Inverse Problems in Science and Engineering
Vol. 16, No. 7, October 2008, 927–951

Taylor & Francis
Taylor & Francis Group

A two-phase parameter estimation method for radiative transfer problems in paper industry applications

Per Edström*

Department of Engineering, Physics and Mathematics, Mid Sweden University, Härnösand, Sweden

(*Received 20 March 2007; final version received 7 March 2008*)

A two-phase method for estimation of the scattering and absorption coefficients and the asymmetry factor (σ_s, σ_a and g) in the radiative transfer problem is presented. The first phase parameterizes σ_s and σ_a through g via a simplified model and performs – at a relatively low cost – a scalar optimization over g. It is shown that this gives such a good starting point that the second phase can be accurately performed by a simple Gauss-Newton method. It is also shown that as a part of the first phase can be used on its own when only σ_s and σ_a are wanted, and it is noted that this gives higher accuracy than the commonly used Kubelka–Munk method when using standardized paper industry reflectance factor measurements. The parameter estimation problem is shown to be non-trivial and ill-conditioned, and its character is analysed. It is discussed that as standard optimization methods are so sensitive to the choice of starting point for this problem that it is hard to find a starting point that gives convergence at all. The new two-phase method is illustrated by application to relevant paper industry problems, and efficiency and sensitivity measures are given.

Keywords: radiative transfer; integro-ordinary differential equations; inverse problems; paper industry applications

AMS Subject Classifications: 45J05; 45Q05; 65R32; 65Z05; 85A25

1. Introduction

There are plenty of rather crude methods for estimating only the scattering and absorption parameters, σ_s and σ_a, for turbid media. Most of them solve very approximate forward problems like the diffusion approximation or the Eddington approximation, but they are generally sufficiently accurate for the given application. This also holds for the Kubelka–Munk model [1–3], which is in widespread industrial use (there are, however, applications where its accuracy is not sufficient [4–6]). Yang and co-workers recently proposed a revised Kubelka–Munk theory [7–10]. However, this was debated by Edström [11], who also discussed different modelling strategies and their feasibility. Murphy [12] recently proposed an interesting method for estimating optical properties of coatings, based on an extended Kubelka–Munk approach. Methods for also estimating the asymmetry factor g are scarce, and even fewer are efficient and accurate. Most efforts come from medical

*Email: per.edstrom@miun.se

ISSN 1741–5977 print/ISSN 1741–5985 online

DOI: 10.1080/17415970802080066
http://www.informaworld.com

applications, while the industrial side has shown less interest so far. Van Gemert and Star [13] presented an attempt, but it includes approximations and boundary conditions that give it limited use. Prahl et al. [14] introduced the inverse adding–doubling method, but reported problems finding suitable starting points. One recent interesting approach of high accuracy was presented by Joshi et al. [15], but it suffers from very long computation times. In spite of the reported problems in the literature, this work aims to use only simple implementations of standard optimization methods for this parameter estimation problem, and still achieve good efficiency.

Estimating only the scattering and absorption parameters from reflectance measurements is straightforward in cruder two-flux problem formulations like Kubelka–Munk. However, more accurate estimations of more parameters is an outstanding issue in more general radiative transfer problem formulations. Siewert [16] presented a semi-analytical approach to a problem similar to the one considered in this work, however only the azimuthally averaged case, based on results proposed by McCormick [17]. Siewert [18] also presented an analytical method for a certain class of problems with only isotropic scattering.

Discrete ordinate solution methods are among the most used and most accurate solution methods for the forward radiative transfer problem today. Stamnes et al. [19,20] presented in a series of papers the successive development of a stable discrete ordinate algorithm, and provided a software package, DISORT. Thomas and Stamnes [21] also wrote a textbook on radiative transfer in the atmosphere. Siewert adopted a similar approach [22], but used Green's functions for the particular solution. Barichello and Siewert [23] also presented a two-dimensional Fourier transform procedure for the case with an incident single-point beam. In a recent paper, Edström [24] presented a systematic review of the stability enhancing and speed increasing steps used in modern discrete ordinate radiative transfer algorithms. Edström also presented the solution method DORT2002, which is adapted to light scattering simulations in paper and print, but which is also designed for methodical numerical experiments.

In this work, the radiative transfer problem has an angle-resolved formulation, and a many-flux solution procedure is used. The parameter estimation problem is formulated as a least-squares optimization problem, and a two-phase method for its solution is implemented and evaluated. The successful recovery of σ_s, σ_a and g is illustrated by application to relevant paper industry problems in the visible range.

The properties of the radiative transfer parameter estimation problem have not yet been fully investigated. In order to interpret the results correctly, some knowledge of the influence of measurement errors or noise on the parameters is needed. This knowledge can also be used for the design of experiments with minimal influence from measurement errors. In this work, the curvature of the problem (as defined and illustrated in Section 5.1) is investigated and a perturbation analysis is performed, and the results of numerical experiments are discussed in the light of those findings.

This work will frequently distinguish between two problems that apply to both the forward and the inverse settings. The *full* problem comprises angle-resolved data together with scattering, absorption and asymmetry parameters. The *d/0* problem involves standardized reflectance factor data as used in the paper industry [25–28], and scattering and absorption parameters only (the expression d/0 means diffuse illumination, usually accomplished by an integrating sphere covered on the inside with a white diffusing powder, and detection at 0° polar angle relative to the normal to the sample surface.) How to handle anisotropy is a new and open question for the paper industry, and refined

measurement and simulation methods are needed to resolve this matter. In the field of applied paper optics, such issues are now starting to be studied. Some paper industries, and several institutes, have built so-called goniophotometers for such studies, but the instruments are of various constructions. An initiative for standardization is being discussed among European paper industries and institutes. This work is a step in that direction.

Section 2 gives a problem formulation and presents the solution methods used in the forward problem. In Section 3, the inverse problem is formulated, the optimization methods used in the inverse calculations are discussed, and specific solution methods to the d/0 and full inverse problems are suggested. The sensitivity analysis is performed in Section 4, numerical experiments are presented in Section 5 on problem characteristics and Section 6 is on method performance. Section 7 gives a brief discussion of the findings.

2. The forward problem

The radiative transfer problem considers the propagation of radiation in a turbid medium. The problem in this work is studied in a plane-parallel geometry, where the horizontal extension of the medium is assumed to be large enough to give no boundary effects at the sides. At the top and bottom boundary surfaces, boundary conditions are assumed to be time- and space-independent. The radiation is assumed to be monochromatic, and the scattering is assumed to be conservative, i.e. without change in frequency between incoming and outgoing radiation. The medium is treated as a continuum of scattering and absorption sites. Polarization effects are ignored, and what is left is then a scalar intensity, which is the variable to solve for.

2.1. *Formulation*

Edström [24] states the equation of radiative transfer as

$$u\frac{\mathrm{d}I(\tau,u,\varphi)}{\mathrm{d}\tau} = I(\tau,u,\varphi) - \frac{1}{4\pi}\frac{\sigma_s}{\sigma_s+\sigma_a}\int_0^{2\pi}\int_{-1}^{1} p(u',\varphi';u,\varphi)I(\tau,u',\varphi')\mathrm{d}u'\,\mathrm{d}\varphi'. \tag{1}$$

The unknown intensity I at optical depth τ is considered as non-interacting beams of radiation in all directions. The scattering and absorption coefficients of the medium are denoted by σ_s and σ_a, and the phase function p specifies the probability distribution of scattering from incident direction (u',φ') to direction (u,φ), where u is cosine of polar angle, and φ is azimuthal angle. The shape of the phase function is controlled by a parameter called the asymmetry factor, g, ranging from complete forward scattering ($g=1$) over isotropic scattering ($g=0$) to complete backward scattering ($g=-1$). Different phase functions have been proposed to describe physically different types of scattering. This work considers the Henyey–Greenstein [29] phase function. It should not be seen as a real phase function, but is a one-parameter analytical approximation of widespread use. It is given by

$$p(\cos\Theta) = \frac{1-g^2}{(1+g^2-2g\cos\Theta)^{3/2}}, \tag{2}$$

where Θ is the scattering angle. It is thus evident that the Henyey–Greenstein phase function is dependent on the scattering angle Θ only, and not on the specific directions of incident and scattered radiation.

2.2. *Solution methods*

DORT2002 [24] is a radiative transfer solution method adapted to light scattering in paper and print. It is used to solve the *full* radiative transfer problem in this work. DORT2002 has been successively improved and evaluated [30], and has also been successfully applied to real paper industry problems [4–6]. DORT2002 uses Fourier analysis on the azimuthal angle to turn the integro-differential Equation (1) into a number of uncoupled equations, one for each Fourier component of the unknown intensity, which are then discretized using numerical quadrature. This yields a system of first-order linear differential equations for each Fourier component, and the natural solution procedure gives an eigenvalue problem.

In the DORT2002 solution method, the azimuthally averaged intensity is given by the 0-th Fourier component. This quantity is all that is needed to calculate the reflectance factor measures R_0 and R_∞ that result from using the d/0° instrument geometry, which is standardized in the paper industry [25–28] (R_0 is measured for a single sample with a black cavity as a background, and R_∞ is measured on an opaque pad of identical samples). To solve this *d/0* problem, DORT2002 also contains a specialized solution method with the specific purpose of performing such calculations at the highest possible speed. This gives a significant reduction in computation time compared to the full method [30]. Since the standardized d/0° instrument geometry is specifically taken into account in DORT2002, a higher accuracy is achieved then with the Kubelka–Munk solution method that is normally used in the paper industry [4,6]. Although only a side effect in the context of this work, this actually makes DORT2002 an alternative to Kubelka–Munk when higher accuracy is needed.

It should be noted that the focus of this work is the inverse method. Therefore, other forward solution methods than DORT2002 could well be considered, and as long as they provide the same accuracy, the conclusions in this work still hold. Replacing DORT2002 would probably increase computation times though, since it very efficiently provides intensities and reflectances for the paper and printing applications studied here.

3. The parameter estimation problem

The full parameter estimation problem consists in determining σ_s, σ_a and g from angle-resolved intensity measurements in chosen directions, $I(u_i, \varphi_i)$. Such measurements are available from special goniophotometers. The full parameter estimation problem can thus be defined as

$$\textit{find } \sigma_s, \sigma_a \text{ and } g \textit{ given } \text{the measurements } I(u_i, \varphi_i)\,. \tag{3}$$

In the d/0 problem, only σ_s and σ_a are estimated from standardized reflectance factor measurements, R_0 and R_∞, using a chosen *ad hoc* value of g. Standardized reflectance factor measurements with d/0° instrument geometry are abundant in the paper industry. They use two measurements of the same sample, one over a black cavity and another over

a thick pile of identical samples. The d/0 parameter estimation problem can thus be defined as

$$\textit{find } \sigma_s \text{ and } \sigma_a \textit{ given} \text{ the measurements } R_0 \text{ and } R_\infty. \tag{4}$$

The parameter estimation problem is to find parameter values that minimize some distance measure between real measurements and model predictions. The full problem is usually over-determined with noisy measurements. The noise can be controlled by averaging over several measurements, but the parameter estimation is nevertheless a nonzero-residual problem. The d/0 parameter estimation problem uses two measurements to determine two parameters, which makes it a zero-residual problem regardless of noise.

3.1. *Formulation*

One way to introduce the distance measure to minimize is through an objective function that sums squared errors, such as

$$F(x) = \frac{1}{2}\|f(x)\|_2^2 = \frac{1}{2}\sum_i f_i(x)^2 = \frac{1}{2}\sum_i \{M_i(x) - b_i\}^2, \tag{5}$$

where i denotes the respective quantity (be it intensity in a given direction or reflectance factor over a given background), and x is the vector of parameters to be determined. This formulation is statistically optimal if the measurement errors are normally distributed, which may reasonably be considered to be the case here. In this work, the model predictions $M_i(x)$ are given by DORT2002 simulations, and b_i are given by goniophotometer or d/0° instrument geometry measurements. An obvious formulation of the parameter estimation problem is then

$$\min_x F(x), \tag{6}$$

or the explicit least-squares formulation

$$\min_x \frac{1}{2}\|f(x)\|_2^2 = \min_x \frac{1}{2}\sum_i f_i(x)^2. \tag{7}$$

If one defines the set of permissible parameter combinations as

$$S = \{(\sigma_s, \sigma_a, g) : \sigma_s > 0, \quad \sigma_a > 0, -1 < g < 1\}, \tag{8}$$

one can state the parameter estimation problem in various ways, which makes it possible to use different optimization methods to find a solution. Of course, since the problem is constrained by

$$x \in S, \tag{9}$$

one needs to deal with this separately if unconstrained formulations are used.

It should be pointed out that it is assumed throughout this work that the model output – or the measurement – is given uniquely from the parameters. This is a quite reasonable assumption, especially for applied problems, and proof has been given for several reasonable cases [31], although not generally for the entire problem class.

A quite general proof for the uniqueness of the solution to the inverse problem – given the existence of a solution to the forward problem – has been given by Choulli and Stefanov [32]. There is still more work to be done to achieve full and general knowledge regarding existence and uniqueness for the general radiative transfer problem and its inverse.

3.2. *Solution methods*

A few simple implementations of standard optimization methods were implemented as described below, and comparison was also made to some Matlab optimization toolbox functions as examples of commercial solvers. In the following, k designates the iteration index.

3.2.1. *Newton*

With an unconstrained minimization formulation like (6), a classical approach is Newton's method. The typical Newton iteration consists of determining a search direction, p_k, by solving

$$\nabla^2_{xx} F(x_k) p_k = -\nabla_x F(x_k), \tag{10}$$

and then updating the solution estimate through

$$x_{k+1} = x_k + \alpha_k p_k. \tag{11}$$

The step size parameter, α_k, is normally chosen in a line search procedure. The constant step size $\alpha_k = 1$ gives the pure form of Newton's method. Near a non-singular minimum, the Hessian $\nabla^2_{xx} F(x_k)$ will be positive definite, and the convergence will be quadratic. The good convergence comes at the cost of expensive calculation of the Hessian. On the other hand, far from such a local minimum, the Hessian may be singular or the search direction may not be a decent direction because the Hessian is not positive definite. Thus, unless a very good starting point is available, the convergence may initially be very poor. The convergence also depends on the condition of the Hessian in the solution.

In this work, a Newton method was implemented with a forward difference scheme for the gradient and the Hessian. The constraints (9) were kept by limiting p_k in the (rare) cases where $x_k + p_k \notin S$. Then an Armijo condition [33] was used for choosing step size in the line search. The convergence criterion was $\|\nabla_x F(x_k)\| < \sqrt{\epsilon_{\text{mach}}}(1 + |F(x_k)|)$, where $\varepsilon_{\text{mach}}$ is the floating point relative accuracy of the machine, as suggested, for example, by Nash and Sofer [34]. The algorithm was as follows.

Algorithm 1 (`Newton`)
Choose starting point x
repeat until convergence or failure
Convergence and failure check
Calculate gradient G by forward differences
Calculate Hessian H by forward differences
Calculate search direction p by solving $Hp = -G$ through factorization
Set step length $\alpha = 1$, then half α until Armijo condition fulfilled
Update $x = x + \alpha p$
end.

From a purely numerical point of view, different finite difference intervals, h, were used for the gradient (about the square root of machine precision) and Hessian (a few decades larger) approximations for the most accurate results [35].

3.2.2. *Quasi-Newton*

Newton's method computes the full Hessian in each iteration, which is expensive. An alternative is to continuously update an approximation of the Hessian, and thereby approximate the Newton direction. Quasi-Newton methods do this by building curvature information from the successive iterates x_k and the corresponding gradients $\nabla_x F(x_k)$ to formulate a quadratic model problem. The idea is to avoid the second derivative calculations, while still maintaining good convergence since the Hessian is progressively approached. There are a large number of Hessian updating methods, but the BFGS method [36–39] is considered to be the best for general purposes. Quasi-Newton methods have the advantage that they always give descent. However, a possible drawback is that inaccurate derivative information may accumulate errors.

In this work, a quasi-Newton method was implemented exactly as the Newton method in Section 3.2.1, but with the Hessian approximated through a BFGS update. The starting estimate of the Hessian was, for positive definiteness, calculated through the Jacobian as described for Gauss-Newton in Section 3.2.3. The algorithm was as follows.

Algorithm 2 (`Quasi-Newton`)
Choose starting point x
Calculate Gauss-Newton Jacobian J at start by forward differences
Calculate Hessian by $H = J^T J$
repeat until convergence or failure
  Convergence and failure check
  Calculate gradient G by forward differences
  Calculate search direction p by solving $Hp = -G$ through factorization
  Set step length $\alpha = 1$, then half α until Armijo condition fulfilled
  Update $x = x + \alpha p$
  Update Hessian H by BFGS method
end.

3.2.3. *Gauss-Newton*

With a least-squares formulation like (7), a specialized method is Gauss-Newton [40]. The Gauss-Newton iteration is based on the idea of linearizing f around the point x_k to obtain the linear least-squares problem

$$\min_{p_k} \frac{1}{2} \| f(x_k) + J(x_k) p_k \|_2^2, \tag{12}$$

where J is the Jacobian of f. This is then solved for p_k, for example by using a QR factorization.

Formally, although not normally used in solution methods, the search direction is found through the normal equations

$$J(x_k)^T J(x_k) p_k = -J(x_k)^T f(x_k). \tag{13}$$

Through this the Gauss-Newton method can be compared with Newton's method. The search direction Equation (13) is actually the same as Equation (10), where the gradient is given by

$$G(x_k) = J(x_k)^T f(x_k) \tag{14}$$

and the Jacobian is used to approximate the Hessian through

$$H(x_k) \approx J(x_k)^T J(x_k). \tag{15}$$

Gauss-Newton saves computations – by not computing the Hessian – possibly at the expense of decreased convergence rate. Near a solution the convergence is quadratic for zero-residual problems if J has full rank in the solution. The difference between the Hessian and its Gauss-Newton approximation can be shown to be $\sum_i f_i(x)\, f_i''(x)$, see Equation (29). Hence, if the residual $f_i(x)$ or curvature $f_i''(x)$ is large, or if the Jacobian is ill-conditioned in an iteration or in the solution, Gauss-Newton may converge slowly or not at all. In such cases the Newton method will be better. On the other hand, if the measurements are noisy, Gauss-Newton may have a regularizing effect by omitting the term $\sum_i f_i(x) f_i''(x)$.

The d/0 case uses two distinct measurements, R_0 and R_∞, to find two well-defined parameters, σ_s and σ_a. It can thus be noted from Equation (13) that since J therefore is square and of full rank, the problem of finding the search direction can be reduced. In this case, the linearization of the problem (7) does not even need to be formulated as a least-squares problem (12), but can be reduced to solving – with zero residual – the linear system

$$J(x_k)p_k = -f(x_k) \tag{16}$$

through Gaussian elimination. However, the full problem is over-determined, so the least-squares formulation is needed there.

In this work, a Gauss-Newton method was implemented with a forward difference scheme for the Jacobian, and an Armijo condition was used for choosing step size in the line search. The constraints (9) were kept by limiting p_k in the (rare) cases where $x_k + p_k \notin S$. The algorithm was as follows.

Algorithm 3 (`Gauss–Newton`)
Choose starting point x
repeat until convergence or failure
Convergence and failure check
Calculate Jacobian J by forward differences
Calculate p by solving $\min\|f + Jp\|_2^2$ through QR factorization
Set step length $\alpha = 1$, then half α until Armijo condition fulfilled
Update $x = x + \alpha p$
end.

3.2.4. *Matlab optimization toolbox functions*

The functions `lsqnonlin` and `fmincon` provided in Matlab optimization toolbox were tested as commercial examples of constrained optimization methods. The function `lsqnonlin` solves non-linear least-squares problems using the Gauss-Newton method with

a mixed quadratic and cubic line search procedure. The function `fmincon` uses a sequential quadratic programming (SQP) method [41], which is related to Newton's method. At each iteration, a positive definite quasi-Newton approximation of the Hessian is calculated using the BFGS method. This is then used to generate a quadratic programming (QP) subproblem whose solution is used to form a search direction for a line search procedure.

3.3. *Solution method to the d/0 parameter estimation problem*

The Kubelka–Munk model [1–3] is simple enough to give closed formulas for the inverse calculations, so no optimization is needed. Its use is prescribed in paper industry standards, and it is well suited for use as starting point in the d/0 parameter estimation problem in this work. However, the Kubelka–Munk parameters s and k are not the physically objective scattering and absorption coefficients used in general radiative transfer theory. Therefore, some translation is needed for comparison. No exact translation exists, but it is generally regarded that it is adequate to use the translation suggested by Mudgett and Richards [42,43], complemented with the compensation for anisotropic single scattering of van de Hulst [44], i.e.

$$s = \frac{3}{4}\sigma_s(1-g), \tag{17}$$

$$k = 2\sigma_a. \tag{18}$$

In the d/0 parameter estimation problem, σ_s and σ_a were estimated from standardized reflectance factor measurements, R_0 and R_∞. The algorithm was as follows, where x^* is the algorithm solution.

Algorithm 4 (`Inverse DORT2002 d/0`)
Choose g (free user input)
Calculate starting point x using Kubelka–Munk
Apply Algorithm 3 (`Gauss-Newton`), using DORT2002 d/0 as $M(x)$
$x^* = (\sigma_s, \sigma_a)^T$.

Since there is no information on anisotropy from d/0 measurements, the asymmetry factor g is a free choice. Choosing $g = 0$ would harmonize most with the Kubelka–Munk model, which assumes perfectly diffused light throughout, but recent work [4,6,14,15] suggests that other values of g may be appropriate. Any choice of g works for the parameter estimation to converge, but different (more or less *ad hoc*) choices of g affect the scattering coefficient in an unsatisfactory way.

3.4. *Two-phase solution method to the full parameter estimation problem*

3.4.1. *Rationale*

The obvious way to attack the full parameter estimation problem would, just as in the d/0 case, be to find a suitable starting point and just apply a standard or commercial solution method. However, it turns out that the full problem has a character that makes it very sensitive to the choice of starting point. In fact, in most cases the solution methods – standard or commercial – do not converge unless the starting point is very close to

the optimum. This is not a problem in the d/0 case, where the Kubelka–Munk model provides a natural and adequate starting point, and where the choice of starting point is not crucial anyway. Unfortunately, there is no equally simple way to devise a starting point for the full problem. On the contrary, while σ_s and σ_a may be approximated with simpler models, g is inaccessible and yet is essential for convergence. Therefore, a direct attack on the full problem seems unfeasible.

However, since the d/0 problem is more well-behaved, and since g is a free parameter there, it should be possible to parameterize that case and run a scalar optimization on g to find a suitable starting point. This also relieves the user from having to supply an initial value of g through guesswork, and without knowledge. Thus, a two-phase approach should be viable.

3.4.2. *The first phase*

Algorithm 4 (`Inverse DORT2002 d/0`) actually provides a parameterization, giving $\sigma_s(g)$ and $\sigma_a(g)$. Using this parameterization, the objective function of the full problem can be evaluated from the single scalar parameter g. Of course, this does not span the entire parameter space but only a curve, so a solution to the full problem cannot be expected. However, it is likely that this curve in parameter space passes that solution close enough to provide a good starting point to the full problem. This starting point is given by the solution to a scalar optimization problem over g, using the objective function of the full problem and the parameterization mentioned above. Preferably, the scalar optimization problem should be solved cheaply. Since g is limited to the interval $]-1, 1[$ and the highest accuracy is not needed, a derivative free golden section search method with a relaxed convergence criterion (parameter tolerance of 10^{-2}) was used (with the entire admissible $]-1, 1[$ as start interval for g). The algorithm was as follows.

Algorithm 5 (`Phase 1`)
repeat until convergence of golden section search
  Calculate g (two endpoints of new golden section interval)
  Calculate $(\sigma_s, \sigma_a)^T$ by Algorithm 4 (`Inverse DORT2002 d/0`)
  Evaluate objective function of full problem by DORT2002
end
$x^* = (\sigma_s, \sigma_a, g)^T$.

This approach needs access to standardized reflectance factor measurements, R_0 and R_∞, apart from the obvious angle-resolved intensity measurements. However, this is not a problem since the necessary d/0 instruments are in widespread use in the paper industry, and such measurements are both fast and cheap.

3.4.3. *The second phase*

Thanks to the special parameterization and scalar optimization in the first phase, the starting point for the second phase is now already close to the optimum (see the numerical experiments in Section 6.2.1). This makes it possible to use a simple Gauss-Newton method for the second phase and still expect good convergence properties.

In the full parameter estimation problem, σ_s, σ_a and g were estimated from angle-resolved intensity measurements (and standardized reflectance factor measurements). The algorithm for the full two-phase solution method was as follows.

Algorithm 6 (`Inverse DORT2002`)
Calculate starting point x by Algorithm 5 (`Phase 1`)
Apply Algorithm 3 (`Gauss-Newton`), using DORT2002 as $M(x)$
$x^* = (\sigma_s, \sigma_a, g)^T$.

4. Sensitivity analysis

To study the sensitivity for perturbation of the solution to the parameter estimation problem, some notation is needed. By defining

$$F(x,b) = \frac{1}{2}\|f(x,b)\|_2^2 = \frac{1}{2}\sum_i f_i(x,b)^2 = \frac{1}{2}\sum_i \{M_i(x) - b_i\}^2, \tag{19}$$

where $f_i(x,b)$ are the residuals, $M_i(x)$ the model predictions and b_i the measurements, the non-linear least-squares problem can be stated

$$\min_x F(x,b), \tag{20}$$

with the solution $\hat{x}$. Likewise, the perturbed problem can be stated

$$\min_x F(x, b+\delta b), \tag{21}$$

with the solution $\hat{x} + \delta x$. This can also be stated as

$$\min_{\delta x} F(\hat{x}+\delta x, b+\delta b), \tag{22}$$

and a Taylor expansion around $(\hat{x}, b+\delta b)$ gives

$$\begin{aligned} F(\hat{x}+\delta x, b+\delta b) &= F(\hat{x}, b+\delta b) + \delta x^T \nabla_x F(\hat{x}, b+\delta b) \\ &\quad + \frac{1}{2}\delta x^T \nabla_{xx}^2 F(\hat{x}, b+\delta b)\delta x + O\big(\|\delta x\|^3\big) \\ &\equiv g(\delta x) + O\big(\|\delta x\|^3\big). \end{aligned} \tag{23}$$

Using the approximation $g(\delta x)$, solving the perturbed problem (22) is now equivalent to finding

$$\min_{\delta x} g(\delta x), \tag{24}$$

which can be done by solving

$$\nabla_{\delta x} g(\delta x) = 0. \tag{25}$$

This is in turn equivalent to solving

$$\nabla_x F(\hat{x}, b+\delta b) + \nabla_{xx}^2 F(\hat{x}, b+\delta b)\delta x = 0 \tag{26}$$

for δx. But - denoting by $J(\hat{x})$ the Jacobian of $f(x, b)$ evaluated in $\hat{x}$

$$\nabla_x F(\hat{x}, b + \delta b) = J(\hat{x})^T f(\hat{x}, b + \delta b) = J(\hat{x})^T (f(\hat{x}, b) - \delta b) = J(\hat{x})^T \delta b, \tag{27}$$

since the optimality conditions for the unperturbed problem gives

$$J(\hat{x})^T f(\hat{x}, b) = 0. \tag{28}$$

In addition to this, it holds that

$$\nabla^2_{xx} F(\hat{x}, b + \delta b) = J(\hat{x})^T J(\hat{x}) + \sum_i f_i''(\hat{x}) \{f_i(\hat{x}, b) - \delta b_i\}, \tag{29}$$

where $f_i''(\hat{x})$ is the Hessian of $f_i(x, b)$ evaluated in $\hat{x}$. Inserting this into Equation (26) gives

$$-J(\hat{x})^T \delta b + \left(J(\hat{x})^T J(\hat{x}) + \sum_i f_i''(\hat{x}) \{f_i(\hat{x}, b) - \delta b_i\} \right) \delta x = 0. \tag{30}$$

Thus, the solution $\hat{x} + \delta x$ to the perturbed problem (21) is approximately given by solving Equation (30) for δx. This gives the formal solution

$$\begin{aligned} \delta x &= \left(J(\hat{x})^T J(\hat{x}) + \sum_i f_i''(\hat{x}) f_i(\hat{x}, b) - \delta b_i \} \right)^{-1} J(\hat{x})^T \delta b \\ &= \left(J(\hat{x})^T J(\hat{x}) + \sum_i f_i''(\hat{x}) f_i(\hat{x}, b) \right)^{-1} J(\hat{x})^T \delta b + O\left(\|\delta b\|^2\right) \\ &\equiv P \delta b + O\left(\|\delta b\|^2\right). \end{aligned} \tag{31}$$

The relative change of parameter x_i for measurement perturbations is thus given by

$$\frac{\delta x_i}{x_i} = \frac{1}{x_i} \sum_j P_{ij} \delta b_j + O\left(\|\delta b\|^2\right) = \frac{1}{x_i} \sum_j P_{ij} b_j \frac{\delta b_j}{b_j} + O\left(\|\delta b\|^2\right), \tag{32}$$

and so the relative sensitivity of parameter x_i for change in measurement b_j is approximately given by

$$\kappa_{ij} \equiv \frac{\delta x_i}{x_i} \Big/ \frac{\delta b_j}{b_j} = \frac{1}{x_i} P_{ij} b_j. \tag{33}$$

Hence, if the element κ_{ij} of the sensitivity matrix κ is large, even small perturbations in measurement j may result in large changes in parameter i. This may be the case for nonzero-residual problems with large curvature and for problems with a large residual, but it may also happen for zero-residual problems with ill-conditioned Jacobian or Hessian in the solution.

5. Numerical experiments: characterization of the parameter estimation problem

To keep the tables and figures as clean as possible, no units are given there. Throughout this work, scattering and absorption coefficients have the unit $\mathrm{m}^2\,\mathrm{kg}^{-1}$, grammages ($w$) have the unit $\mathrm{kg\,m}^{-2}$, and times have the unit s. Reflectance factors and asymmetry factors

are dimensionless. Asterisks indicate target values for reflectances, and value at final iteration for other quantities.

All reflectance measurements and scattering and absorption data used are taken from typical commercial paper or printed paper products in the visible range [45]. Thus all test cases are relevant paper industry problems, and they are typically used for design of new products and for online process control as well as for trade purposes and for colour management. Values of g in the visible range are not abundant in the literature. Since two-flux methods like Kubelka–Munk assume a perfectly isotropic single-scattering process (corresponding to $g = 0$) values of g close to zero are relevant to study. The few references relevant to paper in the visible range indicate g values in the range 0.6–0.8 [15,46], which is therefore used in this work.

In order to characterize the parameter estimation problem itself, not including any methods to solve it, some investigations were done to illustrate the curvature of the problem and the sensitivity of the solution. This was, of course, done separately for the d/0 and full inverse problems.

5.1. *Curvature of the problem*

The curvature of the problem was studied by plotting the objective function $F(x)$ for different test cases. For the d/0 problem, $F(x)$ was plotted as a function of the scattering and absorption parameters, while illumination and other circumstances followed paper industry standards [25–28]. For the full problem, $F(x)$ was plotted as a function of two of the parameters σ_s, σ_a and g while keeping the third fixed, alternating the parameter to be fixed. The simulated sample was here illuminated by perfectly diffuse light of intensity 0.3 and by a beam (or collimated light) of intensity 1.0 with a polar angle of incidence of 60°, and was placed on a black background. A well-behaved problem has a smooth and convex surface with one distinct minimum corresponding to the solution. Typical reasons for ill-conditioning in the solution include the existence of several local minima and lack of smoothness or convexity, which makes a global minimum hard to find. They also include a very flat surface, which gives poor convergence rate and a sensitive solution.

The left pane of Figure 1 indicates that cases with low opacity (defined as $Op = 100 \cdot R_0/R_\infty$) give well-conditioned problems with non-sensitive solutions. This is seen since the objective function surface is smooth and locally convex, and there is one distinct global minimum. Cases with high opacity (right pane of Figure 1), on the other hand, seem to have an objective function surface that is flat in one or more directions, which shows that those cases give ill-conditioned problems with poor convergence (hard to find iteration steps in the optimization that give sufficient descent in the flat areas) or sensitive solutions (a small change in target value can give a large change in the parameter solution). This was also numerically investigated by changing R_0 by 0.1%, and it was noted that the relative change in the parameter solution was a factor 7 larger in the high opacity case.

Figure 2 shows that the full problem is not convex, and indicates that problems with higher asymmetry factor g are more ill-conditioned. Although the objective function surface is smooth, the non-quadratic curvature and the local flatness along lines or curves indicate that the problem is ill-conditioned, possibly with poor convergence or sensitive solutions. It is also evident that there are ridges that will keep

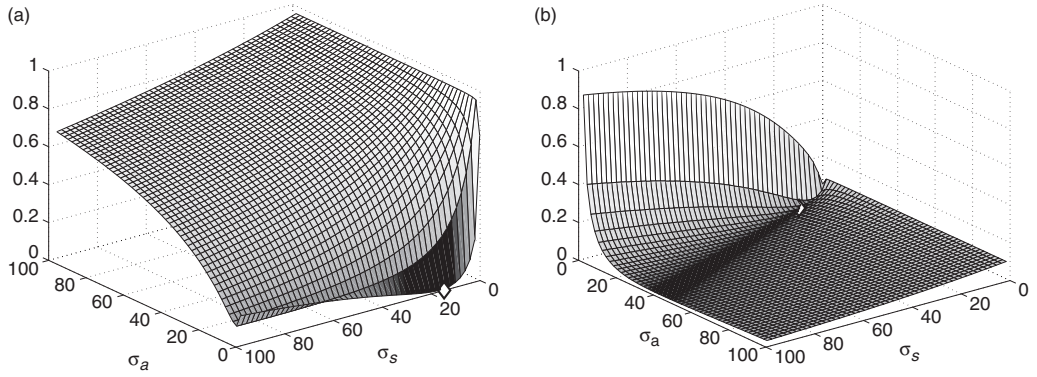

Figure 1. Two test cases for the d/0 problem. (a) ($\sigma_s = 14.7$, $\sigma_a = 0.03$, $g = 0$, $w = 0.1$, opacity = 50.1%): The objective function surface is smooth and locally convex with one distinct local minimum. The problem should be well conditioned with a non-sensitive solution. The solution is, however, close to a boundary, which may give rise to problems. (b) ($\sigma_s = 14.0$, $\sigma_a = 5.6$, $g = 0$, $w = 0.1$, opacity = 95.5%): The objective function surface is smooth, but is locally flat along a line, possibly giving poor convergence or a sensitive solution. The diamonds indicate the points of convergence.

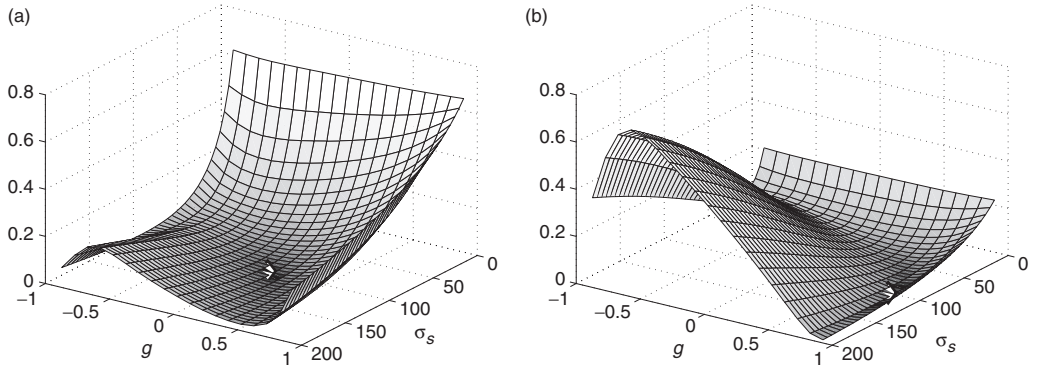

Figure 2. Two test cases for the full problem. Objective function surfaces as a function of σ_s and g with σ_a fixed. (a) ($\sigma_s = 100$, $\sigma_a = 20$, $g = 0.05$, $w = 0.03$). (b) ($\sigma_s = 100$, $\sigma_a = 20$, $g = 0.75$, $w = 0.03$). The objective function surface is smooth but non-convex, and is locally flat along a curve (especially for $g = 0.75$), possibly giving poor convergence or a sensitive solution. A poor starting point may lead to divergence because of the curvature. The diamonds indicate the points of convergence.

the optimization algorithm away from the area with the optimum, unless a sufficiently good starting point is provided. Of course, this figure does not give the whole picture. Allowing all three parameters to be varied simultaneously is the only way to investigate the real properties, but this is not easy to illustrate in a flat figure. Investigations show, however, that the ill-conditioning actually is a larger problem than indicated by Figure 2, especially when noise is included.

It is expected from application experience that cases with high opacity should give more ill-conditioned problems. In the standardized paper industry measurements used here, high opacity means that the two reflectance measurements are very similar

and in the end indistinguishable. This will of course give rise to sensitivity, and it is well known in the paper industry that parameter estimation is problematic for samples with high opacity. The conditioning of the problem also depends on the asymmetry factor, g, since higher g globally flattens the objective function surface, thereby making the problem more ill-conditioned and the solution more sensitive. It is also expected from theory and application experience that cases with high asymmetry factor should give more ill-conditioned problems. In atmospheric research, cases with strongly forward peaked scattering are common, and special methods have been developed to handle those situations. The extreme case of $g=1$ is a singularity that makes the phase function (the kernel in the integral) a Dirac delta function, which of course makes parameter estimation impossible. It should be emphasized that the ill-conditioning is not introduced by any simulation model used, but is inherent in the problem.

5.2. *Sensitivity of the solution*

The sensitivity of the d/0 solution was studied by generating contour plots of how σ_s and σ_a depend on R_0 and R_∞. Areas where the contour lines are close together are more sensitive, since a small change in a target value there may give a large change in the parameter solution. Studying the phase space plot gives visual information on which parameters are sensitive to what measurements, and to what extent in different areas. Generating contour plots in the full problem to illustrate the sensitivity of the solution is not as relevant as in the d/0 case. The d/0 case uses standardized measurements, so the contour plots are commonly and repeatedly useful. The full problem can have intensity measurements in any direction, and each direction will have its own contour plot – with its own different character. However, it is still possible to generate one, if the detailed analysis would be interesting in a specific case.

The sensitivity of the solution was also studied numerically, by calculating the sensitivity matrix κ introduced in Equation (33) for different cases. The elements κ_{ij} give quantitative information on the relative sensitivity of parameter x_i for change in measurement b_j, where $x=(\sigma_s,\ \sigma_a)^T$ and $b=(R_0,\ R_\infty)^T$ in the d/0 setting, and $x=(\sigma_s, \sigma_a, g)^T$ and b is a vector of intensities in different directions in the full problem.

The phase space plot and the sensitivity matrix can be used to interpret the results from the parameter estimation, by giving information on the influence of measurement errors or noise on the parameters. They can also be used to design better experiments and measurements by devising measurements, i.e. specific measurement directions, with minimal influence from noise and errors. It should be pointed out that the sensitivity is not introduced or affected by any simulation model used, but is a property of the problem itself.

For the d/0 problem, it can be seen from the left pane of Figure 3 that the scattering coefficient σ_s increases rapidly with R_0 but decreases slightly with R_∞, and that the rate of change is larger for strongly absorbing samples, i.e. the contour lines are closer together in the lower left part of the figure. This means that a small error in R_0 can cause large deviations in σ_s, and that the relative size of the deviation is larger for highly absorbing samples. It is also obvious that σ_s is highly sensitive to measurement errors in regions close to the line $R_0=R_\infty$, since the contour lines are very close together in this region. Once again, this illustrates the sensitivity in cases with high opacity. The absorption coefficient σ_a shows

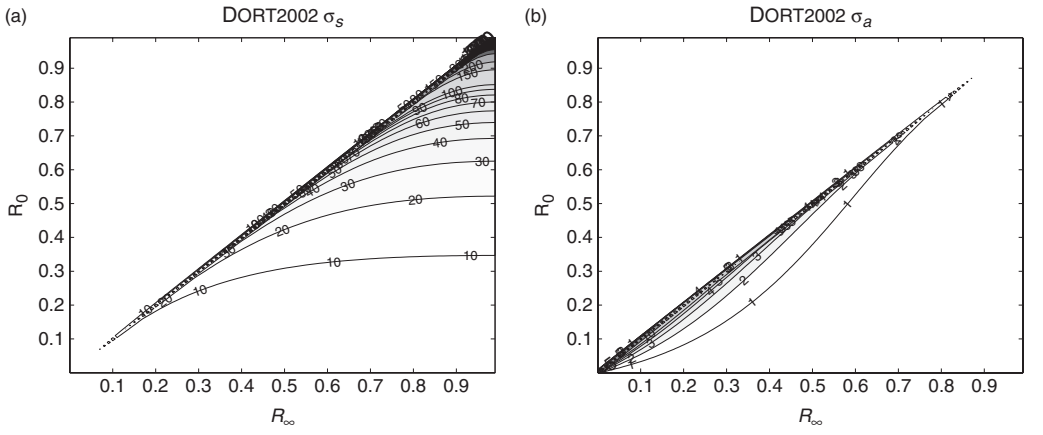

Figure 3. Contour plots showing how σ_s (a) and σ_a (b) depend on R_0 and R_∞ for $g = 0$ and $w = 0.1$. Darker grayscale corresponds to higher value. The parameters increase rapidly with R_0, but decrease slightly with R_∞. The problem is highly sensitive in regions near the line $R_0 = R_\infty$, i.e. in cases with very high opacity.

a similar dependence on the reflectances, i.e. it increases with R_0 and decreases with R_∞, and it is highly sensitive to measurement errors for strongly absorbing samples and in regions close to the line $R_0 = R_\infty$.

The relative sensitivities given in Table 1 for the d/0 problem show, for example, that σ_s is about five times more sensitive in the first and last cases (opacity around 95%) than in the middle case (opacity around 50%). Contrary to this reasoning, it is seen that the relative sensitivity of σ_a for change in R_∞ is much larger in the middle case. To explain this, the sensitivity matrix κ can be analysed in detail from relations (31). It is evident that the sensitivity depends on the condition given by $J^T J$, the curvature given by the Hessians f_i'', and the residuals f_i. Tests show that the residual in the solution is so small in the tested cases, that $J^T J$ dominates. However, in some cases the curvature is larger, as for σ_a in the middle case of Table 1 (see also left pane of Figure 1), which then shows as a higher sensitivity. The relative sensitivities given in Table 2 for the full problem show, for example, that g is about 10 times more sensitive than σ_s and σ_a. It is also evident that the relative sensitivities are much larger for large polar angles – in this particular case. If this were representative for a setting of repeated measurements, the obvious advice would be to design the experiments to use measurements with smaller polar angles to reduce the sensitivity for measurement errors and noise. However, one still should not use too few measurements, since this would increase the sensitivity to individual measurement errors.

6. Numerical experiments: comparison of optimization methods for parameter estimation

6.1. *The d/0 parameter estimation problem*

Algorithm 4 (`Inverse DORT2002 d/0`) was applied to a set of d/0 parameter estimation test problems, but with different optimization methods in place of Algorithm 3 (`Gauss-Newton`). The methods were then compared with respect to speed and accuracy. Three different test cases were used, each with two different values of g. The test problems and the results are found in the following tables.

Table 1. Relative sensitivity of parameter i ($1=\sigma_s$, $2=\sigma_a$) for change in measurement j ($1=R_0$, $2=R_\infty$) in three test cases.

	$R_0^*=0.21$, $R_\infty^*=0.22$ $g=0$, $w=0.1$	$R_0^*=0.44$, $R_\infty^*=0.88$ $g=0$, $w=0.1$	$R_0^*=0.21$, $R_\infty^*=0.22$ $g=0.8$, $w=0.1$,
κ_{ij}	6.7941 −5.8564	1.7347 −0.1408	6.2974 −5.5234
	6.7941 −7.4877	1.7332 −16.5015	6.3093 −6.9964

Table 2. Relative sensitivity of parameter i ($1=\sigma_s$, $2=\sigma_a$, $3=g$) for change in measurement j (1–5 representing polar angles 84.3°, 72.5°, 60.0°, 45.6° and 25.8°) for the $g=0.05$ case.

	0.0375	−0.2098	4.7044
	0.0241	−0.1395	2.6550
κ_{ij}	0.0144	−0.0898	1.3215
	0.0075	0.0565	0.4599
	0.0021	−0.0329	−0.1644

Note: Only azimuthal angle 0° (the same direction as the incident beam) is presented here.

From the results in Tables 3–8, it can be noted that the optimization problem is not at all trivial. One case did not converge for `Newton`, and two cases led to a singularity for `Quasi-Newton`. One case converged to a non-global minimum for `fmincon`, and in another it went out of bounds. Thus, the nature of the problem demands that the optimization methods handle this efficiently.

As can be seen, only the two Gauss-Newton methods – the simple implementation `Gauss-Newton` and the Matlab function `lsqnonlin` – converged for all test cases. In addition to this, the two Gauss-Newton methods were by far the fastest in all cases, and had good accuracy. On average, `Gauss-Newton` was almost twice as fast as `lsqnonlin`, probably because the Matlab functions are written to be general and robust. However, it was not expected that `Newton` and `Quasi-Newton` would be so much slower than `Gauss-Newton`. Based on the numerical tests, Algorithm 3 (`Gauss-Newton`) was the obvious choice for use in Algorithm 4 (`Inverse DORT2002 d/0`).

As noted in Section 2.2 for the forward d/0 problem, the standardized d/0° instrument geometry is specifically taken into account in DORT2002. This gives a higher accuracy than with the Kubelka–Munk method for the inverse d/0 problem too. Therefore, it can be noted again that, although only a side effect in the context of this work, this actually makes it reasonable to recommend DORT2002 instead of Kubelka–Munk when higher accuracy is needed.

6.2. *The full parameter estimation problem*

Two test cases were used when investigating the full parameter estimation problem. The measurements were synthetic, i.e. they were generated from a model, so that the 'true'

Table 3. Test case: ($R_0^* = 0.42$, $R_\infty^* = 0.67$, $g = 0$, $w = 0.1$).

Method	σ_s^*	σ_a^*	F^*	Iterations	Func. evals	Time
Newton	14.7600	0.30546	2.0402e−012	5	43	0.1456
Quasi-Newton	14.7602	0.30546	2.6768e−014	8	32	0.1143
Gauss-Newton	14.7602	0.30546	1.1648e−016	3	12	0.0383
lsqnonlin	14.7603	0.30547	3.2175e−012	3	12	0.0602
fmincon	14.7603	0.30546	8.7934e−015	19	60	0.1962

Table 4. Test case: ($R_0^* = 0.21$, $R_\infty^* = 0.22$, $g = 0$, $w = 0.1$).

Method	σ_s^*	σ_a^*	F^*	Iterations	Func. evals	Time
Newton	13.9901	5.58910	4.6862e−010	22	461	1.1559
Quasi-Newton	14.0087	5.59738	8.8701e−013	16	77	0.2011
Gauss-Newton	14.0080	5.59710	1.0888e−014	3	12	0.0371
lsqnonlin	14.0054	5.59595	2.0481e−011	4	15	0.0643
fmincon	10.0409	3.81456	6.8173e−005	6	21	0.0815

Table 5. Test case: ($R_0^* = 0.44$, $R_\infty^* = 0.88$, $g = 0$, $w = 0.1$).

Method	σ_s^*	σ_a^*	F^*	Iterations	Func. evals	Time
Newton	14.7246	0.02810	3.9438e−015	20	284	0.7482
Quasi-Newton	14.7246	0.02810	8.8381e−019	9	35	0.0959
Gauss-Newton	14.7246	0.02810	4.8130e−015	3	12	0.0356
lsqnonlin	14.7247	0.02810	1.3776e−011	3	12	0.0598
fmincon	14.7254	0.02811	3.1652e−010	17	55	0.1889

Table 6. Test case: ($R_0^* = 0.42$, $R_\infty^* = 0.67$, $g = 0.8$, $w = 0.1$).

Method	σ_s^*	σ_a^*	F^*	Iterations	Func. evals	Time
Newton	80.6102	0.29729	4.2290e−011	355	5842	14.0188
Quasi-Newton	Did not converge – reached singularity					
Gauss-Newton	80.6072	0.29727	7.4665e−014	3	12	0.0466
lsqnonlin	80.6071	0.29727	4.4042e−019	4	15	0.0751
fmincon	80.6066	0.29726	3.1009e−012	21	67	0.2209

Table 7. Test case: ($R_0^* = 0.21$, $R_\infty^* = 0.22$, $g = 0.8$, $w = 0.1$).

Method	σ_s^*	σ_a^*	F^*	Iterations	Func. evals	Time
Newton	83.1915	4.99682	1.2075e−008	58	1009	2.5279
Quasi-Newton	83.7219	5.03223	6.4586e−011	20	133	0.3536
Gauss-Newton	83.6853	5.02978	2.2660e−014	3	12	0.0366
lsqnonlin	83.4505	5.01432	5.4672e−009	3	12	0.0550
fmincon	Optimization went out of bounds ($\sigma_a < 0$)					

Table 8. Test case: ($R_0^* = 0.44$, $R_\infty^* = 0.88$, $g = 0.8$, $w = 0.1$).

Method	σ_s^*	σ_a^*	F^*	Iterations	Func. evals	Time
`Newton`	Solution was not found after 1000 iterations					
`Quasi-Newton`	Did not converge – reached singularity					
`Gauss-Newton`	79.6665	0.02677	1.2130e−016	5	18	0.0520
`lsqnonlin`	79.6679	0.02678	3.1272e−010	3	12	0.0546
`fmincon`	79.6659	0.02677	1.4196e−011	21	67	0.4052

parameter values were known. The difference between the cases was the asymmetry factor, which took on the values of $g = 0.05$ and $g = 0.75$. The other parameters were $\sigma_s = 100\,\mathrm{m^2\,kg^{-1}}$ and $\sigma_a = 20\,\mathrm{m^2\,kg^{-1}}$. The simulated sample had a grammage of $0.03\,\mathrm{kg\,m^{-2}}$, was illuminated by perfectly diffuse light of intensity 0.3 and by a beam (or collimated light) of intensity 1.0 with a polar angle of incidence of 60°, and was placed on a black background. Intensities were measured at six polar angles and six azimuthal angles, as described in Tables 9 and 10. Gaussian noise was added to the measurements, with zero mean and three different standard deviations (0%, 1% and 5% of the intensity). The noise levels were chosen based on the fact that reflectance factor measurements are required by standards to have errors not exceeding 1%. Angle-resolved measurements are not standardized, but somewhat higher errors should be expected there.

6.2.1. *The first phase*

Algorithm 5 (`Phase 1`) was applied to the first phase of the test problems, to determine a good starting point for the second phase. Since the first phase only consists of a scalar optimization that uses a derivative free golden section search method with a relaxed convergence criterion (parameter tolerance of 10^{-2}), only a few iterations were needed for convergence. Since the objective function only used the d/0 forward problem, it was also cheap, and convergence was fast. Note that the user does not have to supply a starting point for the first phase, since the golden section search uses the entire admissible $]-1, 1[$ as start interval for g. The convergence in the two test cases is illustrated in Table 11, where no noise was included in the measurements. Tests showed that noise had little influence on the convergence here, since the convergence criterion is not so firm. It is clear from Table 11 that in just a few cheap iterations the first phase gives a very good starting point for the second phase.

6.2.2. *The second phase*

Algorithm 6 (`Inverse DORT2002`) was applied to the test problems, but with different optimization methods in place of Algorithm 3 (`Gauss-Newton`). It should be noted that no starting point is needed from the user. The methods were then compared with respect to speed and accuracy. All three noise levels were used. Since the first phase provided such a good starting point, Algorithm 6 (`Inverse DORT2002`) most often converged in just a few iterations, even when using only simple implementations for the optimization methods. Exceptions to this were noted, possibly due to the influence of noise. The convergence in the test cases is illustrated in Tables 12–17.

Table 9. Angle-resolved intensity measurements without noise for the $g = 0.05$ case for the different polar angles θ and azimuthal angles φ (zero being the same direction as the incident beam).

	φ					
θ (°)	0°	60°	120°	180°	240°	300°
84.3	0.2576	0.2534	0.2462	0.2432	0.2462	0.2534
72.5	0.2204	0.2175	0.2124	0.2103	0.2124	0.2175
60.0	0.1928	0.1907	0.1872	0.1856	0.1872	0.1907
45.6	0.1712	0.1699	0.1675	0.1664	0.1675	0.1699
25.8	0.1537	0.1530	0.1518	0.1512	0.1518	0.1530

Table 10. Angle-resolved intensity measurements without noise for the $g = 0.75$ case for the different polar angles θ and azimuthal angles φ (zero being the same direction as the incident beam).

	φ					
θ (°)	0°	60°	120°	180°	240°	300°
84.3	0.3459	0.1918	0.1505	0.1440	0.1505	0.1918
72.5	0.2152	0.1434	0.1116	0.1056	0.1116	0.1434
60.0	0.1407	0.1086	0.0878	0.0832	0.0878	0.1086
45.6	0.0977	0.0845	0.0724	0.0692	0.0724	0.0845
25.8	0.0719	0.0680	0.0628	0.0610	0.0628	0.0680

Table 11. Convergence of Algorithm 5 (`Phase 1`) with no noise in the measurements.

Case	σ_s^*	σ_a^*	g^*	F^*	Func. evals	Time
$g = 0.05$	99.8949	20.0032	0.0491377	7.78257e−008	6	1.44267
$g = 0.75$	99.5200	20.0003	0.7489060	6.86673e−008	10	2.73832

Table 12. Test case: noise standard deviation 0% and $g = 0.05$.

Method	σ_s^*	σ_a^*	g^*	F^*	Iterations	Func. evals	Time
`Quasi-Newton`	99.9999	20	0.05	2.7379e−016	2	18	3.9662
`Gauss-Newton`	100	20	0.05	1.5488e−022	2	18	3.6155
`lsqnonlin`	100.1303	20.0257	0.0500	3.2993e−010	1	14	3.9414
`fmincon`	99.8949	20.0032	0.0491	7.7197e−008	1	14	3.7324

Table 13. Test case: noise standard deviation 0% and $g = 0.75$.

Method	σ_s^*	σ_a^*	g^*	F^*	Iterations	Func. evals	Time
`Quasi-Newton`	99.9999	20	0.75	3.8000e−015	3	26	6.4360
`Gauss-Newton`	100	20	0.75	5.3074e−017	2	22	5.4780
`lsqnonlin`	99.8821	19.9814	0.7499	3.6637e−009	1	18	4.6177
`fmincon`	99.52	20.0003	0.7487	3.2048e−008	2	22	5.4998

Table 14. Test case: noise standard deviation 1% and $g = 0.05$.

Method	σ_s^*	σ_a^*	g^*	F^*	Iterations	Func. evals	Time
`Quasi-Newton`	96.6564	19.3549	0.0549	4.2597e−005	9	48	9.1010
`Gauss-Newton`	96.6707	19.3577	0.0549	4.2597e−005	3	23	4.4370
`lsqnonlin`	128.7825	25.6931	0.0563	5.0109e−005	2	18	3.5425
`fmincon`	128.3294	25.6943	0.0540	5.0694e−005	3	22	4.2755

Table 15. Test case: noise standard deviation 1% and $g = 0.75$.

Method	σ_s^*	σ_a^*	g^*	F^*	Iterations	Func. evals	Time
`Quasi-Newton`	95.6613	20.0528	0.7380	2.1582e−005	13	83	19.6580
`Gauss-Newton`	95.6647	20.0528	0.7380	2.1582e−005	3	25	6.0557
`lsqnonlin`	96.5066	20.1835	0.7383	2.1764e−005	4	29	7.1169
`fmincon`	89.2461	19.2873	0.7318	3.0648e−005	3	25	6.0612

Table 16. Test case: noise standard deviation 5% and $g = 0.05$.

Method	σ_s^*	σ_a^*	g^*	F^*	Iterations	Func. evals	Time
`Quasi-Newton`	102.5256	19.7510	0.1033	0.001870	6	54	10.1423
`Gauss-Newton`	102.9187	19.8251	0.1033	0.001870	12	127	23.2268
`lsqnonlin`	97.1308	18.7376	0.1028	0.001871	4	26	4.9741
`fmincon`	91.4524	17.6786	0.1019	0.001875	13	62	11.4656

Table 17. Test case: noise standard deviation 5% and $g = 0.75$.

Method	σ_s^*	σ_a^*	g^*	F^*	Iterations	Func. evals	Time
`Quasi-Newton`	82.0474	20.1069	0.6972	0.000394	15	97	23.1320
`Gauss-Newton`	82.0482	20.1070	0.6972	0.000394	5	33	8.1717
`lsqnonlin`	82.1929	20.1367	0.6972	0.000394	5	33	8.1972
`fmincon`	67.2058	18.7646	0.6574	0.000472	14	69	16.9475

As can be seen, with no noise the methods are very similar with respect to accuracy and time. With increasing noise levels, the two commercial methods (the Matlab functions `lsqnonlin` and `fmincon`, especially the latter) tend to converge to other points further from the 'true' value but with comparable objective value. This is probably due to the 'flatness' of the objective function in combination with the multiple convergence criteria of the commercial solvers. With increasing noise levels, `Quasi-Newton` tends to need relatively more iterations and more time. With the exception of one case, `Gauss-Newton` was both the fastest and the most accurate. Based on the numerical tests, Algorithm 3 (`Gauss-Newton`) was the obvious choice for use in Algorithm 6 (`Inverse DORT2002`).

Two other comments should be given regarding the Newton methods. It may be noted that the pure Newton method is not included in the presented investigation of Algorithm 6 (`Inverse DORT2002`). This is because it often had problems, not giving descent directions or not having a positive definite Hessian. Although there are methods for handling this, it would lead away from the aim of using simple implementations. For `Quasi-Newton`, it should be noted that the initial Hessian approximation is not done with finite differences, since that often gives a Hessian that is not positive definite. Instead, the initial Hessian was approximated using the Jacobian from `Gauss-Newton`. The guaranteed positive definiteness is then preserved in the BFGS updates.

A final remark on Algorithm 6 (`Inverse DORT2002`) should be made, to emphasize the successful two-phase approach. The initial attempts were to attack the full problem directly. It seemed to be impossible to get convergence, and careful investigations were made. It turned out that the starting point (for which there is no obvious way to choose g) was a much greater obstacle than the noise. The quadratic model approximations in the initial points were too far from the real problem, due to its curvature, for the minimization to work properly. Gauss-Newton would find local minima far away. The Newton methods would, in addition to this, generate poor Hessian approximations due to noisy measurements, leading to poor search directions, small step lengths and bad convergence. The idea of introducing the first step with a parameterized scalar optimization using the d/0 problem not only solved the problem of finding a suitable starting point, but actually provided the key for this efficient two-phase solution method for the full parameter estimation problem.

It is appropriate to give a comment on the conditioning in the solution to the full parameter estimation problem. On average the condition number of the Hessian in the solution was 10^7, which shows that the problem is indeed ill conditioned there. This can be compared with the d/0 problem, where the condition number was 2–3 decades smaller. This shows that although the d/0 problem also is ill conditioned in the solution, the introduction of the third parameter g in the parameter estimation problem increased this ill-conditioning.

7. Discussion

Calculating material parameters from reflectance measurements is an outstanding issue in general radiative transfer problems. Finding a feasible starting point can in itself be a great problem in many applications. To the author's knowledge, the kind of two-phase approach for the inverse radiative transfer problem presented in this work has not been published before. It is not uncommon to first estimate σ_s and σ_a from some measurements and an assumption or guess of g, and then use these parameter values as the starting point for the full problem. However, this is far from optimal, since the values of σ_s and σ_a are fixed from the more or less *ad hoc* choice of g. That first step merely ensures that the triplet σ_s, σ_a and g is compatible. No use is made of the full problem or the rest of the measurements when obtaining the starting point; instead, the starting point and the rest of the problem are treated completely separately. The first phase suggested here manages to actually find the best (in some aspects) starting point by actually combining – through the parameterization – the simpler d/0 problem with the objective function of the full problem, thus also utilizing all measurements. This intelligent construction of the first phase made it fast and simple, while still providing a very good starting point for the full problem.

The reported difficulties of finding a suitable starting point (for which the full problem is very sensitive) are thus eliminated, and the user is relieved from having to supply an initial value of g through guesswork, and without knowledge. The starting point from the first phase then made it possible to use a simple and straightforward Gauss-Newton method in the second phase, which gave fast convergence and accurate results. The successful recovery of σ_s, σ_a and g by this two-phase method was illustrated by application to relevant paper industry problems.

One goal of this work was simplicity in implementation. The numerical tests show that this was indeed possible. The first phase is a simple golden section search method, and the second phase (as well as the d/0 case) is a simple Gauss-Newton method, both of which need no more than a few lines of code. The commercial solvers did not do any better, measuring any of convergence, final residual, iterations, function evaluations or computation time. The simple Gauss-Newton method did as well as or better than all the other tested methods, standard or commercial, in both the d/0 and full inverse problems.

Another purpose of this work was to find some characteristics of the studied parameter estimation problem. The investigations showed that the optimization problem is not at all trivial, and that the nature of the ill-conditioned problem thus puts great demands on the optimization methods. It was also shown that higher opacity or higher asymmetry factor increases the ill-conditioning, and that also estimating the asymmetry factor (not only the scattering and absorption parameters) increases the condition number several decades. However, the relative sensitivities found in the investigations are low, and should not be a problem in paper industry applications. The type of analysis made in this work – using objective function surface plots, phase space plots and sensitivity matrices – give good insight into the character of the problem. Similar studies will be valuable in the design of instruments for angle-resolved measurements and in analysis of the corresponding measurement data. This will be particularly important for the paper industry, where anisotropy is gradually becoming an issue. Feasibility and demands for industrial implementation of angle-resolved measurements are in initial discussions between European paper industries and institutes. While online industrial implementation may lie a few years in the future, the use in research labs is already there and will necessarily increase to attack anisotropy issues. Some Swedish paper industries have already started in laboratory scale.

The Kubelka–Munk model is a simple solution method for the radiative transfer problem, and it is well established in several industrial applications due to its speed and ease of use. Its approximate solutions are sufficiently accurate in many applications, but there are a number of problems and applications where higher accuracy is needed. This can be achieved with more general solution methods for the radiative transfer problem, like DORT2002. To compete with Kubelka–Munk in industrial applications, higher accuracy is not enough. Sufficient speed is essential, as well as fast and accurate parameter estimation methods. The specialized code for forward and inverse standardized d/0° reflectance calculations reported in this work show that radiative transfer based solution methods like DORT2002 are now competitive in paper industry applications. With its higher accuracy, larger range of applicability and comparable speed, DORT2002 could well replace Kubelka–Munk in the paper industry for example, for increased understanding. However, the Kubelka–Munk model will still be useful in applications with lower demands on accuracy. When it is now possible to estimate the asymmetry factor g, the possibility of studying new material phenomena such as different direction dependencies in reflectance

and transmittance arises. This opens up for increased understanding, but also for the design of new paper and printed products.

Acknowledgement

This work was financially supported by T2F ('TryckTeknisk Forskning', a Swedish printing research program), which is gratefully acknowledged.

References

[1] P. Kubelka and F. Munk, *Ein beitrag zur optik der Farbanstriche*, Z. Tech. Phys. 11a (1931), pp. 593–601.
[2] P. Kubelka, *New contributions to the optics of intensely light-scattering materials. Part I*, J. Opt. Soc. Amer. 38 (1948), pp. 448–457.
[3] ——, *New Contributions to the optics of intensely light-scattering materials. Part II*, J. Opt. Soc. Amer. 44 (1954), pp. 330–335.
[4] H. Granberg and P. Edström, *Quantification of the intrinsic error of the Kubelka–Munk model caused by strong light absorption*, J. Pulp Paper Sci. 29 (2003), pp. 386–390.
[5] P. Edström, *Comparison of the DORT2002 radiative transfer solution method and the Kubelka–Munk model*, Nordic Pulp Paper Res. J. 19 (2004), pp. 397–403.
[6] M. Neuman, *Anisotropic reflectance from paper – measurements, simulations and analysis*, Master's Thesis, Umeå University, Sweden, 2005.
[7] L. Yang and B. Kruse, *Revised Kubelka–Munk theory. I. Theory and applications*, J. Opt. Soc. Am. A 21 (2004), pp. 1933–1941.
[8] L. Yang, B. Kruse, and S.J. Miklavcic, *Revised Kubelka–Munk theory. II. Unified framework for homogenous and inhomogenous optical media*, J. Opt. Soc. Am. A 21 (2004), pp. 1942–1952.
[9] L. Yang and S.J. Miklavcic, *A theory of light propagation incorporating scattering and absorption in turbid media*, Opt. Lett. 30 (2005), pp. 792–794.
[10] ——, *Revised Kubelka–Munk theory. III. A general theory of light propagation in scattering and absorptive media*, J. Opt. Soc. Am. A 22 (2005), pp. 1866–1873.
[11] P. Edström, *Examination of the revised Kubelka–Munk theory: considerations of modeling strategies*, J. Opt. Soc. Am. A 24 (2007), pp. 548–556.
[12] A.B. Murphy, *Optical properties of an optically rough coating from inversion of diffuse reflectance measurements*, Appl. Opt. 46 (2007), pp. 3133–3143.
[13] M.J.C. van Gemert and W.M. Star, *Relations between the Kubelka–Munk and the transport equation models for anisotropic scattering*, Lasers Life Sci. 1 (1987), pp. 287–298.
[14] S.A. Prahl, M.J.C. van Gemert, and A.J. Welch, *Determining the optical properties of turbid media using the adding-doubling method*, Appl. Opt. 32 (1993), pp. 559–568.
[15] N. Joshi, C. Donner, and H.W. Jensen, *Noninvasive measurement of scattering anisotropy in turbid materials by nonnormal incident illumination*, Opt. Lett. 31 (2006), pp. 936–938.
[16] C.E. Siewert, *Inverse solutions to radiative-transfer problems based on the binomial or the Henyey–Greenstein scattering law*, J. Quant. Spectrosc. Radiat. Transfer 72 (2002), pp. 827–835.
[17] N.J. McCormick, *Transport scattering coefficients from reflection and transmission measurements*, J. Math. Phys. 20 (1979), pp. 1504–1507.
[18] C.E. Siewert, *Inverse solutions to radiative-transfer problems with partially transparent boundaries and diffuse reflection*, J. Quant. Spectrosc. Radiat. Transfer 72 (2002), pp. 299–313.
[19] K. Stamnes, S.C. Tsay, and I. Laszlo, *DISORT, a General-purpose fortran program for discrete-ordinate-method radiative transfer in scattering and emitting layered media*, NASA report, 2000.
[20] K. Stamnes, S.C. Tsay, et al., *Numerically stable algorithm for discrete-ordinate-method radiative transfer in multiple scattering and emitting layered media*, Appl. Opt. 27 (1988), pp. 2502–2509.

[21] G.E. Thomas and K. Stamnes, *Radiative Transfer in the Atmosphere and Ocean*, Cambridge University Press, Cambridge, 1999.
[22] C.E. Siewert, *A concise and accurate solution to Chandrasekhar's basic problem in radiative transfer*, J. Quant. Spectrosc. Radiat. Transfer 64 (2000), pp. 109–130.
[23] L.B. Barichello and C.E. Siewert, *The searchlight problem for radiative transfer in a finite slab*, J. Comput. Phys. 157 (2000), pp. 707–726.
[24] P. Edström, *A fast and stable solution method for the radiative transfer problem*, SIAM Rev. 47 (2005), pp. 447–468.
[25] ISO 2469, *Paper, board and pulps – measurement of diffuse reflectance factor*, International Organization for Standardization, Geneva, Switzerland, 1994.
[26] ISO 2470, *Paper, board and pulps – measurement of diffuse blue reflectance factor (ISO brightness)*, International Organization for Standardization, Geneva, Switzerland, 1999.
[27] ISO 2471, *Paper and board – determination of opacity (paper backing) – Diffuse reflectance method*, International Organization for Standardization, Geneva, Switzerland, 1998.
[28] ISO 9416, *Paper – determination of light scattering and absorption coefficients (using Kubelka–Munk theory)*, International Organization for Standardization, Geneva, Switzerland, 1998.
[29] L.G. Henyey and J.L. Greenstein, *Diffuse radiation in the galaxy*, Astrophys. J. 93 (1941), pp. 70–83.
[30] P. Edström, *Numerical performance of stability enhancing and speed increasing steps in radiative transfer solution methods,* J. Comput. Appl. Math. (2007).
[31] K.M. Case and P.F. Zweifel, *Linear Transport Theory*, Addison-Wesley, London, 1967.
[32] M. Choulli and P. Stefanov, *An inverse boundary value problem for the stationary transport equation*, Osaka J. Math. 36 (1998), pp. 87–104.
[33] L. Armijo, *Minimization of functions having continuous partial derivatives*, Pacific J. Math. 16 (1966), pp. 1–3.
[34] S.G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw Hill, Singapore, 1999.
[35] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, 1999.
[36] C.G. Broyden, *The convergence of a class of double-rank minimization algorithms*, J. Inst. Math. Applic. 6 (1970), pp. 76–90.
[37] R. Fletcher, *A new approach to variable metric algorithms*, Computer J. 13 (1970), pp. 317–322.
[38] D. Goldfarb, *A family of variable metric updates derived by variational means*, Math. Comp. 24 (1970), pp. 23–26.
[39] D.F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp. 24 (1970), pp. 647–656.
[40] J.J. Moré, *The Levenberg–Marquardt algorithm: implementation and theory*, in *Numerical Analysis, Lecture Notes in Mathematics 630*, G.A. Watson, ed., Springer Verlag, Berlin, 1977, pp. 105–116.
[41] K. Schittkowski, *NLQPL: A FORTRAN- subroutine solving constrained nonlinear programming problems*, Ann. Oper. Res. 5 (1985), pp. 485–500.
[42] P.S. Mudgett and L.W. Richards, *Multiple scattering calculations for technology*, Appl. Opt. 10 (1971), pp. 1485–1502.
[43] ——, *Multiple scattering calculations for technology II*, J. Colloid. Interf. Sci. 39 (1972), pp. 551–567.
[44] H.C. van de Hulst, *Multiple Light Scattering. Tables, Formulas and Applications*, Vol. 2, Academic Press, New York, 1980.
[45] N. Pauler, *Paper Optic*, Lorentzen & Wettre, Kista, Sweden, 2002.
[46] H. Granberg and M.-C. Béland, *Modelling the angle-dependent light scattering from sheets of pulp fibre fragments*, Nordic Pulp Paper Res. J. 19 (2004), pp. 354–359.