IOP Conf. Series: Journal of Physics: Conf. Series **810** (2017) 012043 doi:10.1088/1742-6596/810/1/012043

The problem of numerical precision in spectral line shapes calculations

P Ablewski, P Wcisło and R Ciuryło

Institute of Physics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University in Toruń, Grudziadzka 5, 87-100, Toruń

E-mail: piotra@fizyka.umk.pl

Abstract. Spectral line shapes can be described by the transport-relaxation equation (TRE). When *ab initio* collisional operator is incorporated, the TRE needs to be solved numericaly. We report a pure numerical problems encountered during tests of the iterative approach to solving the TRE. As a reference we have used Voigt profile, which can be easily calculated analytically with error function, as well as numerically by solving TRE with simple collisional operator. Our studies lead us to the conclusions about impact of numerical precion and matrix operators dimensions on the accuracy of the calculations.

1. Introduction

Modern high resolution spectroscopy provides experimental data with extremely high singnal-to-noise ratio [1, 2]. The theoretical analysis of these data should come with an accuracy at comparable level. The realistic theory must take into account many physical aspects and in ideal case be based on interaction potentials obtained with *ab initio* calculations [4, 3]. It is possible to find exact analytical formula describing spectral line shape. However, in general this is complicated or even impossible to solve analytically. In that case there is a need to find a stable and reliable numerical method providing a solution with the smallest posibble errors in a finite number of steps. The standard approach [5, 6] converts the physical problem to the set of linear equations which can be easly solved for high and middle pressures, where collsional width is larger or comparable to Doppler one. It falls down in cases of low pressures [7], where the Doppler effect is dominant. Iterative method of solving TRE [7] enables calculations for a wide range of pressures, including zero-pressure limit.

2. Iterative approach to stationary solution of transport-relaxation equation

According to [6], it is convinient to write TRE in as

$$1 = -i(\omega - \omega_0 - \vec{k}\cdot\vec{v})h(\omega,\vec{v}) - \hat{S}^f h(\omega,\vec{v}), \tag{1}$$

where ω_0 - ω is detuning from the resonant frequency, $\vec{k}$ is the wave wector, $\vec{v}$ is the absorber velocity and $\hat{S}^f$ is the collisional operator describing changes of both internal and translational state under collisions. The function $h(\omega,\vec{v})$ can be used to calculate the line-shape profile as

$$I(\omega) = \frac{1}{\pi} Re \int d^3\vec{v}\ f_m(\vec{v}) h(\omega,\vec{v}), \tag{2}$$

where $f_m(\vec{v})$ is the Maxwellian velocity distribution. In the case of low pressures, where the usual solution of TRE fails, there is a need to modify Eq. (1) to ensure that the iterative process coverges. This can be done by introduction of nonphysical parameter Γ_{num} to Eq. (1)

$$1 = [\Gamma_{num} - i(\omega - \omega_0 - \vec{k} \cdot \vec{v})h(\omega, \vec{v})] - (\Gamma_{num} + \hat{S}^f)h(\omega, \vec{v}). \tag{3}$$

After some mathematical operations and decomposition of operators in finite-dimensional Burnet basis [8, 9], iterative form of the solution of TRE can be written as [7]

$$\mathbf{c}^{n+1}(\omega) = \mathbf{a}(\omega) + \mathbf{A}(\omega) \cdot \mathbf{B} \cdot \mathbf{c}^n(\omega), \tag{4}$$

where $\mathbf{c}^n$ is a vector of solutions in n-th iterative step, starting from $\mathbf{c}^0[0] = 1$ and $\mathbf{c}^0[\neq 0] = 0$, $\mathbf{A}$ is a complex matrix depending on detuning from resonance and Doppler shift, $\mathbf{B}$ is a matrix describing collisions and relaxation process (matrix form of collisional operator decomposed in Burnett basis) and $\mathbf{a}$ is a first column of $\mathbf{A}$ matrix. The dimensions of matrices are $(N_{max} \cdot L_{max}) \times (N_{max} \cdot L_{max})$ corresponding to the dimensions of Burnet function basis used for decompositon of the operators. These functions are enumerated by a pair of indexes n, l such that $n = 0, \cdots, N_{max}$ and $l = 0, \cdots, L_{max}$. Further explanation of decomposition and forms of matrix operators can be found in [7].

3. Results

We have performed tests of the iterative method of solving transport-relaxation equation for Voigt profile, which has a simple analytical formula given by Faddeeva function [10]. In this case, the collisional operator $\hat{S}^f$ acts as a scalar quantity equal to $-\Gamma$, where Γ is a collisional half width at half maximum (HWHM) of the spectral line. Doppler effect is treated independently of collisional operator and taken into account in $\mathbf{A}(\omega)$.

In our calculations the ratio between collisional and Doppler line width was equal to $\frac{\Gamma}{\omega_D} = 1$. We have defined the iterative coverence as $|c^n[0] - c^{n-1}[0]|$ for different dimensions of operators matrices. We have used a single numerical precision `binary32` defined by IEEE 754-2008 standard [11].

As shown in Fig. 1, numerical accuracy of method depends on matrices' dimensions. For small dimensions of matrices, the convergence criterion leads to the proper solution. Further iterations do not destroy the agreement with analytical form of the used profile. For the optimal base dimension in single precision calculations, cnovergence criterion does not lead to the minimum of the difference between analytical and numerical solution. Interesting fact is that this difference stabilizes at the same level during further iterations. In case of too large base, range between minimum and maximum value in $\mathbf{A}$ matrix can not be covered by representation of floating point values in the used arithmetic. This causes the destabilization of iterative process and leads to a wrong solution. It means that it is not possible to arbitrarily extend matrixces' dimensions using the same numerical precision. This effect is caused by accumulation of numerical erros and finite representation of significand in floating point arithmetics.

4. Conclusions

We have performed simple test to check why extending of the dimensions of matrix operators does not lead to the better convergence of the calculations. The main problem is numerical precision which is not able to handle range of the values of perturbative corrections. There is only one remedy for that problem - use of larger precision, which could encode value with more significant digits. In this paper we have formed the research problem such that iterative approach to solving TRE for spectral line shapes calculations needs to put more attention on numerical precision and values of nonphysical parameters.

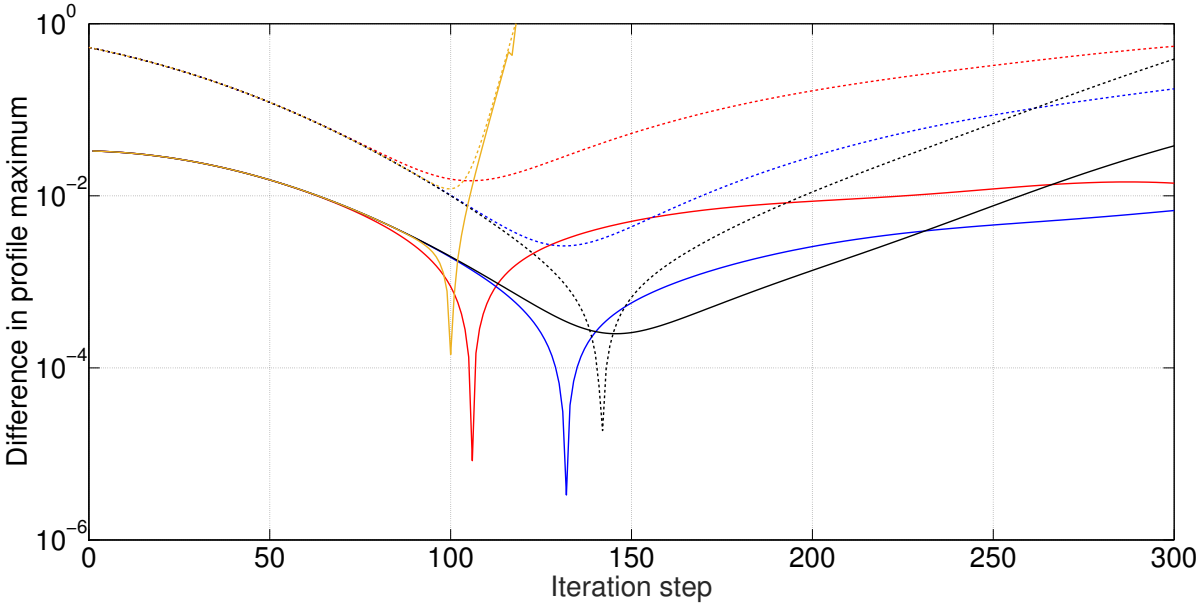

Figure 1. Evolution of the iterative process as a function of the number of steps. Solid lines indicate iterative process covergence defined as $|c^n[0] - c^{n-1}[0]|$ and dashed lines show difference between analytical solution and value obtained by iterative method in n-th step. Colors refer to $L_{max} = N_{max}$ dimensions as following: red - 4, blue - 6, black - 7, yellow - 8.

Acknowledgments

This work is supported by the National Science Centre, Poland, Project No. 2015/19/D/ST2/02195. The research is part of the program of the National Laboratory FAMO in Toruń, Poland. It is also supported by the COST Action CM1405 MOLIM.

References

[1] Cygan A, Lisak D, Wójtewicz Sz, Domysławska J, Hodges J T, Trawiński R S and Ciuryło R 2012 *Phys. Rev. A* **85** 022508
[2] Lin H, Reed Z D, Sirinneau V T and Hodges J T 2015 *J. Quant. Spectrosc. Radiat. Transfer* **161** 11
[3] May D A, Liu W K, McCourt F R W, Ciuryło R, Sanchez-Fortún Stoker J, Shapiro D and Wehr R 2013 *Can. J. Phys.* **91** 879
[4] Blackmore R, Green S and Monchik L 1989 *J. Chem. Phys.* **91** 3846
[5] Blackmore R 1987 *J. Chem. Phys.* **87** 791
[6] Shapiro D A, Ciuryło R, Drummond J R and May A D 2001 *Phys. Rev. A* **65** 012501
[7] Wcisło P, Cygan A, Lisak D and Ciuryło R 2013 *Phys. Rev. A* **88** 012517
[8] Ciuryło R, Shapiro D A, Drummond J R and May A D 2001 *Phys. Rev. A* **65** 012502
[9] Lindenfield M J and Shizgal B 1979 *Chem. Phys.* **41** 81
[10] Gradshteyn I S and Ryzhik I M, 1965 *Table of Integrals, Series and Products* Academic Press. New York
[11] IEEE Computer Society 2008 *IEEE Standard for Floating-Point Arithmetic: IEEE Std 754-2008*