lambda}}^{22}\mathbf{S}_2 - \mathbf{S}_2\tilde{\mathbf{\Lambda}}^{33} = -\tilde{\mathbf{\Lambda}}^{23}, \quad (37)$$

$$\tilde{\mathbf{\Lambda}}^{11}\mathbf{S}_3 - \mathbf{S}_3\tilde{\mathbf{\Lambda}}^{33} = -\tilde{\mathbf{\Lambda}}^{13} - \tilde{\mathbf{\Lambda}}^{12}\mathbf{S}_2. \quad (38)$$

Two algorithms for solving the Sylvester equation are the Bartels-Stewart algorithm and the Hessenberg-Schur algorithm, both of which are given in Golub et al. [32]. The Sylvester equation $\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B} = \mathbf{C}$ has a unique solution if and only if \mathbf{A} and \mathbf{B} do not share any eigenvalues (see Golub et al. [32]), which is guaranteed due to how we have defined the blocks of $\hat{\mathbf{\Lambda}}_n$.

As stated above, the diagonal blocks of (35) are upper triangular, but $\hat{\mathbf{\Lambda}}^{11}$ and $\hat{\mathbf{\Lambda}}^{33}$ are not necessarily strictly diagonally dominant as required in Theorem 1. Fortunately, it is simple to find a similarity transform to make an upper triangular matrix strictly diagonally dominant. Let $\mathbf{Q} \in \mathbb{R}^{m \times m}$ be a diagonal matrix and $\mathbf{\Lambda} \in \mathbb{R}^{m \times m}$ be an upper triangular matrix. Then the product

$$\mathbf{Q}^{-1}\mathbf{\Lambda}\mathbf{Q} = \begin{bmatrix} \lambda_{11} & q_1^{-1}q_2\lambda_{12} & \cdots & q_1^{-1}q_m\lambda_{1m} \\ & \lambda_{22} & \cdots & q_2^{-1}q_m\lambda_{2m} \\ & & \ddots & \vdots \\ & & & \lambda_{mm} \end{bmatrix} \quad (39)$$

is strictly diagonally dominant if we choose $q_m = 1$ and

$$q_i = \frac{\sum_{j=i+1}^m q_j |\lambda_{ij}|}{\gamma |\lambda_{ii}|} \quad (40)$$

for $i = m-1, m-2, \dots, 1$ and $0 < \gamma < 1$. With this in mind, let

$$\bar{\mathbf{V}}_n = \begin{bmatrix} \mathbf{Q}^{11} & & \\ & \mathbf{I}_{m_2} & \\ & & \mathbf{Q}^{33} \end{bmatrix} \iff \bar{\mathbf{V}}_n^{-1} = \begin{bmatrix} (\mathbf{Q}^{11})^{-1} & & \\ & \mathbf{I}_{m_2} & \\ & & (\mathbf{Q}^{33})^{-1} \end{bmatrix} \quad (41)$$

where \mathbf{Q}^{11} and \mathbf{Q}^{33} are diagonal matrices with elements set according to (39) and (40) so that

$$\bar{\mathbf{V}}_n^{-1}\hat{\mathbf{\Lambda}}_n\bar{\mathbf{V}}_n = \mathbf{\Lambda}_n = \begin{bmatrix} \mathbf{\Lambda}^{11} & & \\ & \mathbf{\Lambda}^{22} & \\ & & \mathbf{\Lambda}^{33} \end{bmatrix}. \quad (42)$$

We now have the desired block diagonal matrix $\mathbf{\Lambda}_n$ with strictly diagonally dominant matrices $\mathbf{\Lambda}^{11}$ and $\mathbf{\Lambda}^{33}$. However, we still need $\|\bar{\mathbf{V}}_n^{-1}\bar{\mathbf{V}}_n'\|$ to be small. From (33), (35), and (42), let

$$\tilde{\mathbf{V}}_n\hat{\mathbf{V}}_n\bar{\mathbf{V}}_n = \mathbf{U}_n = \begin{bmatrix} \mathbf{U}^{11} & \mathbf{U}^{12} & \mathbf{U}^{13} \\ \mathbf{U}^{21} & \mathbf{U}^{22} & \mathbf{U}^{23} \\ \mathbf{U}^{31} & \mathbf{U}^{32} & \mathbf{U}^{33} \end{bmatrix}. \quad (43)$$

Also define

$$\frac{1}{d_i} = \left\| \begin{bmatrix} \mathbf{U}^{1i} \\ \mathbf{U}^{2i} \\ \mathbf{U}^{3i} \end{bmatrix} \right\|_{12} \quad (44)$$

for $i = 1, 2, 3$. Now let

$$\check{\mathbf{V}}_n = \begin{bmatrix} d_1 \mathbf{I}_{m_1} & & \\ & d_2 \mathbf{I}_{m_2} & \\ & & d_3 \mathbf{I}_{m_3} \end{bmatrix} \iff \check{\mathbf{V}}_n^{-1} = \begin{bmatrix} d_1^{-1} \mathbf{I}_{m_1} & & \\ & d_2^{-1} \mathbf{I}_{m_2} & \\ & & d_3^{-1} \mathbf{I}_{m_3} \end{bmatrix} \quad (45)$$

so that

$$\check{\mathbf{V}}_n^{-1} \mathbf{A}_n \check{\mathbf{V}}_n = \mathbf{A}_n = \begin{bmatrix} \mathbf{A}^{11} & & \\ & \mathbf{A}^{22} & \\ & & \mathbf{A}^{33} \end{bmatrix}. \quad (46)$$

Note that this last similarity transform does not change \mathbf{A}_n . It only has the effect of scaling \mathbf{V}_n so that $\|\mathbf{V}_n^{-1} \mathbf{V}'_n\|$ is small. This scaling is effective because right-multiplying \mathbf{U}_n by $\check{\mathbf{V}}_n$ normalizes the ‘‘column blocks’’ of \mathbf{U}_n . This makes all elements of \mathbf{U}_n of moderate size. Similarly, left-multiplying \mathbf{U}_n^{-1} by $\check{\mathbf{V}}_n^{-1}$ normalizes the ‘‘row blocks’’ of \mathbf{U}_n^{-1} . Certainly, other choices of $\check{\mathbf{V}}_n$ may be better in this regard, but (46) has been found to be adequate for the electron transport problem.

Finally, we find using (33), (35), (42), and (46) that if we let

$$\mathbf{V}_n = \check{\mathbf{V}}_n \hat{\mathbf{V}}_n \bar{\mathbf{V}}_n \check{\mathbf{V}}_n, \quad (47)$$

then we obtain the necessary similarity transformation for (29).

4.3. The Riccati Method

The Schur method would be expensive if calculated for every mesh point and does not guarantee a smooth $\mathbf{V}(x)$. However, if we already have \mathbf{V}_{n-1} , then

$$\mathbf{V}_{n-1}^{-1} \mathbf{A}_n \mathbf{V}_{n-1} = \tilde{\mathbf{A}}_n = \begin{bmatrix} \tilde{\mathbf{A}}^{11} & \tilde{\mathbf{A}}^{12} & \tilde{\mathbf{A}}^{13} \\ \tilde{\mathbf{A}}^{21} & \tilde{\mathbf{A}}^{22} & \tilde{\mathbf{A}}^{23} \\ \tilde{\mathbf{A}}^{31} & \tilde{\mathbf{A}}^{32} & \tilde{\mathbf{A}}^{33} \end{bmatrix} \quad (48)$$

can be viewed as a perturbation to \mathbf{A}_{n-1} so long as Δx_n is sufficiently small. That is, the off-diagonal blocks should be small. If we can eliminate the off-diagonal blocks through similarity transforms, then we can avoid calculating the Schur decomposition of \mathbf{A}_n .

Let us partition $\tilde{\mathbf{A}}_n$ so that

$$\tilde{\mathbf{A}}_n = \left[\begin{array}{cc|c} \tilde{\mathbf{A}}^{11} & \tilde{\mathbf{A}}^{12} & \tilde{\mathbf{A}}^{13} \\ \tilde{\mathbf{A}}^{21} & \tilde{\mathbf{A}}^{22} & \tilde{\mathbf{A}}^{23} \\ \tilde{\mathbf{A}}^{31} & \tilde{\mathbf{A}}^{32} & \tilde{\mathbf{A}}^{33} \end{array} \right] = \left[\begin{array}{c|c} \mathbf{B}^{11} & \mathbf{B}^{12} \\ \mathbf{B}^{21} & \mathbf{B}^{22} \end{array} \right] = \mathbf{B}_n. \quad (49)$$

Now let

$$\tilde{\mathbf{V}}_n = \begin{bmatrix} \mathbf{I}_{m_1+m_2} & \mathbf{R}_1 \\ & \mathbf{I}_{m_3} \end{bmatrix} \iff \tilde{\mathbf{V}}_n^{-1} = \begin{bmatrix} \mathbf{I}_{m_1+m_2} & -\mathbf{R}_1 \\ & \mathbf{I}_{m_3} \end{bmatrix}. \quad (50)$$

Then we obtain

$$\tilde{\mathbf{V}}_n^{-1} \mathbf{B}_n \tilde{\mathbf{V}}_n = \tilde{\mathbf{B}}_n = \begin{bmatrix} \tilde{\mathbf{B}}^{11} & \\ \tilde{\mathbf{B}}^{21} & \tilde{\mathbf{B}}^{22} \end{bmatrix} \quad (51)$$

if \mathbf{R}_1 solves the algebraic Riccati equation

$$\mathbf{B}^{11} \mathbf{R}_1 - \mathbf{R}_1 \mathbf{B}^{22} = \mathbf{R}_1 \mathbf{B}^{21} \mathbf{R}_1 - \mathbf{B}^{12}. \quad (52)$$

Theorem 2 (Kreiss et al. [11]). *Let $\mathbf{B}^{ij} \in \mathbb{R}^{m_i \times m_j}$ for $i, j = 1, 2$. Also let $\mathbf{R}^{(0)} = \mathbf{O}^{12}$ where \mathbf{O}^{12} is the $m_1 \times m_2$ zero matrix. If $\|\mathbf{B}^{12}\|$ and $\|\mathbf{B}^{21}\|$ are sufficiently small, then the iteration*

$$\mathbf{B}^{11}\mathbf{R}^{(k)} - \mathbf{R}^{(k)}\mathbf{B}^{22} = \mathbf{R}^{(k-1)}\mathbf{B}^{21}\mathbf{R}^{(k-1)} - \mathbf{B}^{12}, \quad k = 1, 2, \dots \quad (53)$$

converges to a locally unique solution of

$$\mathbf{B}^{11}\mathbf{R} - \mathbf{R}\mathbf{B}^{22} = \mathbf{R}\mathbf{B}^{21}\mathbf{R} - \mathbf{B}^{12}. \quad (54)$$

Theorem 2 states that under certain conditions the algebraic Riccati equation (52) can be solved by solving a series of Sylvester equations. We continue by letting

$$\widehat{\mathbf{V}}_n = \begin{bmatrix} \mathbf{I}_{m_1+m_2} & \\ \mathbf{S}_1 & \mathbf{I}_{m_3} \end{bmatrix} \iff \widehat{\mathbf{V}}_n^{-1} = \begin{bmatrix} \mathbf{I}_{m_1+m_2} & \\ -\mathbf{S}_1 & \mathbf{I}_{m_3} \end{bmatrix}. \quad (55)$$

Then we find that

$$\widehat{\mathbf{V}}_n^{-1}\widetilde{\mathbf{B}}_n\widehat{\mathbf{V}}_n = \widehat{\mathbf{B}}_n = \begin{bmatrix} \widehat{\mathbf{B}}^{11} & \\ & \widehat{\mathbf{B}}^{22} \end{bmatrix} \quad (56)$$

provided that \mathbf{S}_1 solves the Sylvester equation

$$\widetilde{\mathbf{B}}^{22}\mathbf{S}_1 - \mathbf{S}_1\widetilde{\mathbf{B}}^{11} = -\widetilde{\mathbf{B}}^{21}. \quad (57)$$

From (49), (51), and (56) we find that

$$\widehat{\mathbf{V}}_n^{-1}\widetilde{\mathbf{V}}_n^{-1}\widetilde{\mathbf{\Lambda}}_n\widetilde{\mathbf{V}}_n\widehat{\mathbf{V}}_n = \widehat{\mathbf{\Lambda}}_n = \begin{bmatrix} \widehat{\mathbf{\Lambda}}^{11} & \widehat{\mathbf{\Lambda}}^{12} & \\ \widehat{\mathbf{\Lambda}}^{21} & \widehat{\mathbf{\Lambda}}^{22} & \\ & & \widehat{\mathbf{\Lambda}}^{33} \end{bmatrix} \quad (58)$$

We now need to zero out the remaining off-diagonal blocks. We can let

$$\bar{\mathbf{V}}_n = \begin{bmatrix} \mathbf{I}_{m_1} & \mathbf{R}_2 & \\ & \mathbf{I}_{m_2} & \\ & & \mathbf{I}_{m_3} \end{bmatrix} \iff \bar{\mathbf{V}}_n^{-1} = \begin{bmatrix} \mathbf{I}_{m_1} & -\mathbf{R}_2 & \\ & \mathbf{I}_{m_2} & \\ & & \mathbf{I}_{m_3} \end{bmatrix} \quad (59)$$

so that

$$\bar{\mathbf{V}}_n^{-1}\widehat{\mathbf{\Lambda}}_n\bar{\mathbf{V}}_n = \bar{\mathbf{\Lambda}}_n = \begin{bmatrix} \bar{\mathbf{\Lambda}}^{11} & & \\ \bar{\mathbf{\Lambda}}^{21} & \bar{\mathbf{\Lambda}}^{22} & \\ & & \bar{\mathbf{\Lambda}}^{33} \end{bmatrix} \quad (60)$$

provided that \mathbf{R}_2 solves the algebraic Riccati equation

$$\bar{\mathbf{\Lambda}}^{11}\mathbf{R}_2 - \mathbf{R}_2\bar{\mathbf{\Lambda}}^{22} = \mathbf{R}_2\bar{\mathbf{\Lambda}}^{21}\mathbf{R}_2 - \bar{\mathbf{\Lambda}}^{12}. \quad (61)$$

Similar to before, we can now let

$$\check{\mathbf{V}}_n = \begin{bmatrix} \mathbf{I}_{m_1} & & \\ \mathbf{S}_2 & \mathbf{I}_{m_2} & \\ & & \mathbf{I}_{m_3} \end{bmatrix} \iff \check{\mathbf{V}}_n^{-1} = \begin{bmatrix} \mathbf{I}_{m_1} & & \\ -\mathbf{S}_2 & \mathbf{I}_{m_2} & \\ & & \mathbf{I}_{m_3} \end{bmatrix}. \quad (62)$$

With this, we obtain

$$\check{\mathbf{V}}_n^{-1} \bar{\mathbf{\Lambda}}_n \check{\mathbf{V}}_n = \mathbf{\Lambda}_n = \begin{bmatrix} \mathbf{\Lambda}^{11} & & \\ & \mathbf{\Lambda}^{22} & \\ & & \mathbf{\Lambda}^{33} \end{bmatrix} \quad (63)$$

if \mathbf{S}_2 solves the Sylvester equation

$$\bar{\mathbf{\Lambda}}^{22} \mathbf{S}_2 - \mathbf{S}_2 \bar{\mathbf{\Lambda}}^{11} = -\bar{\mathbf{\Lambda}}^{21}. \quad (64)$$

Finally, we find using (48), (58), (60), and (63) that

$$\mathbf{V}_n = \mathbf{V}_{n-1} \check{\mathbf{V}}_n \hat{\mathbf{V}}_n \bar{\mathbf{V}}_n \check{\mathbf{V}}_n, \quad (65)$$

gives the necessary similarity transformation for (29).

There are two ways that the Riccati method can fail. The first is if the off-diagonal blocks of (48) are too large, so Theorem 2 does not apply. To fix this, we can simply decrease Δx_n until the off-diagonal blocks are small enough. Typically, we know that Δx_n is too large if too many iterations in (53) are used. The second way the Riccati method can fail is if the block structure of $\mathbf{A}(x)$ changes from x_{n-1} to x_n . That is, if one or more eigenvalues of $\mathbf{A}(x)$ go from being small to large or vice versa, then the size of the submatrices $\mathbf{\Lambda}^{11}$, $\mathbf{\Lambda}^{22}$, and $\mathbf{\Lambda}^{33}$ (i.e. their dimensions m_1 , m_2 , and m_3) change. When this occurs, we simply resort to the Schur method.

4.4. Implementing the Numerical Solution

Before stating the algorithm, a few remarks should be made. First, in Theorem 1, we need to choose constants K_i for $i = 1, 2$ and the degree of smoothness p . For choosing K_i , we simply need $K_i \Delta x \ll 1$, so the choice of K_i is problem dependent. Regardless, the smaller K_i is, the more points the algorithm will use. As for p , it has been found to be sufficient to let $p = 1$. This way, when the derivatives needed in Theorem 1 are calculated, we can use a first order finite difference approximation with only two points. The procedure is summarized in Algorithm 1.

A few more remarks about Algorithm 1 are in order. First, the number of points $N + 1$ changes as the algorithm proceeds. Also, this algorithm leaves open the possibility that there will be an abrupt change in mesh spacing. That is, either $\Delta x_n / \Delta x_{n-1} \ll 1$ or $\Delta x_n / \Delta x_{n-1} \gg 1$. This can give spurious results in the numerical solution, but can be remedied by adding more points so that the ratio $\Delta x_n / \Delta x_{n-1} \in [\frac{1}{2}, 2]$. Lastly in step 9, we use the Schur method but alter how the blocking is done. That is, instead of setting the size of the blocks according to the size of the eigenvalues at x_{n-1} , we set the size of the blocks according to the size of the eigenvalues at x_n . It should also be pointed out that a good approximation to the eigenvalues of $\mathbf{D}(x)$ are its diagonal elements since it is essentially diagonally dominant.

5. Numerical Example

To see how well this numerical solution performs, we need to find a problem for which an exact solution exists. It turns out that if there only exists one species in the atmosphere, then an exact solution is indeed possible. Therefore, for this section we will

Algorithm 1 Implementation of the Upwind Method

- 1: choose a preliminary mesh $a = x_0 < x_1 < x_2 < \dots < x_{N-1} < x_N = b$ that satisfies (28)
 - 2: use the Schur method to find $\mathbf{V}(x_0)$ and $\mathbf{\Lambda}(x_0)$ and calculate $\mathbf{D}(x_0)$
 - 3: set $n = 1$
 - 4: **while** $x_n \neq b$ **do**
 - 5: use the Riccati method to find $\mathbf{V}(x_n)$ and $\mathbf{\Lambda}(x_n)$
 - 6: **if** Riccati iteration (53) does not converge **then**
 - 7: replace Δx_n with $\Delta x_n/\sqrt{2}$, update x_n , and go to step 5
 - 8: **else if** block structure of $\mathbf{\Lambda}(x_n)$ and $\mathbf{\Lambda}(x_{n-1})$ differs **then**
 - 9: use the Schur method to replace $\mathbf{V}(x_{n-1})$ and $\mathbf{\Lambda}(x_{n-1})$ with block structure forced to be the same as $\mathbf{\Lambda}(x_n)$
 - 10: **end if**
 - 11: calculate $\mathbf{D}(x_n)$
 - 12: check smoothness with (26) and (27)
 - 13: **if** not smooth enough **then**
 - 14: replace Δx_n with $\Delta x_n/\sqrt{2}$, update x_n , and go to step 5
 - 15: **end if**
 - 16: accept Δx_n and adjust mesh accordingly
 - 17: calculate finite difference matrices (32) and store
 - 18: replace n with $n + 1$
 - 19: **end while**
 - 20: assemble finite difference matrices (32) and boundary conditions (18) into a linear system
 - 21: solve system to find solution $\mathbf{u}_n \approx \mathbf{y}(x_n)$ for $n = 0, 1, \dots, N$
-

assume all atoms and molecules in the atmosphere are atomic oxygen. We realize that this assumption is far from reality, but it will allow us to find an exact representation of the solution and see how well Algorithm 1 performs.

Let us begin by rewriting (16) using the above simplification. We get

$$\begin{aligned} \mu \frac{\partial I(z, E, \mu)}{\partial z} = n(z) & \left(- [\sigma^{\text{tot}}(E) - f(E)\sigma^{\text{el}}(E)] I(z, E, \mu) \right. \\ & + [1 - f(E)]\sigma^{\text{el}}(E) \sum_{m=0}^{M-1} \frac{2m+1}{2} \chi_m^*(E) P_m(\mu) \int_{-1}^1 P_m(\mu') I(z, E, \mu') d\mu' \\ & \left. + q(z, E, \mu) \right) \end{aligned} \quad (66)$$

where $q(z, E, \mu)$ is the sum of all other terms in (16). We can write it this way because $q(z, E, \mu)$ only contains terms that depend on electron intensities at higher energies, which can be assumed to be known. Let us define the scattering depth to be

$$\tau = \frac{\sigma^{\text{tot}} - f\sigma^{\text{el}}}{\sin \alpha} \int_z^{z_{\text{top}}} n(z') dz' \quad (67)$$

where we have dropped the E dependency. If $\hat{I}(\tau, \mu) = I(z, \mu)$, then our equation becomes

$$\mu \frac{\partial \hat{I}(\tau, \mu)}{\partial \tau} = \hat{I}(\tau, \mu) - \frac{c}{2} \sum_{m=0}^{M-1} (2m+1) \chi_m^* P_m(\mu) \int_{-1}^1 P_m(\mu') \hat{I}(\tau, \mu') d\mu' + \hat{q}(\tau, \mu) \quad (68)$$

where

$$c = \frac{(1-f)\sigma^{\text{el}}}{\sigma^{\text{tot}} - f\sigma^{\text{el}}}, \quad (69)$$

$$\hat{q}(\tau, \mu) = \frac{-q(z, \mu)}{\sigma^{\text{tot}} - f\sigma^{\text{el}}}. \quad (70)$$

Under this change of variables, the boundary conditions become

$$\hat{I}(0, \mu) = I_{\text{top}}(\mu) \text{ for } \mu < 0, \quad \hat{I}(\tau_{\text{max}}, \mu) = 0 \text{ for } \mu > 0. \quad (71)$$

Notice that the top of the atmosphere is at $\tau = 0$ and the bottom of the atmosphere is at $\tau = \tau_{\text{max}}$. Hence, scattering depth (67) is a dimensionless measure of how far an electron has penetrated into the atmosphere.

The full details of how to find the exact solution to (68) are given in Woods [12]. The idea originally comes from Case and Zweifel [33]. We have simply applied those ideas to the electron transport equation. From Woods [12], the exact solution can be found using the boundary element method and is given by

$$\begin{aligned} I(\tau, \mu) = & \int_0^{\tau_{\text{max}}} \int_{-1}^1 F(\tau, \mu; \tau_0, \mu_0) q(\tau_0, \mu_0) d\mu_0 d\tau_0 \\ & + \int_{-1}^0 \mu_0 [F(\tau, \mu; 0, \mu_0) \quad -F(\tau, \mu; \tau_{\text{max}}, \mu_0)] \begin{bmatrix} I_{\text{top}}(\mu_0) \\ I(\tau_{\text{max}}, \mu_0) \end{bmatrix} d\mu_0 \\ & + \int_0^1 \mu_0 [F(\tau, \mu; 0, \mu_0) \quad -F(\tau, \mu; \tau_{\text{max}}, \mu_0)] \begin{bmatrix} I(0, \mu_0) \\ 0 \end{bmatrix} d\mu_0 \end{aligned} \quad (72)$$

where we have dropped the hat notation and $F(\tau, \mu; \tau_0, \mu_0)$ is a fundamental solution.

In order to compare the two solutions, we will use a boundary condition given in Strickland et al. [34]. It is given by

$$I_{\text{top}}(E, \mu) = Q_0 \left(\frac{E}{2\pi E_0^3} e^{-E/E_0} + B \frac{E_0}{E} e^{-E/b} \right) \quad (73)$$

where

$$B = \frac{0.4}{2\pi E_0^2} e^{-1}, \quad (74)$$

$$b = \begin{cases} 0.8E_0, & E_0 < 500 \text{ eV} \\ 0.1E_0 + 350, & E_0 \geq 500 \text{ eV} \end{cases}. \quad (75)$$

Here, E_0 is the characteristic energy and Q_0 is related to the energy flux Q by

$$Q = Q_0 \left(\frac{E_0 + 0.2be^{-1}}{E_0} \right). \quad (76)$$

For a typical aurora, the energy flux is between 0 and 50 erg cm⁻² s⁻¹ (see Arnoldy and Lewis [35] and Solomon et al. [36]). For this reason, we will use $Q = 20, 50$ erg cm⁻² s⁻¹ for each test included here. We will also use characteristic energies of $E_0 =$

5000, 8000, 10000 eV for each energy flux. The maximum difference between the upwind solution and the boundary element solution is given in Table 3 for electron intensities above 65 km and at 10 eV. Although we have chosen $z_{\text{bot}} = 50$ km, we only report the largest relative difference for altitudes above 65 km because below this altitude the relative differences are inflated due to round-off error. For instance, for $E_0 = 10000$ and $Q = 50$, the maximum relative difference is about 13.8. However, this occurs at $z = 55.9$ km where the calculated upwind method intensity is about 1.56×10^{-16} and the boundary element intensity is about 2.31×10^{-15} . The relative difference is somewhat large, but the two intensities are the same for computing purposes. Above 65 km, the intensities are greater than 10^{-4} so that round-off does not present any issues.

Table 3: The maximum relative difference between the upwind and boundary element solution for $z \geq 65$ km at $E = 10$ eV for each combination of E_0 and Q .

Characteristic Energy (eV)	Energy Flux (erg cm ⁻² s ⁻¹)	Maximum Relative Difference
5000	20	1.334×10^{-2}
5000	50	6.066×10^{-3}
8000	20	1.299×10^{-2}
8000	50	5.496×10^{-3}
10000	20	1.993×10^{-3}
10000	50	1.770×10^{-3}

Clearly, Table 3 shows that the upwind method and the boundary element method give close to the same solution. The maximum relative differences are on the order of 10^{-2} , which means that they agree with each other to at least the first two significant digits. Most of the maximum relative differences are on the order of 10^{-3} , which gives agreement to at least three significant digits. This demonstrates that the upwind method is correctly solving the single species problem. The boundary element method solution (72) is exact, and the only error is in the numerical evaluation of its integrals. The upwind method is a finite difference solution, so it is not exact, but its solution has now been verified.

6. Conclusion

We have demonstrated that the electron transport equation (1) is a very stiff boundary value problem whose numerical solution cannot be found without a stiff solver unless unrealistic assumptions are made for the solution domain. To handle the exponential stiffness of the problem, the algorithm derived here incorporates analytic approximations involving a generalized Legendre polynomial approximation, along with an upwind scheme that uses an eigenvalue decomposition to delineate the relevant scales needed to resolve the solution. To demonstrate its effectiveness, in the case of a single species atmosphere, the exact solution can be written in terms of a fundamental solution, and then evaluated using a boundary element method. We showed that the two methods agree to at least two significant digits for all altitudes above 65 km. There is not a simple test for the full multi-species problem, and how this might be done is the subject of future work.

References

- [1] M. H. Rees, *Physics and Chemistry of the Upper Atmosphere*, Cambridge Atmospheric and Space Science, Cambridge Univ. Press, New York, NY, 1989.
- [2] K. Stamnes, O. Lie-Svendsen, M. H. Rees, The Linear Boltzmann Equation in Slab Geometry: Development and Verification of a Reliable and Efficient Solution, *Planet. Space Sci.* 39 (10) (1991) 1435–1463.
- [3] K. Stamnes, Analytic Approach to Auroral Electron Transport and Energy Degradation, *Planet. Space Sci.* 28 (4) (1980) 427–441.
- [4] D. Lummerzheim, M. H. Rees, H. R. Anderson, Angular Dependent Transport of Auroral Electrons in the Upper Atmosphere, *Planet. Space Sci.* 37 (1) (1989) 109–129.
- [5] Q. L. Min, D. Lummerzheim, M. H. Rees, K. Stamnes, Effects of a Parallel Electric Field and the Geomagnetic Field in the Topside Ionosphere on Auroral and Photoelectron Energy Distributions, *J. Geophys. Res.* 98 (11) (1993) 19223–19234.
- [6] D. Lummerzheim, J. Liliensten, Electron Transport and Energy Degradation in the Ionosphere: Evaluation of the Numerical Solution, Comparison with Laboratory Experiments and Auroral Observations, *Ann. Geophys.* 12 (10/11) (1994) 1039–1051.
- [7] B. Lanchester, B. Gustavsson, Imaging of Aurora to Estimate the Energy and Flux of Electron Precipitation, in *Auroral Phenomenology and Magnetospheric Processes: Earth And Other Planets*, vol. 197 of *Geophysical Monograph Series*, Amer. Geophys. Union, Washington, DC, 2012.
- [8] H. S. Porter, F. Varosi, H. G. Mayr, Iterative Solution of the Multistream Electron Transport Equation. I: Comparison with Laboratory Beam Injection Experiments, *J. Geophys. Res.* 92 (A6) (1987) 5933–5959.
- [9] G. P. Mantas, Theory of Photoelectron Thermalization and Transport in the Ionosphere, *Planet. Space Sci.* 23 (2) (1975) 337–354.
- [10] G. P. Mantas, S. A. Bowhill, Calculated Photoelectron Pitch Angle and Energy Spectra, *Planet. Space Sci.* 23 (2) (1975) 355–375.
- [11] H. O. Kreiss, N. K. Nichols, D. L. Brown, Numerical Methods for Stiff Two-Point Boundary Value Problems, *SIAM J. Numer. Anal.* 23 (2) (1986) 325–368.
- [12] M. Woods, Numerical Solution of the Electron Transport Equation, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 2016.
- [13] A. E. Hedin, Extension of the MSIS Thermospheric Model Into the Middle and Lower Atmosphere, *J. Geophys. Res.* 96 (A2) (1991) 1159–1172.
- [14] R. D. Sharma, J. H. Brown, A. Berk, P. K. Acharya, J. Gruninger, J. W. Duff, R. L. Sundberg, User’s Manual for SAMM, SHARC and MODTRAN Merged, Tech. Rep. PL-TR-96-2090, Phillips Lab., Hanscom AFB, MA, 1996.
- [15] R. R. Laher, F. R. Gilmore, Updated Excitation and Ionization Cross Sections for Electron Impact on Atomic Oxygen, *J. Phys. Chem. Ref. Data* 19 (1) (1990) 277–305.
- [16] Y. Itikawa, A. Ichimura, Cross Sections for Collisions of Electrons and Photons with Atomic Oxygen, *J. Phys. Chem. Ref. Data* 19 (3) (1990) 637–651.
- [17] F. Salvat, A. Jablonski, C. J. Powell, ELSEPA - Dirac Partial-Wave Calculation of Elastic Scattering of Electrons and Positrons by Atoms, Positive Ions and Molecules, *Comput. Phys. Commun.* 165 (2) (2005) 157–190.
- [18] S. Pancheshnyi, S. Biagi, M. C. Bordage, G. J. M. Hagelaar, W. L. Morgan, A. V. Phelps, L. C. Pitchford, The LXCat Project: Electron Scattering Cross Sections and Swarm Parameters for Low Temperature Plasma Modeling, *Chem. Phys.* 398 (4) (2012) 148–153.
- [19] W. L. Morgan, Electron Collision Data for Plasma Chemistry Modeling, *Adv. At., Mol., Opt. Phys.* 43 (1) (2000) 79–110.
- [20] Y. Itikawa, Cross Sections for Electron Collisions with Nitrogen Molecules, *J. Phys. Chem. Ref. Data* 35 (1) (2006) 31–53.
- [21] A. V. Phelps, Collision Cross Sections for Electrons with Atmospheric Species, *Ann. Géophys.* 28 (3) (1972) 611–625.
- [22] S. F. Biagi, Monte Carlo Simulation of Electron Drift and Diffusion in Counting Gases Under the Influence of Electric and Magnetic Fields, *Nucl. Instr. Meth. Phys. Res. A* 421 (2) (1999) 234–240.
- [23] L. L. Alves, The IST-LISBON Database on LXCat, *JPCS* 565 (1) (2014) 1–10.
- [24] Y. Itikawa, Cross Sections for Electron Collisions with Oxygen Molecules, *J. Phys. Chem. Ref. Data* 38 (1) (2009) 1–20.
- [25] A. A. Ionin, I. V. Kochetov, A. P. Napartovich, N. N. Yuryshev, *Physics and Engineering of Singlet*

- Delta Oxygen Production in Low-Temperature Plasma, *J. Phys. D: Appl. Phys.* 40 (2) (2007) R25–R61.
- [26] Y. K. Kim, K. K. Irikura, M. E. Rudd, M. A. Ali, P. M. Stone, J. Chang, J. S. Coursey, R. A. Dragoset, A. R. Kishore, K. J. Olsen, A. M. Sansonetti, G. G. Wiersma, D. S. Zucker, M. A. Zucker, Electron-Impact Cross Sections for Ionization and Excitation, Version 3.0, SRD 107, Nat. Inst. of Standards and Technol., 2004.
- [27] M. Hayashi, Nonequilibrium Processes in Partially Ionized Gases, vol. 220 of *NATO ASI Series*, chap. 21, Springer Springer+Business Media, LLC, New York, NY, 333–340, 1990.
- [28] D. C. Cartwright, M. J. Brunger, L. Campbell, B. Mojarrabi, P. J. O. Teubner, Electron Impact Excitation of Nitric Oxide Under Auroral Conditions, *Geophys. Res. Lett.* 25 (9) (1998) 1495–1498.
- [29] W. J. Wiscombe, The Delta-M Method: Rapid Yet Accurate Radiative Flux Calculations for Strongly Asymmetric Phase Functions, *J. Atmos. Sci.* 34 (9) (1977) 1408–1422.
- [30] U. M. Ascher, R. M. Mattheij, R. D. Russell, Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, Soc. Ind. and Appl. Math., Philadelphia, PA, 1995.
- [31] U. M. Ascher, L. R. Petzold, Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations, Soc. Ind. and Appl. Math., Philadelphia, PA, 1998.
- [32] G. H. Golub, S. Nash, C. Van Loan, A Hessenberg-Schur Method for the Problem $AX + XB = C$, *IEEE Trans. Autom. Control* AC-24 (6) (1979) 909–913.
- [33] K. M. Case, P. F. Zweifel, Linear Transport Theory, Addison-Wesley Publishing Company, Inc., Reading, MA, 1967.
- [34] D. J. Strickland, R. E. J. Daniell, J. R. Jasperse, B. Basu, Transport-Theoretic Model for the Electron-Proton-Hydrogen Atom Aurora. II: Model Results, *J. Geophys. Res.* 98 (A12) (1993) 21533–21548.
- [35] R. L. Arnoldy, P. B. Lewis, Correlation of Ground-Based and Topside Photometric Observations with Auroral Electron Spectra Measurements and Rocket Altitudes, *J. Geophys. Res.* 82 (35) (1977) 5563–5572.
- [36] S. C. Solomon, P. B. Hays, V. J. Abreu, The Auroral 6300 Å Emission: Observations and Modeling, *J. Geophys. Res.* 93 (A9) (1988) 9867–9882.