Recent Advances in the Spherical Harmonics Method

Alan H. Karp
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304
karp@hpl.hp.com

May 12, 1997

Abstract

In his classic work on radiative transfer, Chandrasekhar introduced the discrete ordinates method for solving the integro-differential equation of radiative transfer; the spherical harmonics method is a spectral analog of this method. Not surprisingly, there is a close connection between the spherical harmonics and discrete ordinates methods.

The spherical harmonics method raises some interesting questions. How are the boundary conditions defined? How can we evaluate the amount of radiation moving in specific directions if all we know are the moments? What solution method is best? How can the method be applied to more complex models and in more complex geometries? This paper reviews new work on the spherical harmonics method since around 1980 as it applies to these questions.

1 Introduction

We would like to solve the equation of radiative transport for a general, three dimensional scattering medium such as the one illustrated in the left half of Figure 1. The governing equation is

$$\frac{\partial \mathbf{I}}{\partial t} - \frac{1}{\kappa}(\mathbf{\Omega} \cdot \nabla)\mathbf{I} + \mathbf{I} = \mathbf{S} \tag{1}$$

where $\mathbf{I} = \mathbf{I}(t, \mathbf{x}, \mathbf{\Omega})$ is a 4-element vector of the Stokes parameters of the specific intensity, the amount of radiation traveling at a point in space, traveling in a certain direction, at a specific time. Here, κ is the sum of the volume scattering and absorbtion coefficients, and S is the set of sources.

Since this problem is too difficult to be solved, we simplify the situation as illustrated in the right half of Figure 1 with the governing equation

$$\mu \frac{dI}{d\tau} = I - \frac{\omega}{2} \int_{-1}^{1} P(\mu, \mu') I(\tau, \mu') d\mu' + S, \tag{2}$$

where $d\tau = \kappa dz$, μ is the cosine of the zenith angle, ω is the fraction of the radiation that is scattered on each interaction, and $P(\mu, \mu')$ describes how the radiation changes direction when it is scattered. The boundary conditions specify the angular distribution of the radiation incident on both sides.

Even this much simpler problem is hard to solve. Figure 2 shows the scattering phase function for two types of particles and illustrates one of the problems. Representing a function with such a sharp peak (note the logarithmic scale on the ordinate) while retaining the detail in the backscatter direction is difficult.

Chandrasekhar[1] proposed replacing the integral in Equation (2) with a quadrature sum and solving the resulting system of equations for the specific intensity at a number of discrete ordinates, the Discrete Ordinates Method (DOM).

The spherical harmonics method (SHM) is the spectral analogue of the discrete ordinates method. Much in the way we use sines and cosines to represent functions with Fourier transforms, we use spherical harmonics (Legendre polynomials in their simplest form) to represent the specific intensity and scattering phase function. Just as the orthogonality of sines and cosines makes the Fourier transform useful for solving differential equations, the orthogonality of the spherical harmonics makes it easier to solve Equation (2).

This paper will examine what's been going on with the SHM in the recent past. In preparing this review, I did an on-line search of 15 databases in July 1996. Some databases went back as far as 1970, most only as far as 1980. This latter date is a convenient starting point for a talk on "recent" advances because it covers the development of the analytic and linear algebraic methods.

Since I was more interested in finding all relevant articles than in avoiding reading abstracts not relevant to this review, I used rather general search criteria. The initial search found 523 matches; a refinement of the search reduced the count to 470 articles. The topics covered, as determined from the abstracts, are summarized in Table 1. As can be seen, I still had 53 false hits, such as a paper dealing with a spherical harmonics expansion of convective heat flow in the Earth's core. Unfortunately, we can't infer that a large number of false hits guarantees that I didn't miss any relevant papers.

Not surprisingly, the SHM is used most in nuclear engineering and atmospheric sciences. The most interesting case is the use of the SHM in well logging, a means of analyzing the potential productivity of an oil well using sensors put down the bore hole. The second column shows that there is a lot of interest in models that go beyond the basic model of Equation (2). The third column shows that improvements are still being made in the method even for the basic

problem. Note that the numbers don't add up to 470 since a paper describing a 3 dimensional study of nuclear reactors in the P_1 approximation was counted three times.

This review will focus on the "method" in the SHM. Hence, nothing more will be said about specific applications. Furthermore, the papers that simply compare a method against a previously described version of the SHM won't be discussed.

I've also made the rather arbitrary decision to exclude those papers that describe the SHM only in specific orders; I have found generalizing the work in these papers to arbitrary order is not easy. This choice has the unfortunate consequence of eliminating almost all the papers dealing with complex, 3D systems. On the other hand, the excluded papers focus on aspects of the problem not related to the SHM.

Even with all these exclusions, this review encompasses some 30 papers. Hence, the discussion will be brief and often imprecise, leaving out significant details. The reader is referred to the cited literature for more complete information.

2 Spherical Harmonics Method

Long before Chandrasekhar proposed the discrete ordinates method, Jeans[2] wrote down what we now call the spherical harmonics method. The key idea is to expand $I(\tau, \mu)$ and $P(\mu, \mu')$ in orthogonal polynomials such that the integral in Equation (2) is replaced by an orthogonality condition. Write

$$I(\tau, \mu) = \sum_{n=0}^{n} f_n(\tau) P_n(\mu), \tag{3}$$

and

$$P(\mu, \mu') = \sum_{k=0}^{K} \beta_k P_k(\mu) P_k(\mu'), \tag{4}$$

where the β_k are known quantities that depend on the type of scattering particle and $P_k(\mu)$ is the Legendre polynomial that satisfies the standard recurrence relations and orthogonality condition[3]. Substituting Equations (3) and (4) into Equation (2) gives us the system of equations

$$\frac{k+1}{2k+1}\frac{df_{k+1}}{d\tau} + \frac{k}{2k+1}\frac{df_{k-1}}{d\tau} + \left(\frac{\omega\beta_k}{2k+1} - 1\right) f_k = s_k\,. \tag{5}$$

We can write Equation (5) in matrix notation as

$$A\frac{d\mathbf{f}}{d\tau} + C\mathbf{f} = \mathbf{s}, \tag{6}$$

where the components of the vector $\mathbf{f}$ are the f_k, A is a tridiagonal matrix with zeros on the diagonal, and C is a diagonal matrix. The solution to the homogeneous problem can be written as

$$\mathbf{f} = e^{-A^{-1}C\tau}\mathbf{f}_0. \tag{7}$$

Although in general it is difficult to compute a matrix exponential[4], knowledge of the physical problem allows a numerically accurate solution to be obtained[5].

The vector $\mathbf{f}_0$ is determined by the boundary conditions. Recall that the boundary conditions for Equation (2) are expressed in terms of the angular distribution of the incoming radiation. Equation (6) doesn't refer to the specific intensity, so we have to make a choice. If we compute values of $\mathbf{f}_0$ at the boundary that force the intensity to match at specific values of μ, we have the Mark boundary conditions. Alternatively, and more in the spirit of a spectral method, we can force moments of the intensity to match those of the boundary condition. This latter formulation gives us the Marshak conditions. The Mark and Marshak conditions are the most widely used, but others have been proposed. Numerical experiments show little difference among the various options once the expansion includes more than about 20 terms. Regardless of the formulation we choose, we have a two-point boundary problem so that some, typically half, of the moments are determined at each boundary.

There are two kinds of boundary conditions that are normally treated separately. When the Earth's atmosphere is illuminated by the sun, the incoming radiation is unidirectional. If the scattering phase function is highly peaked, the intensity distribution is also highly peaked. Since it is difficult to represent such a function with a polynomial, we usually solve for this component of the radiation field separately, replacing the δ-function angular distribution with the source term

$$s_{k,\mathrm{s}}(\tau) = -\frac{1}{2}Fe^{-\tau/\mu_0}P_k(-\mu_0)\frac{\beta_k}{2k+1}, \tag{8}$$

where F is the magnitude of the solar radiation. The remaining part of the solution is much smoother. Treating the solar radiation this way often lets us use fewer terms in the expansion than we need to approximate the phase function.

The second kind of boundary condition that is treated specially is the ground. We usually assume a Lambert surface which scatters radiation isotropically. The corresponding source term is

$$s_{k,g}(\tau) = \frac{-R_g\beta_k}{2k+1}\int_0^1 e^{-(\tau_0-\tau)/\mu}P_k(\mu)d\mu, \tag{9}$$

where R_g is the reflectivity of the ground. Later, we'll return to the integral in Equation (9), which is hard to compute accurately.

Say that we've computed the f_k by some method. What now? In some situation, we're done since certain moments have physical meaning. For example, f_0 is proportional to the mean intensity, f_1 is the net radiative flux, and

f_2 is proportional to the radiation pressure. In applications where the radiation field is needed only as part of an energy flow calculation, such as in stellar atmospheres or climatology studies, these few moments suffice.

What if we want the intensity at a predetermined set of zenith angles? If we evaluate Equation (3), we get a plot like the wavy line Figure 3. A number of methods have been proposed to evaluate $I(\tau, \mu)$ more accurately at all μ[6], but only the integration of the source function[7] is accurate in all situations. This approach simply solves Equation (2) with a known right hand side. Namely,

$$I(\tau, \mu > 0) = \frac{1}{2} \sum_{k=0}^{K} \beta_k P_k(\mu) f_k \int_{\tau}^{\tau_0} e^{-(t-\tau)/\mu} \frac{dt}{\mu}, \tag{10}$$

with a similar equation for $\mu < 0$. For vertically inhomogeneous atmospheres approximated by piecewise constant layers, the integration is done separately over each layer. The solid line in Figure 3 was computed using Equation (10). The dots in that figure are at the Gauss quadrature points, illustrating the close connection between the SHM and DOM.

Thus far, the description of the SHM has looked only at the azimuthally independent part of the solution. Fortunately, taking the Fourier transform of the azimuthally dependent equation gives us a set of independent equations to be solved. We need only replace the $P_k(\mu)$ in Equations (3) and (4) with the spherical harmonics $Y_k^m(\mu)$ to get equations differing only in some constant terms from the ones described thus far.

3 Methods

Numerical work in solving the spherical harmonics equations largely started in the 1950's. This section describes some of the more recent solution schemes proposed since 1980.

3.1 Linear Algebra Approach

One approach is to use the rules of linear algebra to compute the matrix exponential of Equation (7)[5]. We note that the matrices X and X^{-1} have the same eigenvectors and that the eigenvalues of one are the inverse of the eigenvalues of the other. Hence, we can work with the matrix $C^{-1}A$ from Equation (7) which has some nice properties. In particular, this matrix is tridiagonal with zeros on the diagonal, so we know that its eigenvalues come in $\pm$ pairs.

If we symmetrize $C^{-1}A$ and do an odd-even sort of the rows and columns of the resulting matrix, which just happens to be a similarity transform, we get a matrix with the structure

$$\begin{bmatrix} 0 & B_h \\ B_h^T & 0 \end{bmatrix}, \tag{11}$$

where B_h is bidiagonal. We can now do a singular value decomposition of B_h which allows us to construct the eigenvalues and eigenvectors of $C^{-1}A$ in only $\mathcal{O}(K^3/8)$ operations. Once we know the eigen-decomposition of the matrix, we can write the matrix exponential as $e^{-A^{-1}C\tau} = U^{-1}e^{-\Lambda\tau}U$, where U is the matrix of eigenvectors, and Λ is a diagonal matrix of the eigenvalues.

One problem with all solutions of Equation (2) arises when the gas is conservative, *i.e.*, when $\omega = 1$. Now, the (1,1) element of $C = 0$ leading to a repeated pair of eigenvalues at 0 and making the matrix $A^{-1}C$ defective. An effective approach is to split off the two equations represented by the zero eigenvalues and solve the remaining system as before.

In the linear algebra approach, we instead reduce the matrix to Jordan canonical form, $J = V^{-1}C^{-1}AV$, where V replaces the missing eigenvector with a first order principle vector. Now, the matrix exponential becomes

$$e^{-J\tau} = \begin{bmatrix} 0 & \tau & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & e^{-\lambda_2\tau} & 0 \\ 0 & 0 & 0 & e^{\lambda_2\tau} \end{bmatrix}. \tag{12}$$

A more serious problem is when $1 - \omega = \epsilon$ is very small. Mathematically, the matrix $C^{-1}A$ is diagonalizable, but the eigenvector matrix becomes poorly conditioned. The solution[5] is to reduce the matrix to quasi-Jordan form

$$\exp\left\{-\begin{bmatrix} \lambda_0 & a \\ 0 & -\lambda_0 \end{bmatrix}\tau\right\} = \begin{bmatrix} e^{\lambda_0\tau} & a(e^{\lambda_0\tau} - e^{-\lambda_0\tau})/2\lambda_0 \\ 0 & e^{-\lambda_0\tau} \end{bmatrix}\tau, \tag{13}$$

where only the first two rows and columns are shown for brevity. We can simply write down the missing eigenvector by noting the form of the eigenvectors of the diagonalized system.

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ \epsilon & -\epsilon & 0 & 0 \\ \mathcal{O}(\epsilon^2) & \mathcal{O}(\epsilon^2) & a & b \\ \mathcal{O}(\epsilon^2) & \mathcal{O}(\epsilon^2) & c & d \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & -\epsilon & 0 & 0 \\ \epsilon & 1 & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{bmatrix}, \tag{14}$$

where we ignore terms of $\mathcal{O}(\epsilon^2)$.

If we let J_+ be the part of the Jordan matrix containing the positive eigenvalues of $A^{-1}C$, we can write the system of equations to be solved as

$$\begin{bmatrix} G & I & 0 & 0 \\ -\alpha e^{J_+\tau} & -\beta e^{J_+\tau} & \alpha e^{-J_+\tau} & \beta e^{-J_+\tau} \\ -\gamma e^{-J_+\tau} & -\delta e^{-J_+\tau} & \gamma e^{J_+\tau} & \delta e^{J_+\tau} \\ 0 & 0 & -G & I \end{bmatrix}, \tag{15}$$

where the matrix G comes from the boundary conditions, I is the identity matrix of order $K/2$, and the four matrices denoted by Greek symbols depend only on

the properties of the material. The extension to a set of homogeneous layers is obvious.

Since the growing exponentials in Equation (15) would cause problems for large τ, we *algebraically* scale by the matrix diag$[I \quad e^{-J_+\tau} \quad e^{-J_+\tau} \quad I]$. (I stress the word algebraic because some investigators have gone to great lengths to avoid this transformation.) Note that, after the scaling, the 2×2 blocks in the upper left and lower right decouple as $\tau \to \infty$. In fact, these blocks are closely related to the reflection matrices used in other methods.

3.2 Classical Method

A more standard approach, at least to mathematicians, is to guess a form for the solution and prove that it satisfies the equation of interest. In this case, we write the solution as[8]

$$I(\tau, \mu) = \sum_{k=0}^{K} \frac{2k+1}{2} P_k(\mu) \sum_{j=1}^{J} [A_j e^{-\lambda_j \tau} + (-1)^k B_j e^{-\lambda_j(\tau_0 - \tau)}] g_k(\lambda_j). \qquad (16)$$

Here A_j and B_j are constants determined by the boundary conditions and $g_k(\lambda)$ are the Chandrasekhar polynomials defined by the 3-term recurrence

$$(k+1) g_{k+1}(\lambda) = \frac{2k+1-\omega\beta_k}{\lambda} g_k(\lambda) - k g_{k-1}(\lambda). \qquad (17)$$

The λ_j in Equation (16) are the inverses of the positive roots of $g_{K+1}(\lambda) = 0$. Note that the $g_k(\lambda_j)$ are the eigenvectors referred to in Section 3.1.

We now substitute Equations (16) and (4) into Equation (2), ignoring the source term to get the solution to the homogeneous problem. We end up with the system of equations

$$\sum_{k=0}^{K} \sum_{j=1}^{J} M_{nk} [A_j + (-1)^k B_j e^{-\lambda_j \tau_0}] g_k(\lambda_j) = r_n(0), \qquad (18)$$

$$\sum_{k=0}^{K} \sum_{j=1}^{J} M_{nk} [(-1)^k A_j e^{-\lambda_j \tau_0} + B_j] g_k(\lambda_j) = r_n(\tau_0) \qquad (19)$$

to be solved for the unknowns A_j and B_j.

Next, we need a particular solution that incorporates the various source terms. Particular solutions have been given for solar illumination, diffuse and specular reflective boundaries, and thermal radiation[8, 9, 10], as well as for polarization problems[11] and spherical symmetry[12]. In the later works, these particular solutions take the form

$$I_p(\tau, \mu) = \sum_{k=0}^{K} \frac{2k+1}{2} P_k(\mu) \sum_{j=1}^{J} C_j \lambda_j [A_j(\tau) + (-1)^k B_j(\tau)] g_k(\lambda_j). \qquad (20)$$

The C_j are known, and the functions $A_j(\tau)$ and $B_j(\tau)$ are determined by the source terms.

3.3 Other Methods

Several other solution methods have been proposed since the linear algebra and classical approaches were published. Unfortunately, most of them do not achieve the accuracy and numerical stability achieved in the earlier work. Still, each has an interesting perspective worthy of attention.

Wells and Sidorowich[13] reexamined the closure condition needed for the system of equations (5). Most investigators simply set $f_{K+1} = 0$ to eliminate the extra unknown. Wells and Sidorowich instead use the 3-term recurrence for $P_k(\mu)$ to write Equation (5) as

$$(k+1)\left[\frac{df_{k+1}}{d\tau} + \frac{P_{k+1}}{\mu P_k}(\omega - \beta_k)f_k\right] + k\left[\frac{df_{k-1}}{d\tau} + \frac{P_{k-1}}{\mu P_k}(\omega - \beta_k)f_k\right] = 0. \quad (21)$$

Each of the two terms in Equation (21) is nearly zero in the direction of the sun. Using the second term in place of the equation for f_{K+1} closes the system. The method is accurate to slightly better than 1% in high order for sharply peaked phase functions. This paper shows that trying other closure conditions may lead to improved solutions in specific situations.

Kamiuto[14] formally integrates Equation (5) to get

$$(k+1)f_{k+1}(\xi) + kf_{k-1}(\xi) + \sigma_k \int_{-1}^{1} f_k(\xi)d\xi = r_k(\xi), \quad (22)$$

where the change of variables to $\xi = (2\tau - \tau_0)/\tau_0$ makes the range $[-1, 1]$. The problem is solved by expanding $f_k(\xi)$ in a Chebyshev series

$$f_k(\xi) = \sum_{n=0}^{N} b_{kn}T_n(\xi) \quad (23)$$

and solving the resulting system of $(N+1)(K+1)$ linear equations. Clearly, this method will become computationally inefficient for rapidly varying $f_k(\xi)$ which will happen when τ_0 is large, but results accurate to 3 significant digits were obtained for τ_0 as large as 200 with reasonable computational effort.

There is yet another way to rearrange Equation (5)[15]. Sort the vector $\mathbf{f}$ into vectors of the even components $\mathbf{f}_e$ and the odd components $\mathbf{f}_o$ to give the equations

$$A_e\frac{d\mathbf{f}_e}{d\tau} = C_o\mathbf{f}_o, \quad \text{and} \quad A_o\frac{d\mathbf{f}_o}{d\tau} = C_e\mathbf{f}_e, \quad (24)$$

where the coefficient matrices are simple rearrangements of the elements of the matrices in Equation (5). These two, first order equations can be combined into

the second order equation

$$S\frac{d^2\hat{\mathbf{f}}_e}{d\tau^2} = \hat{\mathbf{f}}_e. \tag{25}$$

The solution to Equation (25) for the moments at any depth is

$$\begin{bmatrix} \mathbf{g}_e(\tau) \\ \mathbf{g}_o(\tau) \end{bmatrix} = \begin{bmatrix} \frac{\cosh[\Lambda(\tau_0-\tau)]}{\cosh(\Lambda\tau_0)} & \frac{\Lambda\sinh(\Lambda\tau)}{\cosh(\Lambda\tau_0)} \\ \frac{\sinh[\Lambda(\tau_0-\tau)]}{\Lambda\cosh(\Lambda\tau_0)} & \frac{\cosh(\Lambda\tau)}{\cosh(\Lambda\tau_0)} \end{bmatrix} \begin{bmatrix} \mathbf{g}_e(0) \\ \mathbf{g}_o(\tau_0) \end{bmatrix}, \tag{26}$$

where

$$\begin{bmatrix} \mathbf{g}_e(\tau) \\ \mathbf{g}_o(\tau) \end{bmatrix} = \begin{bmatrix} V_e^{-1} & 0 \\ 0 & V_o^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{f}_e(\tau) \\ \mathbf{f}_o(\tau) \end{bmatrix}, \tag{27}$$

and $V_e = C_e^{-1/2}U_e$ with a similar term for V_o. Note that Equation (26) properly handles the limit as one of the eigenvalues approaches zero. However, one row of V_e and V_o gets arbitrarily large because the (1,1) element of C_e approaches zero. To avoid this problem the eigenvectors for the small eigenvalues are split off from the rest of the system as in the linear algebra approach. The apparent overflow for large τ in Equation (26) is not a problem in practice; simple algebraic manipulations reduce it to terms involving only decaying exponentials. The results presented are indistinguishable from those computed using the linear algebra and classical approaches.

This paper also suggests a novel scheme for handling the oscillating intensity values obtained from Equation (3). Instead of using that equation, use

$$I(0,\mu) = \sum_{k=0}^{K} f_k(0)\mathrm{sgn}(\mu)P_k(\mu) \tag{28}$$

when $\mu < 0$. Now, there are no negative intensities in the solution, an important result if the solution is to be imbedded in some larger calculation. One unfortunate side effect is that the resulting curve no longer crosses the correct solution at the Gauss quadrature points.

4 Mathematical Relations

Several papers have appeared that examine the mathematical and computational properties of some aspects of the SHM. In particular, I'll review papers on computing an integral arising from reflecting boundaries and various aspects of the Chandrasekhar polynomials.

4.1 Reflecting Boundary

First, let's look at the reflecting boundary condition of Equation (9). It is straightforward to write

$$\int_0^1 e^{-\tau/\mu} P_K(\mu) d\mu = \sum_{k=0}^{K} a_k E_{k+2}(\tau), \tag{29}$$

where E_k is the k-th exponential integral[3]. Unfortunately, the coefficients a_k vary widely in magnitude and differ in sign making the summation numerically unstable. Earlier investigators used Simpson's rule between zeros of $P_K(\mu)$[16] and a 50-point Gauss quadrature[5] to evaluate this integral. While accurate and efficient, neither of these approaches is esthetically pleasing.

More than a decade passed with no further publications in this area until Settle[17] derived a 5-term recurrence relation for a generalization of the integral that includes a factor of μ^r in the integrand. Hence, any reflectivity that can be expressed as a polynomial in μ can be computed. The backward recurrence of

$$\begin{aligned}(n+2)(n+r+3)(2n-1)S_{n+2} + \tau(2n-1)(2n+3)S_{n+1} \\ +(2n+1)(2n^2+2n-3-r)S_n - \tau(2n-1)(2n+3)S_{n-1} + \\ (2n+3)(n-1)(n-r-2)S_{n-2} = 0\end{aligned} \tag{30}$$

is stable allowing efficient evaluation of the integral. When the scattering is isotropic, $r = 0$, and Equation (30) can't be used to evaluate S_0. Fortunately, this term is simply the second exponential integral $E_2(\tau)$.

Settle also examined the behavior of S_n as $n \to \infty$. In the limit of very large n, S behaves like a cosine function in n. In this context, $n = 5$ can be considered *very large* because the asymptotic solution is within a few percent of the exact result.

Another approach[18], taken by Gander, is to note that the integral can be viewed as integrating a polynomial over a finite interval with a positive weight. It is always possible to derive a Gauss-type integration rule in this situation[19]. The idea is to find the coefficients α_k and β_k of the 3-term recurrence of the polynomial that is orthogonal on this interval under the given weight function,

$$\pi_k(x) = (x - \alpha_k)\pi_{k-1}(x) - \beta_k \pi_{k-2}(x), \tag{31}$$

where

$$\alpha_k = \frac{(x\,\pi_k\,,\,\pi_k)}{(\pi_k\,,\,\pi_k)} \quad \text{and} \quad \beta_k = \frac{(\pi_k\,,\,\pi_k)}{(\pi_{k-1}\,,\,\pi_{k-1})}, \tag{32}$$

and $(f, g) = \int_0^1 f(\mu) g(\mu) e^{-\tau/\mu} d\mu$. Unfortunately, until Gander's paper all known methods for finding the 3-term recurrence were numerically unstable. For example, the simple bootstrap method of using π_0 to derive π_1, *etc.* becomes unstable beyond $k = 10$.

Gander's contribution is to note that the coefficients of the 3-term recurrence can be computed in a stable manner using the ORTHPOL package[20]. The coefficients are shown to converge to constants as k gets large. The eigenvalues of the tridiagonal matrix of these coefficients are the Gauss points and the first components of the normalized eigenvectors are the weights.

Gander used these points and weights to evaluate the integral for $\tau = 1$ and k up to 200. The absolute accuracy is good, but as k increases, the value of the integral decreases causing the relative accuracy to suffer. Beyond $k = 100$ the value of the integral is less than 10^{-12} and no significant digits are left. Fortunately, the computed intensity is insensitive to such small contributions.

4.2 Chandrasekhar Polynomials

The Chandrasekhar polynomials defined in Equation (17) are another area of interest. Dehesa, *et al.*[21] derived the differential equation satisfied by the Chandrasekhar polynomials. While their solution is only valid for a certain choice of the scattering coefficients, β_k, with this restriction, they are able to show that these polynomials are hyper-geometric in nature and that the only classical polynomials in the same class are the Jacobi polynomials. These observations lead to interesting conclusions on the distribution of their roots.

Recall that the Chandrasekhar polynomials are the eigenvectors of a matrix that is similar to a symmetric, tridiagonal matrix. This observation allows a number of identities to be derived[22]. For example,

$$C_j \sum_{k=m}^{M} h_k g_k^m(\lambda_i) g_k^m(\lambda_j) = 2\delta_{i,j}, \tag{33}$$

where $h_k = 2k + 1 - \omega\beta_k$, and C_j is a normalization constant. This identity is easily derived from the orthogonality of the eigenvectors of a symmetric matrix. Other identities, such as

$$h_k \sum_{k=1}^{N+1} C_j g_n^m(\lambda_j) g_k^m(\lambda_j) = 2\delta_{n,k}, \tag{34}$$

have less obvious connections to the matrix formulation.

These identities will provide good checks on the correctness of programs used to compute the Chandrasekhar polynomials. They may also be used to reveal interesting properties.

Of course, the Chandrasekhar polynomials are not just mathematical curiosities; we need to compute them. Since they are the eigenvectors of a known matrix, we can use linear algebra techniques, but these methods have two problems. They are slow, needing $\mathcal{O}(N^3)$ operations, and their accuracy leaves something to be desired. To be precise, eigensolvers produce results of high absolute accuracy which leaves small components with high relative errors.

A better approach is to compute the polynomials directly from the 3-term recurrence. Unfortunately, great care is needed[23], especially for high order terms in the Fourier expansion of the azimuthal dependence. When evaluating $g_k^m(\lambda)$ for $|\lambda| > 1$, a range that corresponds to the continuous spectrum of the infinite dimensional problem, the forward recurrence is stable for all k and m. However, when $|\lambda| \approx 1$, the $g_j^m(\lambda)$ get large enough to overflow the numeric range of the computer arithmetic. A simple work-around is to keep the exponent and mantissa in separate words.

When $-1 \leq \lambda \leq 1$, backward recurrence can be used most of the time. However, it can become unstable. The key idea is to use backward recurrence as long as the ratio of successive terms is less than unity, say as far as K_*. Forward recurrence can be used as long as this ratio is greater than unity, say as far as k_*. The remaining terms are found by solving a linear system of order $K_* - k_*$. This difference is not always small. When using a highly peaked phase function appropriate for modeling a cloud, a linear system of order 200 had to be solved.

5 Extensions to the Basic Model

While the basic model presented in Section 2 is useful in many domains, there is a great deal of interest in including other effects. Nuclear engineers want to keep track of neutrons that change energy, the multi-group problem. Atmospheric scientists want to include polarization. Many physical problems are not well represented by plane parallel layers. Everyone would like to find an effective solution scheme for three dimensional problems. Progress has been made in all these areas in the time covered by this review.

5.1 Multi-group

There are situations in which the particles or photons described by Equation (2) have a finite probability of changing energy whenever they interact with the medium. The governing equation can be written as[24]

$$\mu \frac{\partial \mathbf{I}_\nu}{d\tau} + \Sigma \mathbf{I}_\nu = \omega \mathbf{T}(\nu, \nu') \int_{-1}^{1} \mathbf{I}_\nu(\mu') P(\mu, \mu') d\mu' + \mathbf{Q}_n u(\tau, \mu), \tag{35}$$

where there is one component of the vectors $\mathbf{I}$ and $\mathbf{Q}$ for each energy group. The matrix $\mathbf{T}(\nu, \nu')$ gives the probability that a particle with an energy ν changes its energy to ν' on each interaction, and the diagonal matrix Σ gives the probability the particle at a given energy will be absorbed or scattered on each interaction.

As in the classical method, we write the solution as

$$\mathbf{I}(\tau, \mu) = \sum_{n=0}^{N} \frac{2k+1}{2} P_k(\mu) \sum_{j=1}^{J} [A_j e^{-\lambda_j \tau} + (-1)^k B_j e^{-\lambda_j (\tau_0 - \tau)}] \mathbf{T}_k(\lambda_j), \tag{36}$$

where $\mathbf{T}_k = \mathbf{G}_k\mathbf{N}$, and $\mathbf{G}$ is a generalization of the Chandrasekhar polynomials. The vector $\mathbf{N}$ is a null vector of $\mathbf{G}_{K+1}$, *i.e.*, $\mathbf{G}_{K+1}\mathbf{N} = 0$. The λ_j are the eigenvalues of a block, tridiagonal matrix of order $N(K+1)$, where the expansion has $K+1$ terms, and we are dealing with N energies. The diagonal blocks of this matrix are tridiagonal while the sub- and super-diagonal blocks are upper and lower bidiagonal, respectively. Siewert found his solution scheme to be sufficiently fast that he made no effort to use the special structure of the matrix.

5.2 Polarization

The Earth's atmosphere polarizes the radiation passing through it, so there is a lot of interest in including the effect in the radiative transfer problem. The governing equation can be written as[25]

$$\mu\frac{d\mathbf{I}}{d\tau} + \mathbf{I} = \frac{\omega}{4\pi}\int_0^{2\pi}\int_{-1}^{1}\mathbf{P}(\mu,\mu',\phi-\phi')\mathbf{I}(\tau,\mu',\phi')d\mu' d\phi'. \tag{37}$$

The Stokes vector $\mathbf{I} = [I \;\; Q \;\; U \;\; V]^T$, and $\mathbf{P}(\mu,\mu',\phi-\phi')$ is a 4×4 matrix. The Fourier transform that separates the different orders is a bit more complicated because we're dealing with a system of equations. We define

$$\begin{aligned}\phi_1^m(x) &= (2-\delta_{0,m})\mathrm{diag}[\cos mx, \cos mx, \sin mx, \sin mx]\\ \phi_2^m(x) &= (2-\delta_{0,m})\mathrm{diag}[-\sin mx, -\sin mx, \cos mx, \cos mx]\end{aligned} \tag{38}$$

and let

$$\mathbf{I}(\tau,\mu,\phi) = \sum_{k=0}^{K}[\phi_1^m(\phi-\phi_0)\mathbf{I}_1^m(\tau,\mu) + \phi_2^m(\phi-\phi_0)\mathbf{I}_2^m(\tau,\mu)] \tag{39}$$

with a similar expansion for $P(\mu,\mu',\phi,\phi')$.

As in the scalar case, the terms in the Fourier series decouple leaving us with equations of the form

$$\mu\frac{d\mathbf{I}_\alpha^m}{d\tau} + \mathbf{I}_\alpha^m = \frac{\omega}{2}\int_{-1}^{1}\mathbf{K}^m(\mu,\mu')\mathbf{I}_\alpha^m(\tau,\mu')d\mu' + \mathbf{S}_\alpha^m(\tau,\mu) \tag{40}$$

to solve. The spherical harmonics expansion becomes

$$\mathbf{F}(\tau,\mu) = \sum_{k=m}^{M}\mathbf{P}_k^m(\mu)\mathbf{G}_k^m(\lambda)\mathbf{M}(\lambda)e^{-\lambda\tau}, \tag{41}$$

where $\mathbf{M}$ is a 4-element null-vector of $\mathbf{G}_{K+1}^m$. The choice of boundary conditions is clear for polarization problems since the Marshak condition give correct results only when $m = 0$. Why this happens is a topic yet to be studied.

The matrices $\mathbf{G}_k^m$ are a generalization of the Chandrasekhar polynomials with the recurrence relation

$$\mathbf{A}_k^m \mathbf{G}_{k-1}^m + \mathbf{B}_k^m \mathbf{G}_k^m + \mathbf{C}_k^m \mathbf{G}_{k+1}^m = \mathbf{G}_k^m / \lambda. \tag{42}$$

The matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ in Equation (42) have a simple structure if we apply an even-odd sort of their rows and columns with $\mathbf{A}$ and $\mathbf{C}$ becoming block diagonal and $\mathbf{B}$ becoming block skew diagonal. We end up needing to find the complex eigenvalues of a matrix of order $2K$. Particular solutions for a variety of source terms have been derived[11].

5.3 Spherical Symmetry

The radiative transfer equation becomes

$$\mu \frac{\partial I}{\partial \tau} + \frac{1-\mu^2}{\tau} \frac{\partial I}{\partial \mu} + I = \frac{\omega}{2} \int_{-1}^{1} I(\tau, \mu') P(\mu, \mu') d\mu' + S \tag{43}$$

in spherical symmetry. Using the expansions in Equations (3) and (4) gives us the spherical harmonics equations

$$a_k \frac{df_{k+1}}{d\tau} + b_k \frac{df_{k-1}}{d\tau} = c_k \frac{f_{k-1}}{\tau} - d_k \frac{f_{k+1}}{\tau} - (1 - \omega \beta_k) f_k + s_k, \tag{44}$$

where a_k, b_k, c_k, and d_k are known functions of k. The fundamental solutions to this equation are the modified spherical Bessel functions $i_k(x)$ and $k_k(x)$. In the limit of large k, $i_k(x) \sim \sinh(x)$ and $k_k(x) \sim e^{-x}$.

While a number of solution schemes have been proposed, all find that the linear system to be solved for the unknown coefficients gets worse as K/r_0 increases, where r_0 is the inner radius of the spherical shell. Perhaps the form of the modified spherical Bessel functions in the limit of large k is responsible for this instability.

Li and Tong[26] looked at this problem when the scattering phase function as only two terms in its expansion, β_0 and β_1, including the effects of thermal emission and radiating boundaries. They present an analytic form for the solution in the P_1 problem and use an ordinary differential equation (ODE) solver for higher order terms. They were able to compute problems having thin shells and shells or solid spheres with optical thicknesses up to 200 up to the P_{11} approximation.

Siewert and Thomas[27] express the solution in terms of the Bessel functions as $I_k(\lambda_j \tau)$ and $K_k(\lambda_j \tau)/\tau$. They derive particular solution for emission and specular and diffuse reflection at the boundaries. Tine, *et al.*[28] write the solution in terms of the functions $h_k^{\pm} = I_{-k-1/2} \pm I_{k+1/2}$. They, too, note the instability in high order. Could the reason be that $h_k(0) \sim e^k$ as $k \longleftarrow \infty$?

An interesting trick, called the sphere to plane transform, can be used when we only need the first couple of moments. For example, to couple the heat flow

equation with the radiation transport we need only compute the radiative flux, f_1. The form of the governing equation (43) is changed by writing

$$f_1(\tau) = \frac{1}{\tau}\int_{-1}^{1} \mu\Phi(\tau,\mu)d\mu + \frac{1}{\tau^2}\int_{-1}^{1} \mu^2\Phi(\tau,\mu)d\mu, \tag{45}$$

where $\Phi(\tau,\mu)$ is a solution to the plane parallel problem.

5.4 Multiple Dimensions

Well, after a long, roundabout trip, we're ready to return to general geometries as expressed in Equation (1), except that we'll ignore the time dependence and polarization. If we make the spherical harmonics expansion for $I(\mathbf{x},\Omega)$ and $P(\Omega,\Omega')$, we get

$$\sum_{k=0}^{K}\sum_{m=-k}^{k} Y_k^m \left\{ -\frac{1}{\sigma}(\Omega\cdot\nabla)f_k^m + \left[1 - \frac{\omega\beta_k}{2k+1}\right] f_k^m \right\} = 0. \tag{46}$$

We now multiply by $Y_q^p(\Omega)$ and integrate over Ω to get a set of equations to solve. The form of the resulting coefficients depends on the coordinate system.

In two dimensions, we can write

$$(\Omega\cdot\nabla)I = \mu\frac{\partial I}{\partial z} + \sqrt{1-\mu^2}\cos\phi\frac{\partial I}{\partial z}, \tag{47}$$

where

$$\left[\mu\frac{\partial I}{\partial z}\right]_{km} = a_{km}^{-}\frac{\partial f_{k-1}^m}{\partial z} + a_{km}^{+}\frac{\partial f_{k+1}^m}{\partial z}. \tag{48}$$

The x-term has derivatives of $f_{k\pm1}^{m\pm1}$. The resulting system of equations is

$$A\frac{\partial \mathbf{f}}{\partial z} + B\frac{\partial \mathbf{f}}{\partial x} + E\mathbf{f} = S. \tag{49}$$

In the Spherical Harmonics Spatial Grid method[29], Evans now discretizes Equation (49) in space and solves the resulting linear system with the conjugate gradient method. The quantities of interest for atmospheric modelling are the directional fluxes, $F_x = \sqrt{4\pi/3}f_{11}$, $F_z = \sqrt{4\pi/3}f_{10}$, and the mean intensity, $\bar{I} = \sqrt{4\pi}f_{00}$.

The most recent paper covered by this review[30], dated 9 July 1996, describes the Spherical Harmonics Discrete Ordinates Method for 3D problems. This method is very similar to the method of successive orders of scattering used in the 1950's, but it accelerates the iteration. It takes its name from the fact that the integration of the source function is done over a set of discrete ordinates, but the source function is evaluated in the spherical harmonics form.

The method is quite simple. Start with a guess for the spherical harmonics moments, $f_k^m(\mathbf{x})$. Then iterate

1. $J(\mathbf{x}) = \sum_{k,m} Y_k^m [\omega \beta_k f_k^m/(2k+1) + S_k^m]$
2. $I(s) = I(0)e^{-\tau(s)} + \int_0^s e^{-[\tau(s)-\tau(s')]} J(s')\kappa(s')ds'$
3. $f_k^m(\mathbf{x}) = \int_{-1}^1 I(\mathbf{x},\mu) Y_k^m(\mu) d\mu$

until converged. The integration step is carried out over a 3D grid.

The iteration can be written as $\mathbf{J}^{(n)} = A\mathbf{J}^{(n-1)} + B$. Hence, the convergence rate depends on the largest eigenvalue of A. Evans shows that the convergence is accelerated by replacing $\mathbf{J}$ with

$$\mathbf{J}'^{(n)} = \mathbf{J}^{(n)} + a[\mathbf{J}^{(n)} - \mathbf{J}^{(n-1)}] \tag{50}$$

on every second iteration. Here a depends on the residuals of this and the preceding iteration.

Since the integration is carried out in specific directions, we can monitor the optical depth of the grid blocks. When they become too large, the grid can be refined until the solution is sufficiently accurate.

6 Conclusions

We have seen that the spherical harmonics method is widely used in a variety of disciplines. It is well understood for plane geometry and provides efficient, numerically stable solutions. The basic method has been extended to include such things as conduction, polarization, and energy changes.

There are still some open issues. For example, applying the SHM in spherical symmetry doesn't work very well. It's bad enough that we don't have a method that works in high order or with solid spheres; we don't even know why the current approaches become unstable.

There are also problems with the methods used for 3D problems. We know that the radiation falls off exponentially with distance from some point, yet our solution schemes assume a polynomial representation. (That's all a spatial discretization does.) Hence, these methods may become intractable with large optical thicknesses and highly peaked scattering phase functions.

Throughout this review, I've noted where there are unsolved problems. I can only hope that those students looking for a good thesis topic were paying attention.

References

[1] S. Chandrasekhar, *Radiative Transfer*, Dover, New York, 1960.

[2] J. H. Jeans, *The Equations of Radiative Transfer of Energy*, Monthly Notices of the Royal Astronomical Society, 78 (1917), pp. 28–36.

[3] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Function*, U.S. Government Printing Office, Washington, D.C., 1964. Seventh Printing, May 1968, with corrections.

[4] C. Moler and C. V. Loan, *Nineteen Dubious Ways to Exponentiate a Matrix*, SIAM Review, 20 (1978), p. 801.

[5] A. H. Karp, J. Greenstadt, and J. A. Fillmore, *Radiative Transfer through an Arbitrarily Thick, Scattering Atmosphere*, J. Quant. Spectrosc. Radiat. Transfer, 24 (1980), pp. 391–406.

[6] A. H. Karp, *Computing the Angular Dependence of the Radiation of a Planetary Atmosphere*, J. Quant. Spectrosc. Radiat. Transfer, 25 (1981), pp. 403–412.

[7] V. Kourganoff, *Basic Methods in Transfer Problems*, Clarendon Press, Oxford, 1952.

[8] M. Benassi, R. D. M. Garcia, A. H. Karp, and C. E. Siewert, *A High-Order Spherical Harmonics Solution to the Standard Problem in Radiative Tranfer*, Astrophys. J., 280 (1984), pp. 853–864.

[9] N. J. McCormick and C. E. Siewert, *Particular Solutions for the Radiative Transfer Equation*, J. Quant. Spectrosc. Radiat. Transfer, 46 (1991), pp. 519–522.

[10] L. B. Barichello, R. D. M. Garcia, and C. E. Siewert, *A Spherical Harmonics Solution for Radiative Transfer Problems with Reflecting Boundaries and Internal Sources.* preprint, 1996.

[11] C. E. Siewert and N. J. McCormick, *A Particular Solution for Polarization Calculations in Radiative Transfer*, J. Quant. Spectrosc. Radiat. Transfer, 50 (1993), pp. 531–540.

[12] C. E. Siewert and J. R. Thomas, Jr., *On Coupled Conductive-Radiative Heat Transfer in a Sphere*, J. Quant. Spectrosc. Radiat. Transfer, 45 (1991), pp. 63–72.

[13] W. H. Wells and J. J. Sidorowich, *Computational Techniques for Radiative Transfer by Spherical Harmonics*, J. Quant. Spectrosc. Radiat. Transfer, 33 (1985), pp. 347–363.

[14] K. Kamiuto, *Chebyshev Collocation Method for Solving the Radiative Transfer Equation*, J. Quant. Spectrosc. Radiat. Transfer, 35 (1986), pp. 329–336.

[15] Y. Takeuchi, *Use of Spherical Harmonics in the Solution of the Radiative Transfer Problem*, J. Quant. Spectrosc. Radiat. Transfer, 39 (1988), pp. 237–245.

[16] J. V. Dave, *A Direct Solution of the Spherical Harmonics Approximation to the Transfer Equation for a Plane Parallel, Nonhomogeneous Atmosphere*, J. Atmos. Sci., 32 (1975).

[17] J. J. Settle, *Some Properties of an Integral that Occurs in Problems of Radiative Transfer*, J. Quant. Spectrosc. Radiat. Transfer, 52 (1994), pp. 195–206.

[18] M. J. Gander, *Numerical Evaluation of a Particular Integral.* unpublished, January 1996.

[19] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Computer Science and Applied Mathematics, Academic Press, San Diego, second ed., 1984.

[20] W. Gautschi, *ORTHPOL – A Package of Routines for Generating Orthogonal Polynomials and Gauss-type Quadrature Rules*, ACM Trans. Math. Software, 20 (1994).

[21] J. S. Dehesa, A. Zarzo, R. J. Yanez, B. Germano, and P. E. Ricci, *Orthogonal Polynomials and Differential Equations in Neutron Transport and Radiative Transfer Theories*, J. Computational and Applied Math., 50 (1994), pp. 197–206.

[22] C. E. Siewert and N. J. McCormick, *Some Identities for Chandrasekhar Polynomials.* preprint, 1996.

[23] R. D. M. Garcia and C. E. Siewert, *On Computing the Chandrasekhar Polynomials in High Order and High Degree*, J. Quant. Spectrosc. Radiat. Transfer, 43 (1990), pp. 201–205.

[24] C. E. Siewert, *A Spherical Harmonics Method for Multigroup or Nongray Radiation Transport*, J. Quant. Spectrosc. Radiat. Transfer, 49 (1993), pp. 95–106.

[25] R. D. M. Garcia and C. E. Siewert, *A Generalized Spherical Harmonics Solution for Radiative Tranfer Models the Include Polarization Effects*, J. Quant. Spectrosc. Radiat. Transfer, 36 (1986), pp. 401–423.

[26] W. Li and W. Tong, *Radiative Heat Transfer in Isothermal Spherical Media*, J. Quant. Spectrosc. Radiat. Transfer, 43 (1990), pp. 239–251.

[27] C. E. Siewert and J. R. Thomas, Jr., *A Particular Solution for the P_n Method in Spherical Geometry*, J. Quant. Spectrosc. Radiat. Transfer, 46 (91), pp. 459–462.

[28] S. Tine, S. Aiello, A. Bellini, and C. C. Pestellini, *A Solution to the Problem of Radiative Transfer in Inhomogeneous Media Using the SHM*, J. Quant. Spectrosc. Radiat. Transfer, 47 (1992), pp. 95–102.

[29] K. F. Evans, *Two Dimensional Radiative Transfer in Cloudy Atmospheres: The Spherical Harmonic Spatial Grid Method*, J. Atmos. Sci., 50 (1993), pp. 3111–3124.

[30] ——, *The Spherical Harmonics Discrete Ordinate Method for Three Dimensional Atmospheric Radiative Transfer.* submitted to J. Atmos. Sci., July 1996.

List of Tables

Table 1: Summary of literature search.

Applications	202	Extensions	208	Method	298
Nuclear	64	2D/3D	78	Low order	113
Non-gas/neutron	42	Cyl/Sph/Tor	48	Comparison	76
Atmosphere	37	Cond/Conv	25	Method	76
Heat transfer	21	Multigroup	22	Math. rel.	24
Astronomy	19	Inverse	15	BC	9
Fluids	18	Polarization	10		
Oil wells	1	Time dep.	10	False hits	53

List of Figures

Figure 1: Left: General form of problem to be solved. Right: Simplified form.

Figure 2: Examples of scattering phase functions.

Figure 3: Evaluating the intensity from the expansion in spherical harmonics (oscillating line) and by integration of the source function (solid line). The dots are at the Gauss quadrature points.