MODELLING AND SIMULATION IN CHEMICAL ENGINEERING

Coimbra, June 30 - July 4, 2003

Alírio E. Rodrigues, Paula de Oliveira, José Almiro Castro,
José Augusto Ferreira, Maria do Carmo Coimbra

22

CENTRO INTERNACIONAL de MATEMÁTICA

The Workshop Modelling and Simulation in Chemical Engineering took place in Coimbra in July 2003, integrated in a thematic term devoted to Mathematics and Engineering which was supported by the Centro Internacional de Matemática (CIM).

Its main objective was to bring together mathematicians and chemical engineers to improve the understanding of the problems of process engineering and the mathematical tools available to solve them. To enhance the dialogue among theoretical research, computational aspects and reactive flow behaviour short courses and plenary conferences were given covering topics like Mathematical Modelling and Chemical Engineering Systems, Packages and Numerical Methods to Solve P.D.E's, and Optimization Techniques. Twenty contributed talks were also presented.

The workshop was attended by about fifty researchers working in numerical simulation of Chemical Engineering problems. We believe that the texts included in this publication will give a reasonable overview of the state of the art as far as the main challenges posed in our days by the Numerical Simulation of Reactive Flows are concerned.

Coimbra, Julho de 2003
The Organizing Committee

Alírio E. Rodrigues
Paula de Oliveira
José Almiro Castro
José Augusto Ferreira
Maria do Carmo Coimbra

Contents

Courses

Plenary talks

Contributed talks

T.M. Mata, C.A.V. Costa, *Computer modelling and simulation in chemical processes pollution prevention*

F.J.M. Neves, D.C. Silva, J.I.L.C. Tourais, N.M.C. Oliveira, *Global simulation and optimization of a chemical plant*

I.S. Pop, C.J. Van Duijn, *Micro-scale analysis of crystal dissolution and precipitation in porous media*

P. Portugal, H. Jorge, R.M. Quinta-Ferreira, *Sequential method for kinetic models discrimination*

A. Prechtel, F. Radu, P. Knabner, *Modelling and numerical simulation of variably saturated flow and coupled reactive, biogeochemical transport*

P.A. Quadros, N.M.C. Oliveira, C.M.S.G. Baptista, *Modelling of heterogeneous reactions: simultaneous mass transfer and chemical reaction*

R. Robalo, C. Sereno, A. Rodrigues, *Moving finite elements method: application to moving boundary systems*

D.C.M. Silva, N.M.C. Oliveira, *Model-based optimization of discontinuous chemical polymerization systems*

J.L. Soares, *Convex programming tools for disjunctive programs*

Splitting Techniques for Advection-Diffusion-Reaction Equations

Willem Hundsdorfer
CWI, Amsterdam, The Netherlands

Abstract

Notes for minicourse at the Workshop on Modeling and Simulation in Chemical Engineering – Coimbra, 2003. In these notes some classical and modern splitting techniques are reviewed for transport-chemistry problems, modeled as time-dependent advection-diffusion-reaction equations. The material is largely based on the forthcoming book [6], where a more detailed exposition and additional numerical tests can be found.

1 Introduction

Many physical, chemical and biological models take the form of advection-diffusion-reaction problems. Problems of this type occur for instance in the description of transport-chemistry in the atmosphere, surface- and ground-water and we shall consider the equations with such applications as reference.

Consider a concentration $u(x,t)$ of a certain chemical species, with space variable x and time t. If the species is carried along by a flowing medium with velocity $a(x,t)$, with diffusion coefficient $d(x,t)$ and with sources, sinks and chemical reactions described by $f(u,x,t)$, then the mass conservation law leads to the advection-diffusion-reaction equation

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\Big(a(x,t)u(x,t)\Big) = \frac{\partial}{\partial x}\Big(d(x,t)\frac{\partial}{\partial x}u(x,t)\Big) + f\big(u(x,t),x,t\big)\ . \tag{1.1}$$

We shall consider the equation in a spatial interval $\Omega \subset \mathbb{R}$ with time $t \geq 0$. An initial profile $u(x,0)$ will be given and we also assume that suitable boundary conditions are provided.

If we consider multiple spatial variables, say $\underline{x} \in \mathbb{R}^2$ or $\mathbb{R}^3$, then we get

$$\frac{\partial}{\partial t}u + \nabla\cdot(\underline{a}u) = \nabla\cdot(Du) + f(u,\underline{x},t) \tag{1.2}$$

with velocity $\underline{a} = \underline{a}(\underline{x},t)$ and diffusion (tensor) $D = D(\underline{x},t)$. Moreover, with r chemical species we will have a vector $u(x,t) = (u_1(x,t), u_2(x,t), ..., u_r(x,t))^T$ containing concentration values of the chemical species. Each chemical component might have a different velocity or diffusion coefficient. The coupling between the chemical components is then provided by the reaction term. Although more general problems often occur – where, for example, $\underline{a}$ and D also depend on u – the system (1.2) already has many applications, for instance in pollution problems [14, 18].

One might want to apply different time stepping methods to the different parts of the equations. For example, the chemistry can be very stiff, which calls for an implicit ODE method. On the other hand, if the advection is discretized in space with a flux-limiter, then explicit methods seem much more suitable for that part of the equation. Moreover, use of an implicit method to the full equation will lead to a huge algebraic system, with coupling over the species as well as over the space.

1.1 Spatial discretizations and the "method" of lines

We shall consider the discretization of spatial and temporal derivatives as separate processes. If the spatial derivatives are replaced by difference quotients on some spatial mesh Ω_h, we end up with a large system of ordinary differential equations

$$w'(t) = F(t, w(t)) \tag{1.3}$$

where each component $w_i(t)$ of the vector $w(t)$ approximates the PDE solution $u(x,t)$ in a grid point x_i or a surrounding cell. Such an ODE system is often called the *semi-discrete system* since the space is discretized but time is still continuous. We will consider (1.3) for $t \geq t_0$ with given initial condition $w(t_0)$.

Although spatial discretization will not be considered here, some brief remarks are in order. *Advection*: The advection just gives a translation of the solution along the steamlines (characteristics) described by the velocities. This usually results in large solution gradients. Standard second-order differences $\big(u_x(x_i) \approx (2h)^{-1}(u(x+h) - u(x-h)\big)$ then may give large numerical oscillations. Therefore more complicated discretizations, possibly with limiters to suppress oscillations, are then advisable; see for instance [7, 6]. The use of an explicit time stepping method will lead to a stability restriction $\tau/h \leq C$ where τ is the time step and $C > 0$ will depend on the size of a and the methods used.
Diffusion: Diffusion has a smoothing effect on the solution. The standard second-order differences $\big(u_{xx}(x_i) \approx h^{-2}(u(x+h) - 2u(x) + u(x-h)\big)$ usually provide a good discretization. For diffusion terms an explicit time stepping scheme will lead to a time step restriction of the form $\tau/h^2 \leq C$. On a fine mesh such a restriction is not acceptable (too many steps) and therefore implicit methods should then be used.
Reaction: Chemical reaction often have very different time scales, where only the slower scales need to be resolved; the fast scales typically correspond to radicals which are important for the process but not in the final output. However, with an explicit time stepping method also the fast scales need to be resolved for numerical stability reasons. Hence also for reaction terms, implicit methods may be more advisable.

After a discretization in space, a semi-discrete system (1.3) is obtained which then is discretized in time. This separate treatment of time and space is often called the *method of lines*. It is not a 'method' in the numerical sense, but rather a way to construct and analyze numerical schemes.

Due to the different operators and time scales of the processes it is often advisable to use a splitting of F into easier parts, say

$$F(t, w) = F_1(t, w) + F_2(t, w)\,, \tag{1.4}$$

such that the individual processes $v'(t) = F_j(t, v(t))$ are easier to solve than (1.3). Some examples are given below, but for the moment we may think of splitting a two-dimensional equation into two one-dimensional equations; or splitting a reaction-diffusion (or reaction-advection) equation into a diffusion (or advection) equation and a separate reaction equation. The latter type of splitting, also applies to general spatial discretizations on unstructured grids. Here the advantage lies in the fact that the reaction will only have a coupling over the chemical components whereas advection-diffusion will only give a coupling over space.

More general, one can consider a multiple splitting

$$F(t, w) = F_0(t, w) + F_1(t, w) + \cdots + F_s(t, w) \tag{1.5}$$

where F_0 will stand for the contribution of non-stiff terms, suitable for explicit treatment, and the other terms $F_1, \ldots, F_s$ are stiff and need implicit time integration.

Example 1.1. Standard second-order spatial discretization for (1.1) on a uniform mesh leads to the semi-discrete system

$$w_j' = \frac{1}{2h}\Big(a_{j-\frac{1}{2}}(w_{j-1} + w_j) - a_{j+\frac{1}{2}}(w_j + w_{j+1})\Big)$$
$$+ \frac{1}{h^2}\Big(D_{j-\frac{1}{2}}(w_{j-1} - w_j) - D_{j+\frac{1}{2}}(w_j - w_{j+1})\Big) \; + \; f(w_j, x_j, t)\,.$$

Here $w_j = w_j(t)$ is viewed as an approximation to $u(x_j, t)$ or the average value over the cell $[x_j - \frac{1}{2}h, x_j + \frac{1}{2}h]$ at time t, with grid points $x_j = x_0 + jh$. If u is a vector of r chemical components, then each w_j is also a vector in $\mathbb{R}^r$.

Boundary conditions will lead to a system with dimension $\sim rh^{-1}$. As mentioned above, the reaction term is usually stiff and also the diffusion term usually makes explicit time stepping costly. On the other hand, with an implicit method for the whole system we get in each time step a large algebraic system with coupling over space and chemical components.

If we put the advection and diffusion terms in F_1 and reaction in F_2 then this simultaneous coupling is avoided. Advection may also be put separately in F_0, which can then be treated explicitly. This is in particular attractive if more complicated (nonlinear, limited) discretizations are used for the advection. Moreover, for multi-dimensional problems on a Cartesian mesh, the diffusion terms in different directions may be further split to get simple one-dimensional subsystems. ◇

2 Time splitting methods

In this section we shall discuss some methods were the equation is split into several parts, which are all solved independently on the time intervals $[t_n, t_{n+1}]$. Such methods are usually called *(time) splitting* methods or *fractional step* methods. In case the splitting is such that different physical processes are separated, the term 'operator splitting' is also commonly used. If a multi-dimensional problem is split into 1-dimensional sub-problems, this is called 'dimensional splitting'.

2.1 First-order splitting

Consider an ODE system, linear for simplicity,

$$w'(t) = Aw(t),$$

with $A = A_1 + A_2$, arising for example from a linear PDE with homogeneous boundary conditions or periodicity conditions. We have

$$w(t_{n+1}) = e^{\tau A} w(t_n). \tag{2.1}$$

If we are only able, or willing, to solve the 'sub-problems' $v'(t) = A_1 v(t)$ and $v'(t) = A_2 v(t)$, then (2.1) can be approximated by

$$w_{n+1} = e^{\tau A_2} e^{\tau A_1} w_n, \tag{2.2}$$

which is the simplest splitting method. In actual computations the terms $e^{\tau A_k}$ will, of course, be approximated by some suitable ODE method.

Replacing (2.1) by (2.2) will introduce an error, the so-called *splitting error* for this particular splitting. Inserting the exact solution into (2.2) gives $w(t_{n+1}) = e^{\tau A_2} e^{\tau A_1} w(t_n) + \tau\rho_n$ with local truncation error ρ_n. Note that $\tau\rho_n$ is the error introduced per step. We have

$$e^{\tau A} = \Big(I + \tau(A_1 + A_2) + \frac{1}{2}\tau^2(A_1 + A_2)^2 + \cdots\Big),$$

$$e^{\tau A_2} e^{\tau A_1} = \Big(I + \tau(A_1 + A_2) + \frac{1}{2}\tau^2(A_1^2 + 2A_2A_1 + A_2^2) + \cdots\Big).$$

Hence the local truncation error equals

$$\frac{1}{\tau}\Big(e^{\tau A} - e^{\tau A_2} e^{\tau A_1}\Big) w(t_n) = \frac{1}{2}\tau[A_1, A_2] w(t_n) + \mathcal{O}(\tau^2), \tag{2.3}$$

with $[A_1, A_2] = A_1A_2 - A_2A_1$ the commutator of A_1 and A_2. We see that (2.2) will be a 1st-order process, unless A_1 and A_2 commute. Note that we assume here tacitly that terms like $A_1A_2w(t)$ are $\mathcal{O}(1)$, which seems reasonable only if there are no boundary conditions or the PDE solution satisfies certain compatibility conditions.

For general nonlinear ODE systems

$$w'(t) = F_1(t, w(t)) + F_2(t, w(t)),$$

we can apply (2.2) if the terms e^{tA_k} are interpreted as *solution operators*. Written out, we solve subsequently

$$\begin{aligned}
&\tfrac{d}{dt} w^*(t) = F_1(t, w^*(t)) \qquad \text{for } \ t_n \le t \le t_{n+1} \ \text{ with } \ w^*(t_n) = w_n\,, \\
&\tfrac{d}{dt} w^{**}(t) = F_2(t, w^{**}(t)) \quad \text{for } \ t_n \le t \le t_{n+1} \ \text{ with } \ w^{**}(t_n) = w^*(t_{n+1})\,,
\end{aligned}$$

giving $w_{n+1} = w^{**}(t_{n+1})$ as the next approximation. If $w_n = w(t_n)$ we now get the local truncation error

$$\frac{1}{2}\tau\Big[\frac{\partial F_1}{\partial w} F_2 - \frac{\partial F_2}{\partial w} F_1\Big](t_n, w(t_n)) + \mathcal{O}(\tau^2),$$

similar to (2.3). This formula can be derived by Taylor expansions of $w^{**}(t_{n+1})$ and $w^*(t_{n+1})$ around $t = t_n$.

2.2 Higher-order and multi-component splittings

In (2.2) one starts in all steps with A_1. Interchanging the order of A_1 and A_2 after each step will lead to more symmetry and often to better accuracy. Carrying out two half steps with reversed sequence gives the following splitting, due to Strang [12],

$$w_{n+1} = \Big(e^{\frac{1}{2}\tau A_2} e^{\frac{1}{2}\tau A_1}\Big)\Big(e^{\frac{1}{2}\tau A_1} e^{\frac{1}{2}\tau A_2}\Big) w_n = e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\frac{1}{2}\tau A_2} w_n. \tag{2.4}$$

By a series expansion and some tedious calculations it follows that the local truncation error is given by

$$\frac{1}{24}\tau^2\Big([A_2, [A_2, A_1]] - 2[A_1, [A_1, A_2]]\Big) w(t_n) + \mathcal{O}(\tau^4). \tag{2.5}$$

Due to symmetry, the truncation error will only contain even order terms.

If we work with constant step sizes, then (2.4) will require almost the same amount of computational work as (2.2), since for constant τ we can write the total process (2.4) as

$$w_n = e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\tau A_2} \cdots e^{\tau A_1} e^{\frac{1}{2}\tau A_2} w_0.$$

In general, with variable step sizes it will be more expensive, of course.

Generalization to nonlinear systems is straightforward, we get

$$\begin{aligned}
&\tfrac{d}{dt} w^*(t) = F_2(t, w^*(t)) && \text{for } t_n \le t \le t_{n+1/2} \text{ with } w^*(t_n) = w_n\,,\\
&\tfrac{d}{dt} w^{**}(t) = F_1(t, w^{**}(t)) && \text{for } t_n \le t \le t_{n+1} \text{ with } w^{**}(t_n) = w^*(t_{n+1/2})\,,\\
&\tfrac{d}{dt} w^{***}(t) = F_2(t, w^{***}(t)) && \text{for } t_{n+1/2} \le t \le t_{n+1} \text{ with } w^{***}(t_{n+1/2}) = w^{**}(t_{n+1})\,,
\end{aligned}$$

giving $w_{n+1} = w^{***}(t_{n+1})$ as the approximation on the new time level.

With regard to stability of the splitting schemes, there are not that many practical, pertinent results available. However, as a rule, if the individual steps are treated in a stable manner then the whole process will be stable.

Higher-order splittings are possible, but such splittings will contain negative coefficients or negative time steps (Sheng, 1989; Goldman & Kaper, 1996; see [6]). For example, let

$$S_\tau = e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\frac{1}{2}\tau A_2}$$

be the 2nd-order Strang splitting operator. Then a 4th-order splitting is given by

$$w_{n+1} = S_{\theta\tau}\, S_{(1-2\theta)\tau}\, S_{\theta\tau}\, w_n\,,$$

with $\theta = (2 - \sqrt[3]{2})^{-1} \approx 1.35$. Here we have $1 - 2\theta < 0$, so that a step with negative time has to be taken. For partial differential equations with boundary conditions such splittings with negative time steps seem of limited value. We note, however that they are frequently used for time reversible problems, which arise for instance with certain mechanical problems, see [11].

With more splitting components, for example $A = A_1 + A_2 + A_3$, then the first-order splitting (2.2) can be generalized to

$$w_{n+1} = e^{\tau A_3} e^{\tau A_2} e^{\tau A_1} w_n.$$

Likewise, the Strang splitting (2.4) leads to the 2nd-order formula

$$w_{n+1} = e^{\frac{1}{2}\tau A_3} e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\frac{1}{2}\tau A_2} e^{\frac{1}{2}\tau A_3} w_n.$$

Note that this is just a repeated application of (2.4): first approximate $e^{\tau A}$ by $e^{\frac{1}{2}\tau A_3} e^{\tau(A_1+A_2)} e^{\frac{1}{2}\tau A_3}$, and then approximate $e^{\tau(A_1+A_2)}$ in the same fashion. Application to more components and nonlinear systems carries over in the same way.

2.3 Solving the fractional steps

To solve the sub-steps, one may select a method such as Euler or Trapezoidal Rule. If these are applied with the same step size τ that is used for the splitting itself, a specific (classical) splitting method arises. Numerous examples are found in Yanenko [17], Mitchell & Griffiths [9] and Marchuk [8].

For instance, first-order splitting combined with backward Euler gives the first-order method

$$\begin{aligned} w^*_{n+1} &= w_n + \tau F_1(t_{n+1}, w^*_{n+1}), \\ w_{n+1} &= w^*_{n+1} + \tau F_2(t_{n+1}, w_{n+1}). \end{aligned} \tag{2.6}$$

If F_1 and F_2 contain discretized space derivatives in x and y direction, respectively, this method is called the 1st-order LOD method (locally one dimensional) method. It is obvious that we can generalize this method for $F = F_1 + F_2 + ... + F_s$.

The 2nd-order LOD method is obtained by combining Strang splitting with the trapezoidal rule (or, likewise, the implicit midpoint rule),

$$\begin{aligned} w^*_{n+1} &= w_n + \tfrac{1}{2}\tau\Big(F_1(t_n, w_n) + F_1(t_n + (\tfrac{1}{2}+c)\tau, w^*_{n+1})\Big), \\ w_{n+1} &= w^*_{n+1} + \tfrac{1}{2}\tau\Big(F_2(t_n + (\tfrac{1}{2}-c)\tau, w^*_{n+1}) + F_2(t_{n+1}, w_{n+1}))\Big), \\ w^*_{n+2} &= w_{n+2} + \tfrac{1}{2}\tau\Big(F_2(t_{n+1}, w_{n+1}) + F_2(t_{n+1} + (\tfrac{1}{2}+c)\tau, w^*_{n+2})\Big), \\ w_{n+2} &= w^*_{n+2} + \tfrac{1}{2}\tau\Big(F_1(t_{n+1} + (\tfrac{1}{2}-c)\tau, w^*_{n+2}) + F_1(t_{n+1}, w_{n+2}))\Big). \end{aligned} \tag{2.7}$$

Note that here Strang splitting is applied on the interval $[t_n, t_{n+2}]$. For c we can take for example $c = 0$ or $c = \frac{1}{2}$. What is best will depend on the problem, and there is no choice that seems preferable a priori. This is due to the fact that the intermediate vectors w^*_{n+j} are not a consistent approximation to the full problem at some given time level. Again, generalization to more F-components is straightforward. Method (2.7) is known as Yanenko's method; see [17].

With the above splitting methods all sub-problems are treated in the same fashion and with the same time step. In general, it seems better to solve the fractional steps with a method that is suited for that particular sub-step, possibly with a sub-time step $\bar{\tau} \leq \tau$. Here one may chose, for example, an implicit or explicit Runge-Kutta method, depending whether the sub-problem $w'(t) = F_j(t, w(t))$ is stiff or non-stiff, with an appropriate $\bar{\tau}$. This latter approach is the one we recommend for general problems.

2.4 Boundary corrections

The major difficulties with splitting methods occur for problems were the boundary conditions are important. If we consider a PDE problem with boundary conditions, then these are physical conditions for the whole process and boundary conditions for the sub-steps (which may have little physical meaning) are missing.

Therefore one may have to reconstruct boundary conditions for the specific splitting under consideration. For example, consider a linear semi-discrete problem $w'(t) = Aw(t) + g(t)$, were $g(t)$ contains the given boundary conditions. Suppose that

$$Av + g(t) = \Big(A_1 v + g_1(t)\Big) + \Big(A_2 v + g_2(t)\Big),$$

with $g_k(t)$ containing the boundary conditions relevant to A_k. The exact solution satisfies

$$w(t_{n+1}) = e^{\tau A} w(t_n) + \int_0^\tau e^{(\tau - s)A} g(t_n + s) ds.$$

If we consider 1st-order splitting, with inhomogeneous terms $\tilde{g}_1, \tilde{g}_2$, then

$$w_{n+1} = e^{\tau A_2} e^{\tau A_1} w_n + e^{\tau A_2} \int_0^\tau e^{(\tau-s)A_1} \tilde{g}_1(t_n + s) ds + \int_0^\tau e^{(\tau-s)A_2} \tilde{g}_2(t_n + s) ds.$$

Even with commuting matrices, $A_1 A_2 = A_2 A_1$, and constant boundary terms we will get a splitting error if we take $\tilde{g}_k = g_k$. An exact formula for this case is obtained by choosing

$$\tilde{g}_1(t_n + s) = e^{-sA_2} g_1(t_n + s), \quad \tilde{g}_2(t_n + s) = e^{(\tau-s)A_1} g_2(t_n + s).$$

Note that this correction for g_1 requires a *backward* time integration with A_2, and this may not be feasible with an implicit ODE method, due to the fact that the implicit algebraic relations need no longer be well defined with negative step size. One might replace e^{-sA_2} by some explicit polynomial approximation $P(-sA_2)$, but the effect of this on stability and accuracy is unclear.

As a rule of thumb, it can be said that the treatment of the boundaries should coincide as much as possible with the scheme in the interior of the domain. Examples for specific LOD (and ADI) methods can be found in Mitchell & Griffiths [9, Ch. 2]. A general analysis of boundary conditions for splitting methods is, at present, still lacking. Therefore we conclude this subject with an example.

Example. Consider the model advection-reaction equation

$$u_t + u_x = u^2, \qquad 0 \le x \le 1, \quad 0 \le t \le 1/2$$

with given initial value at $t = 0$ and Dirichlet condition at $x = 0$, derived from the exact solution

$$u(t, x) = \frac{\sin(\pi(x - t))^2}{1 - t \sin(\pi(x - t))^2}.$$

Here spatial discretization is performed with 4th-order central differences in the interior and 3rd-order one-sided approximations at the boundaries. The advection step is solved with the classical 4th-order Runge-Kutta method at Courant number $\tau/h = 2$, and the 'reaction' $u_t = u^2$ is solved exactly. Since the nonlinear term is nonstiff, splitting is not really necessary in this example, but it is instructive to consider the errors.

Let us consider :
(*i*) Simple splitting (with reaction followed by advection) where in the advection step the given boundary values are used;
(*ii*) Strang splitting where after each time step the order of the fractional steps is reversed, also with the given boundary conditions;
(*iii*) The same splitting as in (*i*) but with corrected boundary conditions

$$u^{**}(t, 0) = \frac{u(t, 0)}{1 - (t_{n+1} - t)u(t, 0)} \quad \text{for } t \in [t_n, t_{n+1}].$$

The errors in the L_2-norm, together with the estimated orders of convergence, are given in the following table.

	Simple splitting	Strang splitting	Corrected bd.
$\tau = 1/\ 20$	$0.26\ 10^{-1}$	$0.14\ 10^{-1}$	$0.88\ 10^{-3}$
$\tau = 1/\ 40$	$0.14\ 10^{-1}$ (*0.94*)	$0.48\ 10^{-2}$ (*1.58*)	$0.91\ 10^{-4}$ (*3.27*)
$\tau = 1/\ 80$	$0.72\ 10^{-2}$ (*0.96*)	$0.17\ 10^{-2}$ (*1.54*)	$0.13\ 10^{-4}$ (*2.80*)
$\tau = 1/160$	$0.36\ 10^{-2}$ (*0.98*)	$0.58\ 10^{-3}$ (*1.52*)	$0.22\ 10^{-5}$ (*2.57*)

Table 2.1. L_2-errors (and estimated orders) for (4.1) at $t = 1/2$ with $\tau = 2h$.

Note that the simple splitting with boundary corrections is more accurate than its Strang type counterpart. The convergence rate of the scheme with boundary corrections is less than 4, but this is due to order reduction of the Runge-Kutta method, it is not caused by the splitting procedure. A similar order reduction can be observed with Strang splitting: in the absence of boundary conditions it has (at least) order 2, but in the above table an order 1.5 behaviour can be observed. ◇

3 IMEX, ADI and AMF methods

With time splitting by the fractional step approach we have to solve sub-problems that are not consistent with the full model. As we saw this creates difficulties with boundary conditions, and similar problems arise with interface conditions. Also, stationary solutions of the problem are not stationary solutions of the fractional step methods. Moreover in the time splitting approach multi-step schemes cannot be used in a natural fashion. In this section some alternatives to time splitting will be briefly reviewed.

3.1 The θ-IMEX method

Suppose that the semi-discrete system is of the form

$$w'(t) = F(t, w(t)) = F_0(t, w(t)) + F_1(t, w(t)) \tag{3.1}$$

where F_0 is a term suitable for explicit time integration, for instance discretized advection, and F_1 requires an implicit treatment, say discretized diffusion or stiff reactions.

We consider the following simple method

$$w_{n+1} = w_n + \tau F_0(t_n, w_n) + (1-\theta)\tau F_1(t_n, w_n) + \theta\tau F_1(t_{n+1}, w_{n+1}), \tag{3.2}$$

with parameter $\theta \geq \frac{1}{2}$. Here the explicit Euler method is combined with the implicit θ-method. Such mixtures of implicit and explicit methods are called IMEX schemes. Note that in contrast to the time splitting methods there are no intermediate results which are inconsistent with the full equation.

Insertion of the exact solution in the scheme gives the truncation error

$$\frac{1}{\tau}\Big(w(t_{n+1} - w(t_n)\Big) - (1-\theta)F(t_n, w(t_n)) - \theta F(t_{n+1}, w(t_{n+1})) -$$

$$-\theta\Big(F_0(t_{n+1}, w(t_{n+1})) - F_0(t_n, w(t_n))\Big) = \big(\frac{1}{2} - \theta\big)\tau w''(t_n) + \theta\tau\varphi'(t_n) + \mathcal{O}(\tau^2)$$

where $\varphi(t) = F_0(t, w(t))$. If F_0 denotes discretized advection and nonstiff terms, smoothness of w will also imply smoothness of φ, independent of boundary conditions or small mesh widths h. Therefore the structure of the truncation error is much more favourable than with the time splitting methods considered in the preceding section. For example, with a stationary solution $w(t) = w(0)$ we now have a zero truncation error. However, with methods of this IMEX type it is stability that needs a careful examination.

Let us consider the scalar, complex test equation

$$w'(t) = \lambda_0 w(t) + \lambda_1 w(t), \tag{3.3}$$

and let $z_j = \tau\lambda_j$, $j = 0, 1$. In applications to PDEs these λ_j will represent eigenvalues of the two components F_0 and F_1, found by inserting Fourier modes. One would hope that having $|1 + z_0| \le 1$ (stability of the explicit method) and $\mathrm{Re}\, z_1 \le 0$ (stability of the implicit method) would be sufficient to guarantee stability of the IMEX scheme, but this is not so in general. Application of the IMEX scheme to this test equation yields $w_{n+1} = Rw_n$ where $R = R(z_0, z_1)$ is given by

$$R = \frac{1 + z_0 + (1 - \theta)z_1}{1 - \theta z_1}\ . \tag{3.4}$$

Stability for the test equation thus requires $|R| \le 1$.

First, consider the set

$$\mathcal{D}_0 = \{z_0 : \text{the IMEX scheme is stable for any } z_1 \in \mathbb{C}^-\}. \tag{3.5}$$

So, here we insist on A-stability with respect to the implicit part. Using the maximum principle, it follows by some straightforward calculations that $z_0 = x_0 + iy_0$ belongs to this set iff

$$\theta^2 y_0^2 + (2\theta - 1)(1 + x_0)^2 \le 2\theta - 1.$$

Plots are given in Figure 3.1. If $\theta = 1$ we reobtain the stability region of the explicit Euler method, but for smaller values of θ the set start to shrink and for $\theta = \frac{1}{2}$ it reduces to the line segment $[-2, 0]$ on the negative axis.

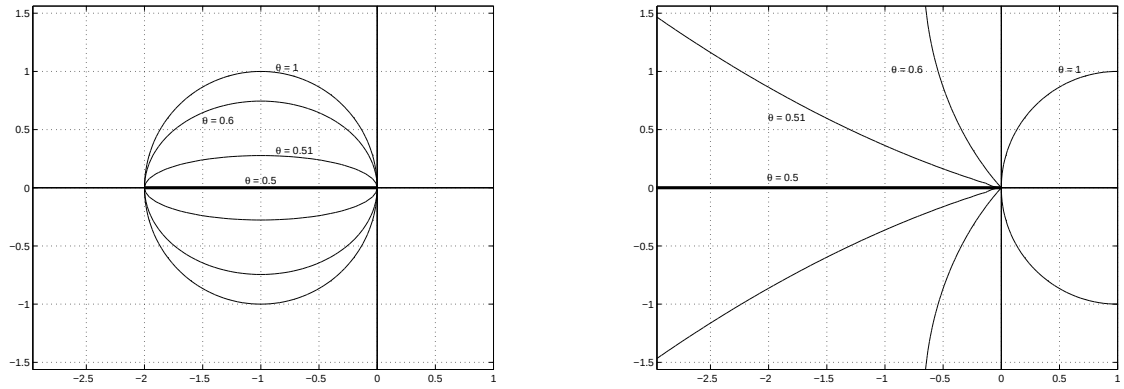


Fig. 3.1. Boundaries of regions $\mathcal{D}_0$ (left) and $\mathcal{D}_1$ (right) for the θ-IMEX method (3.2) with $\theta = 0.5, 0.51, 0.6$ and 1.

Alternatively, one can insist on using the full stability region of the explicit method $\mathcal{S}_0 = \{z_0 : |1 + z_0| \le 1\}$, but then z_1 has to be restricted to the set

$$\mathcal{D}_1 = \{z_1 : \text{the IMEX scheme is stable for any } z_0 \in \mathcal{S}_0\}. \tag{3.6}$$

It easily follows that $z_1 \in \mathcal{D}_1$ iff

$$1 + (1 - \theta)|z_1| \le |1 - \theta z_1|,$$

see the right plot in Figure 3.1. Again it is only for $\theta = 1$ that we get the stability region of the implicit θ-method. If $\theta = \frac{1}{2}$ the set $\mathcal{D}_1$ equals the negative real line $\mathbb{R}^-$.

Note that the implicit θ-method with $\theta > \frac{1}{2}$ is *strongly* A-stable (that is, A-stable with damping at ∞) whereas the trapezoidal rule, $\theta = \frac{1}{2}$, is 'just' A-stable. Apparently, using a strongly implicit method gives better stability properties within an IMEX formula.

On the other hand, the above criteria are rather strict. For instance, if we take z_0 such that $|\rho + z_0| \le \rho$ with $\rho < 1$, then the method with $\theta = \frac{1}{2}$ will be stable if $z_1 = x_1 + iy_1 \in \mathbb{C}^-$ is within the hyperbole $\rho^2 y_1^2 + 4\rho^2(1 - \rho) \le 4(1 - \rho)(\rho - x_1)^2$. Therefore, the IMEX method with $\theta = \frac{1}{2}$ should not be discarded, only extra care should be given to stability when applying this method.

In the above the values of λ_0 and λ_1 have been considered as independent, which is a reasonable assumption if F_0 and F_1 act in different directions, for instance if $F_0 \approx a(\partial/\partial x)$ (horizontal coupling) and $F_1 \approx d(\partial^2/\partial z^2)$ (vertical coupling) or F_1 a reaction term (coupling over chemical species).

Different results are obtained if there is a dependence between λ_0 and λ_1. Then the implicit treatment of λ_1 can stabilize the process so that we do not even need $z_0 \in \mathcal{S}_0$. Consider for example the 1D advection-diffusion equation $u_t + au_x = du_{xx}$ with periodicity in space and with second-order spatial discretization. If advection is treated explicitly and diffusion implicitly, then the relevant eigenvalues (Fourier decomposition) are

$$z_0 = i\nu \sin 2\phi, \qquad z_1 = -4\mu \sin^2 \phi \tag{3.7}$$

with $\nu = a\tau/h$, $\mu = d\tau/h^2$ and $0 \le \phi \le \pi$. A straightforward calculation shows that $|R| \le 1$ iff

$$1 - 8(1 - \theta)\mu s + 16(1 - \theta)^2\mu^2 s^2 + 4\nu^2 s(1 - s) \le 1 + 8\theta\mu s + 16\theta^2\mu^2 s^2$$

where $s = \sin^2 \phi$. This holds for all $s \in [0, 1]$ iff

$$\nu^2 \le 2\mu \quad \text{and} \quad 2(1 - 2\theta)\mu \le 1. \tag{3.8}$$

So for any $\theta \ge \frac{1}{2}$ we now just have the condition $\nu^2 \le 2\mu$, that is $a^2\tau \le 2d$.

Finally we note that the above IMEX method with $\theta = 1$ could be viewed as a time splitting method where we first solve $v'(t) = F_0(t, v(t))$ on $[t_n, t_{n+1}]$ with forward Euler and then $v'(t) = F_1(t, v(t))$ with backward Euler. This explains the favourable stability results with this method. However, the structure of the truncation error is very different from the time splitting methods. This is due to interference of the first-order splitting error with the first-order Euler errors.

In the following subsections we shall consider several generalizations of (3.2). Such generalizations are necessary for practical problems since the explicit Euler method is not well suited for advection, and also first-order accuracy is often not sufficient. Moreover, we may want additional splittings of the implicit terms to resolve the implicit relations more efficiently.

3.2 IMEX multi-step methods

As mentioned already, in the time splitting approach multi-step schemes cannot be used in a natural fashion. Straightforward use of a multi-step scheme with step size τ to solve the sub-problems $v'(t) = F_j(t, v(t))$, $t_n \le t \le t_{n+1}$ leads to inconsistencies since the available past values $w_{n-1}, w_{n-2}, \cdots$ are approximations to the whole problem, not to the particular sub-problem at hand. Here we shall consider an other approach to combine implicit and explicit multi-step methods.

One of the most popular implicit methods is the second-order BDF2 method

$$\frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} = \tau F(t_{n+1}, w_{n+1})$$

where the left hand side is the 2-step backward differentiation formula, hence the name BDF. Along with w_0, the starting value w_1 should be known. It can be computed by a one-step method, for instance Euler. The popularity of this implicit BDF method is due to its stability and damping properties. These are very useful properties for diffusion equations.

Convection equations are often treated more efficiently by an explicit method, such as

$$\frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} = 2\tau F(t_n, w_n) - \tau F(t_{n-1}, w_{n-1}),$$

to which we shall refer as the explicit BDF2 method. The stability region of this explicit method is plotted in Figure 3.2.

With advection-diffusion-reaction problems, explicit advection and implicit diffusion-reaction can then be combined through the IMEX formula

$$\frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} = 2\tau F_0(t_n, w_n) + \tau F_0(t_{n-1}, w_{n-1}) + \tau F_1(t_{n+1}, w_{n+1}), \tag{3.9}$$

where F_0 contains convective terms only and F_1 denotes discretized diffusion together with reaction.

The above can be generalized as follows: consider a fully implicit multistep method

$$\sum_{j=0}^{k} \alpha_j w_{n+1-j} = \tau \sum_{j=0}^{k} \beta_j \Big(F_0(t_{n+1-j}, w_{n+1-j}) + F_1(t_{n+1-j}, w_{n+1-j}) \Big), \tag{3.10}$$

with implicit treatment of advection and diffusion-reaction. We can handle the advection explicitly by applying an extrapolation formula

$$\varphi(t_{n+1}) = \sum_{j=1}^{k} \gamma_j \varphi(t_{n+1-j}) + \mathcal{O}(\tau^q) \tag{3.11}$$

with $\varphi(t) = F_0(t, w(t))$. This leads to the method

$$\sum_{j=0}^{k} \alpha_j w_{n+1-j} = \tau \sum_{j=1}^{k} \beta_j^* F_0(t_{n+1-j}, w_{n+1-j}) + \tau \sum_{j=0}^{k} \beta_j F_1(t_{n+1-j}, w_{n+1-j}), \tag{3.12}$$

with new coefficients $\beta_j^* = \beta_j + \beta_0 \gamma_j$. Methods of this implicit-explicit multistep type were introduced by Crouzeix [2] and Varah [13].

Accuracy of the IMEX multistep methods is easy to establish.

Theorem 3.1. Assume the implicit multistep method has order p and the extrapolation procedure has order q. Then the IMEX method has order $r = \min(p, q)$.

Proof. With $\varphi(t) = F_0(t, w(t))$, the local truncation error can be written as

$$\frac{1}{\tau}\sum_{j=0}^{k}\Big(\alpha_j w(t_{n+1-j}) - \tau\beta_j w'(t_{n+1-j})\Big) + \beta_0\Big(\varphi(t_{n+1}) - \sum_{j=1}^{k}\gamma_j\varphi(t_{n+1-j})\Big)$$

$$= C\tau^p w^{(p+1)}(t_n) + \mathcal{O}(\tau^{p+1}) + \beta_0 C'\tau^q\varphi^{(q)}(t_n) + \mathcal{O}(\tau^{q+1}),$$

with constants C, C' determined by the coefficients of the multistep method and the extrapolation procedure. □

Note that in this truncation error only total derivatives arise, and therefore the error is not influenced by large Lipschitz constants (negative powers of the mesh width) in F_0 or F_1.

Stability results for the IMEX multistep methods are quite complicated, even for the simple test problem (3.3). We consider here two classes of 2-step IMEX methods. Let $\mathcal{S}_0, \mathcal{S}_1$ be the stability regions of the explicit and implicit method, respectively.

The first class is based on the BDF2 method,

$$\begin{aligned}\tfrac{3}{2}w_{n+1} - 2w_n + \tfrac{1}{2}w_{n-1} = 2\tau F_0(t_n, w_n) - \tau F_0(t_{n-1}, w_{n-1}) + \\ + \theta\tau F_1(t_{n+1}, w_{n+1}) + 2(1-\theta)\tau F_1(t_n, w_n) - (1-\theta)\tau F_1(t_{n-1}, w_{n-1})\end{aligned} \tag{3.13}$$

with parameter $\theta \geq 0$. The order is 2 and the implicit method is A-stable for $\theta \geq \frac{3}{4}$. With $\theta = 1$, $F_0 = 0$ we reobtain the fully implicit BDF2 method. If $\theta = \frac{3}{4}$ the implicit method is 'just' A-stable (equivalent with the trapezoidal rule).

We also consider the following class of IMEX methods, based on the two step Adams formulas,

$$\begin{aligned}w_{n+1} - w_n = \tfrac{3}{2}\tau F_0(t_n, w_n) - \tfrac{1}{2}\tau F_0(t_{n-1}, w_{n-1}) + \\ + \theta\tau F_1(t_{n+1}, w_{n+1}) + (\tfrac{3}{2} - 2\theta)\tau F_1(t_n, w_n) + (\theta - \tfrac{1}{2})\tau F_1(t_{n-1}, w_{n-1}),\end{aligned} \tag{3.14}$$

again with order 2. Here the implicit method is A-stable if $\theta \geq \frac{1}{2}$. If $\theta = \frac{1}{2}$ the implicit method reduces to the trapezoidal rule.

In the Figure 3.2 the stability regions $\mathcal{S}_0$ of the explicit methods are plotted together with the regions $\mathcal{D}_0$, defined as in (3.5). We see from the figure that here $\mathcal{D}_0$ is really smaller than $\mathcal{S}_0$ and if the implicit method is just A-stable, the region $\mathcal{D}_0$ reduces to a line. Formulas for the boundary of $\mathcal{D}_0$ can be found in Frank et al. [3] In that paper also results on the set $\mathcal{D}_1$, see (3.6), are presented. It seems that, as a rule, if $z_0 \in \mathcal{S}_0$ and $z_1 < 0$, then the IMEX scheme is stable. Moreover, if the implicit method is strongly A-stable then the IMEX scheme is stable for z_1 in a wedge $\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \leq \alpha\}$, with positive angle α. These results were not proven for arbitrary IMEX schemes, only for some specific schemes in the above BDF2 and Adams2 class.

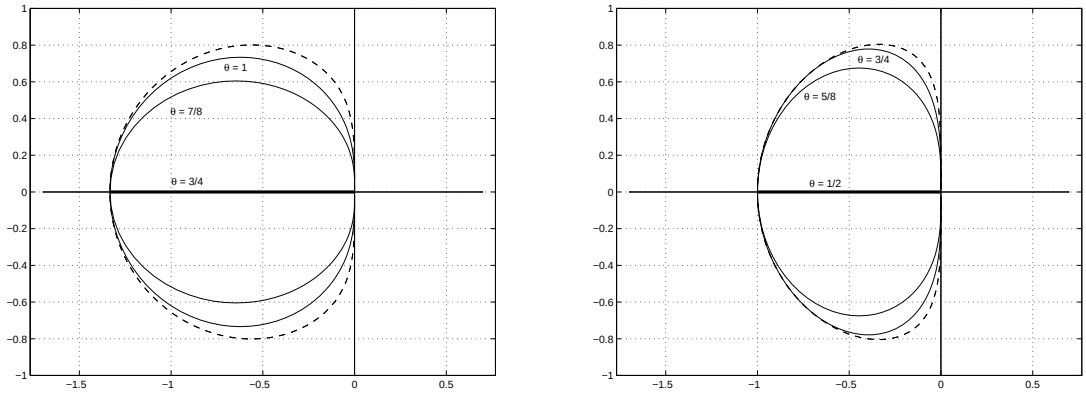


Fig. 3.2. Explicit stability regions $\mathcal{S}_0$ (dashed) and regions $\mathcal{D}_0$ for the IMEX BDF2 methods (left) and ADAMS2 methods (right).

With these regions $\mathcal{D}_0$, z_0 and z_1 are considered as independent. As said before, this holds for example if F_0 represents horizontal advection and F_1 stands for vertical diffusion plus reaction (for air pollution problems these are the most relevant terms, the other processes, such as horizontal diffusion, are small and they can be lumped into F_0). Results for 1D advection-diffusion equations can be found in Varah [13] and Ascher et al. [1]. More general stability results of this type, valid for noncommuting operators, are given in Crouzeix [2].

3.3 Douglas ADI methods

The acronym ADI stands for alternating direction implicit. Originally these methods were developed for dimension splitting with two- and three-dimensional parabolic problems from numerical oil reservoir models, see Peaceman [10]. We will use the name ADI for more general splittings in which all internal stages are consistent with the whole problem.

A familiar ADI method is the second-order Peaceman-Rachford method

$$\begin{aligned} w^*_{n+1/2} &= w_n + \tfrac{1}{2}\tau F_1(t_n, w_n) + \tfrac{1}{2}\tau F_2(t_{n+1/2}, w^*_{n+1/2}), \\ w_{n+1} &= w^*_{n+1/2} + \tau F_1(t_{n+1/2}, w_{n+1}) + \tfrac{1}{2}\tau F_2(t_{n+1}, w^*_{n+1/2}). \end{aligned} \tag{3.15}$$

This could be viewed as a Strang splitting with alternative use of forward and backward Euler, in a symmetrical fashion to obtain second order, but it seems more natural to consider this ADI method as a method of its own. Note that the intermediate value $w^*_{n+1/2}$ is consistent with the whole equation, unlike with the LOD methods. On the other hand, this ADI method does not have a natural extension for more than two components F_j. Therefore we consider a related ADI method that does allow more components.

Suppose we have a decomposition

$$F(t, v) = F_0(t, v) + F_1(t, v) + \cdots + F_s(t, v). \tag{3.16}$$

It will be assumed here that the term F_0 is nonstiff, or mildly stiff, so that this term can be treated explicitly. The other terms will be treated implicitly, in a sequential fashion.

The θ-IMEX method regarded at the beginning of this section can be generalized as follows,

$$\left.\begin{aligned} v_0 &= w_n + \tau F(t_n, w_n), \\ v_j &= v_{j-1} + \theta\tau\Big(F_j(t_{n+1}, v_j) - F_j(t_n, w_n)\Big) \qquad (j = 1, 2, \cdots, s), \\ w_{n+1} &= v_s, \end{aligned}\right\} \tag{3.17}$$

with internal vectors v_j. In case $F_0 = 0$ this is the first-order Douglas-Rachford ADI method if $\theta = 1$, and the second-order Brian-Douglas ADI method if $\theta = \frac{1}{2}$; see [6, 9] for references. This method is also known as the method of Stabilizing Corrections [8]. Note that all internal vectors v_j are consistent with $w(t_{n+1})$ and therefore the accuracy for problems where the boundary conditions are influential is often better than with the time splitting schemes considered in the previous section. In particular, stationary solutions $\bar{w}$ of $w'(t) = F(w(t))$, that is $F(\bar{w}) = 0$, are also stationary solutions of the ADI method, as can be seen by considering consecutive v_j.

Observe that in this ADI method the implicit terms also allow a splitting, which is not the case with the IMEX multistep methods. However, as with the IMEX methods, stability of the method should be carefully examined. The most simple test problem is

$$w'(t) = \lambda_0 w(t) + \lambda_1 w(t) + \cdots + \lambda_s w(t). \tag{3.18}$$

Let $z_j = \tau\lambda_j$, $j = 0, 1, ..., s$. Then the ADI method yields a recursion $w_{n+1} = Rw_n$ with $R = R(z_0, z_1, ..., z_s)$ given by

$$R = 1 + \Big(\prod_{j=1}^{s}(1 - \theta z_j)\Big)^{-1} \sum_{j=0}^{s} z_j. \tag{3.19}$$

Obviously, stability for the test problem requires $|R| \le 1$.

Consider the wedge $\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \le \alpha\}$ in the left half-plane. We consider here stability under the condition that $z_j \in \mathcal{W}_\alpha$, $j \ge 1$. If F_j is a discretized advection-diffusion operator and λ_j an eigenvalue in the Fourier decomposition, then $\alpha < \frac{1}{2}\pi$ means that advection is not allowed to dominate too much. For pure diffusion we have $z_j = \tau\lambda_j \in \mathcal{W}_0$, the line of non-positive real numbers. As before, $z_0, z_1, \ldots, z_s$ are assumed to be independent of each other.

Theorem 3.2. Suppose $z_0 = 0$ and $s \ge 2$, $1 \le r \le s - 1$. For any $\theta \ge \frac{1}{2}$ we have

$$|R| \le 1 \text{ for all } z_i \in \mathcal{W}_\alpha,\ 1 \le i \le s \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,, \tag{3.20}$$

$$\left.\begin{array}{l} |R| \le 1 \text{ for all } z_1, \ldots, z_{s-r} \in \mathcal{W}_\alpha \\ \qquad \text{and } z_{s-r+1}, \ldots, z_s \le 0 \end{array}\right\} \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-r}\,\frac{\pi}{2}\,. \tag{3.21}$$

Proof. Necessity in (3.20) is easy to show: if we take all $z_j = -te^{i\alpha}$, $j \ge 1$, then for $t \to \infty$ we get

$$R = 1 - \frac{ste^{i\alpha}}{\theta^s t^s e^{is\alpha} + \mathcal{O}(t^{s+1})} = 1 - \frac{s}{\theta^s} t^{1-s} e^{i\alpha(1-s)}(1 + \mathcal{O}(t^{-1})),$$

and consequently $\mathrm{Re}(R) > 1$ if t is sufficiently large and $\alpha(1 - s) > \frac{1}{2}\pi$.

To illustrate necessity in (3.21), consider $s = 3$ and $z_3 \le 0$. Since R is fractional linear in z_3, it follows that we have $|R| \le 1$ for all $z_3 \le 0$ iff this holds with z_3 equal to 0 or ∞. This amounts to verification of the inequalities

$$\Big|\, 1 + \frac{z_1 + z_2}{(1 - \theta z_1)(1 - \theta z_2)} \,\Big| \le 1, \qquad \Big|\, 1 - \frac{1}{\theta(1 - \theta z_1)(1 - \theta z_2)} \,\Big| \le 1.$$

For the first inequality we know from (3.20) that $\alpha \le \frac{1}{2}\pi$ is necessary and sufficient, but for the second inequality it can be shown as above that we need $\alpha \le \frac{1}{4}\pi$. The proof of the other results is technical; these can be found in [4, 5]. □

Note that in (3.21), with $r = 1$ we get the same angles α as for $r = 0$. Moreover, it is somewhat surprising that there is no difference between $\theta = \frac{1}{2}$ and $\theta = 1$. In [5] also results are given for $|1 + z_0| \leq 1$, and then the having $\theta = \frac{1}{2}$ or $\theta = 1$ makes a difference. If $\theta = 1$ the above statements remain the same. If $\theta = \frac{1}{2}$ we now need $\alpha = 0$, as we saw already with the θ-IMEX method.

In the following figure the boundary of the stability region $|R| \leq 1$ is plotted for two special choices, namely $z_0 = 0$, $z_j = z$ $(1 \leq j \leq s)$ and $z_0 = 0$, $z_j = z$ $(1 \leq j \leq s-1)$, $z_s = \infty$. Plots for the method with $\theta = \frac{1}{2}$ look very similar. Also drawn, as dotted curved lines, are contour lines of $|R|$ at 0.1, 0.2,,...,0.9. From this it is seen that we have little damping in general. If there are two z_j with large values then $|R|$ will be close to 1. The same holds if we are outside the region of stability, where we may have $|R| > 1$ but very close to 1. Consequently, there may be a very slow instability.

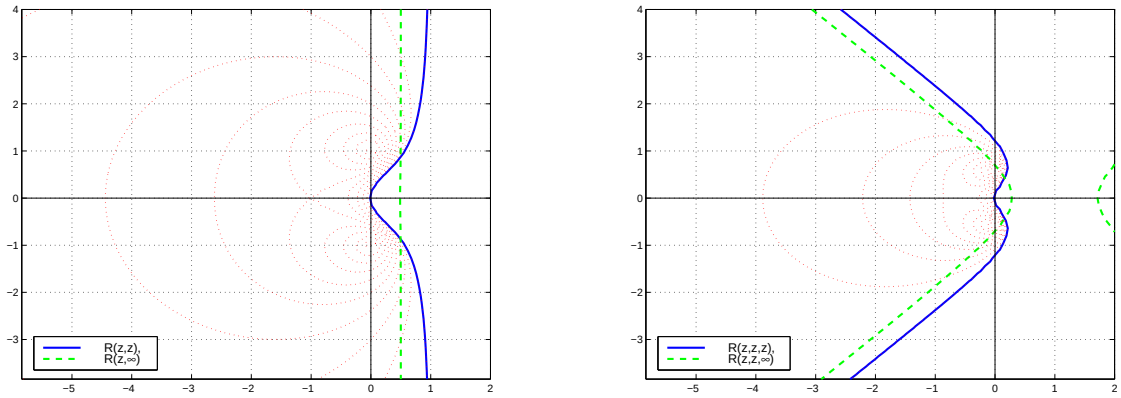


Fig. 3.3. Regions of stability $|R| \leq 1$ for $\theta = 1$, $z_0 = 0$, with equal $z_j = z$ or $z_s = \infty$. Left picture $s = 2$, right picture $s = 3$.

In conclusion, if we consider $\alpha = \frac{1}{2}\pi$, then the essential condition for stability is $z_1 \in \mathcal{W}_{\pi/2}$ and $z_2, \ldots, z_s \leq 0$, so only one of the implicit term should have eigenvalues that are large in modulus and not near the negative real axis. If this is violated, instability can be expected. This instability will be quite slow and therefore difficult to detect before it is too late.

Example. To illustrate the slow onset of instability, we consider the following advection equation with a simple linear reaction term,

$$u_t = a u_x + b u_y + G\, u, \quad (x, y) \in [0, 1]^2,\ 0 \leq t. \tag{3.22}$$

The velocities are given by $a(x, y, t) = 2\pi(y - \frac{1}{2})$, $b(x, y, t)) = 2\pi(\frac{1}{2} - x)$. Further,

$$u = u(x, y, t) = \begin{pmatrix} u_1(x, y, t) \\ u_2(x, y, t) \end{pmatrix}, \qquad G = \begin{pmatrix} -k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix}.$$

We take $k_1 = 1$. The second reaction constant k_2 can be used to vary the stiffness of the reaction term, and is taken here as 2000. Note that the matrix G has eigenvalues 0 and $-(k_1 + k_2)$, and we have a chemical equilibrium if $u_1/u_2 = k_2/k_1$.

The initial condition is chosen as

$$u_1(x,y,0) = c, \quad u_2(x,y,0) = (1-c) + 100\,k_2^{-1}\exp(-80((x-\tfrac{1}{2})^2 - 80(y-\tfrac{3}{4})^2),$$

with $c = k_2/(k_1+k_2)$. After the short transient phase, where most of the Gaussian pulse is transferred from u_2 to u_1, this is purely an advection problem, and the velocity field gives a rotation around the center of the domain. At $t = 1$ one rotation is completed. The exact solution is easily found by superimposing the solution of the reaction term onto the rotation caused by the advection terms.

Dirichlet conditions are prescribed at the inflow boundaries. At the outflow boundaries we use standard upwind discretization, in the interior second-order central differences are used. We consider splitting with F_1, F_2 the finite difference operators for advection in the x and y direction, respectively, and with F_3 defined by the linear reaction term. All three terms are treated implicitly. The corresponding eigenvalues λ_1, λ_2 will be close to the imaginary axis whereas $\lambda_3 = 0$ or $-(k_1+k_2)$. The test has been performed on a fixed 80×80 grid, and with $\tau = 1/160$.

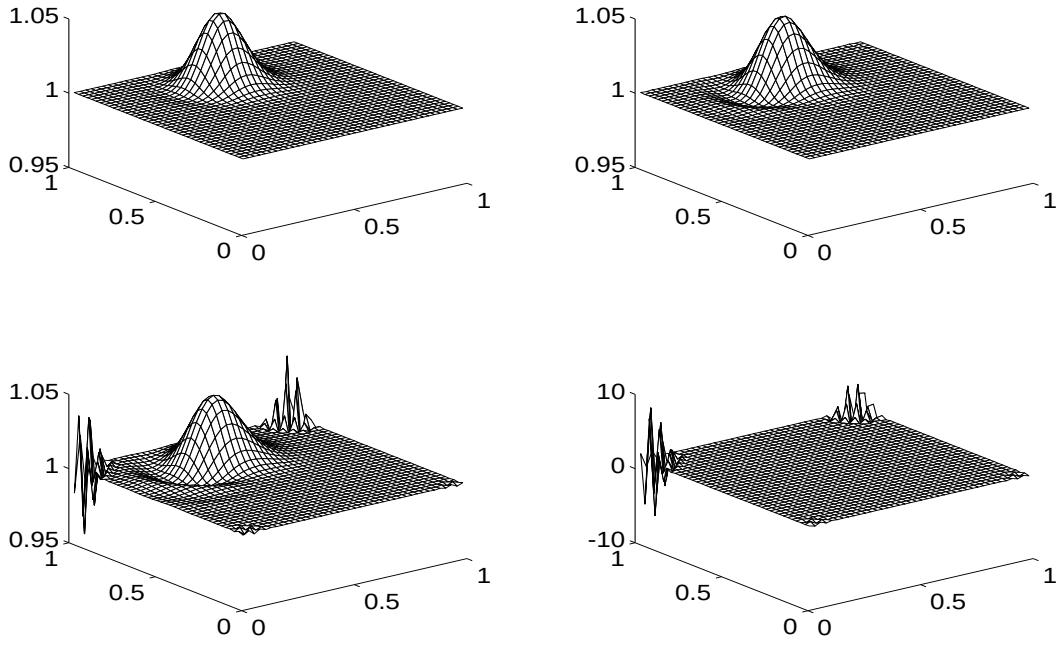


Fig. 3.4. Numerical solutions advection-reaction problem (3.22) at $t = 1, 2, 3, 4$.

The numerical solution of the first component u_1 for the scheme with $\theta = \frac{1}{2}$ is given in in Figure 3.4 at time $t = 1$ (top left), $t = 2$ (top right), $t = 3$ (bottom left) and $t = 4$ (bottom right; different scale). There are some smooth oscillations in the wake of the Gaussian pulse, but these are caused by the spatial discretization with central differences. The instabilities occur near the corners where both advection speeds, in x and y direction, are large. The build up of the instabilities is very slow, and therefore it will be difficult to detect this with error estimators. To some extend the slowness of the instability can be attributed to the fact that they occur near an outflow boundary, but related tests have shown that it is mainly caused by the fact that we have amplification factors only slightly larger than 1 in modulus.

Finally it should be noted that the advection treatment here, implicit with central differences, is only justified for problems with smooth solutions. If steep gradients may arise some upwinding or flux limiting is to be preferred. The experiment here merely serves as an illustration of the theoretical results on the stability of the ADI method with $s = 3$. $\diamond$

3.4 Rosenbrock methods with approximate matrix factorization (AMF)

With the above ADI method we still still are dealing with the explicit Euler method for F_0. To allow a second-order explicit method we first consider a linearization of this ADI method. In the following only autonomous equations are considered.

As starting point we consider the linearized θ-method

$$w_{n+1} = w_n + (I - \theta\tau A)^{-1}\tau F(w_n) \tag{3.23}$$

where A approximates the Jacobian matrix $F'(w_n)$. This is a so-called Rosenbrock method. It has order 1 if $\theta \neq \frac{1}{2}$ and order 2 if $\theta = \frac{1}{2}$ and $A - F'(w_n) = \mathcal{O}(\tau)$.

We consider the form where in the Jacobian approximation the nonstiff term is omitted and the rest is factorized in approximate fashion, that is

$$w_{n+1} = w_n + (I - \theta\tau A_s)^{-1} \cdots (I - \theta\tau A_2)^{-1}(I - \theta\tau A_1)^{-1}\tau F(w_n) \tag{3.24}$$

with $A_j \approx F_j'(w_n)$. The order of this approximate factorization method is 1 in general. For second-order we need $\theta = \frac{1}{2}$ and $F_0 = 0$. If the problem is linear this approximate factorization method is identical to the Douglas ADI method. Hence the linear stability properties are the same.

A 2-stage generalization of the above approximate factorization method is given by

$$\begin{aligned} w_{n+1} &= w_n + \tfrac{3}{2}k_1 + \tfrac{1}{2}k_2, \\ Mk_1 = \tau F(t_n, w_n), \qquad & Mk_2 = \tau F(t_n + c\tau, w_n + k_1) - 2k_1, \end{aligned} \tag{3.25}$$

where $M = \prod_{j=1}^{s}(I - \theta\tau A_j)$, $A_j \approx F_j'(w_n)$ and θ is a free parameter. The order of this method is 2 (in the classical ODE sense). If $F_0 = 0$ and $F_1 = F$ this is a well-known Rosenbrock method that has the special property that the order is not influenced by the Jacobian approximation. This Rosenbrock method is A-stable for $\theta \geq \frac{1}{4}$. On the other hand, if $F = F_0$ we now get a second-order explicit Runge-Kutta method.

The above method has been proposed in Verwer et al. [15], and in that paper the scheme was applied successfully on some 3D atmospheric transport-chemistry. problems. There operator splitting was used with F_0 advection, F_1 diffusion and F_2 reaction, and the free parameter was taken as $\theta = 1 + \frac{1}{2}\sqrt{2}$ to have optimal damping (L-stability). The eigenvalues of F_1 and F_2 were close to the negative real axis, and therefore stability problems were not expected, and indeed did not occur.

It is for such problems, where the structure of the eigenvalues can be well predicted in advance, that these approximate factorization methods seem suited. For general applications values θ in the range $[\frac{1}{2}, 1]$ seem more suitable than $\theta = 1 + \frac{1}{2}\sqrt{2}$, because the latter value gives relatively large error constants.

The above Rosenbrock methods are formulated here for autonomous problems. A nonautonomous problem $w'(t) = F(t, w(t))$ can be written as $v'(t) = G(v(t))$ with $v = (t, w)^T$ and

$G(v) = (1, F(t,w))^T$, and so the methods can be applied to this artificial autonomous problem. Then t is formally also considered as an unknown, but it is easily seen that the approximation t_n found with this method still equals $n\tau$. When reformulated on the original level, in terms of w_n, the methods will now also involve approximations to the derivatives $F_t(t,w)$. For example, with $A_j \approx \partial_w F_j(t_{n+\theta}, w_n) \in \mathbb{R}^{m\times m}$, $b_j \approx \partial_t F_j(t_{n+\theta}, w_n) \in \mathbb{R}^m$ and

$$B_j = \begin{pmatrix} 1 & 0 \cdots 0 \\ (I-\theta\tau A_j)^{-1}\theta\tau b_j & (I-\theta\tau A_j)^{-1} \end{pmatrix} \in \mathbb{R}^{(m+1)\times(m+1)},$$

the factorized Rosenbrock scheme (3.23) then reads

$$\begin{pmatrix} t_{n+1} \\ w_{n+1} \end{pmatrix} = \begin{pmatrix} t_n \\ w_n \end{pmatrix} + B_s \cdots B_2 B_1 \begin{pmatrix} \tau \\ \tau F(t_n, w_n) \end{pmatrix}.$$

We will have $t_{n+1} = t_n + \tau$, as it should be, and the computation of w_{n+1} can be written in the more transparent recursive form

$$dv_0 = \tau F(t_n, w_n)\,, \quad dv_j = (I-\theta\tau A_j)^{-1}\big(\theta\tau^2 b_j + dv_{j-1}\big) \ \ (1 \le j \le s)\,, \quad w_{n+1} = w_n + dv_s\,.$$

Note. It is also possible to linearize a multistep method and then use approximate factorization. Such methods can be found in Warming & Beam [16]. Runge-Kutta methods of the IMEX type have been studied recently by many authors, see the references in [6]; if such methods are applied in a linearized form, they are similar to the above factorized Rosenbrock methods with $s = 1$.

Remark: Modified Newton Iterations. Instead of the above techniques, one could also use a well-known fully implicit method and then try to modify the Newton process such that the computational ease is comparable to the IMEX or approximate factorization methods. The advantage is that if the iteration converges, then the theoretical properties of the fully implicit method are valid.

Consider a generic implicit relation

$$w_{n+1} = W_n + \theta\tau F(w_{n+1})\,, \tag{3.26}$$

where W_n contains the information up to t_n. This may be for instance Backward Euler ($\theta = 1$, $W_n = w_n$), the Trapezoidal Rule ($\theta = \frac{1}{2}$, $W_n = w_n + \frac{1}{2}\tau F(t_n, w_n)$) or the BDF2 method ($\theta = \frac{2}{3}$, $W_n = \frac{4}{3}w_n - \frac{1}{3}w_{n-1}$). Then the Newton iteration to solve the implicit relation will look like

$$u_{i+1} = u_i - M^{-1}\Big(u_i - \theta\tau F(u_i) - W_n\Big)\,, \quad i = 0, 1, 2, ... \tag{3.27}$$

with initial guess u_0. Standard modified Newton would be $M = I - \theta\tau A$ with $A \approx F'(v_0)$. For systems of multi-dimensional PDEs this leads to a very big linear algebra problem that has to be solved by a preconditioned conjugate gradient or multigrid method for example.

As an alternative one can consider approximate factorization inside the Newton process,

$$M = \prod_{j=1}^{s} (I - \theta\tau A_j) \tag{3.28}$$

with $A_j \approx F_j'(v_0)$, but now we have to look at convergence of the iteration.

When applied to the scalar test equation this iteration process has a convergence factor

$$S = 1 - \Big(\prod_{j=1}^{s}(1-\theta z_j)\Big)^{-1}\Big(1-\theta\sum_{j=0}^{s} z_j\Big) \tag{3.29}$$

and for the iteration to converge we need $|S\,| < 1$. This looks very similar to the stability factor with the Douglas ADI method. Indeed, the statements given previously for $|R\,| \le 1$ with the z_j in wedges are also valid for the convergence factor, see [5].

In the next figure the boundaries of the convergence region are plotted for special choices of z_j with $z_0 = 0$, similar to Figure 3.3. The dotted curved lines are the contour lines for $|S\,|$ with all z_j equal. If the z_j assume large negative values, then $|S\,|$ is close to 1 and thus the convergence will be very slow. Moreover divergence may occur if $s \ge 3$ and two or more of the z_j are close to the imaginary axis.

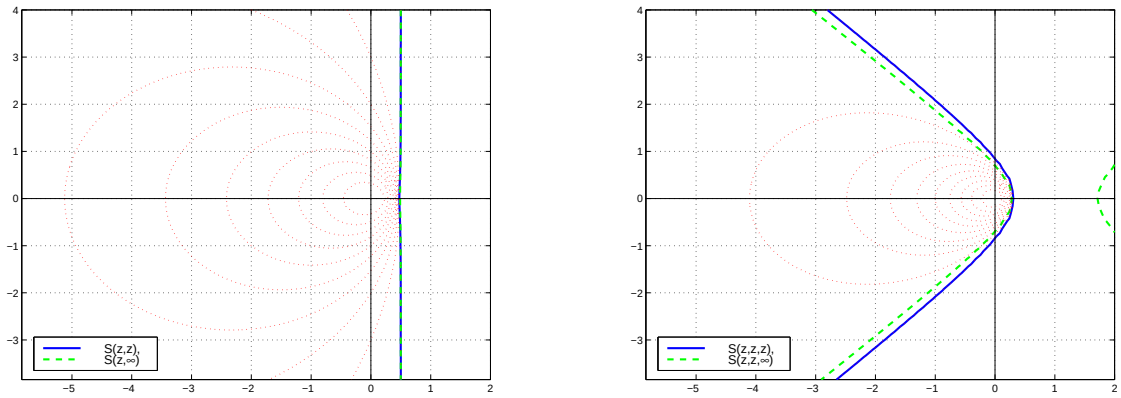


Fig. 3.5. Regions of convergence $|S\,| < 1$ for $\theta = 1$ with equal $z_j = z$ or $z_s = \infty$. Left picture $s = 2$, right picture $s = 3$.

In conclusion it can be said that the convergence of such a modified Newton iteration with approximate factorizations is often very poor, so it is not an approach that is recommended for general equations. Of course, there are special cases, especially with smooth solutions (no high Fourier harmonics), where this approach may work well. However the class of problems where the iteration does not diverge seems close to the class where the Rosenbrock schemes with approximate factorizations are be stable, see Figures 3.3 and 3.5. In those cases the simpler Rosenbrock schemes with approximate factorizations will be more efficient, and with such Rosenbrock schemes smoothness of the solution is not required.

3.5 Numerical illustration

In this section some numerical illustrations are given for the schemes applied to a simple 2D advection-diffusion-reaction equation (see [6] for more realistic problems). We shall refer to the 1-stage scheme (3.23) as ROS1 and to the 2-stage scheme (3.25) as ROS2, and for both schemes parameter values $\theta = \frac{1}{2}$ and 1 are considered.

We consider here the following 2D equation, on spatial domain $\Omega = [0,1]^2$ and $t \in [0,1]$,

$$u_t + \alpha\big(u_x + u_y\big) = \epsilon\big(u_{xx} + u_{yy}\big) + \gamma u^2(1-u)\,, \tag{3.30}$$

with traveling wave solution

$$u(x,y,t) = \Big(1 + \exp(a(x+y-bt)+c)\Big)^{-1}. \tag{3.31}$$

Here $a = \sqrt{\gamma/4\epsilon}$ determines the smoothness of the solution, $b = 2\alpha + \sqrt{\gamma\epsilon}$ is the velocity of the wave and $c = a(b-1)$ a shift parameter. Initial and Dirichlet boundary conditions are prescribed so as to correspond with this solution. Due to the time-dependent boundary conditions, the semi-discrete problem is non-autonomous and the Rosenbrock methods are applied to the extended autonomous form.

For this scalar test example splitting is not really necessary, but the structure of the equations is similar to many real-life problems where splitting cannot be avoided with present day computer (memory) capacities. In Verwer et al. [15] application the ROS2 method can be found for a large scale 3D problem from atmospheric dispersion.

Reaction-diffusion test. First we consider the above test equation with $\alpha = 0$. To give an illustration of the convergence behaviour of the various methods we take $\gamma = 1/\epsilon = 10$, which gives a relatively smooth solution.

For this smooth problem the spatial derivatives are discretized with standard second-order finite differences. Let $D^{(x)}(t,u) = A^{(x)}u + g^{(x)}(t)$ stand for the finite difference approximation of ϵu_{xx} with the associated time-dependent boundary conditions for $x = 0$ and $x = 1$. Likewise $D^{(y)}(t,u)$ approximates ϵu_{yy} with boundary conditions at $y = 0$, $y = 1$. Further, $G(t,u)$ represents the reaction term $\gamma u^2(1-u)$ on the spatial grid. We consider the following two splittings with $s = 3$ and $F_0 = 0$,

$$(A) \quad \cdots \qquad F_1 = D^{(x)}, \quad F_2 = D^{(y)}, \quad F_3 = G,$$

and

$$(B) \quad \cdots \qquad F_1 = G, \quad F_2 = D^{(x)}, \quad F_3 = D^{(y)}.$$

Since the reaction term in (3.30) with $\gamma = 10$ is not stiff, we also consider here the case where this term is taken explicitly,

$$(C) \quad \cdots \qquad F_0 = G, \quad F_1 = D^{(x)}, \quad F_2 = D^{(y)}.$$

The spatial grid is uniform with mesh width h in both directions. The errors in the L_2-norm are calculated at output time $T = 1$ with $\tau = h = 1/N$, $N = 10, 20, 40, 80$. In the Figure 3.6 these errors are plotted versus τ on a logarithmic scale. The results for the ROS1 scheme are indicated by dashed lines with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$. Likewise, the results for the ROS2 scheme are indicated by solid lines with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$.

For comparison, results of the well-known fractional step (LOD) method of Yanenko are included, indicated by dotted lines with stars. With this method fractional steps are taken with the implicit trapezoidal rule $v_j = v_{j-1} + \frac{1}{2}\tau F_j(t_n, v_{j-1}) + \frac{1}{2}\tau F_j(t_{n+1}, v_j)$, with $v_0 = w_n$. After each step the order of the F_j is interchanged to achieve symmetry and second order (in the classical ODE sense), see formula (2.7) with $c = \frac{1}{2}$. If an explicit term F_0 is present, the implicit trapezoidal rule is replaced by its explicit counterpart for the fractional step with F_0.

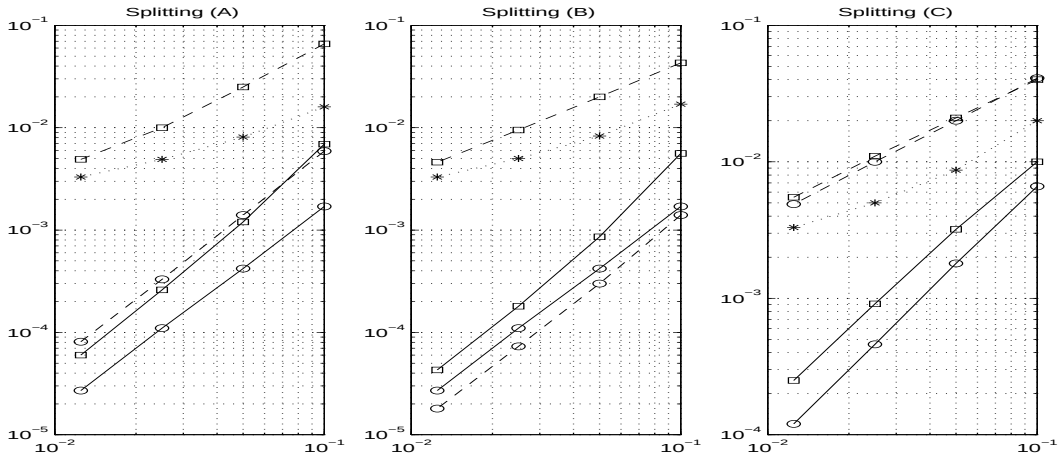


Fig. 3.6. L_2-errors versus $\tau = h$ for the splittings (A), (B) and (C). Methods ROS1 (dashed lines) and ROS2 (solid lines) with $\theta = \frac{1}{2}$ (circles) and $\theta = 1$ (squares). Results for Yanenko's method are indicated with stars.

It is known that Yanenko's method needs boundary corrections to obtain second-order convergence for initial-boundary value problems, otherwise the order of convergence can be lower; see e.g. [6]. In the present test we get convergence with order $\frac{1}{2}$ approximately. The test was repeated with boundary corrections, but still the results were less accurate than with the second-order ROS schemes. Finally we note that boundary corrections were also attempted on the Douglas scheme, similar to formula (101) in Mitchell & Griffiths [9]. In the above test this did lead to smaller errors, reduction with a factor ranging between 1.2 and 2, but the convergence behaviour did not change fundamentally. Since boundary corrections have to be derived for each individual problem, it is a favourable property of the stabilizing correction schemes that such corrections are not necessary to get a genuine second-order behaviour.

Advection-diffusion-reaction test. To illustrate the improved stability behaviour of the 2-stage scheme ROS2 over ROS1 if a substantial explicit term is present, we now consider the test equation with a advection term with $\alpha = -1$ that will be taken explicitly. Further we choose $\gamma = 100$ and $\epsilon = 0.01, 0.001$ which gives solutions that have a steep gradient, relative to the mesh widths used here.

The splitting is such that F_0 contains the convective terms, F_1, F_2 diffusion in x and y direction, respectively, and F_3 the nonlinear reaction term. The convective terms are discretized with third-order upwind-biased differences (4-point stencil). For the diffusion terms standard second-order differences are used as before.

The results with $\epsilon = 0.01$ are given in the Figures 3.7, 3.8. In the plots of Figure 3.7 the solutions $h = 1/40$ and $\tau = 1/80$ are found, represented as contour lines at the levels 0.1, 0.2,...,0.9, with solid lines for the numerical solution and dotted lines for the exact solution. Quantitative results are given in Figure 3.8, where the L_2-errors are plotted as function of the time step for a 40×40 and 80×80 grid with $\tau = h, \frac{1}{2}h$ and so on. As in Figure 3.6 results for ROS1 are indicated with dashed lines, for ROS2 with solid lines, and with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$.

It is obvious that the 2-stage schemes ROS2 give much better results than the corresponding

1-stage schemes ROS1. To achieve a level of accuracy comparable to the ROS2 schemes we need much smaller time steps with the ROS1 schemes, see Figure 3.8. This is primarily due to the more stable treatment of the explicit advection term with the ROS2 schemes. The explicit 2-stage Runge-Kutta method underlying ROS2 is stable for third-order advection discretization up to Courant number 0.87 (experimental bound). On the other hand, some of the eigenvalues associated with this discretization are always outside the stability region of the explicit Euler scheme. In this test it is the (implicit) diffusion part that provides a stabilization for the smaller step sizes. (In fact, for $\epsilon = 0.01$ similar results were obtained with second-order central advection discretization, but not anymore with $\epsilon = 0.001$). Further we note that instabilities do not lead to overflow since the solutions are pushed back to the range [0,1] by the reaction term, but the resulting numerical solutions are qualitatively wrong.

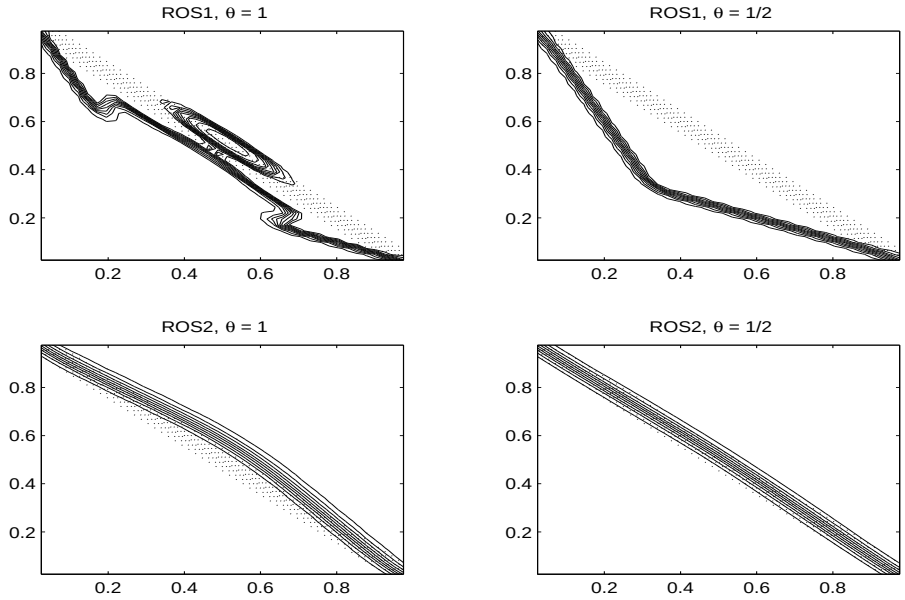


Fig. 3.7 Contour lines numerical solutions for $\epsilon = 0.01$ with $h = 1/40$, $\tau = 1/80$.

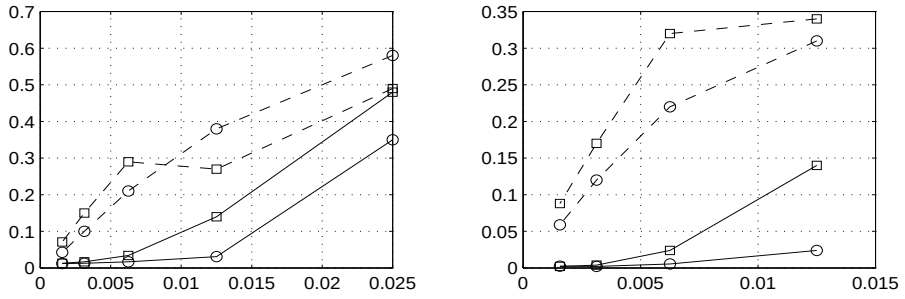


Fig. 3.8. L_2-errors versus time step τ on 40×40 grid (left) and 80×80 grid (right) for $\epsilon = 0.01$. Various methods indicated as in Figure 3.6.

Decreasing the value of the diffusion coefficient ϵ gives a clearer distinction between the methods. Results with $\epsilon = 0.001$ are given in the Figures 3.9 and 3.10. The grids chosen are 80×80 and 160×160, since the 40×40 grid gives quite large spatial errors with this small ϵ. The results are essentially the same as above: the 1-stage schemes ROS1 need much smaller time steps than the ROS2 schemes to obtain reasonable solutions.

For more realistic problems with stiff reaction terms, nonlinear advection discretizations with flux limiters are recommended to avoid oscillations, and this fits easily into the present framework due to the explicit advection treatment.

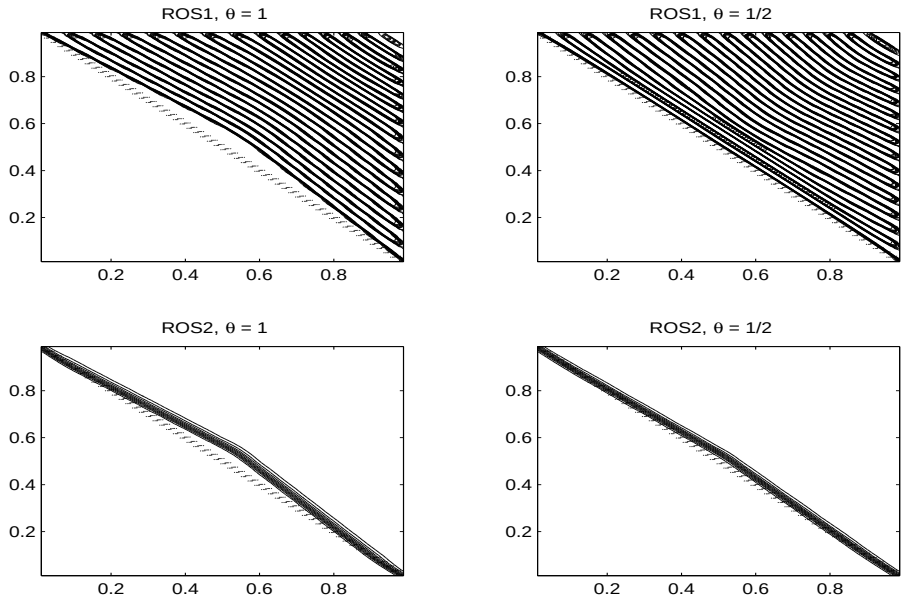


Fig. 3.9 Contour lines numerical solutions for $\epsilon = 0.001$ with $h = 1/80$, $\tau = 1/160$.

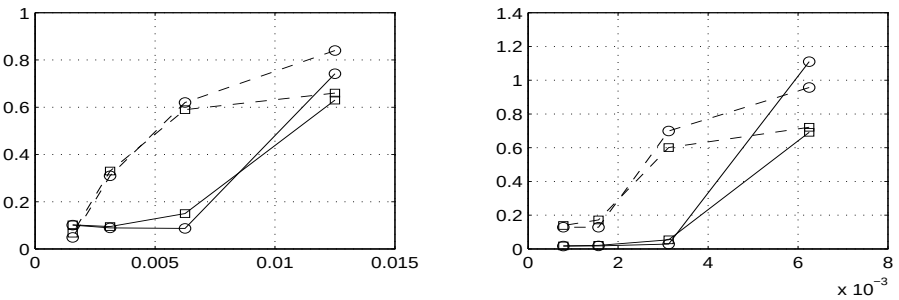


Fig. 3.10. L_2-errors versus time step τ on 80×80 grid (left) and 160×160 grid (right) for $\epsilon = 0.001$. Various methods indicated as in Figure 3.6.

References

[1] U.M. Ascher, S.J. Ruuth, B. Wetton, *Implicit-explicit methods for time-dependent PDE's.* SIAM J. Numer. Anal. 32 (1995), pp. 797–823.

[2] M. Crouzeix, *Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques.* Numer. Math. 35 (1980), pp. 257–276.

[3] J. Frank, W. Hundsdorfer, J.G. Verwer, *On the stability of implicit-explicit linear multistep methods.* Appl. Numer. Math. 25 (1997), pp. 193–205.

[4] W. Hundsdorfer, *A note on stability of the Douglas splitting method.* Math. Comp. 67 (1998), pp. 183–190.

[5] W. Hundsdorfer, *Stability of approximate factorizations with θ-methods.* BIT 39 (1999), pp. 473–483.

[6] W. Hundsdorfer, J.G. Verwer, *Numerical Solution of Advection-Diffusion-Reaction Equations.* Springer Series in Computational Mathematics 33, Springer Verlag, 2003.

[7] R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems.* Cambridge Texts in Applied Mathematics, Cambridge University Press, 2002.

[8] G.I. Marchuk, *Splitting and alternating direction methods.* In: *Handbook of Numerical Analysis I.* Eds. P.G. Ciarlet, J.L. Lions, North-Holland, Amsterdam, 1990, pp. 197–462.

[9] A.R. Mitchell, D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations.* John Wiley & Sons, Chichester, 1980.

[10] D.W. Peaceman, *Fundamentals of Numerical Reservoir Simulation.* Elsevier, Amsterdam, 1977.

[11] J.M. Sanz-Serna, M.P. Calvo, *Numerical Hamiltonian Problems.* Applied Mathematics and Mathematical Computation, Vol. 7, Chapman & Hall, London, 1994.

[12] G. Strang, *On the construction and comparison of difference schemes.* SIAM J. Numer. Anal. 5 (1968), pp. 506–517.

[13] J.M. Varah, *Stability restrictions on second order, three-level finite-difference schemes for parabolic equations.* SIAM J. Numer. Anal. 17 (1980), pp. 300–309.

[14] J.G. Verwer, W.H. Hundsdorfer, J.G. Blom, *Numerical time integration for air pollution models.* Surveys on Mathematics for Industry 10 (2002), pp. 107–174.

[15] J.G. Verwer, E.J. Spee, J.G. Blom, W. Hundsdorfer, *A second order Rosenbrock method applied to photochemical dispersion problems.* SIAM J. Sci. Comput. 20 (1999), pp. 1456–1480.

[16] R.F. Warming, R.M. Beam, *An extension of A-stability to alternating direction methods.* BIT 19 (1979), pp. 395–417.

[17] N.N. Yanenko, *The Method of Fractional Steps.* Springer, Berlin, 1971.

[18] Z. Zlatev, *Computer Treatment of Large Air Pollution Models.* Kluwer, Dordrecht, 1995.

Modelling and Simulation in Chemical Engineering

Alírio E. Rodrigues

Laboratory of Separation and Reaction Engineering

Departamento de Engenharia Química

Faculdade de Engenharia da Universidade do Porto

Rua Dr Roberto Frias, 4200-465 Porto

A short note

José Almiro Abrantes de Menezes e Castro (1953-2002)

This is an opportunity to dedicate this presentation to José Almiro. I met him in 1978 when I was coming on saturday to teach System Dynamics and Design courses at the University of Coimbra and he was starting his academic carreer. Later he went to the University of Leeds where he worked with Prof C. Mc Greavy and defended a PhD thesis on "Aspects of modeling chemical processes for adaptive control" (1983) .

He contributed a lot for the use of modeling and simulation tools not only in academia but also in industry. The last time I met him was during Chempor'2001, September 2001 in Aveiro. I was asked to organize this Workshop as a CIM event "Mathematics and Chemical Engineering" (2003) in collaboration with José Almiro and Paula Oliveira (Department of Mathematics, University of Coimbra). The deadline to submit the proposal was approaching and I tried to call José Almiro. After several trials I learned that he was leaving the Hospital where he knew he had a big fight awaiting him. Our phone conversations occurred immediately after and at a time he was optimistic but unfortunately he could not win the difficult battle.

I hope his example will flourish in Coimbra.

INDEX

Following the advise of Dwig Prater given to Jim Wei [1] on how to present a communication ("tell them what you will tell them, then you tell them and finally you tell them what you have just told them") my talk will be organized as follows:

Back to the origins

Unit operation- first paradigm of Chemical Engineering

George E. Davis (1850-1906) made a proposal (without success) in 1880 for the creation of the "Society of Chemical Engineers in London". In 1887 he gave a series of lectures on Operation of Chemical Processes at Manchester Technical School and published the "Handbook of Chemical Engineering" (1901). His approach, in terms of unit operations, emphasizes the importance of experimentation at pilot scale and safety rules; he uses the term "chemical engineering" to designate the profession then emerging which corresponds in a certain way to today's chemical engineer.

Lewis Mills Norton (1855-1893), Professor of Industrial and Organic Chemistry at MIT, taught in 1888 the first 4 years course in Chemical Engineering -"Course X". The first chemical engineer to complete that course in 1891 was William Page Bryant; his job was in insurance auditor for the Boston Board of Fire Underwriters. Any similarity with current market situation is merely coincidence.

George Davis

Lewis Norton

William H. Walker (1869–1934), Warren K. Lewis (1882–1975), and Arthur D. Little are the pioneers who defined chemical engineering as a profession with proper approach and *training* methods.

William H. Walker (1869-1934)

Arthur D. Little

Warren K. Lewis (1882-1975)

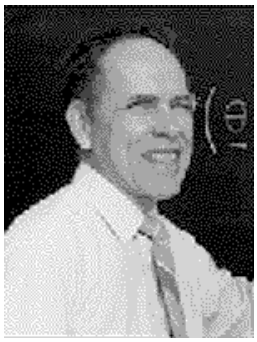
Arthur D. Little was the first to use the term "unit operations" in a report (1915) to the president of MIT. He created in 1886 the company later known as Arthur D. Little, Inc. In 1900 he is associated with William H. Walker (BSc in Chemistry, Penn State and PhD Organic Chemistry, Gottingen) but Walker left to restructure the curriculum of ChE at MIT and in 1908 the Research Laboratory of Applied Chemistry is created. Students worked in real problems given by industry which

was also paying the grants! Warren K. Lewis, who attended the ChE programme at MIT and PhD in Organic Chemistry from the University of Breslau, became a staff member of MIT in 1908; his ability to theorize engineering problems and his strong character contributed to strengthen the programme. The teaching of unit operations became quantitative. In 1916 three units of the School of Chemical Engineering Practice were opened where students worked 8 weeks on experimental work. This period is condensed in the teaching manual: *The Principles of Chemical Engineering* (1923) by Walker, Lewis, and William H. McAdams. Walker, sometimes considered the father of ChE goes back to consulting. Lewis continues in MIT in collaboration with Standard Oil Company (later Exxon) and with Gilliland is the inventor of FCC of petroleum (gasoline for WWII) [2].

Engineering Science Movement- second paradigm of Chemical Engineering

This movement is illustrated by the book of Bird, Stewart e Lighfoot "Transport phenomena" [3] which treats in parallel transfer processes of momentum, heat and mass, which are after all a major portion of the ChE activity. The transport phenomena approach was initiated by Kramers (Delft University of Technology) where Bird spent a semester and knew the lecture notes *Physische Transportverschijnselen.*

Bird



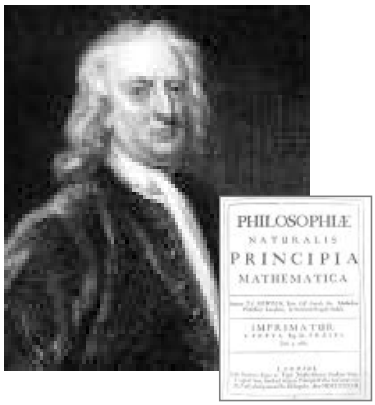
Stewart

Lighfoot

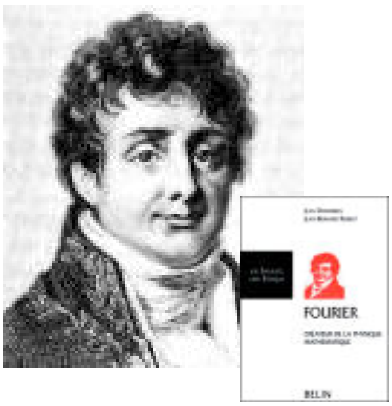
Momentum transfer – Newton's law: $\tau_{yx} = -\mu \frac{dv_x}{dy}$

Heat transfer by conduction – Fourier's law : $q_y = -k \frac{dT}{dy}$

Mass transfer by diffusiono – Fick's law: $j_y = -D_{AB} \frac{d\rho_A}{dy}$ at constant ρ



Newton



Fourier

Fick

The above laws assume infinite velocity of propagation of the signal. This problem is eliminated following the proposal of James Clerk Maxwell for momentum transfer

$$\sigma_{yx} + \tau \frac{\partial \sigma_{yx}}{\partial t} = -\mu \frac{\partial u_x}{\partial y} \qquad [1]$$

where τ is the time constant and the shear stress σ_{yx} in a fluid or solid body σ_{yx};

$$j + \tau \frac{\partial j}{\partial t} = -D \frac{\partial c}{\partial x} \qquad [2]$$

for mass transfer [4]and

$$q + \tau \frac{\partial q}{\partial t} = -k \frac{\partial T}{\partial x} \qquad [3]$$

for heat transfer (VC equation of Vernott and Cattaneo) [5] . In homogeneous substances the relaxation time is 10^{-8}-10^{-14} s and Fourier's law works for normal heating processes. But in biological systems τ is of the order of 10-30 s and CV equation applies.

Philosophy of process modelling

"Le Génie Chimique c'est pas de la plomberie" (P. Le Goff)

I remember Professor Pierre Le Goff when I was a student in Nancy. He said that a chemical engineer when solving a problem writes [6]:

- conservation equations (mass, energy, momentum, electric charge)
- equilibrium law at the interface (s)

- constitutive laws (for example, ideal gas law)
- kinetic laws of transport (heat/mass) and reaction
- initial and boundary conditions
- optimization criteria

This methodology has been useful to analyze problems at various scales involved in ChE :

- pore scale (catalyst, adsorvent) : 1nm – 1000nm
- particle scale : 10 μm- 1 cm
- reactor/separator scale: 1m- 10m

R. Aris

According to Aris [7] a "mathematical model" or simply model "is a complete and consistent set of mathematical equations which are supposed to correspond to some entity – its prototype - which can be a physical, biological, social…entity although here we deal with physicochemical systems". A process model is a relation between "outputs" and "inputs" (feed conditions, design parameters, process adjustable parameters; Shinnar) in view of : i) scale-up from lab to industrial scale; ii) prediction of process dynamics and iii) optimisation of operating conditions.

Models: simplification of reality ; to "better" know the reality

The detail of mathematical description can be guided by objectives which can seem contradictory:

a) Simplification of reality - idealization

In an excellent paper Levenspiel [8], a pioneer of Chemical Reaction Engineering mention Denbigh [9] : "In science it is always necessary to abstract from the complexity of the real world, and in its place to substitute a more or less idealized situation that is more amenable to analysis". This idealization leads to the creation of new models, simplified, which are a "digital impression" of our profession. Examples are: i) boundary layer theory; ii) model of film heat transfer; h; iii) model of film mass transfer; k; iv) theory of residence time distribution (RTD) and tracer technology.

b) Detailed model to "better" know the reality

An example is the Maxwell-Stefan model for multicomponent diffusion [10].

The driving force is the gradient of chemical potential, μ which is for ideal gas:

$$\mu_i = \mu_i^o + RT \ln p_i \qquad [4]$$

The diffusive flux due to this gradient is balanced by friction forces:

$$fu_i = -\frac{d\mu_i}{dz} \qquad [5]$$

where f is the friction coefficient and u_i is the velocity of species i.

J.C. Maxwell (1831-1879)

The flux is:

$$N_i = u_i C_i = -\frac{RT}{f}\frac{\partial \ln p_i}{\partial \ln C_i}\frac{dC_i}{dz} \qquad [6]$$

where the "corrected diffusivity" is $D_i^0 = \frac{RT}{f}$ and the thermodynamic factor is $\frac{\partial \ln p_i}{\partial \ln C_i}$.

For binary systems if the change in partial pressure of species 1 is -dp$_i$ over distance dz, the force acting in 1 per volume is $-\frac{dp_1}{dz}$; if the concentration of 1 is C_1 the force per mole of species 1 is $-\frac{1}{C_1}\frac{dp_1}{dz}$ and for an ideal gas

$$-\frac{RT}{p_1}\frac{dp_1}{dz} = -RT\frac{d\ln p_1}{dz} \quad \text{or}$$

$$-RT\frac{d\ln p_1}{dz} = -\frac{d\mu_1}{dz} \qquad [7]$$

This force is balanced by the friction between species 1 and 2, proportional to the difference of velocities and to the concentration of component 2 , expressed by x_2. The balance of forces acting in species 1 is

$$-\frac{d\mu_1}{dz} = \frac{RT}{D^{MS}}x_2(u_1 - u_2) \qquad [8]$$

or:

$$-\frac{1}{P}\frac{dp_1}{dz} = \frac{1}{D^{MS}}x_1x_2(u_1 - u_2) \qquad [9]$$

For n components:

$$d_i = \nabla x_i = \sum_{j=1}^{n}\frac{x_iN_j - x_jN_i}{c_tD_{ij}^{MS}} \qquad [10]$$

Strategy of modelling

A philosophy of modelling can be based in 4 points [11]:

a. Start with simple models; obtain from such models information which remains valid for more complex models (10$ approach of Levenspiel [8]: "Always start by trying the simplest model and then only add complexity to the extent needed ").

b. The validity of a model is not just a result of a "good fit"; more important is the capability to predict the system behavior under operating conditions different from those used to get model parameters.

c. Good results can only be obtained if the model "well" represents the

d. Use models to obtain useful design parameters and their dependence on operating conditions; use independent experiments if possible to get model parameters.

Levenspiel

In short: model development is a task to be carefully done to avoid waste of energy in the next simulation step. "Keep things as simple as possible, but not simpler" (Einstein)

The "art" of modelling

The "art" of modelling uses some techniques (tricks) such as: adimensionalization and scaling, averaging, appropriate choice of independent variables.

Scaling and dimensionless groups

Chemical engineers have some habits as normalization of variables to get scales between 0 and 1. As a consequence of that mathematical operation dimensionless groups appear with a physical meaning.

An example: diffusion/reaction in an isothermal porous catalyst with slab geometry. The mass balance in steady state for irreversible reaction of order n is:

$$D_e \frac{d^2 c_i}{dz^2} - k c_i^n = 0 \qquad [11]$$

with boundary conditions (symmetry condition in the center and surface condition)

$$z = 0, \frac{dc_i}{dz} = 0$$
$$z = \ell, c_i = c_{iS} \qquad [12]$$

The normalization of space variable and concentration variable by:

$$x = z / \ell$$
$$f_i = c_i / c_{iS}$$

leads to

$$\frac{d^2 f_i}{dx^2} - \ell^2 \frac{k c_{iS}^{n-1}}{D_e} f_i^n = 0 \qquad [13]$$

which shows the dimensionless group governing the reaction/diffusion problem:

$$\ell^2 \frac{k c_{iS}^{n-1}}{D_e} = \phi^2 = Da_{II} \qquad [14]$$

where ϕ is the Thiele modulus.

The physical meaning of the dimensionless group is: $\phi^2 = Da_{II}$ = reaction rate/diffusion rate= diffusion time constant/ reaction time constant. Two extreme cases:

a) reaction rate << diffusion rate – concentration profile inside the catalyst is almost equal to the surface concentration; the catalyst works on "chemical regime";

b) b) reaction rate >> diffusion rate – the catalyst works in "diffusional regime" .

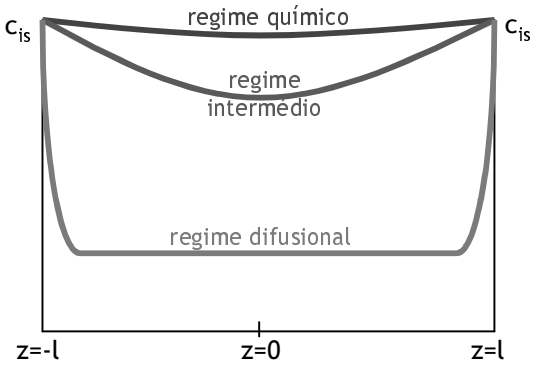


Figure 1. Chemical regime and diffusional regime in an isothermal catalyst

Averaging

Let us illustrate this technique with the LDF model (linear driving force) of Glueckauf [12].

For a spherical "homogeneous" adsorbent particle the mass conservation equation is:

$$\frac{\partial \bar{q}_i}{\partial t} = D_h \frac{1}{R^2} \frac{\partial}{\partial R}\left(R^2 \frac{\partial \bar{q}_i}{\partial R} \right) \qquad [15]$$

with boundary conditions (symmetry at the center and equilibrium with the fluid concentration at the surface through the adsorption equilibrium isotherm $f(c_i)$):

$$R = 0, \frac{\partial \bar{q}_i}{\partial R} = 0$$
$$R = R_p, q_{iS} = f(c_i) \qquad [16]$$

The averaging operation consists on multiplying both members by R^2dR, and integrate over the particle volume (between 0 and R_p) and introduce average concentration $\langle q_i \rangle$; the result is:

$$\frac{\partial \langle q_i \rangle}{\partial t} = \frac{3D_h}{R_p} \left. \frac{\partial \bar{q}_i}{\partial R} \right|_{R_p} = \frac{3D_h}{R_p} \frac{q_{iS} - \langle q_i \rangle}{\alpha R_p} = \frac{15 D_h}{R_p^2} \left(q_{iS} - \langle q_i \rangle \right) \qquad [17]$$

$$\frac{\partial \langle q_i \rangle}{\partial t} = \frac{15 D_h}{R_p^2}\left(q_{iS} - \langle q_i \rangle\right) = k_h\left(q_{iS} - \langle q_i \rangle\right) \qquad [18]$$

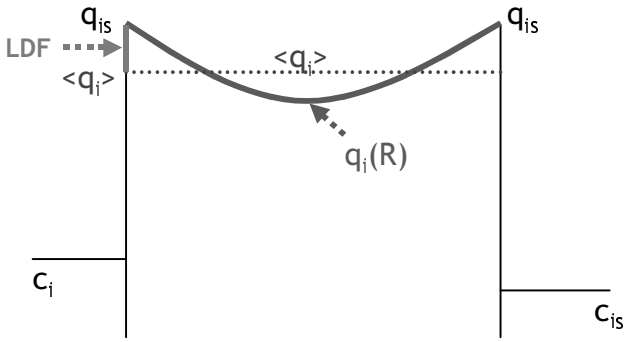


Figure 2. LDF model

Choice of variables

To illustrate this technique let us consider the equilibrium model of an isothermal adsorption column with plug fluid flow of a diluted stream (*trace system*). Model equations are the mass balance of the solute in a bed volume element and the equilibrium law at the interface fluid/solid:

$$u_0 \frac{\partial c_i}{\partial z} + \varepsilon \frac{\partial c_i}{\partial t} + (1-\varepsilon)\frac{\partial q_i^*}{\partial t} = 0$$
$$q_i^* = f(c_i) \qquad [19]$$

For an adsorption isotherm of " constant separation factor" type and normalizing the dependent variables, $\hat{c}_i = \frac{c_i}{c_{i0}}, \hat{q}_i = \frac{q_i}{q_{i0}}$ we get:

$$u_i \frac{\partial \hat{c}_i}{\partial z} + \frac{\partial \hat{c}_i}{\partial t} + \frac{1-\varepsilon}{\varepsilon}\frac{q_{i0}}{c_{i0}}\frac{\partial \hat{q}_i}{\partial t} = 0$$
$$\hat{q}_i = \frac{K \hat{c}_i}{1 + (K-1)\hat{c}_i} \qquad [20]$$

A first dimensionless parameter appears: the "capacity parameter" of the adsorption column

$$\xi_m = \frac{1-\varepsilon}{\varepsilon}\frac{q_{i0}}{c_{i0}} \qquad [21]$$

A combination of the independent variables z and t in only one variable T (*throughput parameter* of Vermeulen [13]) defined as the ratio of moles of solute passed through the bed section located at v=Az and the number of moles retained in the adsorbent contained in the volume v, is:

$$T = c_{i0}(V - \varepsilon v)/(1-\varepsilon) q_{i0} v \qquad [22]$$

The new variable $T = \frac{1}{\xi_m}\left(\frac{u_i t}{z} - 1\right)$ allows us to write the mass balance as:

$$\frac{d\hat{q}_i}{d\hat{c}_i} = T \qquad [23]$$

and taking into account the adsorption equilibrium isotherm

$$\frac{d\hat{q}_i}{d\hat{c}_i} = \frac{K}{[1+\hat{c}_i(K-1)]^2} \qquad [24]$$

we get

$$\hat{c}_i = \frac{1-\sqrt{\frac{K}{T}}}{1-K} \qquad K \le T \le 1/K \qquad [25]$$

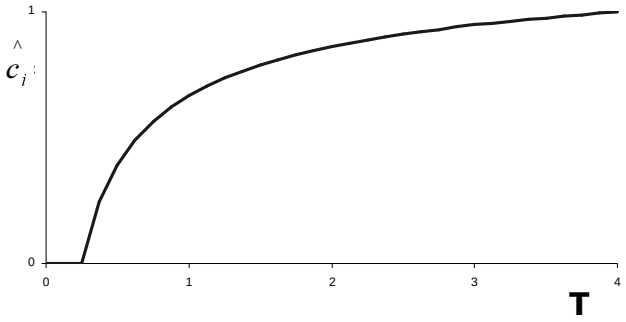


Figure 3. Breakthrough curves for unfavourable isotherms

From model results to real life

Back to the reaction/diffusion problem for first order reaction in isothermal catalyst. The concentration profile is:

$$f_i = \frac{\cosh(\phi x)}{\cosh\phi} \qquad [26]$$

The effectiveness factor of the catalyst (ratio between the observed rate and the reaction rate at reference conditions, e.g., surface) calculated by the Italian or German method (students will recognize Gauss theorem relating divergence and flux…) is:

$$\eta = \frac{\tanh\phi}{\phi} \qquad [27]$$

It is important to know the effectiveness factor to calculate the amount of catalyst one needs to have in the reactor to get a given reactant conversion.

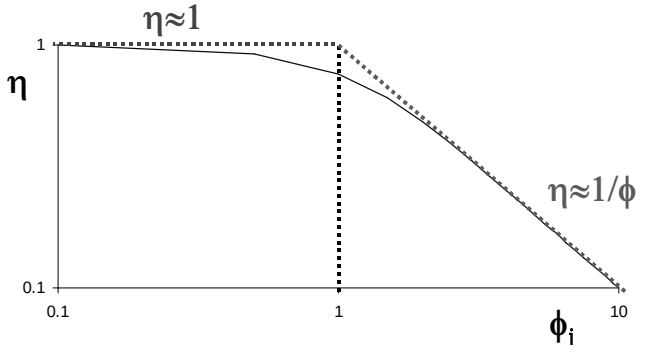


Figure 4. Effectiveness factor *versus* Thiele modulus.

But to know the Thiele modulus the kinetic constant k must be known(and many times it is not …).

Hopefully there are always bright people around to transform theoretical results in practical tools. Weisz and Prater [14] changed the plot $\eta = f(\phi)$ in another more useful $\eta = g(\eta\phi^2)$ where $\eta\phi^2$ does not require the knowledge of k; but only measurable quantities since:

$$\eta\phi^2 = \frac{r_{obs}\ell^2}{c_{iS}D_e} \qquad [28]$$

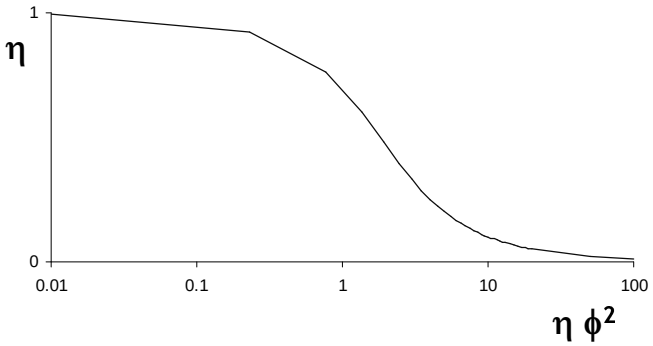


Figure 5. Effectiveness factor *versus* $\eta\phi^2$

Obtaining useful relations between dependent variables

Consider the diffusion/reaction/conduction problem in a non-isothermal catalyst. For slab geometry and first order irreversible reaction conservation equations of mass/energy are:

$$\begin{aligned} &D_e\frac{d^2c_i}{dz^2} - k(T)c_i = 0 \\ &\lambda_e\frac{d^2T}{dz^2} + (-\Delta H)k(T)c_i = 0 \\ &z = 0, \frac{dc_i}{dz} = \frac{dT}{dz} = 0 \\ &z = \ell, c_i = c_{iS}; T = T_s \end{aligned} \qquad [29]$$

Multiplying the first equation by the heat of reaction (-ΔH) and adding the second we get:

$$D_e(-\Delta H)\frac{d^2c_i}{dz^2} + \lambda_e\frac{d^2T}{dz^2} = 0 \qquad [30]$$

Integrating twice we obtain:

$$T - T_s = \frac{D_e(-\Delta H)}{\lambda_e}(c_{iS} - c_i) \qquad [31]$$

This equation was derived by Damkohler [15] and provides a relation between concentration and temperature in a point inside the catalyst. A similar treatment holds for adiabatic catalytic reactors using pseudo-homogeneous models. It is easier to measure temperature than concentrations!

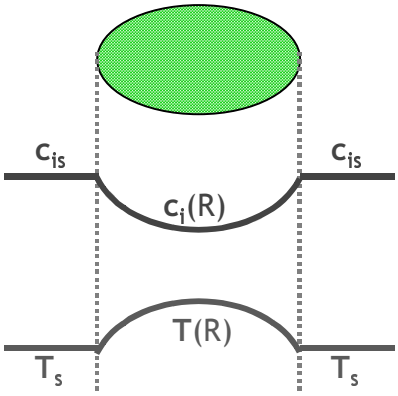


Figure 6. Concentration and temperature profiles in a non-isothermal catalyst.

Models of 10$, 100$ e 1000$ (Levenspiel)

Boundary layer concept

Prandtl

How is the fluid velocity affected when it flows near a solid surface?

The answer was given by Karman-Prandtl (1934). For flow parallel to a flat plate in laminar regime , $Re_x=v\rho x/\mu<2.10^5$ model equations are:

$$v_x\frac{\partial v_x}{\partial x}+v_y\frac{\partial v_x}{\partial y}=\nu\frac{\partial^2 v_x}{\partial y^2}$$

$$\frac{\partial v_x}{\partial x}+\frac{\partial v_y}{\partial y}=0 \qquad [32]$$

$$v_x=v_y=0; y=0$$

$$v_x=v_\infty; y=\infty$$

The velocity distribution in a tube from the laminar sublayer up to the central turbulent core is, in terms of $v^+=\frac{\bar{v}_x}{\sqrt{\tau_0/\rho}}$ versus $y^+=\frac{\sqrt{\tau_0/\rho}}{\nu}y$:

$$v^+=y^+; 5>y^+>0$$

$$v^+=-3.05+5\ln y^+; 30>y^+>5 \qquad [33]$$

$$v^+=5.5+2.5\ln y^+; y^+>30$$

Prandtl (1904) [16] proposed a simpler model for the velocity profiler – linear variation with distance from the solid surface or du/dy=0 and u=0 at y=0 in the viscous layer- and a non viscous layer away from the solid. Von Karman comment about Prandtl: "Prandtl (1875-1953) was an engineer by training. His control of mathematical methods and tricks was limited…However, he had a unique ability to describe physical phenomena in relatively simple terms, to distill the essence of a situation and to drop the unessentials. His greatest contribution is in boundary layer theory".

Film Model for heat transfer, h

When studying heat transfer from a hot fluid flowing around a cold surface W.K. Lewis (MIT, 1916) [17] proposed a linear profile of temperature in the film without additional variation away from the surface.

By analogy with Fourier's law, the total heat flux (J/s) through the film is $\dot{q} = \lambda A \frac{\Delta T}{\Delta y}$ or $\dot{q} = hA\Delta T$ where the film heat transfer coefficient is $h = \frac{\lambda}{\Delta y}$. For isolated spheres $\frac{hd_p}{\lambda} = 2 + 0.6\,\mathrm{Re}^{1/2}\,\mathrm{Pr}^{1/3}$.

T. Sherwood

Film Model for mass transfer

Whitman (1923) [18] used a similar treatment for mass transfer from a fluid to a solid surface and again proposed a linear concentration profile in the film; the total mass flux through the film for species i (mole/s) according to Fick's law is $\dot{N}_i = DA\frac{\Delta c_i}{\Delta y}$ or $\dot{N}_i = kA\Delta c_i$ where $k = \frac{D}{\Delta y}$ is the film mass transfer coefficient. For isolated spheres $\frac{kd_p}{D} = 2 + 0.6\,\mathrm{Re}^{1/2}\,Sc^{1/3}$. The Sherwood number appears in the lhs of equations above; a chemical engineer should remenber some numbers and one

is $Sh_{min}=2$.

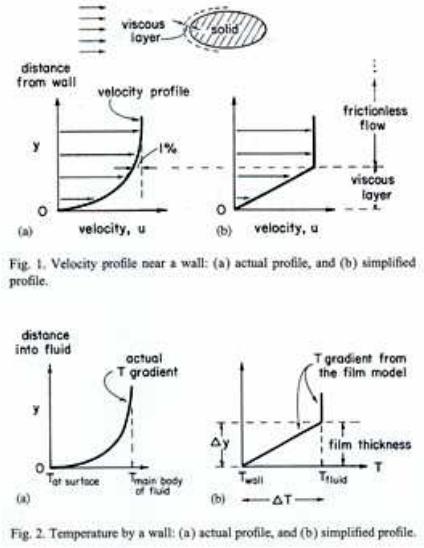


Fig. 1. Velocity profile near a wall: (a) actual profile, and (b) simplified profile.

Fig. 2. Temperature by a wall: (a) actual profile, and (b) simplified profile.

Figure 7. Film Model

Analogies between momentum, heat and mass transfer.

When Sc and Pr are of the order of 1 the analogy of Reynolds between the three transport modes is:

$Ms_m = Ms_h = f/2$ or:

$$\frac{k}{u_m} = \frac{h}{\rho c_p u_m} = f/2 \qquad [34a]$$

When Sc is 1000 (liquids) in laminar flow a small turbulence can affect the transport of heat/mass even if the velocity distribution is not much affected. The analogy of Chilton-Colburn is then applied:

$$j_m = j_h = f/2$$
$$j_m = Ms_m Sc^{2/3}$$
$$j_h = Ms_h \Pr^{2/3} \qquad [34b]$$

Diffusion, convection and reaction in isothermal catalysts, - *intuition is not enough*

The importance of intraparticle convection in the catalyst effectiveness was analysed by Nir and Pismen [19] in 1977 for first order irreversible reaction in isothermal catalysts. The problem was first dealt with by Wheeler in 1954 [20] ; he concluded that intraparticle convection would be important only for gas phase systems at high pressure in catalysts with very large pores. For the reaction A → B in slab catalysts the mass balance is:

$$\frac{d^2 f}{dx^2} - 2\lambda_m \frac{df}{dx} - 4\phi_S^2 f = 0 \qquad [35]$$

with BC: f=1 at x=0 and x=1. Model parameters are: Thiele modulus $\phi_S = \ell\sqrt{\frac{k}{D_e}}$ (ϕ_S^2 ratio between time constants for pore diffusion and reaction); intraparticle Peclet number $\lambda_m = \frac{v_0 \ell}{D_e}$ (ratio between time constants for pore diffusion and convection). The concentration profile inside the catalyst is:

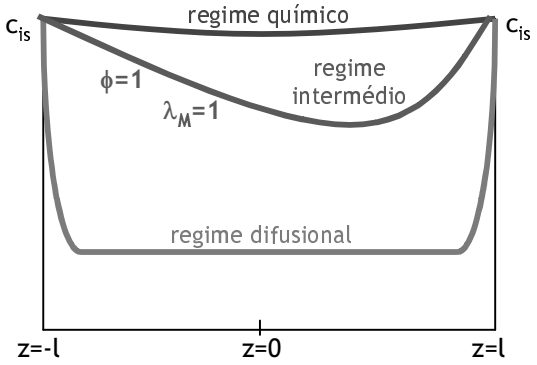


Figure 8. Assymetric concentration profiles in large pore catalysts

$$f = \frac{sh\alpha_2 e^{\alpha_1(2x-1)} - sh\alpha_1 e^{\alpha_2(2x-1)}}{sh(\alpha_2 - \alpha_1)} \qquad [36]$$

where $\alpha_{1,2} = \frac{\lambda_m \pm \sqrt{\lambda_m^2 + 4\phi_S^2}}{2}$.

The effectiveness factor is:

$$\eta_{dc} = \frac{1/\alpha_1 - 1/\alpha_2}{coth\alpha_1 - coth\alpha_2} \qquad [37]$$

When convection is not important, i.e., $\lambda_m = 0$ $\eta_d = \frac{\tanh \phi_S}{\phi_S}$. The effect of convection can be seen in Figure 9 where η_{dc}/η_d is plotted versus λ_m and ϕ_s. In the intermediate region of Thiele modulus (similar reaction and diffusion rates) the effectiveness of the catalyst is improved by convection. The pore convection will apparently increase diffusivity and move the catalyst working regime from diffusional to "chemical" controlled. The message is: intuition is not enough!

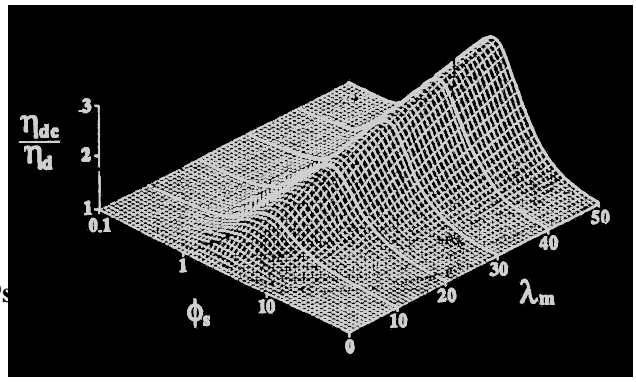


Figure 9. η_{dc}/η_d versus λ_m and ϕ_s

Fluid flow in chemical reactors: Residence Time Distribution (RTD) and tracer technology

Danckwerts (1953) [21] approached the study of fluid flow in reactors in a brilliant and simple way :"introduce a pulse of tracer into the fluid entering the reactor and see when it leaves". The normalized outlet concentration versus time is the Residence Time Distribution (RTD). The study of RTD of flowing fluids and its consequences can be put under the umbrella of tracer technology. This is important for chemical engineers, researchers in the medical field, environment, etc to diagnose the reactor behaviour, drug distribution in the body, etc. When I taught this subject at the University of Virginia students saw the application when in a Department Seminar someone from Merck, Sharp and Dome talked about pharmacokinetics! Danckwerts built a theory based on the characterization of fluid elements of a population inside the reactor (age and life expectation) and leaving the reactor (residence time). Then he introduced the "distribution": relative to each character; the residence time distribution E(t) is then defined as E(t)dt being the fraction of fluid elements leaving the reactor with residence time between *t* and *t+dt*. The next question is how to experimentally have access to E(t). This brings the tracer technology to the center of the arena. The normalized response to an impulse of tracer C(t) is directly related with the RTD,i.e., $C(t) = \tau E(t)$; or the normalized response to a step input of tracer F(t) curve of Danckwerts is $E(t) = \dfrac{dF(t)}{dt}$.

Danckwerts

This is a characteristic of linear systems: the response to an impulse is the derivative of the response to a step input. How this linearity appears in this macroscopic vision of fluid flow where Navier-Stokes $\rho\dfrac{D\mathbf{v}}{Dt} = \rho\mathbf{g} - \nabla P + \mu\nabla^2\mathbf{v}$ applies in a detailed description is a matter of think about.

It is also interesting to note that the RTD is the inverse Laplace transform of the transfer function G(s) , i.e., $E(t) = L^{-1}G(s)$.

This relation allows the calculation of the moments of E(t) from G(s) and its derivatives at s=0 (Van der Laan theorem).

Finally the chemical engineer uses the hydrodynamic characterization to connect with the reaction kinetics obtained in a batch reactor, c_{batch} (t)) and predict the average outlet concentration in a real reactor:

$$< c_S >= \int_0^\infty E(t) c_{batch}(t)\, dt \qquad [38]$$

This result is valid for first-order reactions. For other reaction kinetics it gives the limit when the flow is completely segregated; in the limit of maximum micromixing the Zwietering equation holds.

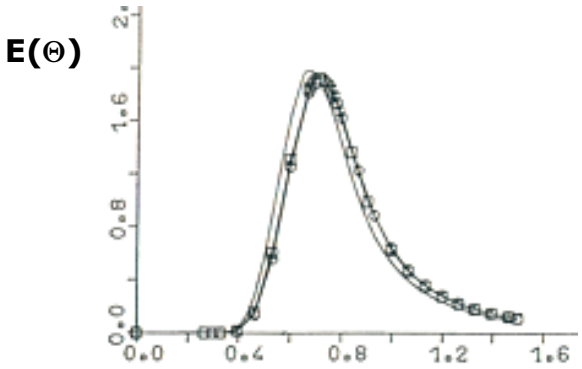


Figure 10. Residence time distribution, RTD

More10$ models: adsorption columns. Physical concepts from simple models

The model assumes isothermal operation, plug fluid flow, infinitely fast mass transfer between fluid and solid phases (instantaneous equilibrium at the interface) and trace system. Model equations are:

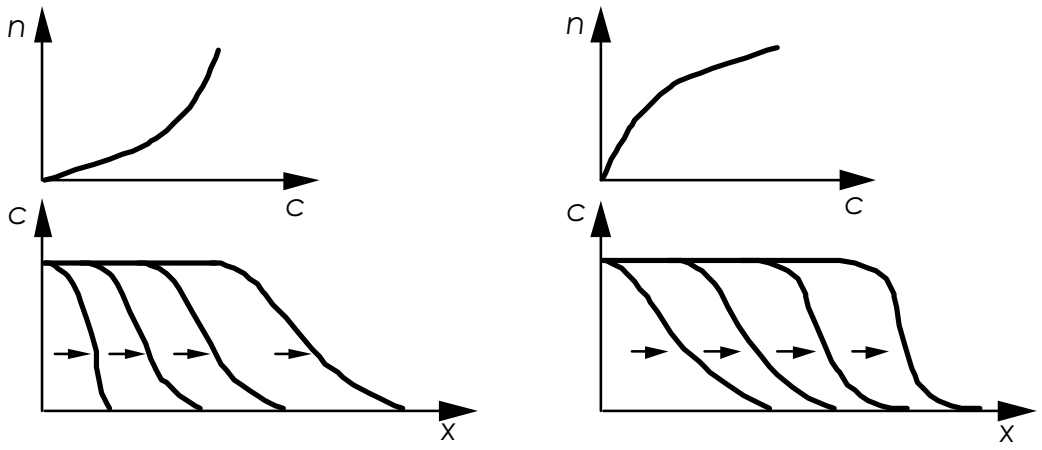
$$u_0 \frac{\partial c_i}{\partial z} + \varepsilon \frac{\partial c_i}{\partial t} + (1-\varepsilon)\frac{\partial \langle q_i \rangle}{\partial t} = 0$$
$$\langle q_i \rangle = q_i^* = f(c_i) \qquad [39]$$

where $<q_i>$ is the average concentration in the adsorbent and $q_i^* = f(c_i)$ is the concentration at the surface in equilibrium with the fluid concentration c_i. Using the cyclic relation between partial derivatives (yes…they are useful!) we get:

$$u_{c_i} = \left(\frac{\partial z}{\partial t}\right)_{c_i} = \frac{u_i}{1 + \frac{1-\varepsilon}{\varepsilon} f'(c_i)} \qquad [40]$$

This is De Vault equation (1943) [22]. Those interested in understanding adsorptive and chromatographic processes will recognize this is the most important result to retain. It shows that adsorption in fixed beds is a phenomenon of propagation of concentration waves. The simplest model shows that the nature of the equilibrium isotherm is the main factor influencing the shape of the breakthrough curve. The physical concepts to be retained are: dispersive waves are formed when isotherms are unfavourable; each concentration propagates with a velocity given by De Vault equation. Compressive waves are formed for favourable isotherms and the

physical limit is a shock which propagates with a velocity $u_{sh} = \dfrac{u_i}{1+\dfrac{1-\varepsilon}{\varepsilon}\dfrac{\Delta q_i}{\Delta c_i}}$, where the slope of the chord linking the feed state and the bed initial state appear instead of the local slope of the equilibrium isotherm.



(a) (b)

Figure 11. (a) Unfavorable isotherms and dispersive fronts (b) Favorable isotherm and compressive front.

1000$ models

Levenspiel [8] summarizes the progress on the study of fluid flow: "In the 19th century there were two approaches to study fluid flow: hydrodynamics (dealt with ideal frictionless fluid; highly mathematical stuff) and hydraulics developed by civil engineers " who amassed mountains of tables of pressure drop and head loss of fluids in open and closed channels of all sort…". At the beginning of the 20th century Prandtl said "Hydrodynamics has little significance for the engineer because of the great mathematical knowledge required for an understanding of it and the negligible possibility of applying its results. Therefore engineers put their trust in the mass of empirical data collectively known as the "science of hydraulics". Prandtl was the genius who patched together these different disciplines with his simple boundary layer theory. The result is modern fluid mechanics.

On the other hand numerical methods for the solution of PDE's exist and the combination of two solid disciplines appears with a new name: "Computational Fluid Dynamics". Twenty years ago I published in ISCRE8 "Residence time distribution in laminar flow through reservoirs from momentum and mass transport equations" [23]. It is a problem of 2-D flow in a reservoir of length L and height H where a stationary laminar flow exists between inlet and outlet (Fig. 12). The formulation is made in terms of vorticity and *stream function*; the flow field is obtained and the RTD is obtained by solving the mass conservation equation :

$$\frac{\partial(u\Omega)}{\partial x}+\frac{\partial(v\Omega)}{\partial y}=\nu\left(\frac{\partial^2\Omega}{\partial x^2}+\frac{\partial^2\Omega}{\partial y^2}\right)$$

$$u=\frac{\partial\psi}{\partial y};v=-\frac{\partial\psi}{\partial x} \qquad [41]$$

$$\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}\right)=-\Omega$$

$$\frac{\partial C}{\partial t}+\frac{\partial(uC)}{\partial x}+\frac{\partial(uC)}{\partial y}=D\left(\frac{\partial^2 C}{\partial x^2}+\frac{\partial^2 C}{\partial y^2}\right)$$

This problem was solved with modern tools (Fluent) recently. It could be another CFD package (CFX, FIDAP, etc).

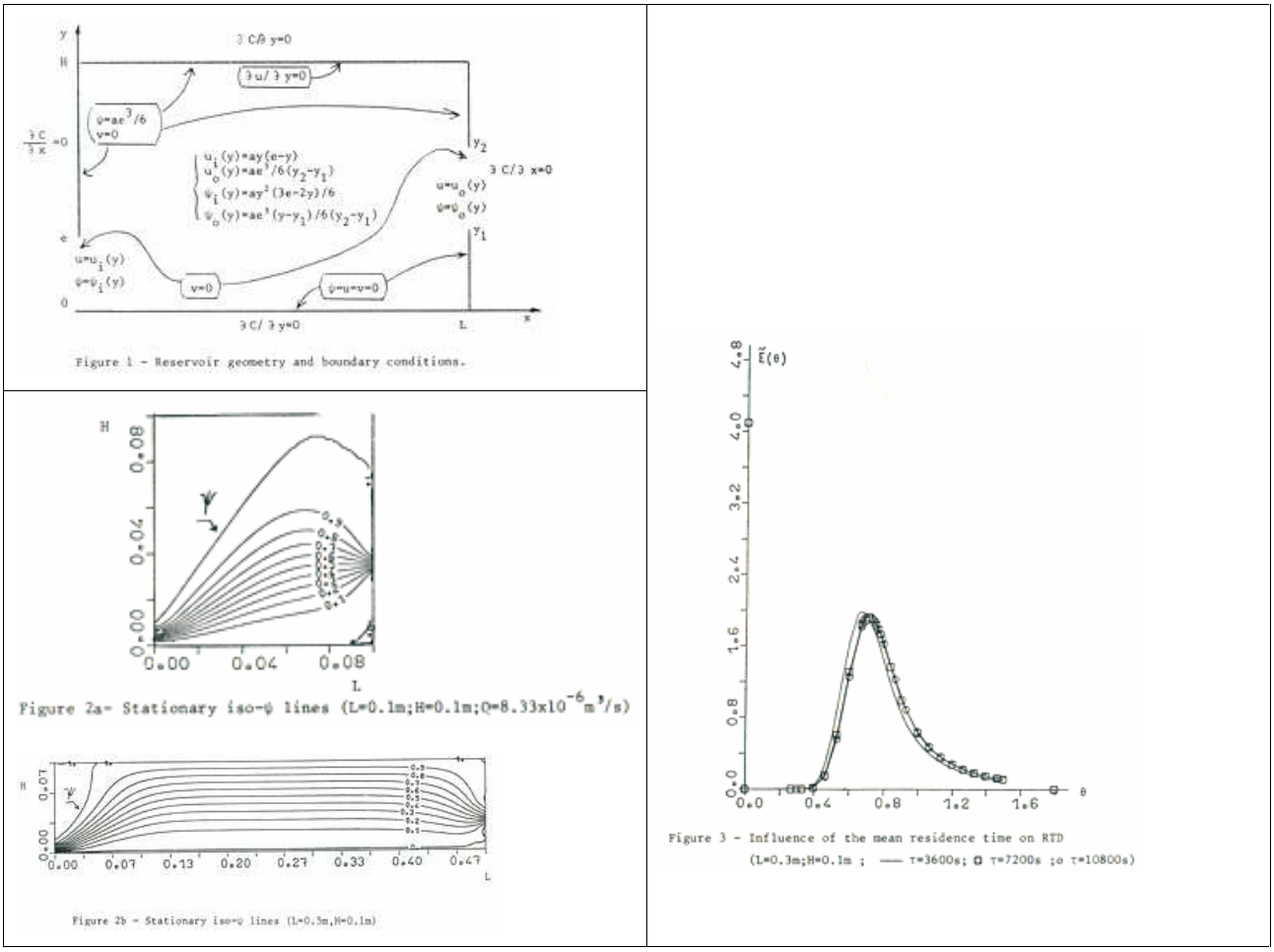
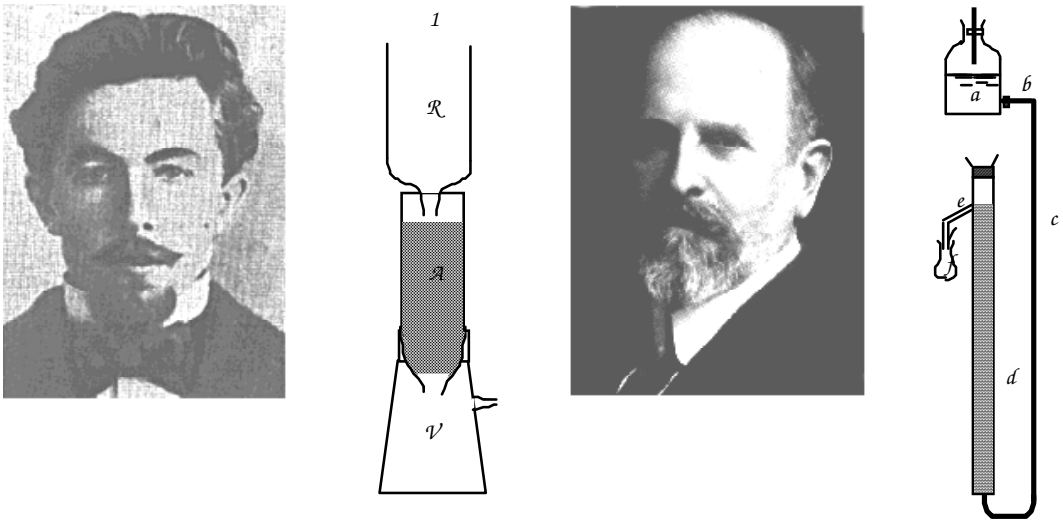


Figure 12. 2-D flow in a reservoir and RTD.

Simulation

Chromatographic processes

The first chromatographic experiment dates from 1903 and was reported in 1905 by Professor of Botany M.Tswett to the Warsaw Society of Natural Sciences: "*On a category of adsorption phenomena and their application to biochemical analysis"* [24]. He coined the term "chromatography" inspired in the experiment : elution of a sample of green leaves extract through a column of calcium carbonate which was separated in a yellow fraction(carotenes) and green fraction (chlorophyll). These studies were rediscovered in 1931 by the Nobel Prize R. Kuhn working on natural pigments. The theory of adsorption chromatography was developed in 1940 by Tiselius and partition chromatography in 1941 by A. J. P. Martin and A. L. M. Synge (all Nobel) (1941). Another vision of history shows David Talbot Day [25], geologist and engineer at the Mineral Resources of the US Geological Survey, who presented in 1900 at the 1st International Petroleum Congress in Paris (1900) one experiment where "*crude oil forced upward through a column packed with limestone changed in color and composition*" . This is the basis of PONA analysis established in 1914 and still used in petroleum industry where the adsorbent is silica-gel.



(1) (2)

Figure 13. Experiments of M. Tswett (1) and D. T. Day (2).

Modeling of chromatographic processes

The factors influencing the behaviour of a fixed bed column can be classified in two categories: equilibrium and kinetic factors (hydrodynamics, heat/mass transfer).

Models can be classified in two groups: I- Chemical kinetics type and II- Physical kinetics type depending on the rate law used[26,27] . The nuclei-model are Thomas [28] for Type I and Rosen [29] for Type II.

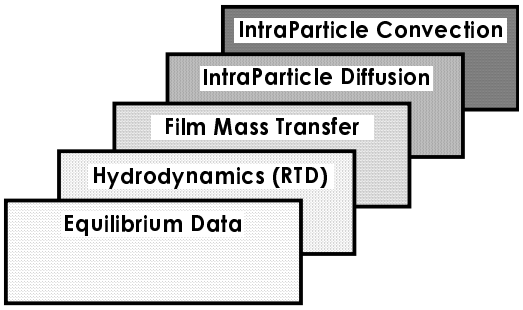


Figure 14. Factors governing the behavior of an adsorptive process

Thomas Model

$$u\frac{\partial c}{\partial x}+\frac{\partial c}{\partial t}+\frac{1-\varepsilon}{\varepsilon}\frac{\partial n}{\partial t}=0 \qquad [42a]$$

$$\frac{\partial n}{\partial t}=k_1[c(n_0-n)-rn(c_0-c)] \qquad [42b]$$

$x = 0$; $c = c_0$ $\forall$ t and $t \le \frac{x}{u}$; $n = 0$ $\forall x$

$$K=\frac{1}{r}=\frac{y^*(1-x^*)}{x^*(1-y^*)} \text{ with } x^*=\left(\frac{c}{c_0}\right)_{eq} \text{ and } y^*=\left(\frac{n}{n_0}\right)_{eq} \qquad [42c]$$

The solution is:

$$\frac{c}{c_0}=\frac{1}{2}+\frac{2}{\pi}\int_0^{\infty} e^A \sin B\frac{d\lambda}{\lambda} \qquad [43]$$

where A and B are functions of the model parameters. The Anzelius/Schumann model [33,34] developed for heat transfer is a simplification when film resistance is considered. The solution is $\frac{c}{c_0}=J(N_{fo},N_{fo}T)$ where the number of film mass transfer units N_{fo} is based on the bed length.

The chromatographic column as a dynamic system

For linear systems the transfer function of the column G(s) is the Laplace transform of the normalized impulse response E(t) and the moments of E(t) are obtained from G(s):

$$\mu_k=\int_0^{\infty} t^k E(t)dt=(-1)^k\left.\frac{d^kG(s)}{ds^k}\right|_{s=0} \qquad [44]$$

For the original model of Martin e Synge [35](The *Plate Theory of Martin and Synge*) the transfer function is:

$$G(s)=\prod_i g_i(s)=(1+\frac{t_r s}{J})^{-J} \qquad [45]$$

where $\tau = Jv_f/U$ is the space time of the fluid phase and $k = mv_s/v_f = (n_s/n_f)_{eq}$ is the ratio of amount of solute in each phase and the stoichiometric time is $t_r = \tau(1+k)$. The outlet normalized concentration is:

$$E(t)=\frac{1}{t_r}\left(\frac{t}{t_r}\right)^{J-1}\frac{J^J}{(J-1)!}e^{-Jt/t_r} \qquad [46]$$

The moments of $E(t)$ are $\mu_0=1$, $\mu_1=\tau(1+k)$ and $\sigma^2 = t_r^2/J$; the peak maximum is $t_{max}=t_r(1-1/J)$ and the peak width at mid-height for high J , is $2\sigma = 2t_r/\sqrt{J}$.

The model (Mixing cells in Cascade with Exchange) [36] is obtained when the mass transfer between phases has finite rate. Model equations are:

$$Uc_0 = Uc_1 + v_f\frac{dc_1}{dt} + k_m S(c_1 - \frac{n_1}{m}) \qquad [47]$$

$$k_m S(c_1 - \frac{n_1}{m}) = v_S\frac{dn_1}{dt} \qquad [48]$$

By putting $t_m = \frac{mv_s}{k_m S} = \frac{m}{k_m a_p}$ the transfer function G(s) is:

$$G(s)=\left\{1+\frac{\tau s}{J}\left(1+\frac{k}{1+t_m s}\right)\right\}^{-J} \qquad [49]$$

The moments of the chromatographic peak are: $t_r = \tau(1+k)$ and $\frac{\sigma^2}{t_r} = \frac{1}{J} + \frac{2k}{1+k}\frac{t_m}{t_r}$.

Van Deemter, Zuiderweg and Klinkenberg viewed the column as a continuous system [37] with

$$G(s)=\exp\left[\frac{Pe}{2}-\frac{Pe}{2}\sqrt{1+\frac{4s\tau(1+M(s))}{Pe}}\right] \qquad [50]$$

where $M(s)=\frac{k}{1+t_m s}$. The variance is obtained by replacing $1/J$ by $2/Pe$.

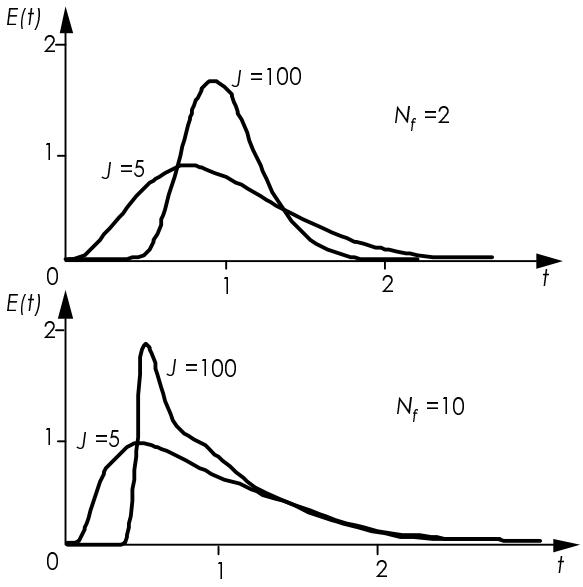


Figure 15. Influence of J and $N_f = k_m a_p \tau\,(1-\varepsilon)/\varepsilon$ on the shape of a chromatographic peak for k=1.

Progressive modelling

A complete treatment details the particle description. When film resistance is included we have:

$$k_f S_p (c - c_s) = V_p \frac{d\langle c \rangle}{dt} \qquad [51]$$

where $(c'/c)_{eq} = \varepsilon_p + \rho_p k_a = m$ and

$$c = \frac{c_s'}{m} + \frac{d}{k_f}\frac{d\langle c \rangle}{dt} \qquad [52]$$

The relation between $\langle c' \rangle$ and c'_s can be obtained by solving the particle mass balance

TABLE 2. Characteristic dimension, $H(s)$ and shape factor.

Geometry	Characteristic Dimension $d=V_p/S_p$	$H(s) = \overline{\langle c' \rangle}/\overline{c'_s}$	Shape Factor μ
slab of thickness $2l$	l	$\frac{\tanh\phi}{\phi}$	1/3
infinite cylinder of radius R	R/2	$\frac{I_1(2\phi)}{\phi I_o(2\phi)}$	1/2
sphere of radius R	R/3	$\frac{1}{\phi}\left(\frac{1}{\tanh 3\phi} - \frac{1}{3\phi}\right)$	3/5
any shape	d	$\frac{1}{1+\mu\phi^2}$	μ
			with $\phi^2 = \frac{d^2 s}{D_e}$

A good approximation is $H(s) = 1/(1+t_d s)$ with $t_d = \mu d^2/D_e$ and then

$$L(s) = \frac{\overline{<c'>}}{\overline{c}} = \frac{m}{1/H(s) + t_d s} = \frac{m}{1 + (t_d + t_e)s} \qquad [53]$$

with $t_e = md/k_f$.

For finite adsorption rate $L(s) \cong m/[1+(t_d+t_e+t_a)s]$ where $t_a = \rho_p K_a/mk_a$. In general $M(s) = k/(1+t_m s)$ with $t_m = t_d + t_e + t_a$ e $G(s) = G_0[s(1+M(s)]$ where $G_0(s)$ is the column transfer function in absence of adsorption.

***Perfusion chromatography*- Importance of intraparticle convection in large-pores**

In chemical engineering there are materials (catalysts, adsorbents, membranes) with large pores (> 1000 A) for transport and smaller pores to provide adsorption capacity and catalytic sites. As mentioned in the section

"Intuition is not enough" my interest in this area started with a problem of measurement of effective diffusivity in large pore catalysts using a chromatographic method and tracer technology. The analysis of results obtained by Ahn [38] with a conventional model led to the conclusion that effective diffusivity was changing with flowrate.

Results were reanalyzed by assuming transport not only by diffusion , D_e but also by convection (pore velocity v_0) and the equivalence with the conventional model where both mechanisms were lumped in an apparent $\tilde{D}_e$ allowed us to show that[39]:

$$\tilde{D}_e = D_e \frac{1}{f(\lambda)} \qquad [54]$$

The apparent diffusivity is augmented by convection and the enhancement factor is $1/f(\lambda)$. This result explains the functioning of perfusion chromatography which appeared in 1990.

Based on the work of Nir and Pismen [19] on diffusion, convection and reaction in large pore catalysts (5000 A) data from Ahn were analyzed.

For a non adsorbable tracer the Para um tracer the "lumped" diffusion/convection model for transient state is:

$$\tilde{D}_e \frac{\partial^2 \bar{c}}{\partial x^2} = \varepsilon_p \frac{\partial \bar{c}}{\partial t} \qquad [55]$$

The particle transfer function is:

$$\tilde{g}_p(s) = \frac{<\bar{c}>}{\bar{c}_s} = \frac{\tanh\sqrt{\tau_d s}}{\sqrt{\tau_d s}} \qquad [56]$$

with an apparent diffusion time constant $\tilde{\tau}_d = \varepsilon_p \ell^2 / \tilde{D}_e$.

The detailed diffusion/convection model is:

$$D_e \frac{\partial^2 \bar{c}}{\partial x^2} - v_o \frac{\partial \bar{c}}{\partial x} = \varepsilon_p \frac{\partial \bar{c}}{\partial t} \qquad [57]$$

and

$$g_p(s) = \frac{\left(e^{2r_2} - 1\right)\left(e^{2r_1} - 1\right)}{\left(e^{2r_2} - e^{2r_1}\right)} \frac{\sqrt{\left(\frac{\lambda}{2}\right)^2 + \tau_d s}}{\tau_d s} \qquad [58]$$

with $r_{1,2} = \frac{\lambda}{2} \pm \sqrt{\left(\frac{\lambda}{2}\right)^2 + \tau_d s}$, $\tau_d = \varepsilon_p \ell^2 / D_e$ e $\lambda = v_0 l / D_e = \tau_d / \tau_c$ (intraparticle Peclet number).

Model equivalence leads to equation [54].

where the enhancement factor for pore diffusivity due to convection is $1/f(\lambda)$ shown in Figure 16, with

$$f(\lambda) = \frac{3}{\lambda}\left[\frac{1}{\tanh\lambda} - \frac{1}{\lambda}\right] \qquad [59]$$

The practical application of this concept is the separation of proteins by HPLC. The pore velocity can be estimated from the equality between bed pressure drop relative to the bed length and particle pressure drop assuming that Darcy's law is valid; the result is: $v_0 = a\, u_0$ where a is the ratio of particle and bed permeabilities.

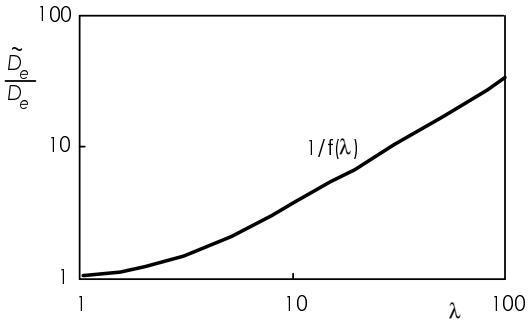


Figure 16. Enhancement factor for diffusivity due to convection, $1/f(\lambda)$

Van Deemter equation for conventional packings is:

$$HETP = A + \frac{B}{u} + \frac{2}{3}\frac{\varepsilon_p(1-\varepsilon_b)b^2}{[\varepsilon_b + \varepsilon_p(1-\varepsilon_b)b]^2}\tau_d u \qquad [60]$$

where ε_p is the particle porosity, ε_b is the interparticle porosity and $b=1+\{(1-\varepsilon_p)/\varepsilon_p\}$ The slope of the equilibrium isotherm is m; or

$$HETP = A + B/u + Cu \qquad [61]$$

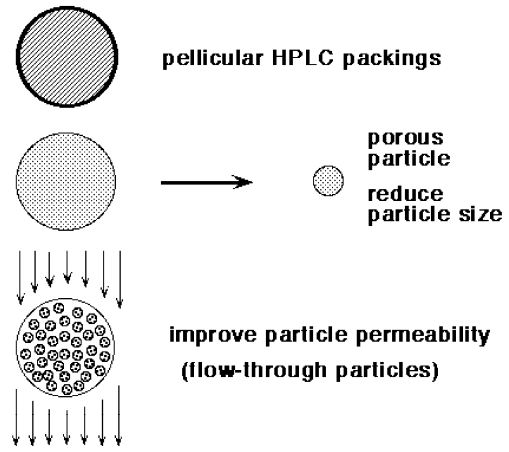


Figure 17 **– How to decrease intraparticle mass transfer resistance?**

For large-pore particles Rodrigues [40,41] derived an extension of the Van Deemter equation:

$$HETP = A + \frac{B}{u} + Cf(\lambda)u \qquad [62]$$

Rodrigues equation

At low velocities $\mathbf{f(\lambda)\approx 1}$ both equations are similar ; at high velocities $f(\lambda)\approx 3/\lambda$ and the last term of Rodrigues equation becomes constant since v_o is proportional to u . The HETP reaches a *plateau* which does not depend on the solute diffusivity but only on

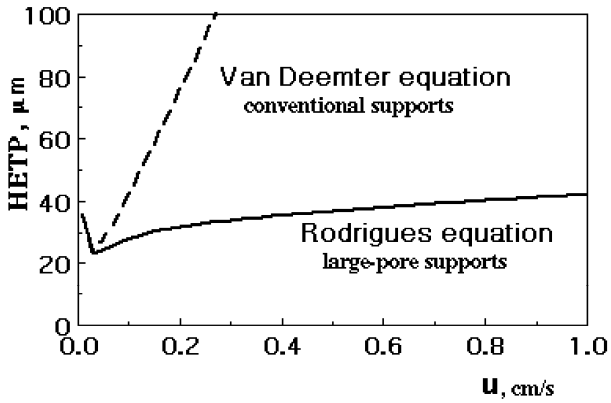
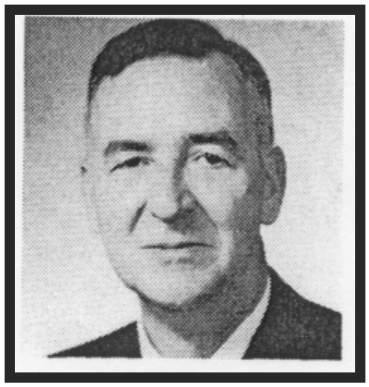


Figure 18. HETP versus u (Van Deemter eq. and Rodrigues eq.)

the particle permeability and pressure gradient (convection-controlled limit). In large-pore supports the column performance is improved since HETP is lower than with conventional supports (the C term of Van Deemter equation is reduced) and the speed of separation is increased without loosing efficiency.

Simulated Moving Bed

The operation of Simulated Moving Bed, SMB is easily understood by analogy with a True Moving Bed (TMB) shown in Fig. 19. The system is divided in four zones each with a specific task; the less retained species is recovered in the raffinate and the more retained in the extract. If we want to separate a binary mixture in the TMB we need to follow the constraints shown in Fig. 20;



D. Broughton

$$\frac{Q_I c_{BI}}{Q_S q_{BI}} > 1 \quad ; \quad \frac{Q_{II} c_{AII}}{Q_S q_{AII}} > 1 \quad \text{and} \quad \frac{Q_{II} c_{BII}}{Q_S q_{BII}} < 1 \quad ;$$

$$\frac{Q_{III} c_{AIII}}{Q_S q_{AIII}} > 1 \quad \text{and} \quad \frac{Q_{III} c_{BIII}}{Q_S q_{BIII}} < 1 \quad ; \quad \frac{Q_{IV} c_{AIV}}{Q_S q_{AIV}} < 1$$

For linear isotherms the separation region is a triangle shown in figure 21.

In order to avoid friction between particles flowing in the true moving bed UOP developed the SMB technology [42] industrially used now for the separation of p-xylene (Parex process), production of HFCS (Sarex process). In the SMB the solid is fixed and the solid movement is simulated by periodically shifting the inlet/outlet positions of streams with a *rotary valve*. Recently the technology was adopted by the paharmaceutical industry for the separation of enantiomers [43]. In this case there are typically 6 columns with valves associated to each column as shown in Figure 23 a) and b. If mass transfer resistance inside particles is important the constraints imposed by the equilibrium theory have to be modified; that is why the concept of separation volume was introduced [44] to illustrate in a 3-D diagram the effect of flowrate in region I where the adsorbent is regenerated (Figure 24).

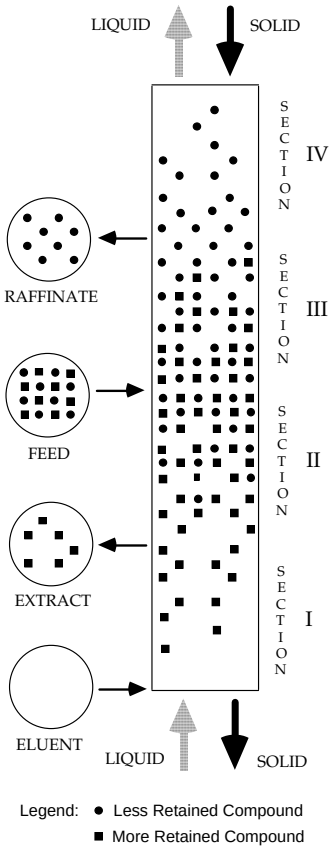


Figure 19. TMB

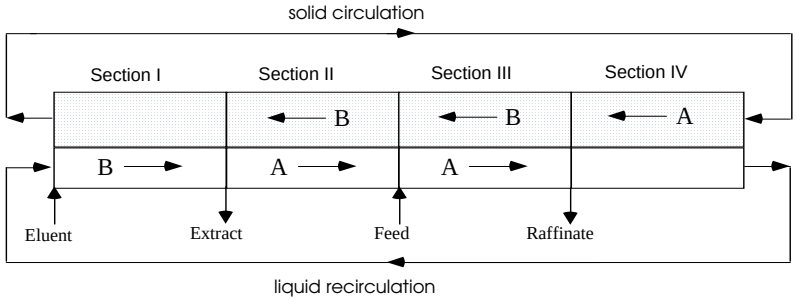
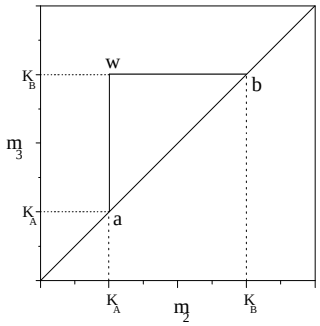


Figure 20. Net fluxes of each species in the TMB zones



$$K_B < m_1 < \infty \qquad (4.7)$$

$$K_A < m_2 < m_3 < K_B \qquad (4.8)$$

$$0 < m_4 < K_A \qquad (4.9)$$

Figure 21. Separation region for linear isotherms

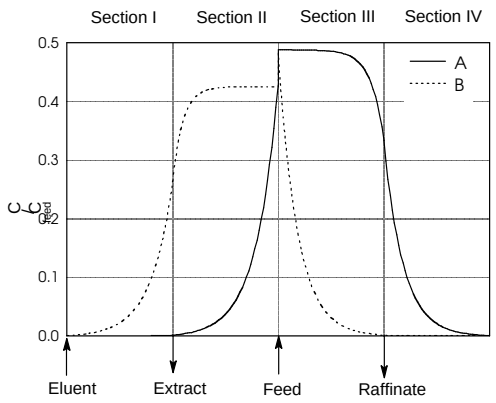


Figure 22. Concentration profiles in TMB
(A – less retained species; B – more retained species)

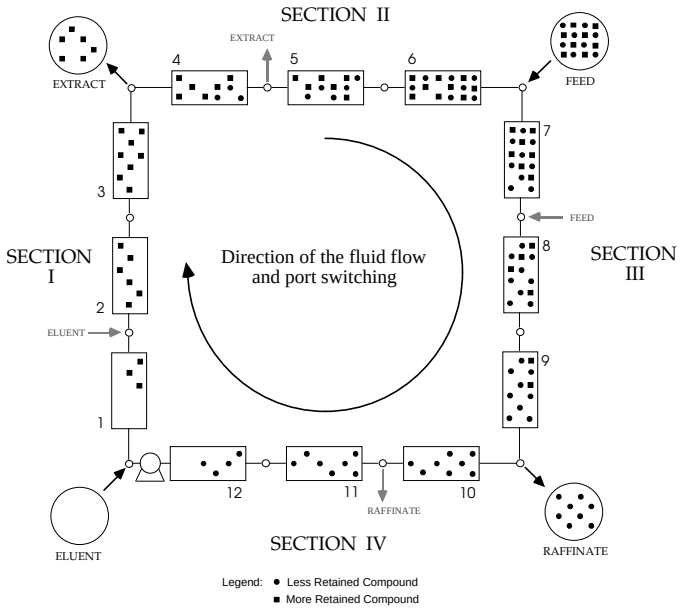


Figure 23a SMB scheme

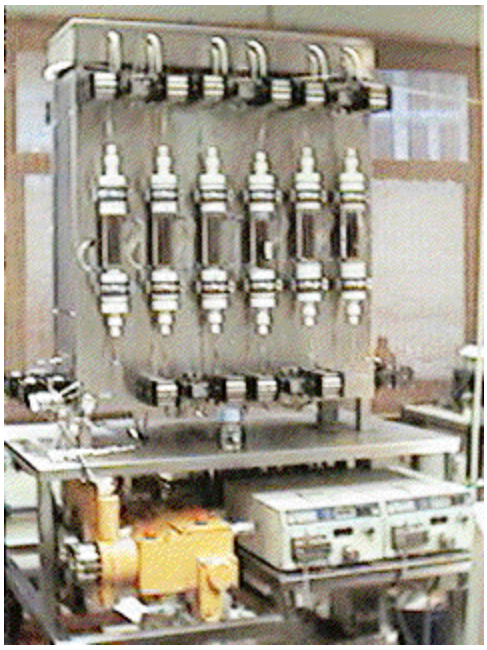


Figure 23b. SMB unit *Licosep 12-26* (Novasep) at the *LSRE*

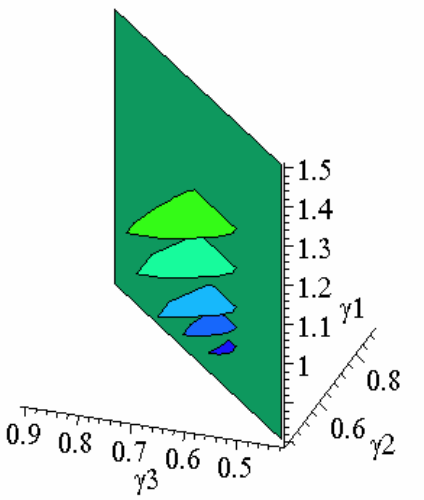


Figure 24. Concept of separation volume

Conclusions

Modelling is the activity which identifies a generation of chemical engineers associated with the Second Paradigm of Chemical Engineering. Simulation tasks can be simplified eventually with the availability of software with friendly user interfaces. The question of validation of results remains and in principle more time will be available to analyze results. The Third Paradigm of Chemical Engineer should come out soon; in the meantime we keep the reflection of Astarita: "the amount of information available grows continuously but the amount of information that any one of us can usefully digest does not grow".

Roger Sargent e A. Westerberg: Development of generic tools: Optimisation, dynamic simulators

References

[1] Wei, J. 'New horizons in reaction engineering', Chem. Eng. Sci., 45(8), 1947-1952 (1990).

[2] http://www.chemheritage.org/EducationalServices/chemach/ce/lwl.html

[3] Bird, B., Stewart, W. and E. Lighfoot, "Transport Phenomena", 2nd edition, J. Wiley (2002)

[4] K. Westerterp, A.E. Kronberg, A. Benneker e V. Dil'man, "Wave concept in the theory of hydrodynamic dispersion- a maxwellian type approach", Trans IchemE, 74, Part A, 944 (1996)

[5] J. Liu, Xu Chen e L.X. Xu, "New thermal wave aspects on burn evaluation of skin subjected to instantaneous heating", IEEE Transactions on Biomedical Engineering, 46 (4) 420 (1999)

[6] P. Le Goff, "Cours de Cinétique Physique et Chimique", ENSIC (1970)

[7] R. Aris,"Mathematical modelling techniques", Dover Pubs, New York (1994)

[8] O. Levenspiel, "Modeling in chemical engineering", *Chem.Eng Sci*., 57, 4691-4696 (2002)

[9] Denbigh,K. , "The thermodynamics of the steady state", Methuen, London (1951)

[10] J.C. Maxwell, "The scientific papers of James Clerk Maxwell",Dover, New York (1952)

[11] A.E. Rodrigues, "Percolation theory I- Basic principles and II- Modeling and design of percolation columns" in AIChEMI Series B: Stagewise and masss transfer operations, pp 7-24, Volume 5 – Chromatography, Percolation, Adsorption and Gas adsorption, edited J.M. Calo and E.J. Henley, AIChE (1984)

[12] Glueckauf, E., 'Theory of Chromatography. Part 10 - Formulae for diffusion into spheres and their applicability to chromatography', Trans. Faraday Soc., 51, 1540-1551 (1955)

[13] Vermeulen, T., 'Separation by adsorption methods', Adv. Chem. Eng., 2, 147-205 (1958)

[14] Weisz, P. e D. Prater, "Interpretation of measurements in experimental catalysis", Adv. in Catalysis, 6, Academic Press (1954)

[15] Damkholer, G., "Ubertemperatur in Kontaktkornern", Zeitschrift fur Physikalische Chemie, 193, 16-28 (1943)

[16] L. Prandtl, On fluid motions with very small friction, Third International Mathematical Congress, Heidelberg pp484-491 (1904)

[17] W.K. Lewis, Ind Eng Chem , 8, 825 (1916)

[18] W.G. Whitman Chemical and Metalurgical Engineering, 24, 147 (1923)

[19] Nir, A. and Pismen, L. (1977) 'Simultaneous intraparticle forced convection, diffusion and reaction in a porous catalyst', Chem. Eng. Sci., 32, 35-41.

[20] Wheeler, A. Reaction rates and selectivity in catalyst pores, Adv in Catalysis, 3 , 250-337 (1951)

[21] P.V. Danckwerts, Continuous flow sytems, Chem Eng Sci , 1 (1953)

[22] De Vault, D. (1943) 'The theory of chromatography', J. Am. Chem. Soc., 65, 532.

[23] Brunier, E., Zoulalian, A., Antonini, G. e A. Rodrigues, "Residence time distribution in laminar flow through reservoirs from momentum and mass transport equations", ISCRE 8, EFCE Publications Series n.37, pp439, The Institution of Chemical Engineers (1984)

[24] Rondest, J.(1972) 'M.Tswett - un botaniste qui fut la providence des chimistes', La Recherche, 24(3), 580.

[25] Heines, S. (1971) 'Chromatography - a history of parallel development', Chemtech, May, 280-285.

[26] Rodrigues, A.E. and Tondeur, D., (eds.) (1981) *Percolation Processes: Theory and Applications*, NATO ASI Series E33, Sijthoff & Noordhoff Pub.

[27] Rodrigues, A.E. (1984) 'Percolation theory I - Basic principles' in Calo, J. and Henley, E. (eds.), *Stagewise and mass transfer operations*, Vol 5, AIChE MI.

[28] Thomas, H. (1944) 'Heterogeneous ion exchange in a flowing system', J. Am. Chem. Soc., 66, 1664.

[29] Rosen, J. (1952) 'Kinetics of a fixed bed system for solute diffusion into spherical particles', 20(3), 387.

[30] Bohart, G. and Adams, E. (1920) 'Some aspects of the behavior of charcoal with respect to diclorine', J. Am. Chem. Soc., 42, 523.

[31] Walter, J. (1945) 'Rate dependent chromatographic adsorption', 13(8), 332.

[32] Klinkenberg, A. (1948) 'Numerical studies of equation describing transient heat and mass transfer in packed solids', Ind. Eng. Chem., 40(10), 1992.]

[33] Anzelius, A. (1926) 'Uber Erwarming vermittels durchstromender Medien', Z. Angew. Math. Mech., 6, 291-294.

[34] Schuman, T. (1929) 'Heat transfer: a liquid flowing through a porous prism', J. Franklin Institute, 208, 305-316.

[35]Martin, A.J.P. and Synge, A.L.M. (1941) 'A new form of chromatogram employing two liquid phases', Biochem. J., 35, 1358.

[36] Villermaux, J. (1972) 'Analyse des processus chromatographiques linéaires à l'aide de modèles phénomélogiques', Chem. Eng. Sci., 27, 1231-1241.

[37] Van Deemter, J., Zuiderweg, F. and Klinkenberg, A. (1956) 'Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography', Chem. Eng. Sci., 5, 271-289.

[38] Ahn, B. (1980) 'Études des Caracteristiques Diffusionelles de Transfer de Matière dans un Reacteur Catalytique à Lit Fixe d'Oxidation Menagée', Ph.D. Thesis, Université de Technologie de Compiègne.

[39] Rodrigues, A.E., Ahn, B. and Zoulalian (1982) 'Intraparticle forced convection effect in catalyst diffusivity measurements and reactor design', AIChE. J., 28, 925-930.

[40] Rodrigues, A.E., Lu, Z.P. and Loureiro, J. (1991) 'Residence time distribution of inert and linearly adsorbed species in fixed-bed containing large-pore supports: applications in separation engineering', Chem. Eng. Sci., in press

[41] A.E. Rodrigues, "An extended Van Deemter equation (Rodrigues equation) for performing chromatographic processes using large-pore, permeable packings", LC-GC, 6(1), 20-29 (1993)

[42] Broughton, D. et al (1970) 'The Parex process for recovery of p-xylene', Chem. Eng. Prog., 66(9), 70.

[43] L. Pais, "Chiral separations by SMB chromatography", Tese de doutoramento, FEUP (1999)

[44] D. Azevedo e A.E. Rodrigues, "Design of a SMB in presence of mass transfer resistances",AIChEJ, 45(5) 956 (1999)

[45] A.E. Rodrigues, "Engenharia Química: os últimos 25 anos", Academia de Engenharia, Lisboa (2000).

Adaptive finite element solutions of time-dependent partial differential equations using moving mesh algorithms

Mike Baines

Dept of Mathematics, University of Reading, UK

Matthew Hubbard, Peter K. Jimack

School of Computing, University of Leeds, UK

presented at the Workshop on Modelling and Simulation in Chemical Engineering, Coimbra, Portugal, June 30th- July 4th 2003

Abstract

A Lagrangian moving finite element algorithm is presented whose mesh velocity is determined by the invariance of the local "mass". The method is applied to second and fourth order nonlinear diffusion equations with moving boundaries in one and two dimensions.

1 Introduction

We consider adaptive finite element solutions of second order and fourth order nonlinear diffusion equations with moving boundaries using a Lagrangian moving finite element method. The method is prompted by recent interest in geometric integration and scale invariance (for references see [6]) which has rekindled interest in the use of adaptive moving meshes in the solution of these equations, suggesting new numerical approaches. The invariance properties combine independent and dependent variables, suggesting that these variables should be treated similarly in numerical work. A natural consequence is to use moving adaptive meshes.

Scale invariance implies a local relationship between the variables which can be used to drive mesh movement. The mechanism is similar to the use of monitor functions to control the movement of the mesh, as in the MMPDE (Moving Mesh Partial Differential Equations) method [9]. It is also related to the Geometric Conservation Law [11] and its associated invariance property

[7]. The local relationship induces a mesh movement which retains the scale invariance properties of the original PDE.

The method is a generalisation of the finite volume approach used in [3, 4] in one dimension. It uses a weighted form of the invariance equation on a patch of elements, as in [1], resulting in a Lagrangian moving finite element method. In order to obtain uniqueness in higher dimensions we exploit the idea of a mesh velocity potential, as proposed in [7]. The link between the method and classical fluid dynamics is discussed in [2].

We describe the Lagrangian moving finite element method and its role in free boundary problems requiring adaptivity. The method is tested against the radial self-similar solution of the two-dimensional Porous Medium Equation (PME) with a free boundary. We also consider applications to problems governed by a fourth order nonlinear diffusion equation with a moving boundary.

We begin by setting up a moving framework for the theory.

2 Fixed and Moving Frames

Consider a scalar PDE in the general form

$$\frac{\partial u}{\partial t} = Lu \tag{1}$$

where $u = u(\mathbf{x}, t)$ in a fixed frame of reference with coordinate $\mathbf{x}$ and L is a multidimensional operator involving space derivatives only.

Instead of working in the fixed frame we take a Lagrangian viewpoint. Define an invertible mapping between fixed labelling coordinates $\mathbf{a}$ at time τ and moving coordinates $\mathbf{x}$ at time t, of the form

$$\mathbf{x} = \hat{\mathbf{x}}(\mathbf{a}, \tau), \quad t = \tau$$

so that

$$u(\mathbf{x}, t) = u(\hat{\mathbf{x}}(\mathbf{a}, \tau), \tau) = \hat{u}(\mathbf{a}, \tau)$$

say, where $\hat{u}, \hat{\mathbf{x}}$ are Eulerian coordinates.

By the chain rule,

$$\frac{\partial \hat{u}}{\partial \tau} = \nabla u \cdot \frac{\partial \hat{\mathbf{x}}}{\partial \tau} + \frac{\partial u}{\partial t}\frac{dt}{d\tau}$$

where $\frac{\partial}{\partial t}$ means differentiation with respect to time t with $\mathbf{x}$ frozen, so that $\frac{\partial u}{\partial t}$ is given by the PDE (1). Hence, writing

$$\dot{u} = \frac{\partial \hat{u}}{\partial \tau}, \qquad \dot{\mathbf{x}} = \frac{\partial \hat{\mathbf{x}}}{\partial \tau},$$

we obtain the form of the differential equation in the moving frame,

$$\dot{u} - \nabla u \cdot \dot{\mathbf{x}} = Lu\,. \tag{2}$$

Clearly, a second equation is required to determine the two unknown Eulerian velocities $\dot{u}$ and $\dot{\mathbf{x}}$.

An integral form similar to (2) may be obtained using Leibnitz' rule on a moving test volume $\Omega(t)$ in the form

$$\frac{d}{dt}\int_{\Omega(t)} u d\Omega = \frac{\partial}{\partial t}\int_{\Omega(t)} u d\Omega + \oint_{\partial\Omega(t)} u\dot{\mathbf{x}} \cdot d\mathbf{\Gamma} = \int_{\Omega(t)} \left(\frac{\partial u}{\partial t} + \nabla \cdot (u\dot{\mathbf{x}})\right) d\Omega \tag{3}$$

giving the integral form in the moving frame (cf. (2)),

$$\frac{d}{dt}\int_{\Omega(t)} u d\Omega - \int_{\Omega(t)} \nabla \cdot (u\dot{\mathbf{x}}) d\Omega = \int_{\Omega(t)} Lu d\Omega \tag{4}$$

where we have made use of the integral form of the PDE (1) in the fixed frame.

For the finite element method we need weak forms. Given a set of suitable test functions w_i moving with the velocity field $\dot{\mathbf{x}}$ a generalisation of the Leibnitz rule (3) gives

$$\begin{aligned}\frac{d}{dt}\int_{\Omega(t)} w_i u d\Omega &= \frac{\partial}{\partial t}\int_{\Omega(t)} w_i u d\Omega + \oint_{\partial\Omega(t)} w_i u\dot{\mathbf{x}} \cdot d\mathbf{\Gamma} \\ &= \int_{\Omega(t)} \left(w_i \frac{\partial u}{\partial t} + u\frac{\partial w_i}{\partial t} + \nabla \cdot (w_i u\dot{\mathbf{x}})\right) d\Omega\end{aligned}$$

Assuming that the test functions w_i are advected with velocity $\dot{\mathbf{x}}$, we have

$$\frac{\partial w_i}{\partial t} + \dot{\mathbf{x}} \cdot \nabla w_i = 0 \tag{5}$$

leading to the integral weak form in the moving frame,

$$\frac{d}{dt}\int_{\Omega(t)} w_i u d\Omega - \int_{\Omega(t)} w_i \nabla \cdot (u\dot{\mathbf{x}}) d\Omega = \int_{\Omega(t)} w_i Lu d\Omega \tag{6}$$

where we have made use of the weak form of the PDE (1) in the fixed frame.

3 A Distributed Conservation Principle

Assume that the problem and the boundary conditions are such that the total mass

$$\int_{\Omega(t)} u d\Omega$$

is conserved in the moving frame. Motivated by the scale invariance of this quantity, we assume that the velocity $\dot{\mathbf{x}}$ of the moving frame is determined locally by the distributed conservation principle

$$\int_{\Omega(t)} w_i u d\Omega = c_i = \text{constant in time.} \tag{7}$$

Differentiating (7) with respect to time,

$$\frac{d}{dt}\int_{\Omega(t)} w_i u d\Omega = 0$$

giving from (6)

$$-\int_{\Omega(t)} w_i \nabla \cdot (u\dot{\mathbf{x}}) d\Omega = \int_{\Omega(t)} w_i L u d\Omega \tag{8}$$

or, after integration by parts,

$$-\oint_{\partial\Omega(t)} w_i u\dot{\mathbf{x}} \cdot d\mathbf{\Gamma} + \int_{\Omega(t)} u\dot{\mathbf{x}} \cdot \nabla w_i d\Omega = \int_{\Omega(t)} w_i L u d\Omega\,. \tag{9}$$

Equation (8) is in effect an equation for the divergence of $u\dot{\mathbf{x}}$. It is insufficient by itself to determine $\dot{\mathbf{x}}$ uniquely but if the vorticity $curl\dot{\mathbf{x}}$ is specified (together with a suitable boundary condition) then, given u, equations (8) or (9) determine the velocity $\dot{\mathbf{x}}$.

For example, suppose that $curl\dot{\mathbf{x}} = curl\mathbf{q}$ is specified. Then there exists a velocity potential ϕ such that

$$\dot{\mathbf{x}} = \mathbf{q} + \nabla\phi \tag{10}$$

so that (8) can be written

$$-\int_{\Omega(t)} w_i \nabla \cdot (u\nabla\phi) d\Omega = \int_{\Omega(t)} w_i \left(Lu + \nabla \cdot (u\mathbf{q})\right) d\Omega \tag{11}$$

and (9) becomes

$$
\begin{aligned}
&- \oint_{\partial\Omega(t)} w_i u \nabla\phi \cdot d\mathbf{\Gamma} + \int_{\Omega(t)} u \nabla\phi \cdot \nabla w_i d\Omega \\
&= \int_{\Omega(t)} w_i L u d\Omega + \oint_{\partial\Omega(t)} w_i u \mathbf{q} \cdot d\mathbf{\Gamma} - \int_{\Omega(t)} u \mathbf{q} \cdot \nabla w_i d\Omega \,. \qquad (12)
\end{aligned}
$$

Equation (12) can be used to determine ϕ, after which $\dot{\mathbf{x}}$ is given by the weak form, from (10),

$$
\int_{\Omega(t)} w_i \dot{\mathbf{x}} d\Omega = \int_{\Omega(t)} w_i \nabla\phi d\Omega + \int_{\Omega(t)} w_i \mathbf{q} d\Omega \,. \qquad (13)
$$

4 A Moving Finite Element Method

A Moving Finite Element method may be constructed using the weak forms (7), (12) and (13).

Linear elements are used for u, $\dot{\mathbf{x}}$, and ϕ, here denoted by upper case U, $\dot{\mathbf{X}}$, and Φ, on a (moving) triangulation of the region. Since $\dot{\mathbf{X}}$ is piecewise linear and W_i is the advected form of $\hat{W}_i$ (cf. (5)), the corresponding functions W_i are the usual linear basis functions on the moving mesh. The support of the integrals in the i'th equation (12) is taken to be the patch of elements $\Pi_i(\dot{\mathbf{X}})$ surrounding the node. The Dirichlet condition $U = 0$ is not imposed strongly at the boundary in the solution of (7), but weakly in the first term of (12).

In effect we solve the ODE system

$$
\frac{d}{dt}\vec{\mathbf{X}} = \vec{\mathbf{F}}(\vec{\mathbf{X}}) \qquad (14)
$$

using the following sequence to evaluate $\vec{\mathbf{F}}(\vec{\mathbf{X}})$:

- Given $\vec{\mathbf{X}}$ recover U from (7)
- Given U calculate $\vec{\Phi}$ from (12)
- Calculate $\vec{\mathbf{F}}(\vec{\mathbf{X}})$ from (13)
- Return

The overall algorithm requires the solution of a single ODE system, where each evaluation of the right-hand side of (14) involves the solution of three (four in the case of a fourth order problem, as outlined below) symmetric linear algebraic systems.

4.1 A Second Order Problem

The Porous Medium Equation in a fixed coordinate system (PME)

$$\frac{\partial u}{\partial t} = \nabla \cdot (u^m \nabla u), \tag{15}$$

is a well-known model equation for gas flows in porous media, spreading liquids etc. It admits compact support solutions with a free boundary for which comparison results are known [6, 10]. In integral form (15) is

$$\frac{d}{dt}\int_{\Omega(t)} u d\Omega = \int_{\Omega(t)} \nabla \cdot (u^m \nabla u) d\Omega = \oint_{\partial\Omega(t)} (u^m \nabla u) \cdot d\mathbf{\Gamma} \tag{16}$$

so that if $u^m \nabla u$ vanishes on the boundary the total mass is conserved, *i.e.*

$$\int_{\Omega(t)} u r^{d-1} dr = \text{constant in time.} \tag{17}$$

Note that for this particular problem (12), with $\mathbf{q}$ set to zero, becomes

$$\begin{aligned} & -\oint_{\partial\Omega(t)} w_i u \nabla\phi \cdot d\mathbf{\Gamma} + \int_{\Omega(t)} u \nabla\phi \cdot \nabla w_i d\Omega \\ = & -\int_{\Omega(t)} u^m \nabla w_i \cdot \nabla u d\Omega \,, \end{aligned} \tag{18}$$

where it is again assumed that $u^m \nabla u$ vanishes on the boundary.

The radially symmetric form of (15) (in d dimensions) is

$$\frac{\partial u}{\partial t} = \frac{1}{r^{d-1}} \frac{\partial}{\partial r} \left(r^{d-1} u^m \frac{\partial u}{\partial r} \right) . \tag{19}$$

Equation (19) is invariant under the scalings

$$t \to \lambda t, \quad r \to \lambda^\beta r, \quad u \to \lambda^{(2\beta-1)/m} u \,. \tag{20}$$

If the boundary conditions are such that the total mass is invariant, then it follows that $\beta = 1/(2+md)$. A source-type self-similar solution,

$$u_{SS} = \begin{cases} \left(\frac{t_0}{t}\right)^{d/(2+md)} \left(1 - \left(\frac{r^2}{Kt^{2/(2+md)}}\right)\right)^{1/m} & r^2 \le Kt^{2/(2+md)} \\ 0 & r^2 > Kt^{2/(2+md)} \end{cases} \tag{21}$$

may be deduced from (20) [6], where t_0, K are constants, which may be used to test numerical results. The function u_{SS} vanishes at the moving boundary and a typical profile is sketched in cross-section in Figure 1 for $m > 1$. For $m = 1$ the slope at the boundary is finite while for $m > 1$ it is infinite. Recall that the global mass is conserved.

For this equation there also exists the following comparison principle [10]: given three sets of initial conditions,

$$u_1(x, y, t_0) \leq u_2(x, y, t_0) \leq u_3(x, y, t_0) \quad \forall\, (x, y) \in \Omega\,, \tag{22}$$

then

$$u_1(x, y, t) \leq u_2(x, y, t) \leq u_3(x, y, t) \quad \forall\, (x, y) \in \Omega,\ t \geq t_0\,. \tag{23}$$

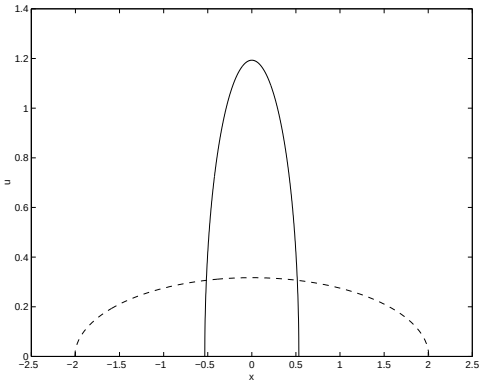


Figure 1: Typical behaviour of the self-similar solution of the PME.

4.2 A Fourth Order Problem

A corresponding fourth order equation (in the fixed coordinate system) is

$$\frac{\partial u}{\partial t} + \nabla \cdot (u^m \nabla(\nabla^2 u)) = 0 \tag{24}$$

which arises in the flow of surface-tension dominated thin liquid films ($m = 3$) and the diffusion of dopant in semiconductors. In integral form it is

$$\frac{d}{dt}\int_{\Omega(t)} u d\Omega = -\int_{\Omega(t)} \nabla \cdot (u^m \nabla(\nabla^2 u)) d\Omega = -\oint_{\partial\Omega(t)} (u^m \nabla(\nabla^2 u)) \cdot d\mathbf{\Gamma} = 0 \tag{25}$$

so that if $u^m\nabla(\nabla^2 u)$ vanishes on the boundary the total mass is constant in time (cf. (17)). The equation (24) may be physically split up into the pair of second order equations

$$u_t + \nabla\cdot(u^m\nabla p) = 0, \qquad p = \nabla^2 u \tag{26}$$

where p is a pressure. In this case, instead of (12), with $\mathbf{q}$ set to zero, we now have

$$\begin{aligned} & -\oint_{\partial\Omega(t)} w_i u\nabla\phi\cdot d\mathbf{\Gamma} + \int_{\Omega(t)} u\nabla\phi\cdot\nabla w_i d\Omega \\ = \; & -\int_{\Omega(t)} u^m\nabla w_i\cdot\nabla\pi d\Omega\,, \end{aligned} \tag{27}$$

again assuming that $u^m\nabla(\nabla^2 u)$ vanishes on the boundary. In (27) π is a weak approximation to the pressure given by

$$\int_{\Omega(t)} w_i\pi d\Omega = -\int_{\Omega(t)} \nabla w_i\cdot\nabla u d\Omega\,, \tag{28}$$

where an additional boundary condition, stating that the normal derivative of u is zero throughout $\partial\Omega(t)$, has been used.

The fourth order radially symmetric equation (24) in d dimensions (in split form) is

$$\frac{\partial u}{\partial t} + \frac{1}{r^{d-1}}\frac{\partial}{\partial r}\left(r^{d-1}u^m\frac{\partial p}{\partial r}\right) = 0, \qquad p = \frac{1}{r^{d-1}}\frac{\partial u}{\partial r} \tag{29}$$

which is invariant under the scalings

$$t\to\lambda t, \quad x\to\lambda^\beta r, \quad u\to\lambda^{(4\beta-1)/m}u\,. \tag{30}$$

Again, if the boundary conditions are such that the total mass (17) is conserved, it then follows that $\beta = 1/(4+md)$.

¿From (30) it follows that for $m=1$ there exists a source-type self-similar solution in the closed form

$$u_{SS} = \begin{cases} \left(\frac{t_0}{t}\right)^{d/(4+d)}\left(1-\left(\frac{r^2}{Kt^{2/(4+d)}}\right)\right)^2 & r^2\le Kt^{2/(4+d)} \\ 0 & r^2 > Kt^{2/(4+d)} \end{cases} \tag{31}$$

where t_0, K are constants, which may be used to test numerical results. More details of this similarity solution may be found in, for example, [8].

5 Numerical Results

5.1 The Second Order Problem

Two sets of results are presented to illustrate the accuracy with which the method approximates the second order problem presented in Section 4.1. The first set is in one dimension ($d = 1$). Figure 2 shows that the rate at which the L^1 error in the approximation decreases is roughly proportional to $(\Delta x)^2$ in the $m = 1$ case, where Δx is taken to be the length of a cell in the initial (uniform) mesh. The order of accuracy is noticeably lower (approximately 1 for the boundary position and 1.4 for the solution values) when $m = 2$ because the exact slope of the solution at the boundary is now infinite.

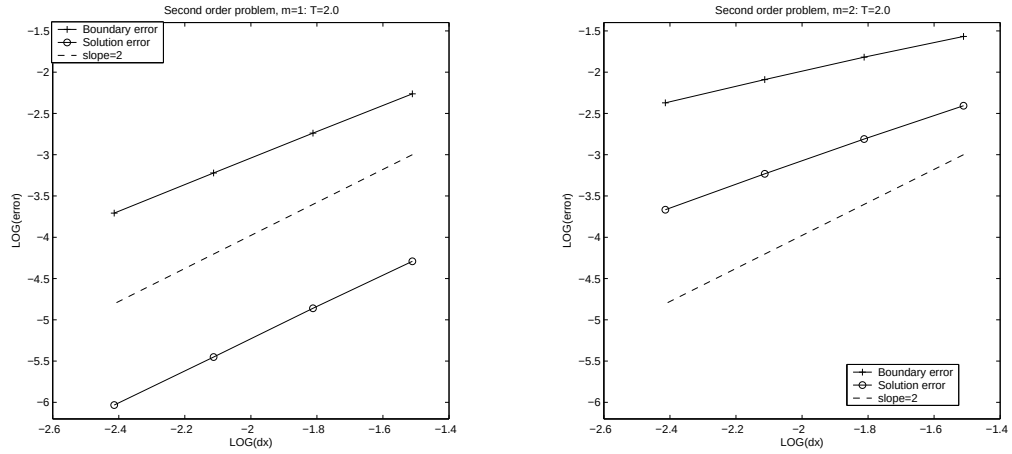


Figure 2: Orders of accuracy in the L^1 norm of approximations to one-dimensional self-similar solutions of the PME for $m = 1$, $T = t - \frac{1}{8}$ (left) and $m = 2$, $T = t - \frac{1}{6}$ (right). The dashed line indicates a slope of 2.

Figure 3 shows snapshots of the evolution of a solution given by equation (21) in the cases $m = 1$ and $m = 2$ in two dimensions ($d = 2$). It is approximated on a genuinely unstructured, but still uniform, 2349 node, 4539 cell, mesh. Further results, presented in [1, 2], give more details and show that the order of accuracy obtained in two space dimensions is the same as in one.

It should be noted that mass is conserved to machine accuracy in all of these calculations, and indeed also for those presented in the next subsection

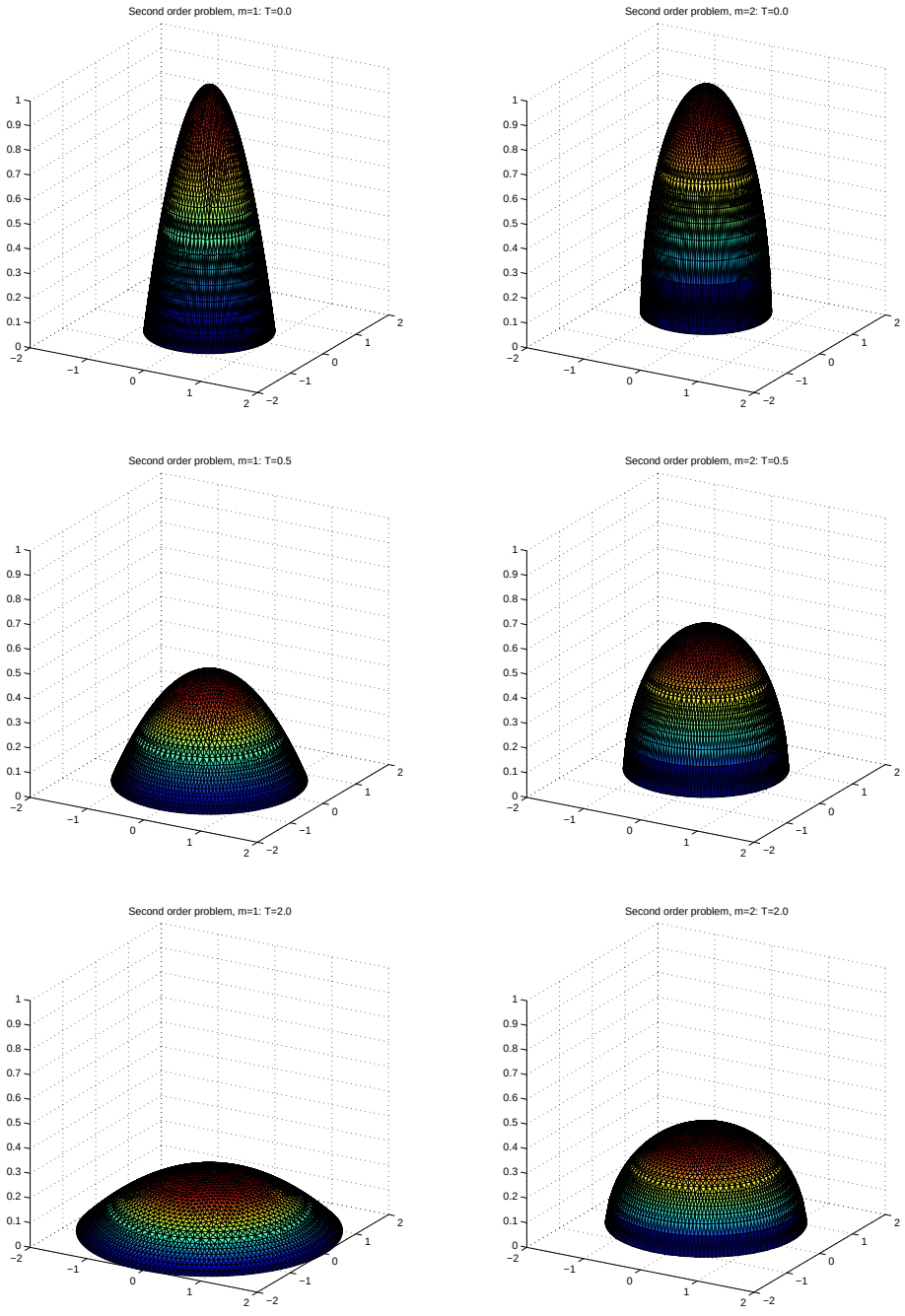


Figure 3: Snapshots of approximations to radially symmetric self-similar solutions (with corresponding triangular meshes) of the PME at three different times for $m = 1$, $T = t - \frac{1}{8}$ (left) and $m = 2$, $T = t - \frac{1}{6}$ (right).

for fourth order problems. Furthermore, although the curl of the mesh velocity field is assumed here to be zero, it is possible to successfully impose a background velocity field $\mathbf{q}$ on the mesh movement equations via the extra vorticity term in (12) [1].

5.1.1 Comparison results

The new scheme is not restricted to modelling self-similar solutions.

We have investigated a comparison property of the approximate solutions which reflects the same property of the exact solution of the PME (see (22,23)). This property holds for the approximate solution derived here, as can be seen in Figures 4 and 5 which show two experiments in which the initial conditions are perturbed. In Figure 4 a random perturbation is applied to the initial solution and its evolution compared with two radially symmetric solutions scaled according to the minimum and maximum perturbations. In Figure 5 a sinusoidal perturbation is applied to the initial position of the boundary and its evolution is found to be sandwiched in a similar manner. In both cases the initial random perturbations are smoothed out very rapidly.

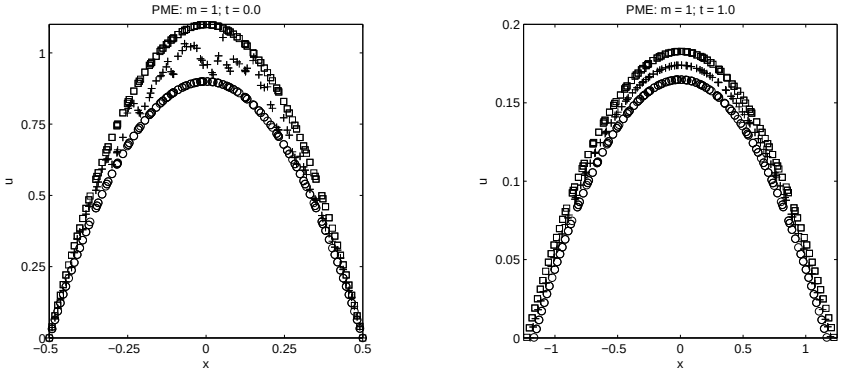


Figure 4: Slices of the initial conditions (left) and approximate solutions at $t = 1$ (right) taken through the origin, illustrating the 'sandwiching' of a randomly perturbed solution to the PME with $m = 1$.

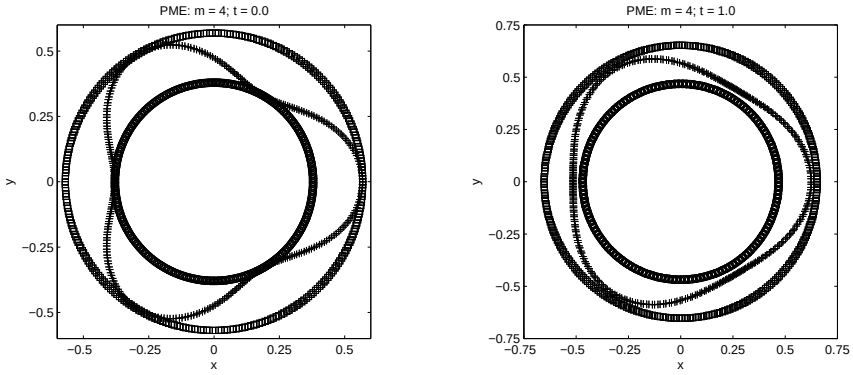


Figure 5: Slices of the initial conditions (left) and approximate solutions at $t = 1$ (right) taken through the origin, illustrating the 'sandwiching' of a sinusoidally perturbed mesh for the PME with $m = 4$.

5.2 The Fourth Order Problem

Similar sets of results are presented for the fourth order problem described in Section 4.2. Figure 6 shows that the L^1 error now decreases in proportion to $(\Delta x)^4$ when $m = 1$ and $d = 1$ (for which the exact self-similar solution (31) exists). Δx is defined as before.

Figure 7 shows snapshots of the evolution of a solution given by equation (31) in two dimensions ($d = 2$), approximated on a uniform unstructured 545 node, 1024 cell mesh. The accuracy of this approximation is comparable to that obtained in one dimension. Note, though, that explicit time-stepping is being used and finer meshes are very expensive to use because the stability of the method appears to require Δt to reduce in proportion to $(\Delta x)^4$.

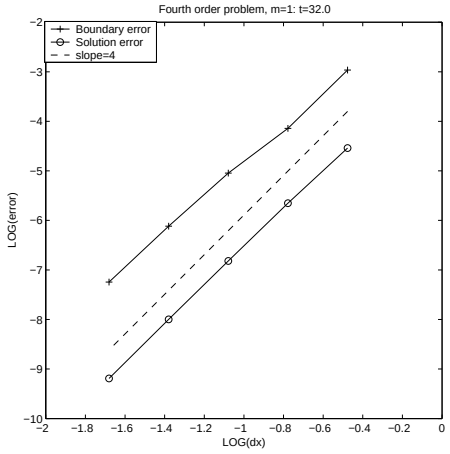


Figure 6: Order of accuracy in the L^1 norm of an approximation to a one-dimensional ($d = 1$) self-similar solution of the fourth order equation with $m = 1$.

6 Conclusions

A Lagrangian moving finite element method has been described which is based on local mass conservation, in line with the scale invariance of problems that exhibit global mass conservation. The method is illustrated on second order and fourth order nonlinear diffusion problems with moving boundaries for which nass is conserved and analytic self-similar solutions exist. Results show that the method is accurate and exhibits approximate scale invariance.

Self-similar solutions have been considered here for the purpose of illustrating the accuracy of the method, but the method can be applied far more generally. For example, we have shown that in the case of the second order problem the comparison principle (22,23), as well as a similar principle for the boundary, is sustained on a numerical level.

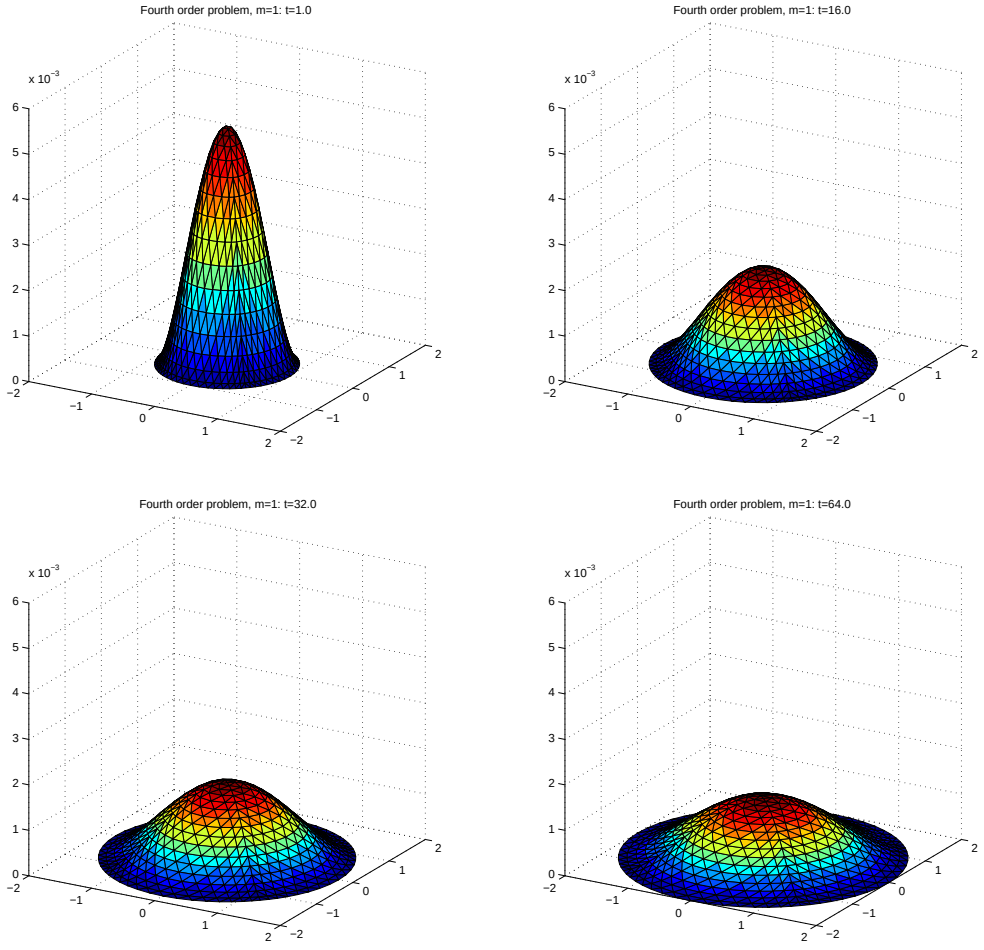


Figure 7: Snapshots of an approximation to a radially symmetric self-similar solution to the fourth order equation at four different times.

References

[1] Baines, M.J., Hubbard, M.E. and Jimack, P.K., A Moving Finite Element Method using Monitor Functions, Report 2003.04, School of Computing, University of Leeds, UK (2003).

[2] Baines, M.J., Hubbard, M.E. and Jimack, P.K., A Lagrangian Moving Finite Element Method using Monitor Functions, in Proceedings of the Third Workshop on Computational Methods in Partial Differential Equations, Hong Kong, (2003).

[3] Blake, K.W., Moving Mesh Methods for Nonlinear Partial Differential Equations, PhD thesis, Dept of Mathematics, University of Reading, UK (2001).

[4] Blake, K.W. and Baines, M.J., Moving Mesh Methods for Nonlinear Partial Differential Equations, Numerical Analysis Report 7/01, Dept of Mathematics, University of Reading, UK (2001).

[5] Budd, C.J, Huang, W. and Russell, R.D., Moving Mesh Methods for Problems with Blow-up, SIAM J. Sci. Comput., 17:305-327 (1996).

[6] Budd, C.J. and Piggott, M., Geometric Integration and its Applications, J. Comp. Appl. Math., 128:399-422 (2001).

[7] Cao, W., Huang, W. and Russell, R.D., A Moving Mesh Method based on the Geometric Conservation Law, SIAM J. Sci. Comput., 24:118-142 (2002).

[8] Diez, J.A., Kondic, L. and Bertozzi, A., Global Models for Moving Contact Lines, Physical Review E, 63:011028 (2000).

[9] Huang, W, Ren, Y. and Russell, R.D , Moving Mesh Partial Differential Equations (MMPDEs) based on the Equidistribution Principle, SIAM J. Num. Anal., 31:709-730 (1994).

[10] Oleinik, O. et al., Istv. Akad. Nauk. SSR Ser. Mat., 22:667-704 (1958).

[11] Thomas, P.D. and Lombard, C.K., The Geometric Conservation Law and its Application to Flow Computations on Moving Grids, AIAA J., 17:1030-1037 (1979).

Numerical analysis of the motion of glass under external pressure.

K. Laevsky, R.M.M. Mattheij
Department of Mathematics and Computer Science,
Eindhoven University of Technology,
PO Box 513, 5600 MB The Netherlands

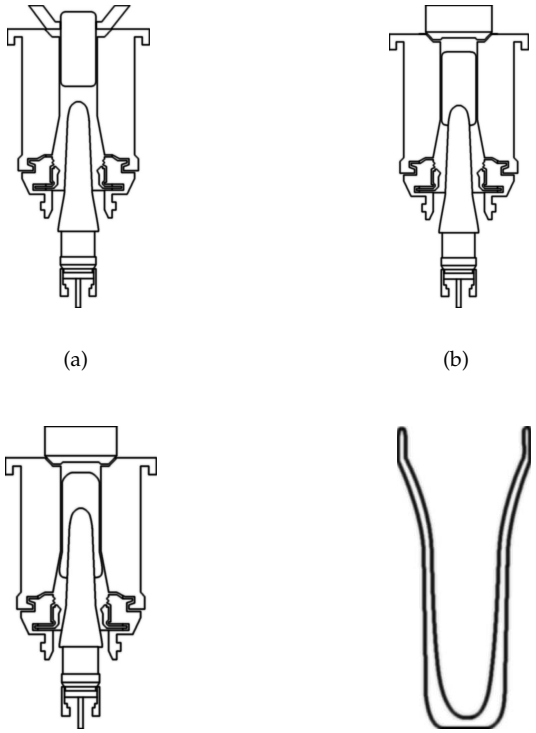
Abstract

We give a mathematical model of the forming of a glass product in a mould under pressure. It turns out the the equations of motion are the Stokes equations. One part of the boundary is given, another part is free. The latter means that the velocity there comes from an external force, in particular from a piston that drives a moving part of the mould (the plunger) into the glass. This provides for an additional (kinematic) boundary condition. The complication here is that the movement of this piston on one hand and the counter force from the glass on the other are coupled. The equation of motion are the Stokes equations. The boundary condition couples these with the motion of the plunger, being an ordinary differential equation. It turns out that the resulting equation for the plunger velocity is stiff, so it should be solved by an implicit method. However, due to the afore mentioned coupling a straightforward implementation of such an implicit scheme is impossible. We give a solution to this problem.

1 Introduction

Glass is a simple material and is available in all sorts of applications. Yet production and forming are matters that still pose questions the answers to them relying more and more on mathematical modelling and simulation, cf.[2], [3], [4]. In this paper we consider the motion of molten glass by pressure, which is an important step in the production of container glass. In particular we model the pressing of a preform or *parison* in a *mould* used in the mechanical production of container glass. In Figure 1.1 we have sketched the various parts making up for the mould. The actual mould consists of the *baffle*, the *blank*, and the *neckring*. Initially the baffle part is removed and the mould is open from above (cf. Figure 1.1a). Once a gob of glass is inside the mould, the baffle is closed and the *plunger* moves up by the force of a piston (cf. Figure 1.1b,c). This parison (see Figure 1.1d), is then blown into it final shape in the next stage (see Figure 1.2).

Although the temperature plays an important role in this modelling [5], it can be shown that during the pressing phase the temperature changes are rather small because of the short time the pressing takes. Hence we consider the problem to be isothermal. We shall model the process assuming all parts of the mould and the plunger to have axisymmetric geometry. An appropriate choice for the coordinate system to be used in order to solve the equations numerically are then cylindrical coordinates. The motion of a fluid can be described by Navier Stokes. By dimension analysis it can be shown that they simplify to Stokes equations, cf. [4] The Stokes equations in cylindrical coordinates can be formulated as follows, cf [1]. Find the velocity field $\mathbf{v} := (u_r(r,z,\varphi),\ u_z(r,z,\varphi),\ u_\varphi(r,z,\varphi))^T$ and pressure field $p := p(r,\ z,\ \varphi)$, which satisfy



(a) (b) (c) (d)

Figure 1.1: Pressing process.

$$\nabla \cdot \sigma(\mathbf{v}, p) = 0, \tag{1.1}$$

$$\nabla \cdot \mathbf{v} = 0, \tag{1.2}$$

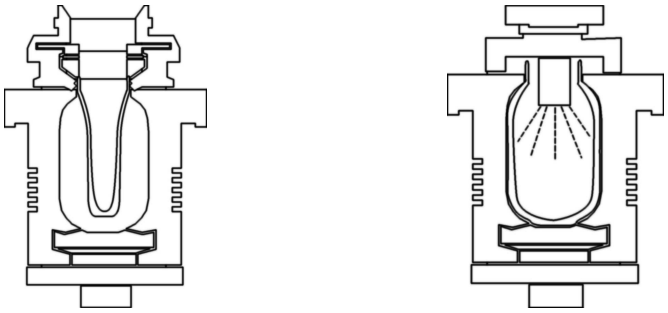
where $\sigma(\mathbf{v}, p)$, the stress tensor, is given by

$$\sigma(\mathbf{v}, p) = -pI + \eta(\nabla \mathbf{v} + \nabla \mathbf{v}^T). \tag{1.3}$$

Here I is the identity tensor.
Using the formula for the gradient in cylindrical coordinates we obtain

$$\sigma = \begin{pmatrix} -p + 2\eta \dfrac{\partial u_r}{\partial r} & \eta\left(\dfrac{\partial u_r}{\partial z} + \dfrac{\partial u_z}{\partial r}\right) & \eta\left(\dfrac{1}{r}\dfrac{\partial u_r}{\partial \varphi} + \dfrac{\partial u_\varphi}{\partial r} - \dfrac{u_\varphi}{r}\right) \\ \eta\left(\dfrac{\partial u_r}{\partial z} + \dfrac{\partial u_z}{\partial r}\right) & -p + 2\eta \dfrac{\partial u_z}{\partial z} & \eta\left(\dfrac{1}{r}\dfrac{\partial u_z}{\partial \varphi} + \dfrac{\partial u_\varphi}{\partial z}\right) \\ \eta\left(\dfrac{1}{r}\dfrac{\partial u_r}{\partial \varphi} + \dfrac{\partial u_\varphi}{\partial r} - \dfrac{u_\varphi}{r}\right) & \eta\left(\dfrac{1}{r}\dfrac{\partial u_z}{\partial \varphi} + \dfrac{\partial u_\varphi}{\partial z}\right) & -p + 2\eta\left(\dfrac{1}{r}\dfrac{\partial u_\varphi}{\partial \varphi} + \dfrac{u_r}{r}\right) \end{pmatrix}. \tag{1.4}$$

Equations (1.2), (1.1), rewritten in terms of cylindrical coordinates, read as



(a) (b)

Figure 1.2: Pressing process.

$$\frac{\partial^2 u_r}{\partial r^2} + \frac{\partial^2 u_r}{\partial z^2} + \frac{1}{r^2}\frac{\partial^2 u_r}{\partial \varphi^2} + \frac{1}{r}\frac{\partial u_r}{\partial r} - \frac{2}{r^2}\frac{\partial u_\varphi}{\partial \varphi} - \frac{u_r}{r^2} = \frac{1}{\eta}\frac{\partial p}{\partial r}, \tag{1.5}$$

$$\frac{\partial^2 u_z}{\partial r^2} + \frac{\partial^2 u_z}{\partial z^2} + \frac{1}{r^2}\frac{\partial^2 u_z}{\partial \varphi^2} + \frac{1}{r}\frac{\partial u_z}{\partial r} = \frac{1}{\eta}\frac{\partial p}{\partial z}, \tag{1.6}$$

$$\frac{\partial^2 u_\varphi}{\partial r^2} + \frac{\partial^2 u_\varphi}{\partial z^2} + \frac{1}{r^2}\frac{\partial^2 u_\varphi}{\partial \varphi^2} + \frac{1}{r}\frac{\partial u_\varphi}{\partial r} + \frac{2}{r^2}\frac{\partial u_r}{\partial \varphi} - \frac{u_\varphi}{r^2} = \frac{1}{\eta r}\frac{\partial p}{\partial \varphi}, \tag{1.7}$$

$$\frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} + \frac{1}{r}\frac{\partial u_\varphi}{\partial \varphi} + \frac{u_r}{r} = 0. \tag{1.8}$$

2 Rotational Symmetry

As was explained in Section 1 both the mould and the plunger are axisymmetric. Since the plunger is moving in vertical direction the velocity, $V_p(t)$ say, we can write

$$\mathbf{v}_p(t) = V_p(t)\mathbf{e}_z := (0,\, V_p(t),\, 0)^T, \tag{2.1}$$

where $\mathbf{e}_z$ is the unit vector in z direction. We may reduce the dimension of the problem and consider (1.1), (1.2) in two-dimensional axisymmetric coordinates. The velocity field then has the components

$$\mathbf{v} := (u_r(r, z, \varphi),\, u_z(r, z, \varphi),\, 0)^T, \tag{2.2}$$

and the pressure field

$$p := p(r,\, z,\, 0). \tag{2.3}$$

From (1.4) we obtain the stress tensor for the axisymmetric case

$$\sigma = \begin{pmatrix} -p + 2\eta \dfrac{\partial u_r}{\partial r} & \eta \left(\dfrac{\partial u_r}{\partial z} + \dfrac{\partial u_z}{\partial r} \right) & 0 \\ \eta \left(\dfrac{\partial u_r}{\partial z} + \dfrac{\partial u_z}{\partial r} \right) & -p + 2\eta \dfrac{\partial u_z}{\partial z} & 0 \\ 0 & 0 & -p + 2\eta \dfrac{u_r}{r} \end{pmatrix}. \tag{2.4}$$

The Stokes equations (1.5)- (1.8) take the following form

$$\frac{\partial^2 u_r}{\partial r^2} + \frac{\partial^2 u_r}{\partial z^2} + \frac{1}{r}\frac{\partial u_r}{\partial r} - \frac{u_r}{r^2} = \frac{1}{\eta}\frac{\partial p}{\partial r}, \tag{2.5}$$

$$\frac{\partial^2 u_z}{\partial r^2} + \frac{\partial^2 u_z}{\partial z^2} + \frac{1}{r}\frac{\partial u_z}{\partial r} = \frac{1}{\eta}\frac{\partial p}{\partial z}, \tag{2.6}$$

$$\frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} + \frac{u_r}{r} = 0. \tag{2.7}$$

Clearly, the pressure p is defined up to a constant. One can should notice singularities in (2.4)-(1.7) when $r = 0$.

3 Boundary Conditions

As we have an axisymmetric problem we obtain a domain Ω, as sketched in Figure 3.1. The boundary $\Gamma := \partial\Omega$ of the domain consists of four parts

$$\Gamma = \Gamma_s \cup \Gamma_m \cup \Gamma_p \cup \Gamma_f, \tag{3.1}$$

where the indices s, m, p, f represent the symmetric, mould, plunger and free boundaries respectively. Let

$$\mathbf{n} = (n_r, n_z, 0)^T, \quad \mathbf{t} = (t_r, t_z, 0)^T \tag{3.2}$$

be the normal and tangent unit vectors respectively for the boundary Γ in the directions as displayed in Figure 3.1. Then we find the following boundary conditions. Because of symmetry, the boundary conditions on Γ_s are

$$\mathbf{v} \cdot \mathbf{n} = 0, \tag{3.3}$$

$$\sigma \mathbf{n} \cdot \mathbf{t} = 0. \tag{3.4}$$

It is easy to see that

$$\mathbf{n} = (-1,\, 0,\, 0)^T, \quad \mathbf{t} = (0,\, -1,\, 0)^T, \quad \sigma\mathbf{n} = (-\sigma_{rr},\, -\sigma_{rz},\, 0)^T \tag{3.5}$$

on Γ_s. Using the expressions for the stress tensor components (2.4) we obtain

$$u_r = 0, \quad \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} = 0. \tag{3.6}$$

Since $u_r \equiv 0$ on Γ_s, it follows that the derivative along Γ_s is also equal to zero, i.e., $\partial u_r/\partial z = 0$. As a result the boundary conditions on Γ_s can be written as

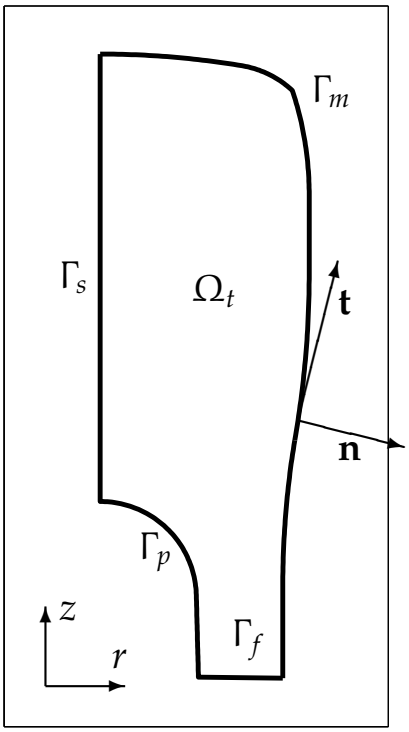


Figure 3.1: Problem domain.

$$u_r = 0, \quad \frac{\partial u_z}{\partial r} = 0. \tag{3.7}$$

For the mould and the plunger we will allow both slip and no slip boundary conditions and everything in between. A partial slip boundary condition for the mould means that the normal component of the velocity should be zero and the tangential component proportional to the tangential stress, i.e.

$$\mathbf{v} \cdot \mathbf{n} = 0, \tag{3.8}$$

$$(\sigma \mathbf{n} + \beta_m \mathbf{v}) \cdot \mathbf{t} = 0, \tag{3.9}$$

where β_m is a friction coefficient. The first equation clearly represents a Dirichlet boundary condition, and the second a Robin boundary condition.

For the plunger which moves with velocity $\mathbf{v}_p$ (see (2.1)), we find

$$(\mathbf{v} - \mathbf{v}_p) \cdot \mathbf{n} = 0, \tag{3.10}$$

$$(\sigma \mathbf{n} + \beta_p (\mathbf{v} - \mathbf{v}_p)) \cdot \mathbf{t} = 0. \tag{3.11}$$

Note that $\mathbf{v}_p$ does not depend on r, z, and β_p is again the friction coefficient. The physical meaning of these conditions is the same as for (3.8), (3.9), with the only difference that here we consider

the velocity relative to $\mathbf{v}_p$, i.e., $\mathbf{v} - \mathbf{v}_p$. Also we are using the fact that $\sigma(\mathbf{v} - \mathbf{v}_p, p) = \sigma(\mathbf{v}, p)$. Let $V_p > 0$ be the absolute velocity of the plunger, then

$$\mathbf{v}_p = V_p \mathbf{e}_z := (0, V_p, 0)^T. \tag{3.12}$$

Actually, the velocity of the plunger V_p is an unknown function of time t, so we should write $V_p(t)$. Nevertheless, for the boundary conditions below and the Stokes problem as such, we view this as just a parameter. Hence, the boundary conditions read as follows

$$\mathbf{v} \cdot \mathbf{n} = V_p \mathbf{e}_z \cdot \mathbf{n}, \tag{3.13}$$

$$(\sigma \mathbf{n} + \beta_p \mathbf{v}) \cdot \mathbf{t} = \beta_p V_p \mathbf{e}_z \cdot \mathbf{t}. \tag{3.14}$$

Finally the boundary conditions at the free boundary Γ_f are defined as the vector relation

$$\sigma \mathbf{n} = -p_0 \mathbf{n}, \tag{3.15}$$

where p_0 is the external pressure. We can take the inner product of (3.15) with $\mathbf{n}$, $\mathbf{t}$ and obtain the boundary conditions in the form of two scalar equations

$$\sigma \mathbf{n} \cdot \mathbf{n} = -p_0, \tag{3.16}$$

$$\sigma \mathbf{n} \cdot \mathbf{t} = 0. \tag{3.17}$$

Note that the velocity field found from (1.1), (1.2) with the boundary conditions (3.3) – (3.17), is independent of the value of p_0. From a physical point of view this can be explained by the incompressibility of the fluid.

4 An Ordinary Differential Equation for the Plunger Velocity

The velocity $V_p(t)$ of the plunger is not known beforehand and in fact coupled to the motion of the glass itself. Indeed, the plunger movement is the result of a certain pressure p_p applied to its bottom. Let $F(t)$ denote the total force on the plunger and m_p be the mass of the plunger. Then

$$\frac{dV_p(t)}{dt} = \frac{F(t)}{m_p}. \tag{4.1}$$

This total force is the sum of

$$F(t) = F_p + F_g(t), \tag{4.2}$$

where F_p remains constant through the whole process and $F_g(t)$ is the force on the plunger from the glass. The constant force can be computed as

$$F_p = S_p p_p = \text{beingsomeconstant}. \tag{4.3}$$

Here S_p is the area of the surface where pressure p_p is applied. The second term $F_g(t)$, is the force on the plunger from the glass. The force from the glass can be expressed in terms of the stress tensor (2.4)

$$F_g(t) = \int_{S(t)} \sigma \mathbf{n} \cdot \mathbf{e}_z \, dS, \tag{4.4}$$

where $\sigma \equiv \sigma(t)$ is the stress tensor, and $S(t)$ the part of the plunger surface which is in contact with the glass at time t. The formula requires integration of the second component of $\sigma \mathbf{n}$ only,

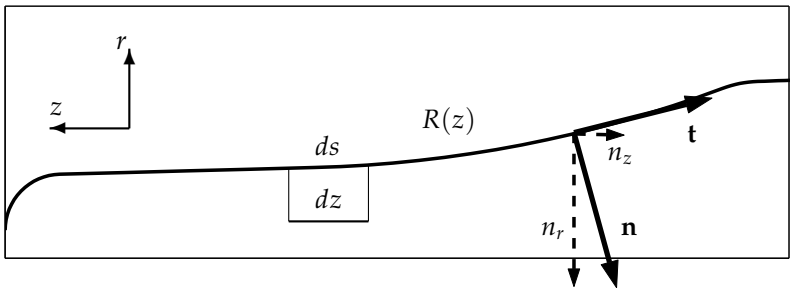


Figure 4.1: Geometry of the plunger.

as the first one will vanish due to such integration because of the axisymmetrical nature of the problem.
Consider Figure 4.1 which depicts one half of the plunger (cf. Figure 4.1) turned by 90 degrees. If z is the axial variable and $R(z)$ denotes the form of the plunger we can derive (cf. [7]

$$dS = 2\pi\, R_p(s)\, ds = 2\pi \sqrt{1 + R_p'(z)^2}\, R_p(z)\, dz, \tag{4.5}$$

where s represents the length over the plunger profile. The two dimensional surface $S(t)$ is related to the interval $[z_0, z_1]$ on the z axis. Then (4.4) can be written as follows

$$F_g(t) = 2\pi \int_{z_0}^{z_1} \sigma\mathbf{n} \cdot \mathbf{e}_z \sqrt{1 + R_p'(z)^2}\, R_p(z)\, dz. \tag{4.6}$$

The values of $\sigma\mathbf{n}$ can be obtained as follows The normal components n_r, n_z (see Figure 4.1) are computed as follows

$$\mathbf{n} = -\frac{1}{\sqrt{1 + R_p'(z)^2}}\, (1,\, R_p'(z),\, 0)^T. \tag{4.7}$$

Using the expressions (2.4) for the stress tensor components, (4.6) reads

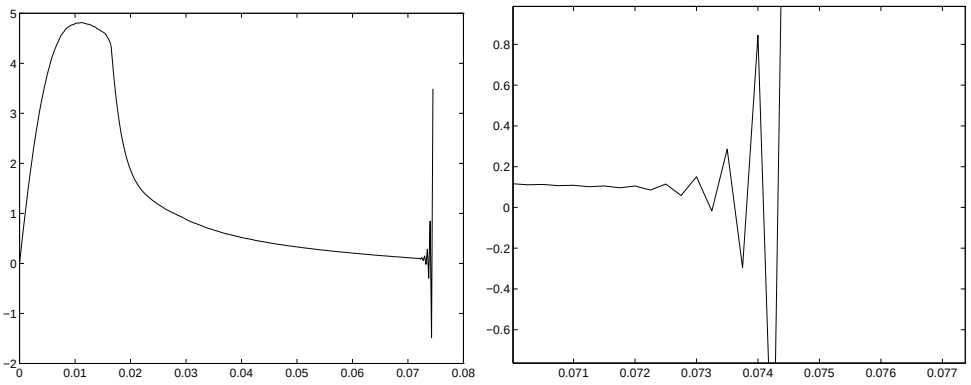
$$F_g(t) = 2\pi \int_{z_0}^{z_1} \left(\left(p - 2\eta \frac{\partial u_z}{\partial z} \right) R_p'(z) + \eta \left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \right) R_p(z)\, dz. \tag{4.8}$$

Now in order to compute the velocity of the plunger $V_p(t)$ as a function of time, one should solve the ordinary differential equation

$$\begin{cases} \dfrac{dV_p(t)}{dt} &= \dfrac{F_g(t)}{m_p} + \dfrac{F_p}{m_p}, \\ \\ V_p(0) &= V_0, \end{cases} \tag{4.9}$$

where V_0 is some initial velocity of the plunger. Note that we can compute $F_g(t)$, once u_r, u_z, p (or $\sigma\mathbf{n}$) are known. The latter are obtained from solving the Stokes equations. In order to solve the Stokes equations one needs some value for the plunger velocity V_p in (3.13) and (3.14). So, at time $t = 0$ we use V_0 from (4.9) and find $F_g(0)$. We can thus perform an explicit integration step in (4.9). In general, suppose we use the Euler forward scheme

$$V_p^{k+1} = V_p^k + \Delta t^k\, \frac{F_g(t^k) + F_p}{m_p}. \tag{4.10}$$



(a) (b)

Figure 4.2: Velocity of the plunger (numerical instabilities).

Having solved the Stokes equations, with the new velocity of the plunger V_p^{k+1}, we can complete the boundary conditions for the Stokes problem at $t = t^{k+1}$. To this end the velocity of the plunger obtained from (4.10) is used. However, as illustrated in Figure 4.2, the algorithm turns out to be unstable. Looking more carefully at Figure 4.2 we detect a phenomenon that looks like stiffness. To overcome this we should take recourse to implicit methods. A fully implicit scheme, however, practically impossible as we do not know the plunger velocity at t^{k+1}; thus we cannot use it for the boundary conditions (3.13), (3.14). Of course, a predictor-corrector scheme for such an implicit integrator will only converge for infeasible small time steps because of stiffness.

5 A Stiffness Phenomenon

In this section we like to investigate the stiffness of the ordinary differential equation (4.9). Clearly, we need to have a closer look at $F_g(t)$, as derived in (4.8). In general it is impossible to compute it exactly so we take recourse to a thin film approximation. Here we shall approach the problem analytically in order to point out the stiffness phenomenon detected in numerical simulation. For a more detailed discussion see [7]. We shall consider a simple, yet meaningful geometry for the mould and the plunger, see Figure 5.1. Let each of them be defined by a parabola, say

$$R_m(z) = d_m\sqrt{z}, \quad R_p(z) = d_p\sqrt{z - z_0}, \tag{5.1}$$

where coefficients d_m, d_p have positive values and z_0 is the position of the plunger.

Let us define $\varepsilon := D/L$ as the ratio between the length scales corresponding to the parison's wall thickness D and the height of the parison L. Since D is smaller than L, ε is a small parameter. The variables can be then scaled as follows

$$r = Dr', \quad z = Lz', \quad u_r = \varepsilon V u_r', \quad u_z = V u_z', \quad p = \frac{\eta V L}{D^2} p', \tag{5.2}$$

where V is the typical flow velocity. Using (5.2) we can make (4.8) dimensionless

$$F_g(t) := 2\pi\eta V L F_g'(t). \tag{5.3}$$

Then (4.8) can be approximated by the following expression

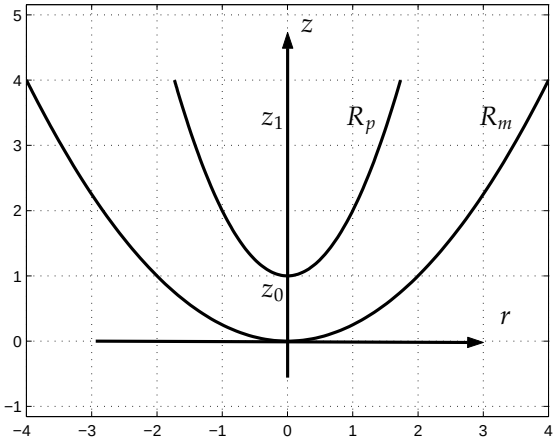


Figure 5.1: Mould and plunger geometries defined by parabolas.

$$\begin{aligned} F_g'(t) &= \int_{z_0'}^{z_1'} \left(\left(p' - 2\varepsilon^2 \frac{\partial u_z'}{\partial z'} \right) {R_p'}'(z') + \left(\varepsilon^2 \frac{\partial u_r'}{\partial z'} + \frac{\partial u_z'}{\partial r'} \right) \right) R_p'(z')\, dz \\ &\approx \int_{z_0'}^{z_1'} \left(p' {R_p'}'(z') + \frac{\partial u_z'}{\partial r'} \right) R_p'(z')\, dz. \end{aligned} \tag{5.4}$$

Using (5.2) it is possible to find the exact solution of the Stokes equations (2.5), (2.6), (2.7) (see [7])

$$\begin{aligned} u_r' &= \frac{1}{r'} \frac{d}{dz'} \int_{r'}^{R_m'} r' u_z'(r', z')\, dr, \\ u_z' &= \frac{1}{4} {r'}^2 \frac{dp'}{dz'} + A(z') \ln r' + B(z'), \end{aligned} \tag{5.5}$$

where $A(z')$ and $B(z')$ can be obtained from the boundary conditions. The eventual dimensional force $F_g(t)$ takes then the following form

$$F_g(t) \approx 2\pi\eta V L\, V_p'(t') \int_{z_0'}^{z_1'} \frac{c_m - c_p}{(b_m - b_p)^2 - (a_m - a_p)(c_m - c_p)}\, dz. \tag{5.6}$$

Here $V_p'(t')$ is the dimensionless velocity of the plunger scaled with V; a_m, a_p, b_m, b_p, c_m, c_p denote

$$\begin{aligned} a_m &= \ln R_m'(z') + s_m / R_m'(z'), \quad a_p = \ln R_p'(z') + s_p / R_p'(z'), \\ b_m &= {R_m'}^2(z')(1 + 2s_m / R_m'(z')), \quad b_p = {R_p'}^2(z')(1 + 2s_p / R_p'(z')) \\ c_m &= {R_m'}^4(z')(1 + 4s_m / R_m'(z')), \quad c_p = {R_p'}^4(z')(1 + 4s_p / R_p'(z')), \end{aligned} \tag{5.7}$$

respectively. Here s_m, s_p are dimensionless parameters similar to the friction coefficients β_m, β_p as defined in Section 3. Note that all defined quantities are dimensionless.

The dimensionless integral in (5.4) can be computed numerically. The graph in Figure 5.2 shows the results of this integration as a function of upper bound z_1' in (5.4). Using the same scaling (4.9) reads

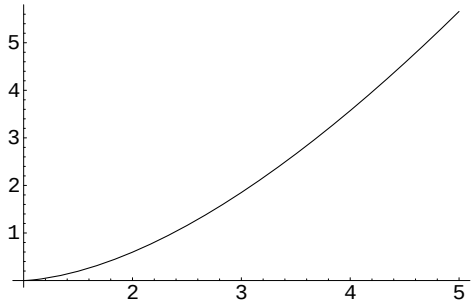


Figure 5.2: Force on the plunger as a function of z_1'

$$\frac{dV_p'}{dt'} = V_p' I(t) \frac{2\pi L^2 \eta}{V m_p} + C, \tag{5.8}$$

where $t = t'\, L/V$, $V_p = V V_p'$, $I(t)$ is the dimensionless integral from (5.4) and C some constant. The typical values for L and V are 10^{-1} m and 10^{-1} s respectively. The mass of the plunger device m_p is of order 1. The viscosity coefficient η for our problem is a large number

$$\eta \approx 10^4 \ \ \mathrm{kg/s\,m}. \tag{5.9}$$

One can see that the coefficient of V_p' on the right-hand side of **??**chapter6/section2: equation8) is large. Indeed, taking $I(t) \approx 1$ (see Figure 5.2) we find

$$I(t)\, \frac{2\pi L^2 \eta}{V m_p} \approx 10^4. \tag{5.10}$$

This clearly indicates that (5.8) is a stiff equation. One should note that η is the dominating quantity. This will also be the case for more complicated geometries. This then shows the inherent stiffness of the plunger motion equation.

6 Uncoupling the Flow Equations and the Plunger Velocity

From the preceding analysis it follows that an explicit method leads to numerical instabilities, for not unduly small time steps. We therefore prefer to use an implicit method instead. However, the right-hand side $F(t)/m_p$ of (4.1) depends on the solution of the Stokes equations. In order to apply an implicit step to (4.1) at time $t = t^k$ we need to know $F_g(t^{k+1})$. In this case we would compute

$$V_p^{k+1} = V_p^k + \Delta t^k\, \frac{F_g(t^{k+1}) + F_p}{m_p}. \tag{6.1}$$

Note that $F_g(t^{k+1})$ resulting from the solution of the Stokes equations with V_p^{k+1}. Clearly, in this way the Stokes equations and the motion of the plunger are coupled. In order to use the implicit scheme (6.1), we could, for example, predict the velocity of the plunger using (4.10) and then use it for the boundary conditions in the Stokes equations. After having solved the latter, let us compute the value for $F_g(t^{k+1})$ and perform (6.1). Unfortunately this does not work because of the explicit prediction step, which sooner or later cause numerical instabilities.

Below we work out how to overcome the stiffness phenomenon for our problem. The crucial role here is played by regarding the velocity of the plunger $V_p(t)$ uncoupled from the parameter V_p in the boundary conditions for the Stokes problem. We shall make use of the following lemma.

LEMMA 6.1 Let $\mathbf{v}_1$, p_1 and $\mathbf{v}_2$, p_2 be the solutions of the Stokes equations (2.5), (2.6) and (2.7) with corresponding plunger velocities V_{p_1} and V_{p_2} respectively. Then $k_1\mathbf{v}_1 + k_2\mathbf{v}_2$, $p_0 + k_1(p_1 - p_0) + k_2(p_2 - p_0)$ is also a solution of these equations with $V_p = k_1 V_{p_1} + k_2 V_{p_2}$.

Proof. From $\nabla \cdot p_0 I = 0$, it follows that

$$\begin{aligned} \nabla \cdot \sigma(k_1\mathbf{v}_1 + k_2\mathbf{v}_2,\ p_0 + k_1(p_1 - p_0) + k_2(p_2 - p_0)) &= \\ k_1 \nabla \cdot \sigma(\mathbf{v}_1, p_1) + k_2 \nabla \cdot \sigma(\mathbf{v}_2, p_2) &= 0. \end{aligned} \tag{6.2}$$

It is simple to see that such a linear combination satisfies Stokes equation. Note that

$$\nabla \cdot (k_1\mathbf{v}_1 + k_2\mathbf{v}_2) = k_1\, \nabla \cdot \mathbf{v}_1 + k_2\, \nabla \cdot \mathbf{v}_2 = 0. \tag{6.3}$$

Likewise such a property can be shown for the boundary conditions. Considering the pressure field relative to p_0, the boundary conditions (3.16), (3.17) are satisfied

$$\begin{aligned} \sigma(k_1\mathbf{v}_1 + k_2\mathbf{v}_2,\ p_0 + k_1(p_1 - p_0) + k_2(p_2 - p_0))\mathbf{n} &= \\ k_1(\sigma(\mathbf{v}_1, p_1)\mathbf{n} + p_0\mathbf{n}) + k_2(\sigma(\mathbf{v}_2, p_2)\mathbf{n} + p_0\mathbf{n}) - p_0\mathbf{n} &= -p_0.\mathbf{n}. \end{aligned} \tag{6.4}$$

This proves the lemma. □

From Lemma 6.1 it follows that we may consider the velocity and pressure fields at some time t as affine functions of V_p, so

$$\begin{aligned} \mathbf{v}(t; V_p) &= V_p\, \mathbf{v}(t; 1), \\ p(t; V_p) &= p_0 + V_p\, (p(t; 1) - p_0). \end{aligned} \tag{6.5}$$

Here $\mathbf{v}(t; \alpha)$, $p(t; \alpha)$ is the solution of the Stokes equations with the velocity of the plunger equal to $\alpha = \text{const}$. As a consequence we deduce from (4.8) that this then also holds for the glass force

$$F_g(t; V_p) = F_0(t) + V_p\ \left(F_g(t; 1) - F_0(t)\right), \tag{6.6}$$

where $F_0(t)$ is the force on the glass due to normal air pressure

$$F_0(t) = 2\pi \int_{z_0}^{z_1} p_0 R_p'(z) R_p(z)\, dz. \tag{6.7}$$

Using (6.6) we can reformulate (4.9) as follows

$$\begin{cases} \dfrac{dV_p(t)}{dt} &= V_p(t)\dfrac{F_g(t; 1) - F_0(t)}{m_p} + \dfrac{F_p + F_0(t)}{m_p}, \\ V_p(0) &= V_0. \end{cases} \tag{6.8}$$

Note that one should use $V_p = 1$ for the boundary conditions (3.13), (3.14). By tracking the free boundary and defining the Stokes problem, the glass force $F_g(t; 1)$ can be computed for the changing domain Ω. As a consequence it makes sense to consider the force as a function of the plunger position, not the time. So we slightly change the notation

$$F_g := F_g(z; V_p), \quad V_p := V_p(z). \tag{6.9}$$

Equation (6.8) should be reformulated as follows

$$\begin{cases} \dfrac{1}{2}\dfrac{dV_p^2(z)}{dz} &= V_p(z)\dfrac{F_g(z;1)-F_0(z)}{m_p}+\dfrac{F_p+F_0(z)}{m_p}, \\ V_p(0) &= V_0. \end{cases} \tag{6.10}$$

Here we used

$$\frac{dV_p(t)}{dt}=\frac{dV_p(z)}{dz}V_p(z). \tag{6.11}$$

By solving these equations for a evolving glass domains, we can obtain a table with plunger positions, and velocity and pressure fields computed for $V_p = 1$ in such domains. Hence, the velocity of the plunger can be considered to be a function of the plunger position, but still being unknown as a function of t.

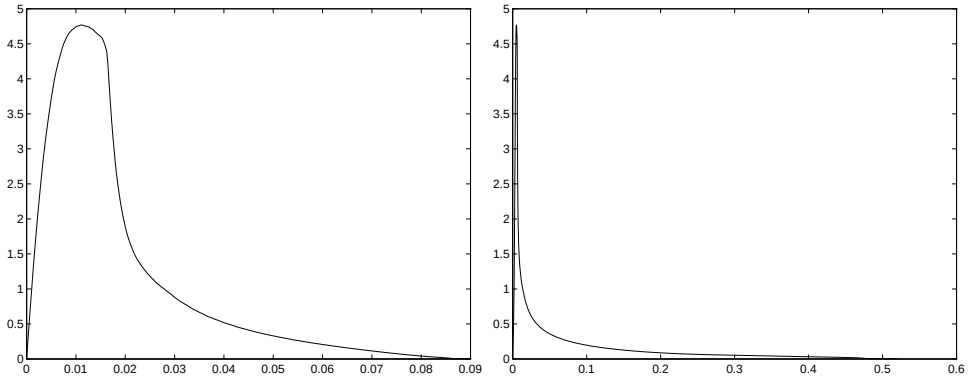
If one applies the Euler explicit method to (6.10),

$$\begin{cases} \dfrac{1}{2}\dfrac{V_p^{k+1\,2}-V_p^{k\,2}}{z^{k+1}-z^k} &= V_p^k\,\dfrac{F_g(z^k;1)-F_0(z^k)}{m_p}+\dfrac{F_p+F_0(z^k)}{m_p}, \\ V_p^0 &= V_0. \end{cases} \tag{6.12}$$

it appears that this approach is identical to one in which the plunger velocity for the boundary conditions at the next time-step were obtained straight from the previous velocity field and pressure field

$$V_p(t+\Delta t)=V_p(t)+\Delta t\,\frac{F_g(t)+F_p}{m_p}. \tag{6.13}$$

The boundary conditions (3.13), (3.14) for the next stationary Stokes problem should use $V_p(t+\Delta t)$. We omit further discussion of (6.13).



(a) As a function of position

(b) As a function of time

Figure 6.1: Velocity of the plunger obtained using implicit scheme.

Now consider the implicit Euler method instead

$$\begin{cases} \dfrac{1}{2}\dfrac{V_p^{k+1\,2}-V_p^{k\,2}}{z^{k+1}-z^k} &= V_p^{k+1}\,\dfrac{F_g(z^{k+1};1)-F_0(z^{k+1})}{m_p}+\dfrac{F_p+F_0(z^{k+1})}{m_p}, \\ V_p^0 &= V_0. \end{cases} \tag{6.14}$$

Although (6.14) is implicit, we just have a quadratic equation for V_p^{k+1}, which can be solved trivially. The result is in Figure 6.1a. We clearly have a stable calculation now. The velocity of the plunger in Figure 6.1a is a function of z. In order to obtain the velocity as a function of t the following approximation can be used

$$\begin{cases} z^{k+1} &= z^k + \Delta t^k \, V_p(z^k), \\ t^{k+1} &= t^k + \Delta t^k, \end{cases} \tag{6.15}$$

where $t^0 = 0$. The final graph is depicted in Figure 6.1b.

References

[1] G.K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge 1987.

[2] P.D. Howell, *Models for Thin Viscous Sheets*, Euro. Jnl of Applied Mathematics 7(1996), pp. 321-343.

[3] C.E. Humphreys, D.M. Burley, M. Cable, *Problems Arising in the Computation of the Flow of Molten Glass During a Pressing Operation*, Zeitschrift für angewandte Mathematik und Mechanik, vol. 76 (1996), 309-312.

[4] K. Laevsky, R.M.M. Mattheij, *Mathematical Modelling of Some Glass Problems*, in: Complex Flows in Industrial Processes (A.Fasano. ed) Birkhäuser, Boston (2000), 191-214.

[5] K. Laevsky, B.J. van der Linden, R.M.M. Mattheij, *Flow and Heat Tarnsfer in Pressing of Glass Products*, in: Computational Mathematics driven by Iindustrial problems (V. Capasso et al. eds.), Springer, Berlin (2000), 267-286

[6] R.M.M. Mattheij, J. Molenaar, *Ordinary Differential Equations in Theory and Practice*, SIAM, Pladelphia, 2002.

[7] S.W. Rienstra, T.D. Chandra, *Analytical Approximations to the Viscous Glass-flow Problem in the Mould-plunger Pressing Process, Including an Investigation of Boundary Conditions*, to appear Journal of Engineering Mathematics, 2001.

Adaptive Numerical Methods for Sensitivity Analysis of Differential-Algebraic Equations and Partial Differential Equations*

Linda Petzold† Yang Cao‡ Shengtai Li§ Radu Serban¶

Abstract

Sensitivity analysis generates essential information for design optimization, parameter estimation, optimal control, model reduction, process sensitivity and experimental design. Recent work on methods and software for sensitivity analysis of DAE and PDE systems has demonstrated that forward sensitivities can be computed reliably and efficiently. However, for problems which require the sensitivities with respect to a large number of parameters, the forward sensitivity approach is intractable and the adjoint (reverse) method is advantageous. Unfortunately, the adjoint problem is quite a bit more complicated both to pose and to solve. Our goal for both DAE and PDE systems has been the development of methods and software in which generation and solution of the adjoint sensitivity system are transparent to the user. This has been largely achieved for DAE systems. We propose a solution to this problem for PDE systems solved with adaptive mesh refinement.

1 Introduction

In recent years, there has been a growing interest in sensitivity analysis for large-scale systems governed by both differential algebraic equations (DAEs) and partial differential equations (PDEs). The results of sensitivity analysis have wide-ranging applications in science and engineering, including optimization, parameter estimation, model simplification, data assimilation, optimal control, uncertainty analysis and experimental design.

This paper has two parts. In the first part, we will outline the basic problem of sensitivity analysis for DAE systems and examine the recent results on numerical methods

*This work was supported by grants: DOE DE-FG03-00ER25430, NSF/ITR ACI-0086061, and NSF/KDI ATM-9873133.

†Department of Computer Science, University of California, Santa Barbara, California.

‡Department of Computer Science, University of California, Santa Barbara, California.

§Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico.

¶Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California.

and software for DAE sensitivity analysis based on the forward and adjoint methods. The second part of the paper will deal with sensitivity analysis for time-dependent PDE systems solved by adaptive mesh refinement (AMR).

2 Sensitivity Analysis for DAE Systems

Recent work on methods and software for sensitivity analysis of DAE systems [12, 26, 23, 24, 27] has demonstrated that forward sensitivities can be computed reliably and efficiently via automatic differentiation[6] in combination with DAE solution techniques designed to exploit the structure of the sensitivity system. The DASPK3.0[23, 24] software package was developed for forward sensitivity analysis of DAE systems with index up to two[7, 3], and has been used in sensitivity analysis and design optimization of several large-scale engineering problems[19, 30]. DASPK3.0 is an extension of the DASPK software [8, 7] developed by Brown, Hindmarsh and Petzold for the solution of large-scale DAE systems. For a DAE depending on parameters,

$$\begin{cases} F(x, \dot{x}, t, p) &= 0 \\ x(0) &= x_0(p), \end{cases} \tag{1}$$

these problems take the form: find dx/dp_j at time T, for $j = 1, \ldots, n_p$. Their solution requires the simultaneous solution of the original DAE system with the n_p sensitivity systems obtained by differentiating the original DAE with respect to each parameter in turn. For large systems this may look like a lot of work but it can be done efficiently, if n_p is relatively small, by exploiting the fact that the sensitivity systems are linear and all share the same Jacobian matrices with the original system.

2.1 The Adjoint DAE

Some problems require the sensitivities with respect to a large number of parameters. For these problems, particularly if the number of state variables is also large, the forward sensitivity approach is intractable. These problems can often be handled more efficiently by the adjoint method [11]. In this approach, we are interested in calculating the sensitivity of an objective function

$$G(x, p) = \int_0^T g(x, t, p) dt, \tag{2}$$

or alternatively the sensitivity of a function $g(x, T, p)$ defined only at time T. The function g must be smooth enough that g_p and g_x exist and are bounded. While forward sensitivity analysis is best suited to the situation of finding the sensitivities of a potentially large number of solution variables with respect to a small number of parameters, reverse (adjoint) sensitivity analysis is best suited to the complementary situation of finding the sensitivity of a scalar (or small-dimensional) function of the solution with respect to a large number of parameters.

In [10] we derived the adjoint sensitivity system for DAEs of index up to two (Hessenberg) and investigated some of its fundamental properties. Here we summarize the main results.

The adjoint system for the DAE

$$F(t, x, \dot{x}, p) = 0$$

with respect to the derived function $G(x, p)$ (2) is given by

$$(\lambda^* F_{\dot{x}})' - \lambda^* F_x = -g_x, \tag{3}$$

where * denotes the transpose operator and prime denotes the total derivative with respect to t.

The adjoint system is solved backwards in time. For index-0 and index-1 DAE systems, the initial conditions for (3) are taken to be $\lambda^* F_{\dot{x}}|_{t=T} = 0$, and the sensitivities of $G(x, p)$ with respect to the parameters p are given by

$$\frac{dG}{dp} = \int_0^T (g_p - \lambda^* F_p)\, dt + (\lambda^* F_{\dot{x}})|_{t=0}(x_0)_p. \tag{4}$$

For Hessenberg index-2 DAE systems, the initial conditions are more complicated, and are described in detail along with an algorithm for their computation in [10].

For a scalar derived function $g(x, T, p)$, the corresponding adjoint DAE system is given by

$$(\lambda_T^* F_{\dot{x}})' - \lambda_T^* F_x = 0, \tag{5}$$

where λ_T denotes $\frac{\partial \lambda}{\partial T}$. For index-0 and index-1 DAE systems, the initial conditions $\lambda_T(T)$ for (5) satisfy $(\lambda_T^* F_{\dot{x}})|_{t=T} = [g_x - \lambda^* F_x]\,|_{t=T}$. We note that the initial condition $\lambda_T(T)$ is derived in such a way that the computation of $\lambda(t)$ can be avoided. This is the case also for index-2 DAE systems. The full algorithm for consistent initialization of the adjoint DAE system is given in [10]. The sensitivities of $g(x, T, p)$ with respect to the parameters p are given for index-0 and index-1 DAE systems by

$$\frac{dg}{dp} = (g_p - \lambda^* F_p)|_{t=T} - \int_0^T (\lambda_T^* F_p) + (\lambda_T^* F_{\dot{x}})|_{t=0}(x_0)_p. \tag{6}$$

Note that the values of both λ at $t = T$ and λ_T at $t = 0$ are required in (6). If $F_p \neq 0$, the transient value of λ_T is also needed. For an index-2 system, if the index-2 constraints depend on p explicitly, an additional term must be added to the sensitivity (6).

If the objective function is of the integral form $G(x, p)$ (2), it can be computed easily by adding a *quadrature variable*, which is equal to the value of the objective function, to the original DAE. For example, if the number of variables in the original DAEs is N, we append a variable x_{N+1} and equation

$$\dot{x}_{N+1} = g(x, t, p).$$

Then $G = x_{N+1}(x, T, p)$. In this way, we can transform any objective function in the integral form (2) into the scalar form $g(x, T, p)$. The quadrature variables can be calculated very efficiently [23] by a staggered method in DASPK3.0; they do not enter into the Jacobian matrix.

From [10] we know that for DAE systems of index up to two (Hessenberg), asymptotic numerical stability in solving the forward problem is preserved by the backward Euler method, but only (for fully-implicit DAE systems) if the discretization of the time derivative is performed 'conservatively', which corresponds to solving an *augmented adjoint* DAE system,

$$\begin{aligned} \dot{\bar{\lambda}} - F_x^* \lambda &= 0, \\ \bar{\lambda} - F_{\dot{x}}^* \lambda &= 0. \end{aligned} \tag{7}$$

It was shown in [10] that the system (7) with respect to $\bar{\lambda}$ preserves the stability of the original system. Note that the augmented system (7) is of (one) higher index than the original adjoint system (5). This is not a problem in the implementation since the newly high-index variables do not enter into the error estimate and it can be shown that basic DAE structures such as combinations of semi-explicit index-1 and Hessenberg index-2 are preserved under the augmentation. Also, the linear algebra is accomplished in such a way that the matrix needed is the transpose of that required for the original system. Thus there are no additional conditioning problems for the linear algebra due to the use of the augmented adjoint system.

2.2 DASPKADJOINT

We have written a new code called DASPKADJOINT which accomplishes the DAE solution along with adjoint sensitivity analysis. The code is described in detail in [22], along with some example problems. Much of the challenge in writing DASPKADJOINT was concerned with handling the complexities of formulation and solution of the adjoint sensitivity system while requiring as little additional information from the user as is mathematically necessary. Here we describe some of the details of the implementation.

In the adjoint system (5) and the sensitivity calculation (6), the derivatives F_x, $F_{\dot{x}}$ and F_p may depend on the state variables x, which are the solutions of the original DAEs. Ideally, the adjoint DAE (5) should be coupled with the original DAE and solved together as we did in the forward sensitivity method. However, in general it is not feasible to solve them together because the original DAE may be unstable when solved backward. Alternatively, it would be extremely inefficient to solve the original DAE forward any time we need the values of the state variables.

The implementation of the adjoint sensitivity method consists of three major steps. First, we must solve the original ODE/DAE forward to a specific output time T. Second, at time T, we compute the consistent initial conditions for the adjoint system. The consistent initial conditions must satisfy the boundary conditions of (3). Finally, we solve the adjoint system backward to the start point, and calculate the sensitivities.

With enough memory, we can store all of the necessary information about the state variables at each time step during the forward integration and then use it to obtain the values of the state variables by interpolation during the backward integration of the adjoint DAEs. For example, we can store x and $\dot{x}$ at each time step during the forward integration and reconstruct the solution at any time by cubic Hermite interpolation[1] during the backward integration. The memory requirements for this approach are proportional to the number of time steps and the dimension of the state variables x, and are unpredictable because the number of time steps varies with different options and error tolerances of the ODE/DAE solver.

To reduce the memory requirements and also make them predictable, we use a two-level checkpointing technique. First we set up a checkpoint after every fixed number of time steps during the forward integration of the original DAE. Then we recompute the forward information between two consecutive checkpoints during the backward integration by starting the forward integration from the checkpoint. This approach needs to store only the forward information at the checkpoints and at a fixed number of times between two checkpoints.

In the implementation we allocated a special buffer to communicate between the forward and backward integration. The buffer is used for two purposes: to store the necessary information to restart the forward integration at the checkpoints, and to store the state variables and derivatives at each time step between two checkpoints for reconstruction of the state variable solutions during the backward integration.

In order to obtain the fixed number of time steps between two consecutive checkpoints, the second forward integration should make exactly the same adaptive decisions as the first pass if it restarts from the checkpoint. Therefore, the information saved at each checkpoint should be enough that the integration can repeat itself. In the case of DASPK3.0, the necessary information includes the order and stepsize for the next time step, the coefficients of the BDF formula, the history information array of the previous k time steps, the Jacobian information at the current time, etc.. To avoid storing Jacobian data (which is much larger than other information) in the buffer, we enforce a reevaluation of the iteration matrix at each checkpoint during the first forward integration.

If the size of the buffer is specified, the maximum number of time steps allowed between two consecutive checkpoints and the maximum number of checkpoints allowed in the buffer can be easily determined. However, the total number of checkpoints is problem-dependent and unpredictable. It is possible that the number of checkpoints is also too large for some applications to be held in the buffer. We then write the data of the checkpoints from the buffer to a disk file and reuse the buffer again. Whenever we need the information on the disk file, we can access it from the disk. We assume that the disk is always large enough to hold the required information.

Another important issue is how to formulate the adjoint DAE and the initial conditions so that the user doesn't have to learn all about the adjoint method and derive these for themselves. The adjoint equations involve matrix-vector products from the left side

[1]We could of course consider basing the interpolant on the interpolating polynomial underlying the BDF formula, but this is more complicated, requires more storage, and it not as smooth.

(vector-matrix products). Although a matrix-vector product $F_x v$ can be approximated via a directional derivative finite difference method, it is difficult to evaluate the vector-matrix product vF_x directly via a finite difference method. The vector-matrix product vF_x can be written as a gradient of the function $vF(x)$ with respect to x. However, N evaluations of $vF(x)$ are required to calculate the gradient by a finite-difference method if we don't assume any sparsity in the Jacobian. Therefore, automatic differentiation (AD) is necessary to improve the computational efficiency. A forward mode AD tool cannot compute the vector-matrix products without evaluation of the full Jacobian. It has been shown [14] that an AD tool with reverse mode can evaluate the vector-Jacobian product as efficiently as a forward mode AD tool can evaluate the Jacobian-vector product. In our implementation with DASPK3.0, we use the AD tool TAMC [14] to calculate the vector-matrix products. Initialization of the adjoint DAE is quite a bit more complicated than in the ODE case. For details, see [9] and [10]. In general, one needs to be able to provide some information about the structure of the problem (i.e. which are the index-1 and index-2 variables and constraints).

3 Sensitivity Analysis for Time-Dependent PDE Systems

Sensitivity methods for steady-state PDE problems have been studied by many authors (see [1, 5, 15, 17, 18]). Here we outline some recent results on adjoint methods for general transient PDE systems. Although many of the results from the steady-state system can be readily extended to the time-dependent PDE system, the time-dependent system has some unique features that must be treated differently. For example, apart from the boundary conditions, we now have initial conditions that must be determined. Two important classical fields that make extensive use of sensitivity analysis are inverse heat-conduction problems [13] and shape design in aerodynamic optimization [29].

Given a parameter-dependent PDE system

$$F(t, u, u_t, u_x, u_{xx}, p) = 0, \tag{8}$$

and a vector of objective functions $G(x, u, p)$ that depend on u and p, the sensitivity problem usually takes the form: find $\frac{dG}{dp}$, where p is a vector of parameters. By the chain rule, the sensitivity $\frac{dG}{dp}$ is given by

$$\frac{dG}{dp} = \frac{\partial G}{\partial u}\frac{\partial u}{\partial p} + \frac{\partial G}{\partial p}. \tag{9}$$

If we treat (8) as a nonlinear system about u and p, say $H(u, p) = F(t, u, u_t, u_x, u_{xx}, p)$, we have the following relationship

$$\frac{\partial H}{\partial u}\frac{\partial u}{\partial p} + \frac{\partial H}{\partial p} = 0.$$

Assuming that $\frac{\partial H}{\partial u}$ is boundedly invertible, the sensitivity $\frac{dG}{dp}$ is given by

$$\frac{dG}{dp} = -\frac{\partial G}{\partial u}\left(\frac{\partial H}{\partial u}\right)^{-1}\frac{\partial H}{\partial p} + \frac{\partial G}{\partial p}. \tag{10}$$

There are two basic methods to calculate $\frac{dG}{dp}$ in (10): forward and adjoint. The forward methods calculate $\frac{\partial u}{\partial p} = \left(\frac{\partial H}{\partial u}\right)^{-1}\frac{\partial H}{\partial p}$ first, which is the solution of the sensitivity PDE for each uncertain parameter. The adjoint methods, however, compute $\frac{\partial G}{\partial u}\left(\frac{\partial H}{\partial u}\right)^{-1}$ first, which is the solution of the adjoint PDE. The sensitivity and adjoint PDEs will be defined later. Although both methods yield the same analytical sensitivities, the computational efficiency may be quite different, depending on the number of objective functions (dimension of G) and the number of sensitivity parameters (dimension of p). The forward method is attractive when there are relatively few parameters or a large number of objective functions, while the adjoint method is more efficient for problems involving a large number of sensitivity parameters and few objective functions. We have studied the forward method in [25] and shown how it is possible to make use of the methods and software for sensitivity analysis of DAEs in combination with an adaptive mesh refinement algorithm for PDEs. In [21], we have studied extensively the adjoint method for PDEs solved with adaptive mesh refinement. Those results are outlined in what follows.

Two approaches can be taken for each method. In the first, called the *discrete* approach, we approximate the PDE by a discrete nonlinear system and then differentiate the discrete system with respect to the parameters. The discrete approach is easy to implement with the help of *automatic differentiation* tools [6, 14]. However, when the mesh is solution or parameter dependent (e.g., for an *adaptive* mesh or moving boundary), or a nonlinear discretization scheme (e.g., upwinding) is used, the discrete approach may not be computationally effective.

It is well-known that the method of lines (MOL) can transform a PDE system into an ODE or DAE system by spatial discretization. Thus the sensitivity calculation methods in [10] can be used if the semi-discretized PDE is obtained. However, we have observed that the *adjoint of the discretization* (AD) may not be consistent with a PDE, and the adjoint variables are not smooth on an adaptive grid. Therefore, if the adaptive region is changing with time (e.g. in adaptive mesh refinement (AMR) [4, 25]), the interpolation for the adjoint variables between different grids will introduce large errors. The AD method cannot be used in this case. On the other hand, one can show [21] that for linear discretization methods applied on a fixed grid with appropriate treatment of the boundary conditions, the sensitivities generated are accurate except in a small boundary layer.

In the second, called the *continuous* approach, we differentiate the PDE with respect to the parameters first and then discretize the sensitivity or adjoint PDEs to compute the approximate sensitivities. The system resulting from the continuous approach is usually much simpler than that from the discrete approach, and is naturally consistent with the adjoint PDE system. Therefore, the adaptive grid method and interpolation can be used

without difficulties. Derivation of the adjoint PDE could be handled by symbolic methods such as MAPLE. However it is very difficult to formulate proper boundary conditions for the adjoint of a general PDE system, and to the best of our knowledge an algorithm for generating the boundary conditions does not exist for a general PDE system. Moreover, the adjoint system may become inadmissible for some objective functionals (see [1, 2]), where the boundary conditions (or initial conditions) for the adjoint PDE system cannot be formulated properly. The discrete approach does not have such difficulties.

We propose an approach to combine the AD method and the *discretization of the adjoint* (DA) method in an efficient manner so that it can be used with AMR. The new approach (called the ADDA method) not only solves the problem for AD on the adaptive grid, it also solves the inadmissibility problem for DA. Both the AD and DA methods are used in this new approach but are applied in different regions.

We developed the ADDA method based on an observation that the discretization from the AD method is consistent with the adjoint PDE (hence it can be replaced with the discretization of the DA method) at the internal points if the mesh and the (linear) discretization are uniform everywhere except at the boundaries. The basic idea of the ADDA method is illustrated in Fig. 1.

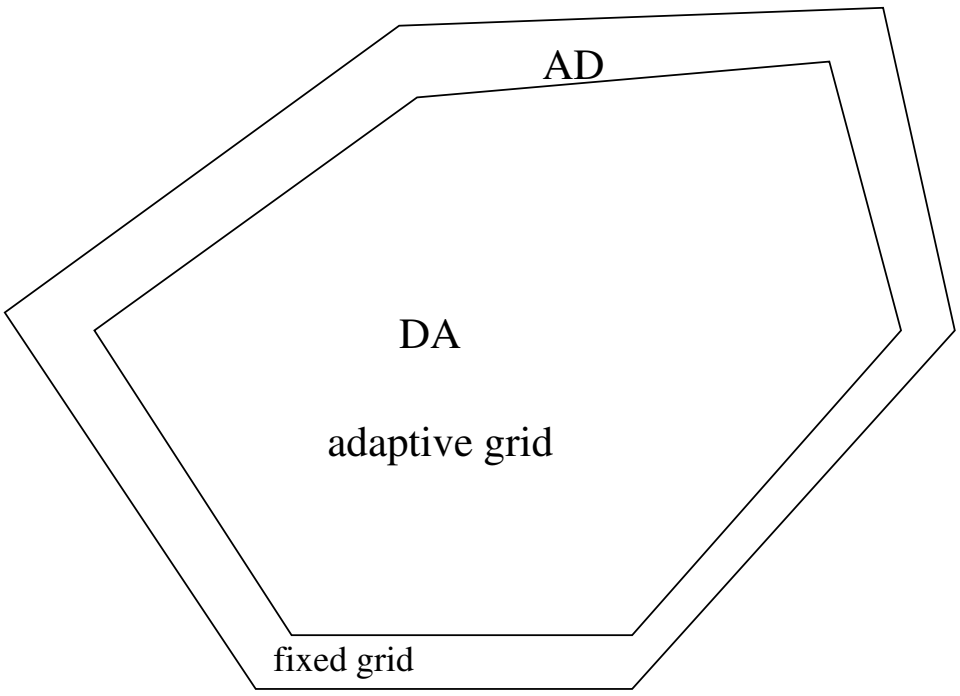


Figure 1: Diagram of the ADDA method

The results of the ADDA method should be equal or close to the results of the AD method on a nonadaptive fine grid. Given a reference nonadaptive fine grid, we first split the whole domain into two zones: boundary buffer zone and internal zone (see

Fig. 1). The boundary buffer zone consists of the boundary points and points that use the boundary points in their discretization. The remainder of the points belong to the internal zone. Since the discretization of the AD method may not be consistent with the adjoint PDE at or near the boundaries, the buffer zone is fixed and never adapted during the entire time integration. In the internal zone, the discretization from the AD method can be replaced with the discretization from the DA method if we assume that the discretization from the AD method is consistent with the adjoint PDE. It turns out [21] that this assumption is not always true for a general discretization and grid. However, if the mesh and discretization of the forward problem are uniform in the internal zone, the adjoint of the discretization is indeed consistent with the adjoint PDE.

After the discretization in the internal zone has been replaced by that from the DA method, the mesh can be adapted to achieve efficiency without loss of the accuracy. The adaptive mesh refinement in the internal zone is invisible to the AD method, which expects that the discretizations for the adjoint system are generated by the AD method on a nonadaptive fine grid. Instead the discretization is accomplished efficiently by the DA method on an adaptive grid. The internal zone looks like a black box to the AD method.

Since the sensitivity calculation is based on the AD method, the initial conditions for the adjoint system must be generated by the AD method. However, the initial conditions generated by the AD method may involve the grid spacing information, due to the objective functional evaluation [21]. A variable transformation [21] is used to eliminate the grid spacing information related to the integration scheme in the objective function evaluation. However, it cannot eliminate the grid spacing information related to the integrand function.

Strictly speaking, the values of the adjoint variables are different on different grids. That is why the sensitivity calculation by the AD method must be performed on a fixed mesh. The initial given mesh, which is the last mesh generated at $t = T$ in the forward adaptive method, may not be the same as the reference nonadaptive fine mesh we seek. Therefore, we must calculate the initial conditions for the ADDA method on the reference mesh first and then project them onto the initial given mesh by interpolation.

The overall algorithm of the ADDA method is as follows: First we obtain the initial conditions for the adjoint system by the AD method on a virtual nonadaptive fine grid. Then we transform and project them to the adaptive grid with a fixed boundary buffer zone. We assume that the discretization has been chosen so that AD is consistent with the adjoint PDE internally. Then the spatial discretization in the boundary buffer zone is generated by the AD method via automatic differentiation, and the discretization in the internal zone is defined by discretization of the adjoint PDE. Finally, an ODE or DAE time solver is used to advance the solution to the next time step. After the adjoint variables have been computed, the sensitivity evaluations of the AD method are used to calculate the sensitivities.

Examples are presented in [21] which demonstrate the effectiveness of the ADDA method.

References

[1] W. K. Anderson and V. Venkatakrishnan, *Aerodynamic design optimization on unstructured grid with a continuous adjoint formulation*, AIAA 97-0643, 35'th Aerospace Science meeting & exhibit, (1997).

[2] E. Arian and M. Salas, *Admitting the inadmissible: adjoint formulation for incomplete cost functionals in aerodynamic optimization*, ICASE Report No. 97-69, 1997.

[3] U. M. Ascher and L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, 1998.

[4] M. J. Berger and J. Oliger, *Adaptive mesh refinement for hyperbolic partial differential equations*, J. Comput. Phys. 53 (1984), 484-512.

[5] J. Borggaard and J. Burns, *A PDE sensitivity equation method for optimal aerodynamic design*, J. Comp. Phys., 136 (1997), 366-384.

[6] C. Bischof, A. Carle, G. Corliss, A. Griewank and P. Hovland, *ADIFOR - Generating derivative codes from Fortran programs*, Scientific Programming 1 (1992), 11-29.

[7] K. E. Brenan, S. L. Campbell and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Second edition, SIAM, 1996.

[8] P. N. Brown, A. C. Hindmarsh and L. R. Petzold, *Using Krylov methods in the solution of large-scale differential-algebraic systems*, SIAM J. Sci. Comput. (1994), 1467-1488.

[9] Y. Cao, S. Li, L. Petzold and R. Serban *Adjoint Sensitivity Analysis for Differential-Algebraic Equations: The Adjoint DAE System and its Numerical Solution*, SIAM J. Sci. Comput. 24(3) (2003), 1076-1089,

[10] Y. Cao, S. Li and L. Petzold, *Adjoint sensitivity analysis for differential-algebraic equations: algorithms and software*, J. Comp. Appl. Math. 149 (2002), 171-192.

[11] R. M. Errico, *What is an adjoint model?*, Bulletin of the American Meteorological Society 78 (1997), 2577-2591.

[12] W. F. Feehery, J. E. Tolsma and P. I. Barton, *Efficient sensitivity analysis of large-scale differential-algebraic systems*, Applied Numerical Mathematics, 25 (1997) 41-54.

[13] Y. Jarny, M.N. Ozisik and J.P. Bardon, *A general optimization method using adjoint equation for solving multidimensional inverse heat conduction*, Int. J. Heat Mass. Transfer 34 (1991) 2911-2919.

[14] R. Giering and T. Kaminski, *Recipes for adjoint code construction*, ACM Trans. Math. Software 24 (1998), 437-474.

[15] M. B. Giles and N.A. Pierce, *Adjoint equations in CFD: duality, boundary conditions and solution behaviour*. AIAA Paper 97-1850, (1997).

[16] M. B. Giles and N.A. Pierce, *An introduction to the adjoint approach to design*, ERCOFTAC Workshop on Adjoint Methods, Toulouse, June 21-23, 1999.

[17] O. Ghattas and J.-H. Bark, *Optimal control of two and three-dimensional incompressible Navier-Stokes Flows*, J. Comp. Phys., 136 (1997), 231-244.

[18] A. Jameson, *Aerodynamic Design via Control Theory,* J. of Scientific Computing, 3 (1988) 233-260.

[19] D. Knapp, V. Barocas, K. Yoo, L. Petzold and R. Tranquillo, *Rheology of reconstituted type I collagen gel in confined compression*, J. Rheology 41 (1997), 971-993.

[20] R. M. Lewis, *Numerical computation of sensitivities and the adjoint approach*, ICASE Technical Report No. 97-61, 1997.

[21] S. Li and L. Petzold, *Adjoint Sensitivity Analysis for Time-Dependent Partial Differential Equations with Adaptive Mesh Refinement*, submitted, J. Comp. Phys.

[22] S. Li and L. Petzold, *Description of DASPKADJOINT: An Adjoint Sensitivity Solver for Differential-Algebraic Equations*, UCSB Technical Report, www.engineering.ucsb.edu/ cse.

[23] S. Li and L. Petzold, *Software and algorithms for sensitivity analysis of large-scale differential-algebraic systems*, to appear, J. Comp. and Appl. Math.

[24] S. Li and L. Petzold, *Design of New DASPK for Sensitivity Analysis*, Technical Report, Dept. of Computer Science, UCSB, 1999.

[25] S. Li, L. Petzold and J. Hyman, *Solution adapted nested grid refinement and sensitivity analysis for parabolic partial differential equations*, to appear in special volume of Lecture Notes in Computational Science and Engineering, Springer, 2001.

[26] S. Li, L. Petzold and W. Zhu, *Sensitivity analysis of differential-algebraic equations: A comparison of methods on a special problem*, Applied Numerical Mathematics 32 (2000), 161-174.

[27] T. Maly and L. R. Petzold, *Numerical methods and software for sensitivity analysis of differential-algebraic systems*, Applied Numerical Mathematics, 20 (1997), 57-79.

[28] G. I. Marchuk, V. I. Agoshkov and V. P. Shutyaev, *Adjoint equations and perturbation algorithms*, CRC Press, Boca Raton, Fl, 1996.

[29] S. K. Nadarajah and A. Jameson, *A comparison of the continuous and discrete adjoint approach to automatic aerodynamic optimization*, AIAA paper 00-0067, 2000.

[30] L. L. Raja, R. J. Kee, R. Serban and L. Petzold, *Dynamic optimization of chemically reacting stagnation flows*, Proc. Electrochemical Society, 1998.

[31] A. Sei and W. W. Symes, *A note on consistency and adjointness for numerical schemes*, Tech. Report TR95-04, Department of Computational and Applied Math., Rice University, 1995.

An Implicit-Explicit Runge-Kutta-Chebyshev Scheme for Diffusion-Reaction Equations

J.G. Verwer and B.P. Sommeijer

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

jan.verwer@cwi.nl (www.cwi.nl/∼janv), ben.sommeijer@cwi.nl

Abstract

This preprint deals with the numerical time integration of diffusion-reaction problems with highly stiff reaction terms. An implicit-explicit (IMEX) extension of the explicit Runge-Kutta-Chebyshev (RKC) scheme is proposed. With respect to stability, the explicit scheme can be positioned in between common explicit and implicit Runge-Kutta schemes. RKC is explicit and thus avoids algebraic system solutions. It does however possess extended real stability intervals with a length proportional to s^2, where s is the number of stages. This implies that the scaled stability interval length, which takes into account the work load per time step, increases linearly with s, rendering RKC an attractive, user-friendly scheme for integrating large-scale semi-discrete parabolic problems. In case of severe stiffness RKC will become inefficient since then a very large number of stages will be needed for reasonable step sizes. By treating the reaction terms implicitly, in this preprint this restriction is removed for diffusion-reaction problems for which severe stiffness emanates from the reaction terms and the reaction Jacobian has a real spectrum.

1 Introduction

This preprint deals with the numerical time integration of parabolic partial differential equations, in particular diffusion-reaction problems with highly stiff reaction terms. We adopt the method of lines approach, thus assuming that the PDE problem including its boundary conditions has already been spatially discretized to a semi-discrete problem on a chosen space grid. This semi-discrete problem, being an initial value problem for a system of ordinary differential equations (ODEs), is denoted as

$$w'(t) = F_D(t, w(t)) + F_R(t, w(t)) \,, \quad t > 0 \,, \qquad w(0) = w_0 \,, \tag{1.1}$$

where F_D represents the semi-discrete diffusion operator and F_R contains the reaction terms. Typically, the dimension of this system is huge, especially for multi-space dimensional PDEs (number of PDE components times number of grid cells) and often this system is nonlinear and stiff. The stiffness rules out easy-to-use standard explicit solvers and the huge dimension with the nonlinearity complicates the use of implicit solvers.

In this preprint we propose an implicit-explicit (IMEX) extension of the explicit Runge-Kutta-Chebyshev (RKC) scheme. This scheme has been designed by van der Houwen & Sommeijer [10] for the numerical time integration of parabolic PDEs. With respect to stability, this scheme can be positioned in between common explicit and implicit schemes. RKC is an explicit Runge-Kutta scheme and thus avoids algebraic system solutions. It does however possess *extended real stability intervals* with a length proportional to s^2, where s is the number of stages. This quadratic dependence is derived from the first kind Chebyshev polynomial. The quadratic dependence is very attractive, since it means that the scaled stability interval length, which takes into account the work load per time step (the number of

stages), increases linearly with s. Therefore RKC is an attractive, user-friendly scheme for integrating large-scale semi-discrete parabolic problems. However, in case of severe stiffness, RKC will of course become inefficient since then a very large number of stages will be needed to achieve stability with reasonable step sizes. In such situations the use of an implicit, unconditionally stable scheme is advocated.

The IMEX extension proposed in this prepint is meant for problems (1.1) with a severely stiff reaction function $F_R(t, w(t))$ and a moderately stiff diffusion function $F_D(t, w(t))$. This extension thus treats the diffusion function $F_D(t, w(t))$ still explicitly and the reaction function $F_R(t, w(t))$ implicitly. With a zero reaction term the original RKC scheme is recovered so that the IMEX extension maintains the attractive feature of the explicit scheme that no algebraic system solutions are required, except those of small dimension (number of PDE components) coming from the reaction function. These small sized algebraic systems can be dealt with by the classic solution approach based on modified Newton iteration and standard LU-decomposition. Note that they are decoupled over the grid and hence the reaction computation can be easily parallelized, as is the case for the explicit diffusion computation. Furthermore, the IMEX extension maintains the recursive Chebyshev nature such that we have stability for the testmodel

$$w'(t) = \lambda_D w(t) + \lambda_R w(t)\,,$$

for *all* real non-positive λ_D and λ_R, as long as $\tau\lambda_D$ lies in the original real stability interval (τ is here the step size). In this sense the IMEX scheme is unconditionally stable for the reaction part, assuming real eigenvalues.

In Section 2 we review the explicit RKC scheme from [10]. The construction of the new IMEX scheme is discussed in Section 3. This new scheme is numerically illustrated in Section 4 for a highly stiff, nonlinear radiation-diffusion problem. We conclude with Section 5 which collects some final remarks and conclusions.

2 The explicit RKC scheme

In this section we review the explicit RKC scheme from [10] in order to prepare the construction of the IMEX scheme. We here closely follow Ch. V of [11] where also more details and references to earlier and additional work and related methods can be found.

2.1 The first-order scheme

To avoid too many technicalities in the beginning, we will start with the most simple form (first-order and undamped). Let T_s be the first kind Chebyshev polynomial $T_s(x) = \cos\big(s \arccos(x)\big)$ of degree s with $x \in [-1, 1]$ and consider the shifted Chebyshev polynomial

$$P_s(z) \;=\; T_s\Big(1 + \frac{z}{s^2}\Big) \quad \text{for} \quad z \in [-2s^2, 0]\,.$$

This polynomial satisfies $|P_s(z)| \le 1$ for $z \in [-2s^2, 0]$ and approximates e^z up to order z^2 for $z \to 0$, i.e., $e^z = P(z) + \mathcal{O}(z^2)$. Consequently, any s-stage first-order consistent explicit Runge-Kutta scheme for systems $w'(t) = F(t, w(t))$,

$$\begin{aligned} W_0 \;&=\; w_n\,, \\ W_j \;&=\; w_n + \tau \sum_{k=0}^{j-1} \alpha_{jk}\, F(t_n + c_k\tau, W_k)\,, \qquad j = 1, \ldots, s\,, \\ w_{n+1} \;&=\; W_s\,, \end{aligned} \tag{2.1}$$

giving the recursion $w_{n+1} = P_s(z)w_n, z = \tau\lambda$, when applied to the stability test equation $w'(t) = \lambda w(t), \lambda \in \mathbb{C}$, has P_s as stability function and $[-\beta, 0]$ as real stability interval with real stability boundary $\beta = 2s^2$. This boundary is optimal, that is, for any consistent scheme (2.1) we have $\beta \le 2s^2$.

The quadratic dependence implies that the scaled boundary β/s which takes into account the number of function evaluations per time step, linearly increases with s and hence for problems with large negative eigenvalues (semi-discrete parabolic PDEs) it may pay to use s-stage schemes (2.1) having P_s as stability function with s large. Within class (2.1) one can conceive different schemes giving P_s as stability function. However, if s and β get large, internal stability (round-off accumulation over the stages within a single step) must be taken into account in addition to the common step-by-step stability governed by P_s. Without internal stability the range of applicable values of s is too limited [9].

The first-order RKC scheme for nonlinear systems $w'(t) = F(t, w(t))$ is internally stable and is constructed as follows. First, all functions $P_j(z), 0 \le j \le s$, satisfying $W_j = P_j(z)w_n$ at the internal stages, similar to $w_{n+1} = P_s(z)w_n$, are supposed to be given by the shifted, first kind Chebyshev polynomial

$$P_j(z) \;=\; T_j\Big(1 + \frac{z}{s^2}\Big)\,. \tag{2.2}$$

Second, these functions are retrieved from the three-term Chebyshev recursion

$$\begin{aligned} &T_0(x) = 1\,, \quad T_1(x) = x\,,\\ &T_j(x) = 2x\,T_{j-1}(x) - T_{j-2}(x)\,, \qquad j = 2, 3, \ldots, s\,, \end{aligned}$$

where arguments may be complex-valued, giving

$$\begin{aligned} &P_0(z) = 1\,, \quad P_1(z) = 1 + \frac{1}{s^2}\,z\,,\\ &P_j(z) = 2P_{j-1}(z) - P_{j-2}(z) + \frac{2}{s^2}\,P_{j-1}(z)z\,, \qquad j = 2, 3, \ldots, s\,. \end{aligned}$$

Third, for systems $w'(t) = F(t, w(t))$, the occurrence of $P_{j-1}(z)$ is associated with a stage value W_{j-1} and the occurrence of $P_{j-1}(z)z$ with $\tau F_j(t_n + c_{j-1}\tau, W_{j-1})$ (and other occurrences likewise). This gives the 1-st order RKC integration formula

$$\begin{aligned} W_0 \;&=\; w_n\,,\\ W_1 \;&=\; W_0 + \frac{\tau}{s^2}\,F(t_n, W_0)\,,\\ W_j \;&=\; 2W_{j-1} - W_{j-2} + \frac{2\tau}{s^2}\,F(t_n + c_{j-1}\tau, W_{j-1})\,, \qquad j = 2, \ldots, s\,,\\ w_{n+1} \;&=\; W_s\,. \end{aligned} \tag{2.3}$$

From (2.2) follows $P_j(z) = e^{c_j z} + \mathcal{O}(z^2)$ defining $c_j = j^2/s^2$.

This scheme obviously belongs to class (2.1) and it can be applied for any (practical) value of s without giving internal stability problems. We owe this to the three-term Chebyshev recursion [10, 18]. In actual application, first a step size τ is selected on the basis of accuracy considerations followed by an adjustment of the number of stages s to provide step-by-step stability. That means that for efficiency the smallest s is chosen such that at each integration step the heuristic stability condition

$$\tau\rho\left(F'(t_n, w_n)\right) \le \beta = 2s^2 \tag{2.4}$$

is satisfied, where ρ denotes the spectral radius and F' the Jacobian matrix which is assumed to have a negative real spectrum (and normal and constant for a rigorous L_2-analysis of stability and convergence [18]). Consequently, the RKC method is applied as an unconditionally stable scheme in the sense that no a priori restriction is laid on the step size τ. Of course, if the problem is excessively stiff leading to a huge spectral radius, the minimal value of s required to satisfy (2.4) may become too large for a feasible computation with this (stabilized) explicit scheme.

Remark 2.1 The real stability interval contains interior points $z \in (-\beta, 0)$ where $|P_s(z)| = 1$. Hence an imaginary perturbation on z might yield instability. For this reason, the P_j given by (2.2) are slightly damped [8] resulting in

$$P_j(z) \;=\; \frac{T_j(\omega_0 + \omega_1 z)}{T_j(\omega_0)}\,, \qquad \omega_1 \;=\; \frac{T_s(\omega_0)}{T_s'(\omega_0)}\,, \tag{2.5}$$

where $\omega_0 > 1$ is a parameter and ω_1 is chosen such that $P_s'(0) = 1$, implying first-order consistency. The real stability interval is determined by the relation $-\omega_0 \le \omega_0 + \omega_1 z \le \omega_0$, giving $\beta = 2\omega_0/\omega_1$. In the interior of the stability interval $P_s(z)$ now alternates between $T_s(\omega_0)^{-1}$ and $-T_s(\omega_0)^{-1}$. A convenient choice for ω_0 is $\omega_0 = 1 + \epsilon/s^2$ with ϵ a small positive number. Expanding at $\omega_0 = 1$ and using $T_s'(1) = s^2, T_s''(1) = \frac{1}{3}s^2(s^2-1)$ then shows $T_s(\omega_0) \approx 1 + \epsilon$ and

$$\beta \;=\; \frac{2\omega_0 T_s'(\omega_0)}{T_s(\omega_0)} \;\approx\; (2 - \tfrac{4}{3}\epsilon)s^2\,.$$

A suitable value for ϵ is 0.05. For practical problems this gives sufficient damping (approximately 5%) and it gives only a minor decrease of the stability boundary to approximately $1.93\,s^2$. Figure 2.1 (borrowed from [11]) illustrates the stability region $\mathcal{S} = \{z \in \mathbb{C} : |P_s(z)| \le 1\}$ for P_5 with and without damping. The effect of damping is that that at the interior of $[-\beta, 0]$ the boundary of $\mathcal{S}$ has no points on the real axis.

Finally, the damping leads to slightly different coefficients in (2.3); see Remark 2.3 and the general RKC formula (2.9) wherein (2.3) is contained. ◇

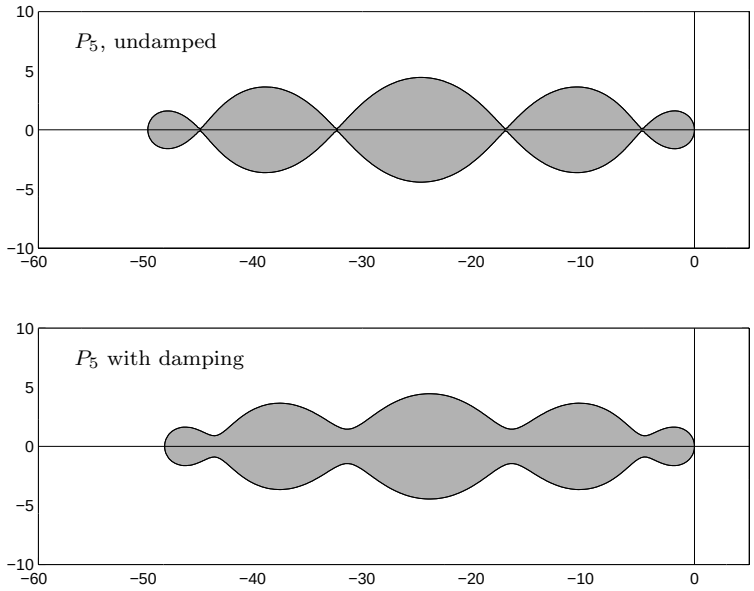


Figure 2.1: Stability region for the first-order shifted Chebyshev polynomial P_5.

2.2 The second-order scheme

In actual computation first-order consistency may be too low. Van der Houwen & Sommeijer [10] therefore have also constructed a second-order RKC scheme with

$$B_s(z) \;=\; \frac{2}{3} + \frac{1}{3s^2} + \Big(\frac{1}{3} - \frac{1}{3s^2}\Big)\,T_s\left(1 + \frac{3z}{s^2-1}\right) \tag{2.6}$$

as stability function which has

$$\beta \approx \tfrac{2}{3}\,(s^2-1)$$

as real stability boundary. This polynomial, due to [4], satisfies $e^z = B_s(z) + \mathcal{O}(z^3)$ and generates about 80% of the optimal stability interval for second-order polynomials, being $\beta \approx 0.814 s^2$. Within the interior of the stability interval $B_s(z)$ alternates between $\approx 1/3$ and 1.

Remark 2.2 For stabilized schemes of order p greater than or equal to two, stability functions with the largest possible real stability boundary are known to exist [15], but explicit analytical expressions like (2.6) are not available. However, there do exist accurate approximations to the optimal boundaries $\beta = c_p(s)s^2$ for $2 \le p \le 11$, see Section 2.5 of [1]. $\diamond$

The damped form of the stage polynomials $B_j, j = 0, \ldots, s$, reads

$$B_j(z) = a_j + b_j T_j(\omega_0 + \omega_1 z)\,, \qquad a_j = 1 - b_j T_j(\omega_0)\,, \tag{2.7}$$

where $\omega_0 = 1 + \epsilon/s^2$ as in (2.5), $\omega_1 = T_s'(\omega_0)/T_s''(\omega_0)$, and

$$b_j \;=\; T_j''(\omega_0)\,/\,(T_j'(\omega_0))^2\,, \qquad j = 2, \ldots, s\,. \tag{2.8}$$

The parameters b_0 and b_1 are still free. Here we put $b_0 = b_1 = b_2$ (for the IMEX extension another choice is made). Using $T_s'(1) = s^2, T_s''(1) = \frac{1}{3}s^2(s^2-1)$ and $T_s'''(1) = \frac{1}{15}s^2\,(s^2-1)\,(s^2-4)$, the boundary β of the damped stability function B_s can now be seen to satisfy

$$\beta \;\approx\; \frac{(\omega_0+1)T_s''(\omega_0)}{T_s'(\omega_0)} \;\approx\; \tfrac{2}{3}(s^2-1)\big(1 - \tfrac{2}{15}\epsilon\big)\,.$$

Taking $\epsilon = 2/13$, we get approximately 5% damping in the interior of the stability interval and a reduction in the stability boundary of about 2% compared to the undamped case. Figure 2.2 (borrowed from [11]) illustrates the stability region $\mathcal{S}$ of B_5 with and without damping.

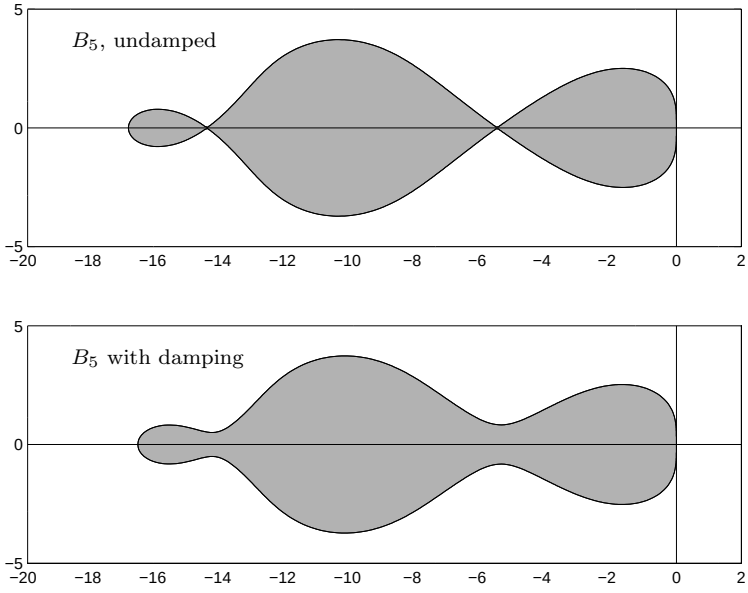


Figure 2.2: Stability region for the second-order shifted Chebyshev polynomials B_5.

The construction of the second-order integration formula for systems $w'(t) = F(t, w(t))$ is a bit more complicated then in the first-order case, but is basically identical:

$$\begin{aligned} W_0 &= w_n\,, \\ W_1 &= W_0 + \tilde{\mu}_1 \tau F_0\,, \\ W_j &= (1-\mu_j-\nu_j)W_0 + \mu_j W_{j-1} + \nu_j W_{j-2} + \tilde{\mu}_j \tau F_{j-1} + \tilde{\gamma}_j \tau F_0\,, \\ w_{n+1} &= W_s\,, \end{aligned} \tag{2.9}$$

where $j = 2, \ldots, s$ and F_k denotes $F(t_n + c_k\tau, W_k)$. Further,

$$\tilde{\mu}_1 \;=\; b_1\omega_1\,, \quad \mu_j \;=\; \frac{2b_j\omega_0}{b_{j-1}}\,, \quad \nu_j \;=\; \frac{-b_j}{b_{j-2}}\,, \quad \tilde{\mu}_j \;=\; \frac{2b_j\omega_1}{b_{j-1}}\,, \quad \tilde{\gamma}_j \;=\; -a_{j-1}\tilde{\mu}_j \tag{2.10}$$

and $c_0 = 0, c_1 = c_2/T_2'(\omega_0) = c_2/(4\omega_0)$,

$$c_j \;=\; \frac{T_s'(\omega_0)}{T_s''(\omega_0)}\,\frac{T_j''(\omega_0)}{T_j'(\omega_0)} \;\approx\; \frac{j^2-1}{s^2-1} \quad (2 \le j \le s-1)\,, \quad c_s \;=\; 1\,.$$

Remark 2.3 By replacing ω_1 by $\omega_1 = T_s(\omega_0)/T_s'(\omega_0)$ and (2.8) by

$$b_j = \frac{1}{T_j(\omega_0)}\,, \qquad j = 0, \ldots, s\,, \tag{2.11}$$

(2.9) becomes just the first-order consistent scheme based on the damped stage functions (2.5). We then have

$$c_0 = 0\,, \qquad c_j = \frac{T_s(\omega_0)}{T_s'(\omega_0)}\,\frac{T_j'(\omega_0)}{T_j(\omega_0)} \;\approx\; \frac{j^2}{s^2} \quad (1 \le j \le s-1)\,, \quad c_s \;=\; 1\,,$$

and $(1 - \mu_j - \nu_j) = 0$ and $\tilde{\gamma}_j = 0$. ◇

Remark 2.4 If the stability function of a Runge-Kutta scheme approximates the exponential e^z with order $p \le 2$, the scheme also has order $p \le 2$ for general problems $w'(t) = F(t, w(t))$. This greatly simplifies the construction of the RKC integration formulas. ◇

Remark 2.5 A variable stepsize code based on the second-order scheme has been developed in [17]. [1] This code also works with a variable amount of stages to minimize computational costs. For that purpose it has been equipped with a spectral radius estimator. In Section 4 the explicit code RKC will be numerically illustrated. ◇

Remark 2.6 Related stabilized explicit methods are the ROCK [3, 2] and DUMKA methods [12, 13]. These have close to optimal real stability boundaries and can possess a higher order (up to order 4). However, the formulas are not known in an explicit analytical form and are therefore less amenable for extension to an IMEX scheme. Numerical comparisons between the 2-nd order RKC code from [17] and a 4-th order ROCK code[2] can be found in [3, 11]. ◇

3 The implicit-explicit Runge-Kutta-Chebyshev scheme

In this section we will construct the IMEX-RKC scheme for the general nonlinear system (1.1).

3.1 The integration formula

For this system, the IMEX-Euler scheme that is obtained from modifying the first stage formula of (2.9) reads

$$W_1 = W_0 + \tilde{\mu}_1\tau F_D(t_n, W_0) + \tilde{\mu}_1\tau F_R(t_n + \tilde{\mu}_1\tau, W_1)\,, \qquad \tilde{\mu}_1 = b_1\omega_1\,, \tag{3.1}$$

where the reaction term is treated implicitly. All subsequent stages of the RKC method (2.9) are modified in a similar manner such that the recursive nature derived from the first kind Chebyshev polynomial is maintained.

Consider the scalar stability test equation

$$w'(t) = \lambda_D w(t) + \lambda_R w(t) \tag{3.2}$$

[1] See `ftp://ftp.cwi.nl/pub/bsom/rkc` or `http://www.netlib.org/ode/` for the source code.
[2] See `http://www.unige.ch/math/folks/hairer/software.html` for the source code.

with λ_D and λ_R standing for eigenvalues of (frozen) Jacobians $F'_D(t,w(t))$ and $F'_R(t,w(t))$, respectively. Applied to this test equation, (3.1) yields

$$W_1 = R_1(z_D, z_R)\, W_0\,, \qquad R_1(z_D, z_R) = \frac{1 + b_1\omega_1 z_D}{1 - b_1\omega_1 z_R}\,. \tag{3.3}$$

As we will see, it is convenient to impose

$$b_1 = \frac{1}{\omega_0}\,, \tag{3.4}$$

so that

$$R_1(z_D, z_R) = \frac{1 + \frac{\omega_1}{\omega_0} z_D}{1 - \frac{\omega_1}{\omega_0} z_R}\,. \tag{3.5}$$

Observe that the choice (3.4) for b_1 differs from the choice made in Section 2.2 beneath formula (2.8). Here we exploit the freedom we have for b_1 (like before, b_0 is still free too and is again set equal to b_2). This choice enables the

Ansatz 3.1 All stage functions $R_j(z_D, z_R), j = 0, 1, \ldots, s$, of the IMEX-RKC scheme are taken to be of the form

$$R_j(z_D, z_R) = a_j + b_j T_j\left(\frac{\omega_0 + \omega_1 z_D}{1 - \frac{\omega_1}{\omega_0} z_R}\right), \qquad a_j = 1 - b_j T_j(\omega_0) \tag{3.6}$$

with b_j copied from (2.7), so that for $z_R = 0$ the R_j reduce to the stage functions (2.7). Of importance is that the argument of the T_j is identical over the stages. $\diamond$

Thus the construction of the IMEX-RKC scheme is based on the rational function expression (3.6). First we write

$$T_j(x) = \frac{-a_j}{b_j} + \frac{R_j}{b_j}\,, \qquad x = \frac{\omega_0 + \omega_1 z_D}{1 - \frac{\omega_1}{\omega_0} z_R}\,,$$

where $R_j = R_j(z_D, z_R)$ and apply the recursion $T_j(x) = 2x\, T_{j-1}(x) - T_{j-2}(x)$. Inserting x gives

$$\begin{aligned} &R_j \cdot (1 - \frac{\omega_1}{\omega_0}\, z_R) = a_j\, (1 - \frac{\omega_1}{\omega_0}\, z_R) + 2\frac{b_j}{b_{j-1}}\, R_{j-1} \cdot (\omega_0 + \omega_1 z_D) - \\ &2\, \frac{b_j}{b_{j-1}}\, a_{j-1}\, (\omega_0 + \omega_1 z_D) + \frac{b_j}{b_{j-2}}\, a_{j-2}\, (1 - \frac{\omega_1}{\omega_0}\, z_R) - \frac{b_j}{b_{j-2}}\, R_{j-2} \cdot (1 - \frac{\omega_1}{\omega_0}\, z_R)\,. \end{aligned}$$

From this relation we can now deduce the IMEX integration scheme for system (1.1) by identifying the occurrence of R_j with W_j and $R_j z_R$ with $\tau F_R(t_n + c_j\tau, W_j)$ and a_j with $a_j W_0$, etc. Using the coefficient expressions (2.10) this gives

$$\begin{aligned} &W_j - \tilde{\mu}_1\tau\, F_{R,j} = (a_j - \mu_j a_{j-1} - \nu_j a_{j-2})\, W_0 + \mu_j W_{j-1} + \nu_j W_{j-2} + \\ &\tilde{\mu}_j\tau F_{D,j-1} + \tilde{\gamma}_j\tau F_{D,0} - \nu_j\tilde{\mu}_1\tau\, F_{R,j-2} - \tilde{\mu}_1\, (a_j - \nu_j a_{j-2})\tau\, F_{R,0}\,, \end{aligned}$$

where $F_{R,j} = F_R(t_n + c_j\tau, W_j)$, etc. Next, using

$$a_j - \mu_j a_{j-1} - \nu_j a_{j-2} = 1 - \nu_j - \mu_j\,,$$

we find the aimed IMEX-RKC integration scheme

$$\begin{aligned} W_0 &= w_n\,, \\ W_1 &= W_0 + \tilde{\mu}_1\tau F_{D,0} + \tilde{\mu}_1\tau F_{R,1}\,, \\ W_j &= (1 - \nu_j - \mu_j)\, W_0 + \mu_j W_{j-1} + \nu_j W_{j-2} + \tilde{\mu}_j\tau F_{D,j-1} + \tilde{\gamma}_j\tau F_{D,0} + \\ &\quad\ [\tilde{\gamma}_j - (1 - \nu_j - \mu_j)\, \tilde{\mu}_1]\tau\, F_{R,0} - \nu_j\tilde{\mu}_1\tau\, F_{R,j-2} + \tilde{\mu}_1\tau\, F_{R,j}\,, \\ w_{n+1} &= W_s\,, \end{aligned} \tag{3.7}$$

where $j = 2, \ldots, s$.

Remark 3.2 If F_R is absent, the explicit scheme (2.9) is recovered. For the diffusion operator F_D the IMEX scheme thus operates in the same way as the explicit scheme. The difference is that (3.7) is implicit in the stiff reaction operator F_R, requiring at each stage the solution of a system of non-linear algebraic equations

$$W_j - \tilde{\mu}_1 \tau F_R(t_n + c_j\tau, W_j) = V_j \,, \tag{3.8}$$

with V_j given and W_j as unknown vector. Because F_R has no underlying spatial grid connectivity, this system consists of a great number (the number of grid points) of decoupled small sized subsystems with dimension the number of coupled PDEs to be solved. Hence the modified Newton method can be used with a common LU-decomposition for the linear solves as is customary in the stiff ODE field. For efficiency reasons it could be advantageous that the coefficient $\tilde{\mu}_1$ is independent of j, since this could enable the use of LU-decompositions identical over the stages. ◇

Remark 3.3 In many diffusion-reaction applications one is interested in transient behaviour and in steady-state solutions w for autonomous problems

$$F_D(w) + F_R(w) = 0 \,.$$

Standard ODE integrators (Runge-Kutta and linear multistep methods) return steady states exactly. This property is shared by all stages of the current IMEX-RKC scheme (3.7). It takes an elementary calculation to prove this. Note that with operator splitting where the subsystems $w'(t) = F_D(w(t))$ and $w'(t) = F_R(w(t))$ are integrated completely decoupled within time steps (time splitting), steady states are not returned exactly. ◇

3.2 Stability properties

We consider (linear test model) stability for equation (3.2). The underlying assumption here, made for the sake of analysis, is that λ_D and λ_R stand for eigenvalues of frozen Jacobians $A_D = F_D'(t, w(t))$ and $A_R = F_R'(t, w(t))$, respectively, with A_D and A_R normal matrices which commute. They then have a common set of orthonormal eigenvectors implying that stability results in the L_2 sense [11] for the constant coefficient linear system $w'(t) = (A_D + A_R)w(t)$ can be retrieved from the scalar equation $w'(t) = (\lambda_D + \lambda_R)w(t)$. Additionally, we suppose that both λ_D and λ_R are real and non-positive and note that for many practical cases this imposes no restriction.

Thus, we will require stability for all possibles values (z_D, z_R) with

$$z_D \in [-\beta, 0] \quad \text{and} \quad z_R \le 0$$

for the IMEX-RKC stability function

$$R_s(z_D, z_R) = a_s + b_s T_s\left(\frac{\omega_0 + \omega_1 z_D}{1 - \frac{\omega_1}{\omega_0} z_R}\right) . \tag{3.9}$$

Because z_R is non-positive, implying

$$\left|\frac{\omega_0 + \omega_1 z_D}{1 - \frac{\omega_1}{\omega_0} z_R}\right| \le |\omega_0 + \omega_1 z_D| \,,$$

it follows trivially that $|R_s(z_D, z_R)| \le 1$ as long as $z_D \in [-\beta, 0]$. Hence with respect to the reaction part the IMEX-RKC scheme is unconditionally stable and the stability with respect to the diffusion part remains unchanged.

For $z_R \to -\infty$ (infinite reaction stiffness) the argument of T_s approaches zero. Hence for the IMEX scheme derived from the first-order formula having $a_s = 0$, it is advocated to choose s odd, giving $R_s(z_D, -\infty) = 0$ for all z_D. This gives optimal damping of stiff components from the reaction term. Likewise, for the IMEX scheme derived from the second-order formula it is advocated to choose s odd, giving $R_s(z_D, -\infty) \approx 2/3$, or s such that $T_s(0) = -1$, giving $R_s(z_D, -\infty) \approx 1/3$. For both cases this also would lead to a strong damping of stiff components from the reaction term.

3.3 Consistency properties

To see the change in consistency properties incurred by the IMEX extension, let us examine how the new stability functions $R_s(z_D, z_R)$ do approximate the exponential $e^z, z = z_D + z_R$, for $z \to 0$. Note that for first- and second-order Runge-Kutta methods the consistency properties of the stability function largely dictate the consistency properties for nonlinear problems, see also Remark 2.4.

First consider the IMEX scheme derived from the first-order explicit RKC formula. For simplicity of presentation we put $\omega_0 = 1$ (no damping). Then the argument x of T_s in (3.9) satisfies

$$x = \frac{1 + \omega_1 z_D}{1 - \omega_1 z_R} = 1 + \omega_1 \tilde{z}\,, \qquad \tilde{z} = \frac{z}{1 - \omega_1 z_R} \tag{3.10}$$

with $\omega_1 = 1/s^2$, so that (3.9) becomes

$$R_s(z_D, z_R) = T_s\left(1 + \frac{\tilde{z}}{s^2}\right).$$

Assuming s sufficiently large and letting $\tilde{z} \to 0$, we can now use a known expansion of T_s [11] giving

$$R_s(z_D, z_R) \approx 1 + \tilde{z} + \frac{1}{6}\,\tilde{z}^2 + \frac{1}{90}\tilde{z}^3 + \cdots.$$

It follows that

$$e^z - R_s(z_D, z_R) \approx \left(\frac{1}{3} - \frac{1}{s^2}\,\frac{z_R}{z}\right) z^2 + \cdots.$$

Compared to the explicit case, the leading term of the local error has become slightly smaller and this small difference vanishes with increasing number of stages.

Next consider the IMEX scheme derived from the second-order explicit RKC formula and assume again $\omega_0 = 1$ (no damping). The argument x of T_s in (3.9) then satisfies (3.10) with $\omega_1 = 3/(s^2 - 1)$ so that we can write

$$R_s(z_D, z_R) \;=\; \frac{2}{3} + \frac{1}{3s^2} + \left(\frac{1}{3} - \frac{1}{3s^2}\right) T_s\left(1 + \frac{3}{s^2 - 1}\,\tilde{z}\right).$$

With s sufficiently large and $\tilde{z} \to 0$ there holds [11]

$$R_s(z_D, z_R) \approx 1 + \tilde{z} + \frac{1}{2}\,\tilde{z}^2 + \frac{1}{10}\tilde{z}^3 + \cdots,$$

giving

$$e^z - R_s(z_D, z_R) \approx \frac{1}{15}z^3 - \frac{3}{s^2 - 1}z_R z + \cdots.$$

This result reveals a reduction of the order from two to one due to the IMEX extension. However, the new leading order term $3z_R z/(s^2 - 1)$ vanishes with increasing number of stages indicating that in actual application the effect of the order reduction will remain small.

4 Numerical results

We will numerically compare the new IMEX scheme (3.7) derived from the second-order explicit RKC scheme (2.9) with this explicit scheme. For that purpose the variable step size code RKC from [17] implementing this explicit scheme is used (see Remark 2.5). The comparison is based on a radiation-diffusion problem from [14]. The following description of this problem, the used spatial discretization, and part of the numerical results (those for the explicit scheme for $Z_0 = 1, 5$) were borrowed from Ch.V of [11].

4.1 A radiation-diffusion problem

The problem consists of two strongly nonlinear diffusion equations with a highly stiff reaction term (an idealization of non-equilibrium radiation diffusion in a material). The dependent variables are E and T, representing, respectively, radiation energy and material temperature. These problems are for instance found in laser fusion applications. The equations are defined on the unit square for $t > 0$,

$$\begin{aligned} E_t &= \nabla \cdot (D_1 \nabla E) + \sigma(T^4 - E)\,, \\ T_t &= \nabla \cdot (D_2 \nabla T) - \sigma(T^4 - E)\,, \end{aligned}$$

with

$$\sigma = \frac{Z^3}{T^3}\,, \qquad D_1 = \frac{1}{3\sigma + |\nabla E|/E}\,, \qquad D_2 = k\,T^{5/2}\,.$$

Here $|\nabla E|$ is the Euclidean norm and $Z = Z(x, y)$ represents the atomic mass number which may vary in the spatial domain to reflect inhomogeneities in the material. The temperature diffusion coefficient $k = 0.005$ and

$$Z(x,y) = \begin{cases} Z_0 & \text{if } |x - \frac{1}{2}| \le \frac{1}{6} \text{ and } |y - \frac{1}{2}| \le \frac{1}{6}\,, \\ 1 & \text{otherwise}\,. \end{cases}$$

The initial values are constant in space,

$$E(x, y, 0) = 10^{-5}\,, \qquad T(x, y, 0) = E(x, y, 0)^{1/4} \approx 5.62\,10^{-2}\,,$$

and the boundary conditions are

$$\begin{aligned} \tfrac{1}{4}E - \tfrac{1}{6\sigma}E_x &= 1 \qquad && \text{at } x = 0\,, \\ \tfrac{1}{4}E + \tfrac{1}{6\sigma}E_x &= 0 \qquad && \text{at } x = 1\,, \\ T_x &= 0 \qquad && \text{at } x = 0, 1\,, \end{aligned}$$

together with homogeneous Neumann conditions for E and T at $y = 0, 1$.

The solution consists of a steep (temperature) front moving to the right. For $Z_0 > 1$ the movement is hampered at the interior region with larger atomic mass number (and corresponding smaller diffusion). The radiation energy E is for the most part almost equal to T^4, except near the front where it is slightly larger with a steeper profile. Figure 4.1 shows contour levels and cross sections of an accurate reference solution of the radiation temperature $E^{1/4}$ and material temperature T at time $t = 3$ for $Z_0 = 10$. More illustrations for different values of the temperature diffusion coefficient ($k = 0, 0.1$) can be found in [14].

The spatial discretization has been performed on a uniform cell centered grid with grid size h by means of second-order central conservative discretization. This gives a semi-discrete system $w'(t) = F_D(w(t)) + F_R(w(t))$ of dimension $2/h^2$, for which the spectral radii of frozen Jacobians F_D', F_R' are estimated as

$$\rho_D = 8h^{-2}\,, \qquad \rho_R = 6000 Z_0^3\,, \tag{4.1}$$

assuming $1 \le Z(x, y) \le Z_0$. Note that we have at each grid point the nonlinear reaction system

$$f_R(E,T) = \begin{pmatrix} Z^3 T^{-3}(T^4 - E) \\ -Z^3 T^{-3}(T^4 - E) \end{pmatrix}\,, \qquad f_R'(E,T) = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix}\,,$$

with

$$\alpha = \frac{Z^3}{T^3}\,, \qquad \beta = Z^3\Big(1 + \frac{3E}{T^4}\Big)$$

and eigenvalues 0 and $-(\alpha + \beta)$. In the expression for $\alpha + \beta$ the term Z^3/T^3 will be the dominating one. Since we a priori know that $1/T^3 \lesssim 5.6\ 10^3$, we can estimate ρ_R as $6000Z^3$. Thus in total we get $\rho = \rho_D + \rho_R$ which is to be maximized over the spatial region. With increasing atomic mass number Z_0, ρ_R thus quickly becomes much larger than ρ_D for realistic grid sizes h. This is the kind of situation where the IMEX scheme will be significantly more efficient than its explicit counterpart.

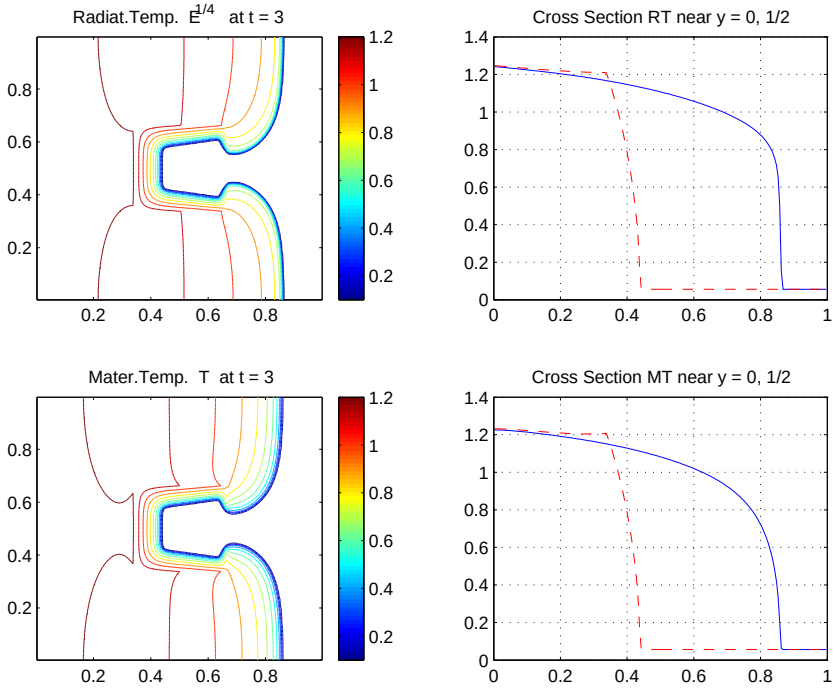


Figure 4.1: Contour levels and cross sections of the radiation temperature $E^{1/4}$ and material temperature T at time $t = 3$ for $Z_0 = 10$. Contour levels: $0.1, 0.2, \ldots, 1.2$.

4.2 Test results for the explicit code RKC

The code RKC [17] works as most other variable step size ODE codes. A difference is that at each time step it minimizes the number of stages s so as to satisfy the stability condition $\tau\rho \leq \beta \approx 0.65s^2$. Variable stepsizes are based on a local error per step criterion (which implies that if all is going well, this second-order code will reduce the numerical integration error by a factor of roughly 5 upon a tolerance reduction factor of 10 [16]). The code uses a tentative initial step size $\tau_0 = 1/\rho$ that is on scale with the dynamics at $t = 0$.

In Table 4.1 temporal L_2-errors are listed for $t = 3$ with various tolerances on 50×50 and 100×100 grids for $Z_0 = 1, 5, 10$. These errors were obtained by comparison with an accurate reference solution. Also given are estimated spatial L_2-errors (obtained by comparison on grids with twice as many grid points in both spatial directions). From the tables we see that with a decreasing local error tolerance *Tol*, the temporal errors quickly become insignificant in comparison to the spatial errors. Hence further decreasing *Tol* makes no sense and for this problem the code thus should work reliably for crude tolerances.

RKC solves the problem in all test cases reliably. However, with respect to efficiency we find the results satisfactory only for $Z_0 = 1$ where ρ_D still dominates. For $Z_0 > 1$ the reaction problem becomes increasingly stiff leading to very high stage numbers s and thus high costs. In this situation the IMEX scheme is expected to do a much better job. Observe that the integration behaviour is more or less independent of the increasing stiffness imposed by Z_0. Also observe that on the finer grid more time steps are used compared to the coarser grid. On the finer grid the front is better resolved, which presumably also steepens up the temporal solution requiring more time steps. The relatively large number of step rejections for the smallest $Tol = 10^{-3}$ is odd; as yet we have no explanation for it.

$Z_0 = 1$	$h = 1/50$	$err_{2,s} = 3.0\ 10^{-2}$	$h = 1/100$	$err_{2,s} = 8.5\ 10^{-3}$
Tol	$err_{2,t}$	Costs	$err_{2,t}$	Costs
10^{-1}	$2.3\ 10^{-2}$	2175 (36+2, 82)	$7.4\ 10^{-3}$	5207 (52+7, 122)
10^{-2}	$3.6\ 10^{-3}$	3020 (68+3, 101)	$3.0\ 10^{-3}$	6393 (101+2 , 78)
10^{-3}	$1.3\ 10^{-3}$	5779 (180+33, 49)	$4.4\ 10^{-4}$	12484 (266+47, 54)
$Z_0 = 5$	$h = 1/50$	$err_{2,s} = 7.8\ 10^{-2}$	$h = 1/100$	$err2, s = 2.7\ 10^{-2}$
Tol	$err_{2,t}$	Costs	$err_{2,t}$	Costs
10^{-1}	$2.3\ 10^{-2}$	11598 (33+3, 459)	$9.1\ 10^{-3}$	15496 (52+7, 395)
10^{-2}	$4.1\ 10^{-3}$	15678 (67+2, 513)	$3.6\ 10^{-3}$	18624 (99+2, 213)
10^{-3}	$1.5\ 10^{-3}$	28980 (173+27, 249)	$4.3\ 10^{-4}$	31868 (254+20, 142)
$Z_0 = 10$	$h = 1/50$	$err_{2,s} = 1.0\ 10^{-1}$	$h = 1/100$	$err_{2,s} = 3.0\ 10^{-2}$
Tol	$err_{2,t}$	Costs	$err_{2,t}$	Costs
10^{-1}	$2.2\ 10^{-2}$	33258 (34+3, 1297)	$9.1\ 10^{-3}$	42805 (53+6, 1052)
10^{-2}	$4.2\ 10^{-3}$	44303 (67+2, 1448)	$1.8\ 10^{-3}$	52842 (100+2, 601)
10^{-3}	$1.5\ 10^{-3}$	81259 (173+26, 702)	$4.3\ 10^{-4}$	89640 (255+19, 402)

Table 4.1: Results for the explicit code RKC for the radiation-diffusion problem with L_2-errors and Costs; $err_{2,t}$ is the temporal error and $err_{2,s}$ is the spatial error. Costs is given as $N_F\,(N_{acc} + N_{rej}, s_{max})$ with N_F total number of function evaluations, N_{acc} number of accepted steps, N_{rej} number of rejected steps, and s_{max} the maximal number of stages per time step.

4.3 Test results for the IMEX scheme

Table 4.2 gives results obtained with a preliminary test version of the IMEX extension of the code RKC. The result are presented in the same way as in Table 4.1, except that the total number of function evaluations N_F has been replaced by the total number of stages N_{stage} ($N_{stage} = N_F$ for the explicit code). The gain due to the IMEX extension is very clear: to a great extent the workload is independent of the stiffness imposed by Z_0, which means high savings in numbers of stages for $Z_0 = 5, 10$ compared to the explicit case. Note also that for all nine test runs the number of accepted and rejected integration steps is nearly the same as in the explicit case.

The (preliminary) IMEX code used the same step size and local error control as the explicit one and thus differed only in the additional solution of the reaction systems (3.8). The additional solution costs for these systems diminish of course the anticipated savings from the lesser amounts of stages. Thus the efficiency of the solution process for systems (3.8) should be as high as possible. As noted in Remark 3.2, it makes sense to use modified Newton iteration in the same way as in the stiff ODE field. The results of Table 4.2 were indeed obtained with a standard modified Newton implementation that evaluates a new Jacobian and performs a new LU-decomposition at each stage of the RKC scheme, and at each grid point. Acceptance for the iterants was thus decided per grid point, allowing the number of iterations to differ over the grid points.

As start vector the accepted iterant of the previous stage was used and the iteration process was terminated as soon as the modified Newton correction was 1% smaller than Tol, or as soon as the modified Newton residual was less than 10^{-9} (both measured in the maximum norm). A seemingly cheaper alternative, based on recomputing the Jacobian and LU-decomposition only once per integration step at the beginning of the step, turned out to be slightly less efficient requiring more integration steps and more iterations. This of course is problem dependent.

Table 4.3 contains CPU times in seconds for the runs with the test case $Z_0 = 10$ on the 100×100 grid (measured on a SUN Workstation Ultra5). For this test case and on this grid, the IMEX code turned out to be about 3.5 to 4 times faster than the explicit code, while it has spent about half of its total elapsed CPU time on solving the reaction systems (3.8) which makes sense. The gain for the IMEX code would

$Z_0 = 1$	$h = 1/50$	$err_{2,s} = 3.0\ 10^{-2}$	$h = 1/100$	$err_{2,s} = 8.5\ 10^{-3}$
Tol	$err_{2,t}$	Costs	$err_{2,t}$	Costs
10^{-1}	$3.3\ 10^{-2}$	1954 (34+3, 76)	$1.1\ 10^{-2}$	4708 (54+3, 115)
10^{-2}	$4.6\ 10^{-3}$	2610 (68+2, 80)	$3.1\ 10^{-3}$	6167 (101+2, 75)
10^{-3}	$1.5\ 10^{-3}$	5209 (182+36, 39)	$4.5\ 10^{-4}$	12306 (268+52, 52)
$Z_0 = 5$	$h = 1/50$	$err_{2,s} = 7.8\ 10^{-2}$	$h = 1/100$	$err_{2,s} = 2.7\ 10^{-2}$
Tol	$err_{2,t}$	Costs	$err_{2,t}$	Costs
10^{-1}	$2.3\ 10^{-2}$	1834 (33+2, 76)	$9.4\ 10^{-3}$	4835 (54+4, 126)
10^{-2}	$4.4\ 10^{-3}$	2590 (67+2, 80)	$2.2\ 10^{-3}$	6144 (100+2, 69)
10^{-3}	$1.4\ 10^{-3}$	4726 (173+24, 39)	$4.9\ 10^{-4}$	10754 (256+25, 49)
$Z_0 = 10$	$h = 1/50$	$err_{2,s} = 1.0\ 10^{-1}$	$h = 1/100$	$err_{2,s} = 3.0\ 10^{-2}$
Tol	$err_{2,t}$	Costs	$err_{2,t}$	Costs
10^{-1}	$2.2\ 10^{-2}$	1816 (32+2, 76)	$1.0\ 10^{-2}$	4601 (54+2, 115)
10^{-2}	$4.5\ 10^{-3}$	2589 (67+2, 80)	$1.9\ 10^{-3}$	6151 (100+2, 69)
10^{-3}	$1.4\ 10^{-3}$	4774 (175+25, 39)	$5.0\ 10^{-4}$	10840 (258+26, 49)

Table 4.2: Results for the IMEX version of the code RKC for the radiation-diffusion problem with L_2-errors and Costs; $err_{2,t}$ is the temporal error and $err_{2,s}$ is the spatial error. Costs is given as N_{stage} (N_{acc}+ N_{rej}, s_{max}) with N_{stage} total number of stages, N_{acc} number of accepted steps, N_{rej} number of rejected steps, and s_{max} the maximal number of stages per time step.

increase with the reaction stiffness and recall that its computational effort (the numbers of stages) is largely determined by the stiffness coming from the diffusion term. For the current problem the 100×100 grid gives a spectral radius $\rho_D = 8.0\,10^4$, see (4.1), which is considerable of course. On the 50×50 grid, giving $\rho_D = 2.0\,10^4$, the IMEX code was about 6.5 to 7 times faster.

Table 4.3 also gives the average and the maximum number of modified Newton iterations, counted over all grid points, all stages and all steps. These numbers are low and in accordance with stiff ODE practice.

	Time RKC	Time IMEX			Iterations	
Tol	Total	Total	Systems	Rest	Average #	Maximum #
10^{-1}	2127	510 (4.2)	228	282	1.03	2
10^{-2}	2630	698 (3.8)	317	381	1.23	2
10^{-3}	4400	1227 (3.6)	562	665	1.36	3

Table 4.3: CPU times with bracketed numbers the speed up and numbers of modified Newton iterations at the grid points for the runs with the test case $Z_0 = 10$ on the 100×100 grid.

5 Final remarks

The original second-order code RKC is fully explicit, stabilized, and requires little memory. This makes it an attractive, user-friendly code for integrating large-scale semi-discrete parabolic problems. Its limitation lies in the stiffness and hence for efficiency reasons RKC is not advocated for severely stiff semi-discrete parabolic problems. By treating reaction terms implicitly and diffusion terms still explicitly (IMEX approach), this limitation has been removed for severely stiff diffusion-reaction problems where severe

stiffness emanates from reaction terms having a Jacobian matrix with a real spectrum. By the IMEX approach the code remains user-friendly and memory usage is still low.

The results and conclusions reported in this preprint are based on ongoing research. The very good comparative results for the radiation-diffusion problem are no doubt promising and justify further work on the subject. In the near future we plan further development of the current preliminary IMEX code and further testing including comparisons with the popular implicit BDF code VODPK [5, 6, 7] and the linearly implicit Rosenbrock code ROWMAP [19] (both use iterative Krylov methods). Furthermore it seems very worthwhile to develop an IMEX version of the RKC scheme that can also handle severely stiff reaction terms having a Jacobian with a complex spectrum.

References

[1] A. Abdulle (2001), *Chebyshev methods based on orthogonal polynomials*. Thesis No. 3266, Dept. Math., Univ. of Geneva.

[2] A. Abdulle, A.A. Medovikov (2001), *Second order Chebyshev methods based on orthogonal polynomials*. Numer. Math. 90, pp. 1–18.

[3] A. Abdulle (2002), *Fourth order Chebyshev methods with recurrence relation*. SIAM J. Sci. Comput. 23, pp. 2042–2055.

[4] M. Bakker (1971), *Analytical aspects of a minimax problem* (in Dutch). Technical Note TN 62, Mathematical Centre, Amsterdam.

[5] P.N. Brown, A.C. Hindmarsh (1989), *Reduced storage matrix methods in Stiff ODE Systems*, J. Appl. Math. Comput. 31, pp. 40–91.

[6] P.N. Brown, C.S. Woodward (2001), *Preconditioning strategies for fully implicit radiation diffusion with material-energy transfer*, SIAM J. Sci. Comput. 23, pp. 499-516.

[7] G.D. Byrne (1992), *Pragmatic experiments with Krylov methods in the stiff ODE setting*, Computational Ordinary Differential Equations, J.R. Cash, I. Gladwell (eds.), Oxford Univ. Press, Oxford, pp. 323-356.

[8] A. Guillou, B. Lago (1961), *Domaine de stabilité associé aux formules d'intégration numérique d'équations différentielles, a pas séparés et a pas liés. Recherche de formules a grand rayon de stabilité*. Ier Congr. Assoc. Fran. Calcul, AFCAL, Grenoble, Sept. 1960, pp. 43–56.

[9] P.J. van der Houwen (1977), *Construction of Integration Formulas for Initial Value Problems*. North-Holland, Amsterdam.

[10] P.J. van der Houwen, B.P. Sommeijer (1980), *On the internal stability of explicit, m-stage Runge-Kutta methods for large m-values*. Z. Angew. Math. Mech. 60, pp. 479–485.

[11] W. Hundsdorfer, J.G. Verwer (2003), *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Series in Computational Mathematics, Vol. 33, Springer, Berlin (to appear in July).

[12] V.I. Lebedev (1994), *How to solve stiff systems of differential equations by explicit methods*. In: *Numerical Methods and Applications*. Ed. G.I. Marchuk, CRC Press, pp. 45–80.

[13] V.I. Lebedev (2000), *Explicit difference schemes for solving stiff problems with a complex or separable spectrum*. Comput. Math. and Math. Phys. 40, pp. 1801–1812.

[14] V.A. Mousseau, D.A. Knoll, W.J. Rider (2000), *Physics-based preconditioning and the Newton-Krylov method for non-equilibrium radiation diffusion*. J. Comput. Phys. 160, pp. 743–765.

[15] W. Riha (1972), *Optimal stability polynomials*. Computing 9, pp. 37–43.

[16] L.F. Shampine (1994), *Numerical Solution of Ordinary Differential Equations*. Chapman & Hall, New York.

[17] B.P. Sommeijer, L.F. Shampine, J.G. Verwer (1997), *RKC: An explicit solver for parabolic PDEs*. J. Comput. Appl. Math. 88, pp. 315–326.

[18] J.G. Verwer, W.H. Hundsdorfer, B.P. Sommeijer (1990), *Convergence properties of the Runge-Kutta-Chebyshev method*. Numer. Math. 57, pp. 157–178.

[19] R. Weiner, B.A. Schmitt, H. Podhaisky (1997), *ROWMAP - a ROW code with Krylov techniques for large stiff ODEs*. Appl. Numer. Math. 25, pp. 303-319

NUMERICAL AND COMPUTATIONAL CHALLENGES IN ENVIRONMENTAL MODELLING

Z. ZLATEV*

Abstract.
Large-scale mathematical models can successfully be used in different environmental studies. These models are described by systems of partial differential equations. Splitting procedures followed by discretization of the spatial derivatives lead to several large systems of ordinary differential equations of order up to 80 millions. These systems have to be handled numerically at up to 250 000 time-steps. Furthermore, many scenarios are often to be run in order to study the dependence of the model results on the variation of some key parameters (as, for example, the emissions). Such huge computational tasks can successfully be treated only if (i) fast and sufficiently accurate numerical methods are used and (ii) the models can efficiently be run on parallel computers.

The mathematical description of a large-scale air pollution model will be discussed in this paper. The principles used in the selection of numerical methods and in the development of parallel codes will be described. Numerical results, which illustrate the ability of running the fine resolution versions of the model on Sun computers, will be given. Applications of the model in the solution of some environmental tasks will be presented. The ideas are fairly general and can be used in the development of some other kinds of environmental models as well as in modelling in some other fields of science and engineering.

Key words. Air pollution modelling, Partial differential equations, Ordinary differential equations, Numerical methods, Cache utilization, Parallel computations, Applications

1. Why are large-scale mathematical models used?. The control of the pollution levels in different highly polluted regions of Europe and North America (as well as in other highly industrialized parts of the world) is an important task for the modern society. Its relevance has been steadily increasing during the last two-three decades. The need to establish reliable control strategies for the air pollution levels will become even more important in the future. Large-scale air pollution models can successfully be used to design reliable control strategies. Many different tasks have to be solved before starting to run operationally an air pollution model. The following tasks are most important:

- describe in an adequate way all important physical and chemical processes,
- apply fast and sufficiently accurate numerical methods in the different parts of the model,
- ensure that the model runs efficiently on modern high-speed computers (and, first and foremost, on different types of parallel computers),
- use high quality input data (both meteorological data and emission data) in the runs,
- verify the model results by comparing them with reliable measurements taken in different parts of the space domain of the model,
- carry out some sensitivity experiments to check the response of the model to changes of different key parameters

and

- visualize and animate the output results to make them easily understandable also for non-specialists.

The solution of the first three tasks will be the main topic of this paper (however, several visualizations will be used to present results from some real-life runs in the end of the paper). The air pollution model, which is actually used here, is the Danish

*National Environmental Research Institute, Frederiksborgvej 399, P. O. Box 358, DK-4000 Roskilde, Denmark (zz@dmu.dk).

Eulerian Model (DEM); see [34], [36]. However, the principles are rather general, which means that most of the results are also valid for other air pollution models.

1.1. Main physical and chemical processes. Five physical and chemical processes have to be described by mathematical terms in the beginning of the development of an air pollution model. These processes are:

- horizontal transport (advection),
- horizontal diffusion,
- chemical transformations in the atmosphere combined with emissions from different sources,
- deposition of pollutants to the surface

and

- vertical exchange (containing both vertical transport and vertical diffusion).

It is important to describe in an adequate way all these processes. However, this is an extremely difficult task; both because of the lack of knowledge for some of the processes (this is mainly true for some chemical reactions and for some of the mechanisms describing the vertical diffusion) and because a very rigorous description of some of the processes will lead to huge computational tasks which may make the treatment of the model practically impossible. The main principles used in the mathematical description of the main physical and chemical processes as well as the need to keep the balance between the rigorous description of the processes and the necessity to be able to run the model on the available computers are discussed in [34].

1.2. Mathematical formulation of a large air pollution model. The description of the physical and chemical processes by mathematical terms leads to a system of partial differential equations (PDEs) of the following type:

$$
\begin{aligned}
\frac{\partial c_s}{\partial t} = & -\frac{\partial (u c_s)}{\partial x} - \frac{\partial (v c_s)}{\partial y} - \frac{\partial (w c_s)}{\partial z} \\
& + \frac{\partial}{\partial x}\left(K_x \frac{\partial c_s}{\partial x} \right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s}{\partial y} \right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial c_s}{\partial z} \right) \\
& + E_s - (\kappa_{1s} + \kappa_{2s}) c_s + Q_s(c_1, c_2, \ldots, c_q), \quad s = 1, 2, \ldots, q,
\end{aligned} \tag{1.1}
$$

where (i) the concentrations of the chemical species are denoted by c_s, (ii) u, v and w are wind velocities, (iii) K_x, K_y and K_z are diffusion coefficients, (iv) the emission sources are described by E_s, (v) κ_{1s} and κ_{2s} are deposition coefficients and (vi) the chemical reactions are denoted by $Q_s(c_1, c_2, \ldots, c_q)$. The CBM IV chemical scheme, which has been proposed in [13], is actually used in the version of DEM (the Danish Eulerian Model; [34], [36]) that will be considered in this paper. It should be mentioned here that the CBM IV scheme is also used in other well-known air pollution models.

1.3. Space domain. The space domain of DEM is a 4800 km x 4800 km square, which contains the whole of Europe together with parts of Africa, Asia, the Arctic area and the Atlantic Ocean. Two discretizations of this domain, a coarse one and a fine one, will be used in this paper. The space domain is divided into 96x96 small, 50 km x 50 km, squares when the coarse discretization is applied. The space domain is divided into 480x480 small, 10 km x 10 km, squares when the fine discretization is applied. Thus, one of the coarse grid-squares contains 25 small grid-squares.

1.4. Initial and boundary conditions. If initial conditions are available (for example from a previous run of the model), then these are read from the file where they are stored. If initial conditions are not available, then a five day start-up period is used to obtain initial conditions (i.e. the computations are started five days before the desired starting date with some background concentrations and the concentrations found at the end of the fifth day are actually used as starting concentrations).

The choice of lateral boundary conditions is in general very important. However, if the space domain is very large, then the choice of lateral boundary conditions becomes less important; which is stated on p. 2386 in [6]: *"For large domains the importance of the boundary conditions may decline"*. The lateral boundary conditions are represented in the Danish Eulerian Model with typical background concentrations which are varied, both seasonally and diurnally. It is better to use values of the concentrations at the lateral boundaries that are calculated by a hemispheric or global model when such values are available.

For some chemical species, as for example ozone, it is necessary to introduce some exchange with the free troposphere (on the top of the space domain).

The choice of initial and boundary conditions is discussed in [14], [34], [36], [37] and [38].

1.5. Applying splitting procedures. It is difficult to treat the system of PDE's (1.1) directly. This is the reason for using different kinds of splitting. A splitting procedure, which is based on ideas proposed in [21] and [22], and which leads to five sub-models, has been proposed in [34] and used after that in many studies involving DEM (as, for example, in [36]). Each of the five sub-models obtained by this splitting procedure is representing one of the major physical and chemical processes discussed in §1.1; i.e. the horizontal advection, the horizontal diffusion, the chemistry (together with the emission terms), the deposition and the vertical exchange.

In the newest version of DEM, which is used here, the horizontal advection was merged with the horizontal diffusion, while the chemical sub-model was combined with the deposition sub-model. This means that the number of sub-models is reduced from five to three:

$$\frac{\partial c_s^{(1)}}{\partial t} = -\frac{\partial (w c_s^{(3)})}{\partial z} + \frac{\partial}{\partial z}\left(K_z \frac{\partial c_s^{(3)}}{\partial z} \right) \tag{1.2}$$

$$\begin{aligned} \frac{\partial c_s^{(2)}}{\partial t} = & -\frac{\partial (u c_s^{(2)})}{\partial x} - \frac{\partial (v c_s^{(2)})}{\partial y} \\ & + \frac{\partial}{\partial x}\left(K_x \frac{\partial c_s^{(2)}}{\partial x} \right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s^{(2)}}{\partial y} \right) \end{aligned} \tag{1.3}$$

$$\begin{aligned} \frac{dc_s^{(3)}}{dt} = & E_s + Q_s(c_1^{(3)}, c_2^{(3)}, \ldots, c_q^{(3)}) \\ & -(\kappa_{1s} + \kappa_{2s}) c_s^{(3)} \end{aligned} \tag{1.4}$$

The first of these sub-models, (1.2), describes the vertical exchange. The second sub-model, (1.3), describes the combined horizontal transport (the advection) and the horizontal diffusion. The last sub-model, (1.4), describes the chemical reactions together with emission sources and deposition terms.

The main principles used to treat the sub-models at a given time-step are the same as the principles discussed in [21], [22] and [34]; see also [39].

Splitting allows us to apply different numerical methods in the different sub-models and, thus, to reduce considerably the computational work and to exploit better the properties of each sub-model. These are the main advantages of using splitting. Unfortunately, there are drawbacks also: the splitting procedure is introducing errors, and it is difficult to control these errors. Some attempts to obtain some evaluation of the splitting errors were recently carried out; see [19] and [9].

1.6. Space discretization. Assume that the space domain is discretized by using a grid with $N_x \times N_y \times N_z$ grid-points, where N_x, N_y and N_z are the numbers of the grid-points along the grid-lines parallel to the Ox, Oy and Oz axes. Assume further that the number of chemical species involved in the model is $q = N_s$. Finally, assume that the spatial derivatives in (1.2) are discretized by some numerical algorithm. Then the system of PDE's (1.2) will be transformed into a system of ODEs (ordinary differential equations):

$$\frac{dg^{(1)}}{dt} = f^{(1)}(t, g^{(1)}), \tag{1.5}$$

In a similar way, the system of PDEs (1.3) can be transformed into the following system of ODEs when the spatial derivatives in the right-hand-side of (1.3) are discretized:

$$\frac{dg^{(2)}}{dt} = f^{(2)}(t, g^{(2)}), \tag{1.6}$$

There are in fact no spatial derivatives in the right-hand-side of (1.4), because the non-linear functions Q_s can be represented as

$$Q_s(c_1, c_2, \ldots, c_q) = -\sum_{i=1}^{q} \alpha_{si} c_i + \sum_{i=1}^{q} \sum_{j=1}^{q} \beta_{sij} c_i c_j, \qquad s = 1, 2, ..., q. \tag{1.7}$$

where α_{si} and β_{sij} are coefficients describing the rates of the chemical reactions (for the CBM IV schemes these coefficients are listed in [34]). By using this observation, it is easy to represent (1.4) as a system of ODEs:

$$\frac{dg^{(3)}}{dt} = f^{(3)}(t, g^{(3)}), \tag{1.8}$$

The components of functions $g^{(i)}(t) \in R^{N_x \times N_y \times N_z \times N_s}$, $i = 1, 2, 3$, are the approximations of the concentrations (at time t) at all grid-squares and for all species. The components of functions $f^{(i)}(t, g) \in R^{N_x \times N_y \times N_z \times N_s}$, $i = 1, 2, 3$, depend on the numerical method used in the discretization of the spatial derivatives.

A simple linear finite element method is used to discretize the spatial derivatives in (1.2) and (1.3). This method is described in [26] and [27]. Its implementation in DEM is discussed in [11].

The spatial derivatives can also be discretized by using other numerical methods:

- Pseudospectral discretization (described in detail in [34]).
- Semi-Lagrangian discretization (can be used only to discretize the first-order derivatives, i.e. the advection part should not be combined with the diffusion part when this method is to be applied), see for example [20].
- Methods producing non-negative values of the concentrations. The method proposed in [4] is often used in air pollution modelling. The method from [17] is based on a solid theoretical foundation.

As mentioned above, there are no spatial derivatives in (1.4), which means that the system of ODEs (1.8) is trivially obtained by (1.4).

Much more details about the methods, which can be used in the space discretization, can be found in [34].

1.7. Time integration. It is necessary to couple the three ODE systems (1.5), (1.6) and (1.8). The coupling procedure is connected with the time-integration of these systems. Assume that the values of the concentrations (for all species and at all grid-points) have been found for some $t = t_n$. According to the notation introduced in the previous sub-section, these values can be considered as components of a vector-function $g(t_n) \in R^{N_x \times N_y \times N_z \times N_s}$. The next time-step, time-step $n+1$ (at which the concentrations are found at $t_{n+1} = t_n + \triangle t$, where $\triangle t$ is some increment), can be performed by integrating successively the three systems. The values of $g(t_n)$ are used as an initial condition in the solution of (1.5). The solution of (1.5) is used as an initial condition of (1.6). Finally, the solution of (1.6) is used as an initial condition of (1.8). The solution of the last system (1.8) is used as an approximation to $g(t_{n+1})$. In this way, everything is prepared to start the calculations in the next time-step, step $n+2$.

The first ODE system, (1.5), can be solved by using many classical time-integration methods. The so-called θ-method (see, for example, [18]) is currently used in DEM. The choice of numerical method is not very critical in this part, because as it will be shown Section 4, it is normally not very expensive.

Predictor-corrector (PC) methods with several different correctors are used in the solution of the ODE system (1.6). The correctors are carefully chosen so that the stability properties of the method are enhanced; see [33]. The reliability of the algorithms used in the advection part was verified by using the well-known rotational test proposed simultaneously in 1968 by [7] and [23]. If the code judges the time-stepsize to be too large for the currently used PC method, then it switches to a more stable (but also more expensive) PC scheme. On the other hand, if the code judges that the stepsize is too small for the currently used PC method, then it switches to more stable (and less expensive) PC scheme. In this way the code is trying both to keep the same stepsize and to optimize the performance.

The solution of (1.8) is much more complicated, because this system is both time-consuming and stiff. Very often the QSSA method is used in this part of the model. The QSSA (quasi-steady-state approximation; see, for example, [15] or [16]) is simple and relatively stable but not very accurate (therefore it has to be run with a small time-stepsize). The QSSA method can be viewed as an attempt to transform dynamically, during the process of integration, the system of ODEs (1.8) into two systems: a system of ODEs and a system of non-linear algebraic equations. These two systems, which have to be treated simultaneously, can be written in the following generic form:

$$\frac{dg_1}{dt} = f_1(t, g_1, g_2), \tag{1.9}$$

$$0 = f_2(t, g_1, g_2). \tag{1.10}$$

In this way we arrive at a system of differential-algebraic equations (DAEs). There are special methods for treating such systems as, for example, the code DASSL (see [5]). Problem-solving environments (such as MATLAB or Simulink) can be used in the preparation stage (where a small chemical systems at one grid-point only is used in the tests). More details about the use of such problem solving environments can be found in [28]. A method based on the solution of DAE for air pollution models was recently proposed in [10].

The classical numerical methods for stiff ODE systems (such as the Backward Euler Method, the Trapezoidal Rule and Runge-Kutta algorithms) lead to the solution of non-linear systems of algebraic equations and, therefore, they are more expensive; [18]. On the other hand, these methods can be incorporated with an error control and perhaps with larger time-steps. The extrapolation methods, [8], are also promising. It is easy to calculate an error estimation and to carry out the integration with large time-steps when these algorithms are used. However, it is difficult to implement such methods in an efficient way when all three systems, (1.5), (1.6) and (1.8), are to be treated successively.

Partitioning can also be used ([1]). Some convergence problems related to the implementation of partitioning are studied in [35].

The experiments with different integration methods for the chemical sub-model are continuing. The QSSA with some enhancements based on ideas from [29] and [30] will be used here. The method is described in [1]. There are still very open questions related to the choice of method for the chemical part. The choice of the improved QSSA method was made in order to get well-balanced parallel tasks.

2. Need for high performance computing in the treatment of large air pollution models. The computers are becoming more and more powerful. Many tasks, which several years ago had to be handled on powerful supercomputers, can be handled at present on PCs or work-stations. However, there are still many tasks that can only be run on parallel computers. This is especially true for the large air pollution models. The size of the computational tasks in some versions of DEM is given in the following two paragraphs in order demonstrate the fact that high performance computing is needed when large air pollution models are to be treated.

2.1. Size of the computational tasks when 2-D versions are used. Only the two systems of ODEs (1.6) and (1.8) have to be treated in this case. Assume first that the coarse 96×96 grid is used. Then the number of equations in each of the two systems of ODEs (1.6) and (1.8) is equal to the product of the grid points (9216) and the number of chemical species (35), i.e. 322560 equations have to be treated at each time-step when any of the systems (1.6) and (1.8) is handled. The time-stepsize used in the transport sub-model (1.6) is 900 s. This stepsize is too big for the chemical sub-model; the time-stepsize used in the latter model is 150 s. A typical run of this model covers a period of one year (in fact, as mentioned above, very often a period of extra five days is needed to start up the models. This means that 35520 time-steps

are needed in the transport sub-model, while six times more time-steps, 213120 time-steps, are needed in the chemical part. If the number of scenarios is not large, then this version of the model can be run on PCs and work-stations. If the number of scenarios is large or if runs over many years have to be performed (which is the case when effects of future climate changes on the air pollution studies is studied), then high performance computations are preferable (this may be the only way to complete the study when either the number of scenarios is very large or the time period is very long).

Assume now that the medium 288×288 grid is used. Since the number of chemical species remains unchanged (35), the number of equations in each of the systems (1.6) and (1.8) is increased by a factor of 9 (compared with the previous case). This means that 2903040 equations are to be treated at each time step when any of the systems (1.6) and (1.8) is handled. The time-stepsize remains 150 s when the chemical part is treated. The time-stepsize has to be reduced from 900 s to 300 s in the transport part. This means that a typical run (one year + 5 days to start up the model) will require 106760 time-steps when (1.6) is treated and 213120 time-steps are needed when (1.8) is handled. Consider the ratio of the computational work when the medium grid is used and the computational work when the coarse grid is used. For the transport sub-model this ratio is 18, while the ratio is 9 for the chemical-sub-model.

Finally, assume that the fine 480×480 grid is used. Using similar arguments as in the previous paragraph, it is easy to show that the number of equations in each of the systems (1.6) and (1.8) is increased by a factor of 25 (compared with the 96x96 grid). This means that 8064000 equations are to be treated at each time step when any of the systems (1.6) and (1.8) is handled. The time-stepsize remains 150 s when the chemical part is treated. The time-stepsize has to be reduced from 900 s to 150 s in the transport part. This means that a typical run (one year + 5 days to start up the model) will require 213520 time-steps for each of the systems (1.6) and (1.8). Consider the ratio of the computational work when the fine grid is used and the computational work when the coarse grid is used. For the transport sub-model this ratio is 150, while the ratio is 25 for the chemical-sub-model. It is clear that this version of the model must be treated on powerful parallel architectures.

2.2. Size of the computational tasks when 3-D versions are used. All three sub-models, (1.5), (1.6) and (1.7), have to be treated in this case. Assume that the number of layers in the vertical direction is n ($n = 10$ is used in this paper). Under this assumption the computational work when both (1.6) and (1.8) is handled by the 3-D versions (either on a coarse grid or on a fine grid) is n times bigger than the computational work for the corresponding 2-D version. The work needed to handle (1.5) is extra, but this part of the total computational work is much smaller than the parts needed to treat (1.6) and (1.8).

The above analysis of the amount of the computational work shows that it is much more preferable to run the 3-D version on high-speed parallel computers when the coarse grid is used. It will, furthermore, be shown that the runs are very heavy when the 3-D version is to be run on a fine grid. In fact, more powerful parallel computers than the computers available at present are needed if meaningful studies with the 3-D version of DEM discretized on a fine grid are to be carried out.

2.3. Exploiting the cache memory of the computer. In the modern computers the time needed for performing arithmetic operations is reduced dramatically (compared with computers which were available 10-15 years ago). However, the reductions of both the time needed to bring the numbers which are participating in the

arithmetic operations from the memory to the place in the computer where the arithmetic operation is to be actually performed and the time needed to store the results back in the memory are much smaller. This is why most of the nowadays computers have different caches. It is much more efficient to use data which is in cache than to make references to the memory. Unfortunately, it is very difficult for the user (if at all possible) to control directly the utilization of the cache. Nevertheless, there are some common rules by the use of which the performance can be improved considerably. The rules discussed in [24] and [25] will be outlined below. These rules have been used in runs on several other computers in [24] and [25]. It will be shown in Section 5 that these rules are performing rather well also when Sun parallel computers are used.

Consider the 2-D versions of DEM. Assume that the concentrations are stored in an array $CONS(N_x \times N_y, N_s)$. Each column of this array is representing the concentrations of a given chemical species at all grid-points, while each row is containing the concentrations of all chemical species at a given grid-point. There are seven other arrays of the same dimension.

There are no big problems when the transport sub-model is run (because the computations are carried out by columns). However, even here cache problems may appear, because the arrays are very long. This will be further discussed in Section 5.

Great problems appear in the chemical part, because when the concentration of some species in a given row is modified, some other species in the same row are participating in the computations, which becomes clear from the pseudo Fortran code given below (with $M = N_x \times N_y$ and $NSPECIES = N_s$).

```
DO J=1,NSPECIES
   DO I=1,M
      Perform the chemical reactions involving
      species J in grid-point I
   END DO
END DO
```

This code is perfect for some vector machines. However, if cache memory is available, then the computations, as mentioned above, can be rather slow, because in step I, $I = 1, 2, \ldots M$, of the inner loop $CONS(I, J)$ is updated, but the new value of the chemical species J depends on some of the other species K, $K = 1, 2, \ldots, J-1, J+1, \ldots, NSPECIES$. Thus, when we are performing the I'th step of the second loop, we have to refer to some addresses in row I of array $CONS(M, NSPECIES)$. The same is true for the seven other arrays of the same dimension. It is intuitively clear that it is worthwhile to divide these arrays into chunks and to carry out the computations by chunks. Assume that we want to use $NCHUNKS$ chunks. If M is a multiple of $NCHUNKS$, then the size of every chunks is $NSIZE = M/NCHUNKS$, and the code given above can be modified in the following way.

```
DO ICHUNK=1,NCHUNKS
   Copy chunk ICHUNK from some of the eight
   large arrays into small two-dimensional
   arrays with leading dimension NSIZE
   DO J=1,NSPECIES
      DO I=1,NSIZE
         Perform the chemical reactions involving
         species J for grid-point I
      END DO
```

```
        END DO
        Copy some of the small two-dimensional
        arrays with leading dimension NSIZE
        into chunk ICHUNK of the corresponding
        large arrays
    END DO
```

Both the operations that are performed in the beginning and in the end of the first loop in the second code are extra. The extra work needed to perform these operations is fully compensated by savings during the inner double loop, which is very time-consuming.

A straight-forward procedure will be to copy the current chunks of all eight arrays in the corresponding small arrays. However, this is not necessary, because some of the arrays are only used as helping arrays in the chemical module. In fact, copies from five arrays are needed in the beginning of the first loop. This means that there is no need to declare the remaining three arrays as large arrays; these arrays can be declared as arrays with dimensions $(NSIZE, NSPCIES)$, which leads to a reduction of the storage needed. The reduction is very considerable for the fine 480×480 grid.

The situation in the end of the first loop is similar; it is necessary to copy back to the appropriate sections of the large arrays only the contents of three small arrays. The number of copies made at the end of the first loop has been reduced from five to three because some information (as, for example, the emissions) is needed in the chemical module (and has to be copied from the large arrays to the small ones), but it is not modified in the chemical module (and, thus, there is no need to copy it back to the large arrays in the end of the first loop).

When the 3-D versions are used, the array $CONS(N_x \times N_y, N_s)$ must be replaced by $CONS(N_x \times N_y, N_s, N_z)$. However, the device described above can be applied, because the computations for each layer can be carried out independently from the computations for the other layers when (1.6) and (1.8) are treated.

It will be shown in Section 4 that the use of chunks leads to considerable savings in computing time in the chemical sub-model.

3. Achieving parallelism. It was explained in the previous section that the discretization of an air pollution model is as a rule resulting in huge computational tasks. This is especially true in the case where the model is discretized on a fine grid. Therefore it is important to prepare parallel codes which run efficiently on modern parallel computers. The preparation of such a code will be discussed in this section.

3.1. Basic principles used in the preparation of the parallel versions. The preparation of a parallel code is by no means an easy task. Moreover, it may happen that when the code is ready the computing centre exchanges the computer which has been used in the preparation of the code with another (hopefully, more powerful) computer. This is why it is desirable to use only standard tools in the preparation of the code. This will facilitate the transition of the code from one computer to another when this becomes necessary. Only standard OpenMP ([31]) and MPI ([12]) tools are used in the parallel versions of DEM.

3.2. Development of OpenMP versions of DEM. The programming for shared memory machines is relatively easy. It is necessary to identify the parallel tasks and to insert in the code appropriate OpenMP directives (which on ordinary sequential machines will be viewed as comments). The parallel tasks in the three sub-models are discussed below.

- **Parallel tasks in the transport sub-model.** This sub-model is mathematically described (after the discretization) by (1.6). It is easy to see that the system of ODEs (1.6) is consisting of $q \times N_z$ independent systems of ODEs, where q is the number of chemical species and N_z is the number of grid-points in the vertical direction. This means that there are $q \times N_z$ parallel tasks. Each parallel task is a system of $N_x \times N_y$ ODEs. In the chemical scheme adopted in DEM there are 35 chemical species, but three of them are linear combinations of other chemical species. N_z is equal to 1 in the 2-D case and to 10 in the 3-D case. Therefore, the actual number of parallel tasks is 32 in the 2-D case and 320 in the 3-D case. The tasks are large and the loading balance in the transport sub-model is perfect. The use of this technique is, thus, very efficient when the number of processors used is a divisor of 32 in the 2-D case and 320 in the 3-D case. Some problems may arise in the 2-D case. If more than 32 processors are available, then it will be necessary to search for parallel tasks on a lower level of the computational process when the 2-D versions are used.
- **Parallel tasks in the chemical sub-model.** This sub-model is mathematically described (after the discretization) by (1.8). It is easy to see that the system of ODEs (1.8) is consisting of $N_x \times N_y \times N_z$ independent systems of ODEs, where N_x, N_y and N_z are the numbers of grid-points along the coordinate axes. The number of parallel tasks is very large (2304000 when the $480 \times 480 \times 10$ grid is used) and the loading balance is perfect. However, the parallel tasks are very small (each parallel task is a system of q ODEs). Therefore, it is necessary to group them in clusters. Moreover, some arrays are handled by rows, which may lead to a large number of cache misses, especially for the fine grid versions. Therefore, chunks are to be used in this part (see the end of the previous version).
- **Parallel tasks in the vertical exchange sub-model.** This sub-model is mathematically described (after the discretization) by (1.5). It is easy to see that the system of ODEs (1.5) is consisting of $N_x \times N_y \times N_s$ independent systems of ODEs. N_x and N_y are the numbers of grid-points along the coordinate axes O_x and O_y. $N_s = q$ is the number of chemical species. The number of parallel tasks is very large (8064000 when the $480 \times 480 \times 10$ grid is used with 35 chemical species) and the loading balance is perfect. However, the parallel tasks are again small (each parallel task is a system of N_z ODEs. Therefore, also in this sub-model it is necessary to group the parallel tasks in an appropriate way. It should also be emphasized that a very long array (its leading dimension being $N_x \times N_y \times N_s$) has to handled by rows. The vertical exchange is not very expensive computationally. Nevertheless, it is desirable to use chunks in the efforts to avoid a large number of cache misses (this is especially true for the fine resolution versions). No chunks are used at present, but there are plans to introduce chunks in the near future.

It is seen from the above discussion that it is very easy to organize the computational process for parallel runs when OpenMP tools are used. Moreover, it is clear that the parallel computations depend on the splitting procedure, but not on the numerical methods that have been selected.

3.3. Development of MPI versions of DEM. The approach used when MPI tools are to be implemented is based in dividing the space domain of the model into p sub-domains, where p is the number of processors which are to be used in the run.

Two specific modules are needed in the MPI versions: (i) a pre-processing module and (ii) a post-processing module.

- **The pre-processing module.** The input data is divided into p portions corresponding to the p sub-domains obtained in the division of the space domain. In this way, each processor will work during the whole computational process with its own set of input data.
- **The post-processing module.** Each processor prepares its own set of output data. During the post-processing the p sets of output data corresponding to the p sub-domains are collected and common output files are prepared for future use.
- **Benefits of using the two modules.** Excessive communications during the computational process are avoided when the two modules are used. It should be stressed, however, that not all communications during the computational process are avoided. Some communications along the inner boundaries of the sub-domains are still needed. However, these communications are to be carried only once per step and only a few data are to be communicated. Thus, the actual communications that are to be carried out during the computations are rather cheap when the pre-processing and the post-processing modules are proper implemented.

It is important to emphasize here that the introduction of p sub-domains leads to a reduction of the main arrays by a factor of p. Consider as an illustrations the major arrays used in the chemical sub-model. The dimensions of these arrays are reduced from $(N_x \times N_y, N_s)$ to $(N_x \times N_y/p, N_s)$. It is clear that this is equivalent to the use of p chunks; see §2.3. Chunks of length $N_x \times N_y/p$ are still very large. Therefore, the second algorithm given in §2.3 has also to be used (in each sub-domain) when the MPI versions are used. However, the reduction of the arrays leads to a reductions of the copies that are to be made in the beginning and in the end of the second algorithm in §2.3. Thus, the reduction of the arrays leads to a better utilization of the cache memory.

The automatic reduction of the sizes of the involved arrays, and the resulting from this reduction better utilization of the cache memory, make the MPI versions attractive also when shared memory machines are available. It will be shown in the next section that on Sun computers the MPI versions of DEM are often performing better than the corresponding OpenMP versions.

4. Moving from different versions to a common model. Only a two years ago several different options of DEM were available. Six versions were mainly used in the runs (three 2-D versions discretized on the 96×96, 288×288 and 480×480 grids respectively together with the corresponding three 3-D versions). Recently, these versions were combined in a common model. A special input file, "save_inform" is used to decide how to run the common model. Eight parameters are to be initialized in "save_inform" before the start of the run. These parameters are:

- NX - the number of grid-points along the Ox axis.
- NY - the number of grid-points along the Oy axis.
- NZ - the number of grid-points along the Oz axis.
- $NSPECIES$ - the number of chemical species involved in the model.
- $NREFINED$ - allows us to use refined emissions when available.
- $NSIZE$ - the size of the chunks to be used in the chemical parts.
- $NYEAR$ - the number of chunks to be used in the chemical parts.

TABLE 5.1
The computers available at the Sun grid.

Computer	Type	Power	RAM	Processors
Bohr	Sun Fire 6800	UltraSparc-III 750 MHrz	48 GB	24
Erlang	Sun Fire 6800	UltraSparc-III 750 MHrz	48 GB	24
Hald	Sun Fire 12k	UltraSparc-III 750 MHrz	144 GB	48
Euler	Sun Fire 6800	UltraSparc-III 750 MHrz	24 GB	24
Hilbert	Sun Fire 6800	UltraSparc-III 750 MHrz	36 GB	24
Newton	Sun Fire 15k	UltraSparc-IIIcu 900 MHrz	404 GB	72

- $PATH$ - the working path where the data attached to the different processors will be stored.

There are several restrictions for the parameters, which can be used at present. The restrictions are listed below:

1. The allowed values for NX and NY are 96, 288 and 480. Furthermore, NX must be equal to NY.
2. The allowed values for NZ are 1 (corresponds to the 2-D versions) or 10 (i.e. only 10 layers are allowed for the 3-D versions).
3. Only one value, 35, is allowed at present for $NSPECIES$.
4. Refined emissions are available only for the 288×288 case and will used when $NREFINED = 1$. If $NREFINED = 0$, then the emissions for the 96×96 grid will be used (simple interpolation will be used when any of the other two grids, the 96×96 grid or the 480×480 grid, is specified).
5. $NSIZE$ must be a divisor of $NX \times NY$.

Many of these restrictions will be removed (or, at least, relaxed) in the near future. It will, for example, be allowed to

- specify a rectangular space domain,
- use more than 10 layers

and

- apply chemical schemes with more than 35 species.

The common model, which can be run as described in this section, will be called UNI-DEM.

5. Numerical results. Some results will be presented in this sections to demonstrate (i) the performance of different versions of UNI-DEM and (ii) the ability of the code to carry out parallel computations in an efficient way. Some information about the computer grid used will be presented before the start of the discussion of the numerical results.

5.1. Description of the grid of Sun computers. Sun computers located at the Danish Centre for Scientific Computing (the Danish Technical University in Lyngby) were used in the runs. The computers and the their characteristics are shown in Table 4.1. All these computers were connected with a 1Gbit/s Switch.

The computers are united in a grid (consisting of 216 processors) so that a job sent without a special demand will be assigned on the computer on which there are sufficiently many free processors. The different computers have processors of different power (therefore, it is in principle possible to use the grid as a heterogeneous architecture, but this option is not available yet).

We have been allowed to use no more than 16 processors, and in the runs in

TABLE 5.2
Computing times (measured in seconds) obtained in the advection part when six options of UNI-DEM are run on 8 processors

$NX \times NY$	2-D ($NZ = 1$)	3-D ($NZ = 10$)
96	450	6240
288	17493	173685
480	110390	986672

TABLE 5.3
Computing times (measured in seconds) obtained in the chemistry part when six options of UNI-DEM are run on 8 processors

$NX \times NY$	2-D ($NZ = 1$)	3-D ($NZ = 10$)
96	2645	21545
288	35433	268721
480	150068	1055289

this section we used only "newton" (i.e. we had always a requirement specifying the particular computer on which the job must be run)

More details about the high speed computers that are available at the Technical University of Denmark can be found in [32].

5.2. Running the MPI options of UNI-DEM. Six MPI options of UNI-DEM have been tested: (i) the 2-D options obtained by using $NX = NY = 96$, $NX = NY = 288$ and $NX = NY = 480$ and (ii) the corresponding 3-D versions. $NSIZE = 48$ was used in all runs. The year was 1997. Refined emissions were used when $NX = NY = 288$, while the emissions given on the coarse grid were used in the other case (using interpolation when $NX = NY = 480$).

The most time consuming parts of the model are the advection part (combined with the diffusion part) and the chemical part (combined with the emissions and the deposition part). Therefore, we shall compare the advection times, the chemistry times and the total times for the six options discussed in this sub-section.

The results shown in Table 5.2 - Table 5.4 allow us to draw the following conclusions:

- The horizontal-advection diffusion part and the chemical part are indeed the most time-consuming parts of the computational process.
- It is possible to run the refined options (even the refined 3-D options) on the computers available at present. However, such runs are very time consuming. Therefore, it is still not possible to use these options in comprehensive studies (where the time-intervals are very long and, moreover, many scenarios are to be run).

5.3. Scalability of the code. It is interesting to see whether increasing of the number of processors used by a factor of k will lead to an reduction of the computing time by a factor approximately equal to k. We selected the most time-consuming option (the option discretized on a $480 \times 480 \times 10$ grid) and run it on 32 processors. The results were compared, see Table 5.5, with the results obtained when 8 processors were used (i.e. we have $k = 4$).

It is clearly seen (from Table 5.5) that the speed ups are rather close to linear.

The conclusion is that more processors and more powerful processors might resolve

TABLE 5.4
Total computing times (measured in seconds) obtained when six options of UNI-DEM are run on 8 processors

$NX \times NY$	2-D ($NZ = 1$)	3-D ($NZ = 10$)
96	5060	33491
288	70407	549801
480	355387	2559173

TABLE 5.5
Comparison of the computing times obtained on 8 processors and 32 processors with UNI-DEM discretized on a $480 \times 480 \times 10$ *grid. The speed-ups obtained in the transition from 8 processors to 32 processors are given in brackets.*

Process	8 processors	32 processors
Hor. adv. + diff.	986672	308035 (3.20)
Chem. + dep.	1055289	268978 (3.92)
Total	2559173	744077 (3.44)

many of the problems mentioned in the previous sub-section. Furthermore, the results in Table 5.5 indicate that some improvements in the advection part are desirable.

6. Some practical applications of DEM. Some results obtained by running DEM with meteorological and emission data for 1997 will be presented in this section. These results will demonstrate the usefulness of using fine resolution version of DEM.

The comparison of the concentration levels that are calculated by the model with the input levels of the emissions used is important. For species like SO_2, NO_2 and NH_3 the calculated by the model pollution levels should reflect the pattern of the emissions used.

We choose to make some comparisons for NO_2 concentrations in an area containing Denmark. The pattern of the corresponding NO_x emissions is seen in Fig. 5.1. It is seen that the largest emissions are in the regions of the three largest Danish cities (Copenhagen, Århus and Odense). This is not a surprise, because the traffic in cities is one of the major sources for the NO_x emissions.

The calculated by the coarse resolution version of the model pattern for the NO_2 concentrations is shown in Fig. 5.2. It is immediately seen that concentrations are smoothed very much when the coarse grid is used (and the pattern calculated by the model is not very similar to the input pattern of the related emissions.

The use of the fine resolution version of DEM calculates a pattern of the NO_2 concentrations which is clearly closer to the pattern of the NO_x emissions. This can be seen by comparing the highest concentration levels in Fig. 5.3 with with the highest emission levels in Fig. 5.1.

The distribution of the NO_2 concentrations in the whole model space domain are shown in Fig. 5.4. (note that the scale used in Fig. 5.4 is different from the scale used in Fig. 5.2 and Fig. 5.3). It is seen that Denmark is located between highly polluted regions in Central and Western Europe and regions in Scandinavia, which are not very polluted.

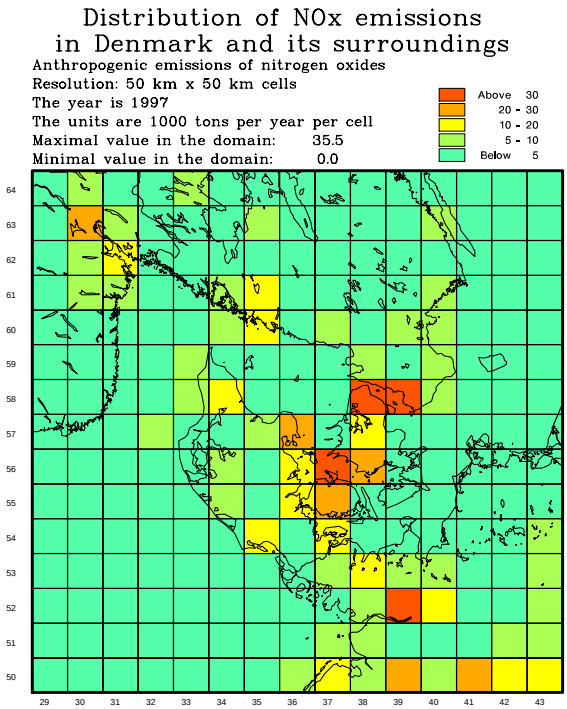


Fig. 5.1. Danish NO_x emissions in 1997

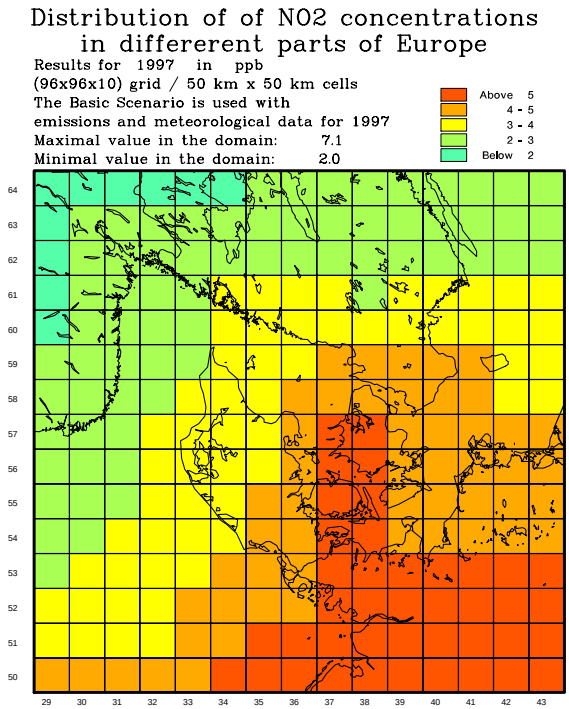


Fig. 5.2. NO_2 pollution in Denmark - coarse resolution

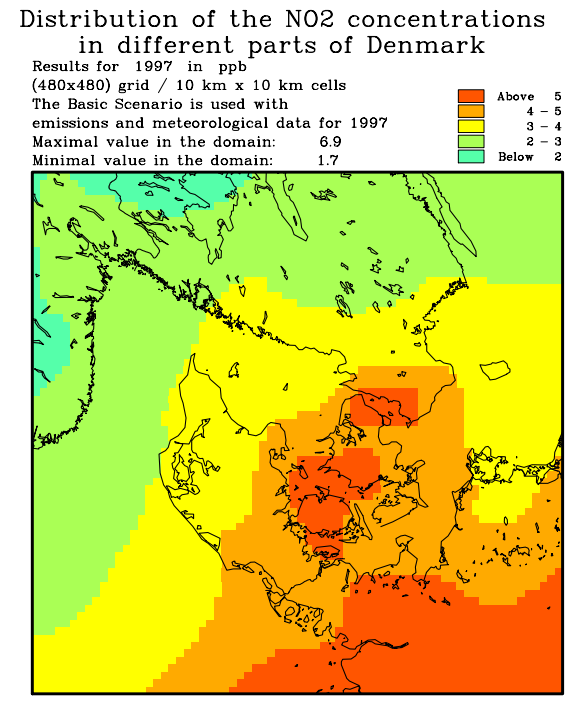


Fig. 5.3. NO_2 pollution in Denmark - fine resolution

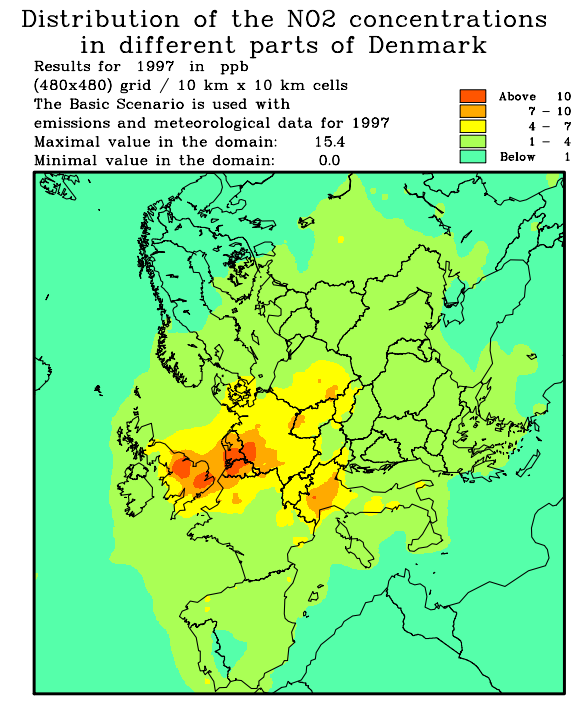


Fig. 5.4. NO_2 pollution in Europe - fine resolution

It should be mentioned here that the results in Fig. 5.2 and Fig. 5.3 are obtained by zooming in Fig. 5.4 to the region containing Denmark (and, as already mentioned, by changing the scale). Zooming might be used to get more details for the distribution of the NO_2 concentrations (or the concentrations of any other of the studied by the model chemical species) in any sub-domain of the space domain of DEM.

The results presented in Fig. 5.3 indicate that the fine resolution version is producing results which are qualitatively better than the results produced by the coarse resolution version (the areas with large emission sources, the the area around Copenhagen and the area around Århus-Odense can easily be seen in this plot). Quantitative validation of the models results can be obtained by comparing concentrations calculated by the model with measurements. Such comparisons were carried out in [2], [3], [14], [40] and [41] for the coarse resolution version. It is still very hard to carry out such extensive studies by using the fine resolution versions, but the results presented in this paper indicate that this will become possible in the near future.

Acknowledgements. A grant (CPU-1101-17) from the Danish Centre for Scientific Computing (DCSC) gave us access to the Sun computers at the Technical University of Denmark. The members of the staff of DCSC helped us to resolve some difficult problems related to the efficient exploitation of the grid of Sun computers.

REFERENCES

[1] V. Alexandrov, A. Sameh, Y. Siddique and Z. Zlatev, *Numerical integration of chemical ODE problems arising in air pollution models*, Environmental Modelling and Assessment, Vol. 2 (1997), 365–377.

[2] C. Ambelas Skjøth, A. Bastrup-Birk, J. Brandt and Z. Zlatev, *Studying variations of pollution levels in a given region of Europe during a long time-period*, Systems Analysis Modelling Simulation, Vol. 37 (2000), 297-311.

[3] A. Bastrup-Birk, J. Brandt, I. Uria and Z. Zlatev, *Studying cumulative ozone exposures in Europe during a seven-year period*, Journal of Geophysical Research, Vol. 102 (1997), 23917-23935.

[4] A. Bott, *A positive definite advection scheme obtained by non-linear renormalization of the advective fluxes*, Monthly Weather Review, Vol. 117 (1989), 1006-1015.

[5] K. Brenan, S. Campbell and L. Petzold, *Numerical solution of initial value problems in differential-algebraic equations*, SIAM, Philadelphia, 1996.

[6] R. A. Brost, *The sensitivity to input parameters of atmospheric concentrations simulated by a regional chemical model*, Journal of Geophysical Research, Vol. 93 (1988), 2371-2387.

[7] W. P. Crowley, *Numerical advection experiments*, Monthly Weather Review, Vol. 96 (1968), 1–11.

[8] P. Deuflhard, (1985). *Recent progress in extrapolation methods for ordinary differential equations.* SIAM Review, Vol. 27 (1985), 505-535.

[9] I. Dimov, I. Farago, A. Havasi and Z. Zlatev, *L-Commutativity of the operators in splitting methods for air pollution models*, Annales Univ. Sci. Budapest, Vol. 44, (2001), 129-150.

[10] R. Djouad and B. Sportisse, *Solving reduced chemical models in air pollution modelling*, Applied Numerical Mathematics, Vol. 40 (2003), 49-61.

[11] K. Georgiev and Z. Zlatev, *Parallel Sparse Matrix Algorithms for Air Pollution Models*, Parallel and Distributed Computing Practices, Vol. 2 (1999), 429-442.

[12] W. Gropp, E. Lusk and A. Skjellum, *Using MPI: Portable programming with the message passing interface*, MIT Press, Cambridge, Massachusetts (1994).

[13] M. W. Gery, G. Z. Whitten, J. P. Killus and M. C. Dodge, *A photochemical kinetics mechanism for urban and regional computer modeling,* Journal of Geophysical Research, Vol. 94 (1989), 12925–12956.

[14] A. Havasi and Z. Zlatev, *Trends of Hungarian air pollution levels on a long time-scale*, Atmospheric Environment, Vol 36 (2002), 4145-4156.

[15] E. Hesstvedt, Ø. Hov and I. A. Isaksen, *Quasi-steady-state approximations in air pollution modelling: comparison of two numerical schemes for oxidant prediction*, International Journal of Chemical Kinetics, Vol. 10 (1978), 971–994.

[16] Ø. Hov, Z. Zlatev, R. Berkowicz, A. Eliassen and L. P. Prahm, *Comparison of numerical techniques for use in air pollution models with non-linear chemical reactions*, Atmospheric Environment, Vol. 23 (1988), 967–983.

[17] W. Hunsdorfer, B. Koren, M. van Loon and J. G. Verwer, *A positive finite difference advection scheme*, J. Comput. Phys., Vol. 117 (1995), 35-46.

[18] J. D. Lambert, *Numerical methods for ordinary differential equations.* Wiley, New York (1991).

[19] D. Lancer and J. G. Verwer, *Analysis of operators splitting for advection-diffusion-reaction problems in air pollution modelling*, J. Comput.Appl. Math., Vol. 111 (1999), 201-216.

[20] M. van Loon, *Testing interpolation and filtering techniques in connection with a semi-Lagrangian method*, Atmospheric Environment, Vol. 27A (1993), 2351-2364.

[21] G. I. Marchuk, *Mathematical modeling for the problem of the environment*, Studies in Mathematics and Applications, No. 16, North-Holland, Amsterdam (1985).

[22] G. J. McRae, W. R. Goodin and J. H. Seinfeld, *Numerical solution of the atmospheric diffusion equations for chemically reacting flows*, Journal of Computational Physics, Vol. 45 (1984), 1–42.

[23] C. R. Molenkampf, *Accuracy of finite-difference methods applied to the advection equation,* Journal of Applied Meteorology, Vol. 7 (1968), 160–167.

[24] W. Owczarz and Z. Zlatev, *Running a large air pollution model on an IBM SMP computer*, International Journal of Computer Research, Vol. 10, No. 4 (2001), 321-330.

[25] W. Owczarz and Z. Zlatev, *Parallel matrix computations in air pollution modelling*, Parallel Computing, Vol. 28 (2002), 355-368.

[26] D. W. Pepper and A. J. Baker, *A simple one-dimensional finite element algorithm with multidimensional capabilities*, Numerical Heath Transfer, Vol. 3 (1979), 81-95.

[27] D. W. Pepper, C. D. Kern and P. E. Long, Jr., *Modelling the dispersion of atmospheric pollution using cubic splines and chapeau functions*, Atmospheric Environment, Vol. 13 (1979), 223–237.

[28] L. F. Shampine, M. W. Reichelt and J. A. Kierzenka, *Solving Index-1 DAEs in MATLAB and Simulink.* SIAM Rev., Vol. 41 (1999), 538-552.

[29] J. G. Verwer and M. van Loon, *An evaluation of explicit pseudo-steady state approximation for stiff ODE systems from chemical kinetics*, J. Comp. Phys., Vol. 113 (1996), 347-352.

[30] J. G. Verwer and D. Simpson, *Explicit methods for stiff ODE's from atmospheric chemistry*, Appl. Numer. Math., Vol. 18 (1995), 413-430.

[31] WEB-site for OPEN MP tools, http://www.openmp.org, 1999.

[32] WEB-site of the Danish Centre for Scientific Computing at the Technical University of Denmark, *Sun High Performance Computing Systems*, http://www.hpc.dtu.dk, 2002.

[33] Z. Zlatev, *Application of predictor-corrector schemes with several correctors in solving air pollution problems*, BIT, Vol. 24 (1984), 700–715.

[34] Z. Zlatev, *Computer treatment of large air pollution models*, Kluwer Academic Publishers, Dordrecht-Boston-London (1995).

[35] Z. Zlatev, *Partitioning ODE systems with an application to air pollution models*, Computers and Mathematics with Applications, Vol. 42 (2001), 817-832.

[36] Z. Zlatev, *Massive data set issues in air pollution modelling*, In: Handbook on Massive Data Sets (J. Abello, P. M. Pardalos and M. G. C. Resende, eds.), pp. 1169-1220, Kluwer Academic Publishers, Dordrecht-Boston-London (2002).

[37] Z. Zlatev, J. Christensen and A. Eliassen, *Studying high ozone concentrations by using the Danish Eulerian Model*, Atmospheric Environment, Vol. 27A (1993), 845-865.

[38] Z. Zlatev, J. Christensen and Ø. Hov, *An Eulerian air pollution model for Europe with nonlinear chemistry*, Journal of Atmospheric Chemistry, Vol. 15 (1992), 1-37.

[39] Z. Zlatev, I. Dimov and K. Georgiev, *Studying long-range transport of air pollutants*, Computational Science and Engineering, Vol. 1, No. 3 (1994), 45-52.

[40] Z. Zlatev, I. Dimov, Tz. Ostromsky, G. Geernaert, I. Tzvetanov and A. Bastrup-Birk, *Calculating losses of crops in Denmark caused by high ozone levels*, Environmental Modelling and Assessment, Vol. 6 (2001), 35-55.

[41] Z. Zlatev, J. Fenger and L. Mortensen, *Relationships between emission sources and excess ozone concentrations*, Computers and Mathematics with Applications, Vol. 32, No. 11 (1996), 101-123.

Modeling drug release from collagen matrices undergoing enzymatic degradation

M. Bause[1], W. Friess[2], P. Knabner[1], I. Metzmacher[2] and F. Radu[1,2],

[1] Institute of Applied Mathematics I, University of Erlangen-Nuremberg, 91058 Erlangen, Germany

[2] Department of Pharmacy, Pharmaceutical Technology and Biopharmacy, University of Munich, 81377 Munich, Germany

Abstract Dense collagen matrices for prolonged release of higher molecular weight drugs such as proteins or polysaccharides offer an alternative to implants based on synthetic polymers [1]. The hydrophilic matrix takes up large quantities of aqueous liquid upon contact with physiological fluids and swells. In order to reduce water uptake and to prolong the release, collagen matrices are chemically or physically crosslinked. A mathematical model was previously developed and tested for description of water penetration, swelling and drug release by assuming Fickian diffusion [2]. However, drug release from collagen matrices is not only governed by diffusion, but also by enzymatic degradation of the protein matrix. Consequently, a mathematical model to describe drug release from collagen matrices undergoing enzymatic degradation was established. For a numerical simulation the resulting equations were discretized, in space, by the mixed Raviart–Thomas finite element method of lowest order and, in time, by the backward Euler scheme. The mixed finite element method locally preserves mass which is an appreciable advantage of the approach. Two and three dimensional simulations were performed. The numerical results are in good agreement with the experimental measurements.

1 Introduction

Dense collagen matrices for prolonged release of higher molecular weight drugs such as proteins or polysaccharides offer an alternative to implants based on synthetic polymers [6]. The hydrophilic matrix takes up large quantities of aqueous liquid upon contact with physiological fluids and swells. In order to reduce water uptake and to prolong the release, collagen matrices are chemically or physically crosslinked. A mathematical model was previously developed and tested for description of water penetration, swelling and drug release by assuming Fickian diffusion [8]. However, drug release from collagen matrices is not only governed by diffusion, but also by enzymatic degradation of the protein matrix. Consequently, a mathematical model to describe drug release from collagen matrices must incorporate also the effect of the enzymatic degradation. A complete model was established, implemented and validated.

The outline of the rest of the paper is as follows. In the next section a mathematical model describing drug release from collagen matrices is presented. In Section 3 the model equations are discretized and a fully discrete numerical scheme based on a backward Euler discretization in time and a hybrid mixed finite element approximation in space is proposed. To confirm our theoretical approach, in Section 4 numerical results computed with realistic, experimentally determinated model parameters are presented and discussed. The last section contains some concluding remarks.

2 Mathematical Model

The hydrolytic degradation of a solid polymer matrix can occur by two extreme mechanisms. In the first one, refered to as *heterogeneous*, degradation is confined to the surface of the device and the undegraded carrier retains its chemical integrity during the process. In the other, called *homogeneous*, hydrolysis involves random cleavage at a uniform rate throughout the bulk of the matrix. While the molecular

weight of the polymer steadily decreases, the carrier can remain essentially intact until the polymer has undergone signifiant degradation, and reaches a critical molecular weight at which solubilization starts. Up to 90% of the matrix can degrade without signifiant mass loss (cf. [11]).

Here we assume *homogeneous* degradation. We also assume that part of the enzyme is adsorbed to the collagen fibers becoming immobile. Consequently, we have to consider free enzyme (E) and adsorbed (or immobilized) enzyme (E_i). The degradation reaction is catalyzed only from E_i. The product of the reaction, the hydrolized collagen, is not able to diffuse in the matrix until the concentration of collagen is less than 10 % (because of the still very large molecular weight). Thus, the processes to be mathematically described are:

- diffusion of the enzyme in the matrix,
- adsorption of the enzyme from the fluid to the collagen fibers,
- enzymatic degradation of the polymer,
- drug release.

To start with, let Ω (the polymer matrix) be a bounded domain in $\mathbb{R}^d$, $d = 1, 2$ or 3 with sufficiently smooth boundary $\Gamma := \partial\Omega$. Let $J = (0, T]$ be some finite interval with final time T. The diffusion of the enzyme in the matrix, considered to be Fickian, is described by

$$\partial_t C_E - \nabla \cdot (D_E(C_K)\nabla C_E) + \partial_t C_{E_i} = 0 \quad \text{in } J \times \Omega\,, \tag{1}$$

$$C_E = C_E^{ext} \quad \text{on } J \times \Gamma \tag{2}$$

with C_E, C_{E_i} denoting the concentrations, expressed in mol per volume, of the free and immobilized enzyme, respectively. In (2), C_E^{ext} is the enzyme concentration in the ambient medium. In (1), the term $\partial_t C_{E_i}$ models a delay in the diffusion process due to sorption. The initial concentration of the enzyme in the matrix is zero, and at the boundary $C_E = C_E^{ext}$ is prescribed. For the diffusion coefficient (D_E) of the enzyme in the matrix we assume a Fujita like dependence (free volume theory, cf. [7]) on the concentration of the collagen (C_K), i.e. $D_E = D_E^0 \exp\left(\beta \frac{C_K^0 - C_K}{C_K^0}\right)$, with D_E^0 denoting the diffusion coefficient of the enzyme in the undegraded matrix and β a dimensionless parameter.

The sorption itself is considered to be either an equilibrium process or a kinetic one, depending on its time scale compared to the diffusion process. In our case, the sorption is described as a kinetic reaction by

$$\partial_t C_{E_i} - k_p(\phi(C_E) - C_{E_i}) = 0\,, \tag{3}$$

where ϕ denotes a sorption isotherm and k_p a rate parameter. The sorption isotherm can admit one of the forms given in Table 1. The type of the sorption and the corresponding dimensionless parameters must be determined experimentally (cf. Section 4).

Having described the transport of the collagenase, including the absorption effects, we now proceed by modeling the enzymatic degradation of the polymer matrix. A direct consequence of the assumption of homogeneous degradation is that the degradation of the polymeric substrate by the environmental fluid is independent of the active agent. The general behaviour of an enzymatically catalyzed degradation process can be summarized by the equations

Enzyme + Substrate → Enzyme-Substrate Complex

and

Enzyme-Substrate Complex → Enzyme + Product .

In our case, collagen represents the substrate whereas the hydrolized collagen is the product. An analysis of these equations indicates that the reaction rate depends on a number of factors. First of

Linear	$\phi_{lin}(c) = K_d c$
Freundlich	$\phi_F(c) = K_d c^\alpha$
Langmuir	$\phi_L(c) = \dfrac{K_d c}{1 + \dfrac{K_d}{MaxSorp} c}$
Freundlich-Langmuir	$\phi_{FL}(c) = \dfrac{K_d c^\alpha}{1 + \dfrac{K_d}{MaxSorp} c^\alpha}$

Table 1: Sorption isotherms.

all, if the concentration of substrate increases, we would expect that the reaction rate also increases. However, because the reaction also depends on the amount of enzyme present, it can only increase until all of the available enzyme is fully employed. This behaviour is called saturation kinetics, due to the fact that as substrate concentrations become large, all of the enzyme is fully utilized (saturated) and reaction rates can no longer increase (cf. [5], p. 90). Such behaviour can mathematically be expressed by the Michaelis–Menten kinetic, which reads as

$$\partial_t C_K = -\mu_{max} \frac{C_K}{K_M + C_K} C_{E_i} \,, \tag{4}$$

where K_M, the Michaelis–Menten constant, is the collagen concentration at that the reaction rate is supposed to be half of the maximum rate (μ_{max}).

To model drug release, we assume (as also done, for instance, in [11]) that the initial load of active agent is composed of two pools: a pool of mobile active agent which is free to diffuse upon hydration of the matrix by the environmental fluid, and some part which is immobilized by the polymer and can diffuse only after the degradation of the matrix. The immobilization of the active agent is due to physical entrapment.

The release of the active agent is then governed by a diffusion equation with a source term due to liberation of the immobilized active agent by matrix degradation. Thus, writing a mass balance, we obtain that

$$\partial_t C_A - \nabla \cdot (D_A(C_K) \nabla C_A) = -\partial_t (C_{A_i}) \,, \tag{5}$$

where C_A, C_{A_i} denote the concentrations of free and immobilized drug, respectively. The active agent incorporated in the polymer matrix is released by a diffusive mechanism, under the influence of a concentration gradient. Due to the degradation process occuring concurrently, the matrix phase through which the diffusion takes place changes continuously as a function of the extend of hydrolysis of the polymer. Therefore, the diffusion coefficient of the active agent within the matrix can not be considered as constant but rather as a function of the fluid and (or) collagen concentration. Again, according to the free volume theory (cf. [7]), a possible form for the diffusion coefficient is given by

$$D_A = D_A^0 \exp\left(\alpha \frac{C_K^0 - C_K}{C_K^0}\right) \,, \tag{6}$$

with D_A^0 denoting the diffusion coefficient of the drug in the undegraded matrix, C_K^0 the initial concentration of collagen and α a dimensionless constant (cf. also [10]).

The diffusion is entangled because of the physical entrapment. In the equation this effect is described by the source term $\partial_t C_{A_i}$. To close the model, we still need a relation between the concentration of the immobilized and free active agent. In [11], it is simply assumed that $C_{A_i} = \sigma C_A$, with σ being a dimensionless constant, denoting the immobilizing capacity of the polymer, equal to the number of hindering crosslinks or entanglements per mole of (fully swollen) substrate. Instead, we assume a more

general functional dependence $C_{A_i} = f(C_A)$, where the precise form of f has to be determined by experimental studies. We remark that the particular choice $f(x) = \sigma x$ leads us to the model presented in [11]. The initial concentration of the free to diffuse active agent can be determined experimentally by measuring the quantity of drug which would remain in the matrix if no degradation occurs (no enzyme available). Dirichlet or Neumann boundary conditions complete the model.

3 Discretization

In this section we shall present a fully discrete numerical scheme approximating the system of partial and ordinary differential equations presented above. The backward Euler scheme is used for the temporal discretization and, in order to ensure local mass conservation, the mixed finite element method is applied for the spatial discretization. More precisely, Raviart–Thomas elements of lowest order are used for the approximation of the fluxes and piecewise constants for the concentrations. The equations (1), (3) and (4) are fully coupled and, therefore, solved simultaneously by a damped version of Newton's method. The linear systems of the Newton iteration are solved by a multigrid algorithm. Having determined the concentration of the collagen at time point t_n we can then solve the drug release equation (5) and proceed to the next time step (cf. Fig. 1).

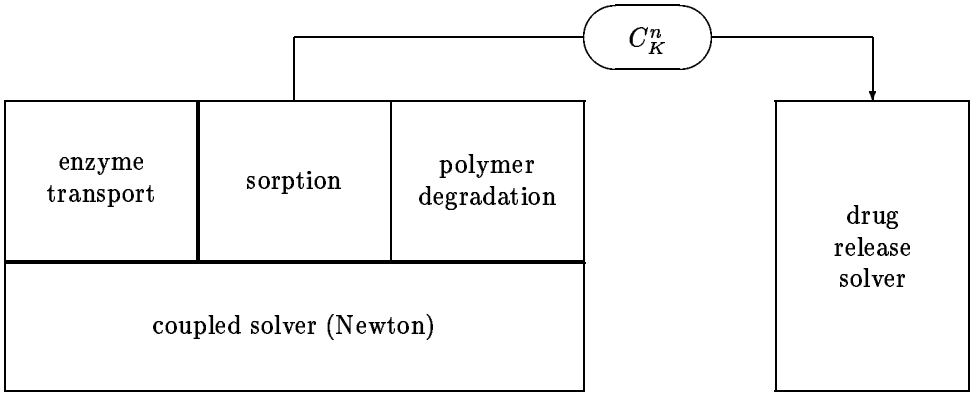


Figure 1: Schematic of the algorithm.

To explain the numerical algorithm, we rewrite the equations (1) and (5) as a first order system of equations by introducing the fluxes $\mathbf{q_E}$ and $\mathbf{q_A}$ of the free enzyme and drug, respectively, as unknowns. Thus we have

$$\partial_t C_E + \nabla \cdot \mathbf{q_E} + \partial_t C_{E_i} = 0\,, \tag{7}$$

$$\mathbf{q_E} = -D_E(C_K)\nabla C_E\,, \tag{8}$$

$$\partial_t C_A + \nabla \cdot \mathbf{q_A} = -\partial_t(C_{A_i})\,, \tag{9}$$

$$\mathbf{q_A} = -D_A(C_K)\nabla C_A\,. \tag{10}$$

In the following we need some more notation. Let $t_0 = 0 < t_1 < \ldots < t_N = T$ be a partition of J with time step size Δt. $L^2(\Omega)$ denotes the space of square integrable functions on Ω and $H(\mathrm{div};\Omega)$ the space of d-dimensional vector functions having all the components and the divergence in $L^2(\Omega)$. Let $\mathcal{T}_h$ be a decomposition of the domain $\Omega \subset \mathbb{R}^d$ into, depending on the dimension d, intervals, triangles or tetrahedrons. We denote by $\mathcal{S}_h$ the set of faces of $\mathcal{T}_h$. We compute approximations in the mixed finite

element spaces $W_h \times V_h \subset L^2(\Omega) \times H(\mathrm{div};\Omega)$ where

$$\begin{aligned} W_h &:= \{p \in L^2(\Omega) \mid \ p \text{ is constant on each element } T \in \mathcal{T}_h\}, \\ V_h &:= \{\mathbf{q} \in H(\mathrm{div};\Omega) \mid \ \mathbf{q}(\mathbf{x})_{|T} = \mathbf{a} + b\mathbf{x} \ \text{ for all } T \in \mathcal{T}_h, \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}\}. \end{aligned} \tag{11}$$

Hence, W_h denotes the space of piecewise constant functions, while V_h is the lowest order Raviart–Thomas (RT_0) space (cf. [4]). Consequently, for the concentrations of enzyme, free and immobilized, collagen and drug we consider determining approximations C_{ET}, C_{E_iT}, C_{KT} and C_{AT} being constant on each element T of the triangulation. This means we have one degree of freedom per concentration and per element. For the given decomposition (triangulation) of the matrix we then relax the continuity constraint on the normal components of the fluxes over interelement faces which is implied by $\mathbf{q} \in H(\mathrm{div};\Omega)$ to requiring only $\mathbf{q} \in H(\mathrm{div};T)$ for all $T \in \mathcal{T}_h$. The continuity of the normal fluxes is now enforced by an additional variational equation involving Lagrange multipliers $(\lambda_{ES}, \lambda_{AS})$ being defined on the faces and piecewise constant there, i.e., $(\lambda_{ES}, \lambda_{AS}) \in \Lambda_h := \{\lambda \in L^2(\mathcal{S}_h) \mid \lambda_{|S} = \text{ constant } \forall S \in \mathcal{S}_h\}$. This approach can be regarded as a hybridization of the problem; cf. [4] for details. After an elimination of internal degrees of freedom, also known as static condensation, the Lagrange multipliers become the unknowns of a global nonlinear system of equations which is solved by Newton's method and a multigrid algorithm as mentioned above. The Lagrange multipliers can further be considered as the degrees of freedom of higher order discretizations of the concentrations (cf. [4] for details) or as approximations of the concentrations on the element faces. This observation can then be exploited to calculate, without any significant extra computational effort, a reliable a posteriori error indicator; cf. [2].

More precisely, the algorithm works in the following way. We represent the unknowns in terms of basis functions of the respective finite element spaces; cf. [2] for details. The expansion coefficient of the discrete enzyme flux at time t_n with respect to the element T and its face S is denoted by $\mathbf{q}^{\mathbf{n}}_{\mathbf{E},\mathbf{TS}}$, similarly for the other variables. Then the nonlinear algebraic problem to be solved at time t_n for the transport of enzyme and polymer degradation reads as: *Given* $C^{n-1}_{ET}, C^{n-1}_{E_iT}, C^{n-1}_{KT}, \lambda^{n-1}_{ES}$. *Find* $C^n_{ET}, C^n_{E_iT}, C^n_{KT}, \mathbf{q}^{\mathbf{n}}_{\mathbf{E},\mathbf{TS}}, \lambda^n_{ES}$ *such that there holds:*

$$C^n_{ET} + C^n_{E_iT} + \frac{\Delta t}{V_T} \sum_{S \subset T} \mathbf{q}^{\mathbf{n}}_{\mathbf{E},\mathbf{TS}} = C^{n-1}_{ET} + C^{n-1}_{E_iT} \qquad \forall T \in \mathcal{T}_h, \tag{12}$$

$$C^n_{E_iT} - C^{n-1}_{E_iT} = \Delta t k_p(\phi(C^n_{ET}) - C^n_{E_iT}) \qquad \forall T \in \mathcal{T}_h, \tag{13}$$

$$C^n_{KT} - C^{n-1}_{KT} = -\Delta t \mu_{max} \frac{C^n_{KT}}{K_M + C^n_{KT}} C^n_{E_iT} \qquad \forall T \in \mathcal{T}_h, \tag{14}$$

$$\sum_{S' \subset T} \mathbf{B}_{\mathbf{TSS'}} \mathbf{q}^{\mathbf{n}}_{\mathbf{E},\mathbf{TS'}} = D_E(C^n_{KT})(C^n_{ET} - \lambda^n_{ES}) \qquad \forall T \in \mathcal{T}_h, \ S \subset T, \tag{15}$$

$$\mathbf{q}^{\mathbf{n}}_{\mathbf{E},\mathbf{TS}} = -\mathbf{q}^{\mathbf{n}}_{\mathbf{E},\mathbf{T'S}} \qquad \forall S \in \mathcal{S}_h, \ T, T' \supset S. \tag{16}$$

Here, V_T denotes the volume of the element T and $B_{TSS'}$ the entries in the stiffness matrix of the RT_0 functions. First, the flux variables are now eliminated locally and the nonlinear fully coupled equations (12)–(14) are solved by a damped Newton procedure for the (piecewise constant) concentrations. In the next step the global system corresponding to equation (16) is solved for the Lagrange multipliers, again, by a Newton method. Then we solve (12)–(14) again using the just computed values for λ^n_{ES}, and so on. The iteration is stopped when a given tolerance is reached.

As explained before, having calculated the collagen concentration at time point t_n, we can proceed with solving the equation of drug release. Similarly to the enzyme, the discrete problem now reads as: *Given* $C^{n-1}_{AT}, C^n_{KT}, \lambda^{n-1}_{AS}$. *Find* $C^n_{AT}, \mathbf{q}^{\mathbf{n}}_{\mathbf{A},\mathbf{TS}}, \lambda^n_{AS}$ *such that there holds:*

$$C^n_{AT} - C^{n-1}_{AT} + \frac{\Delta t}{V_T} \sum_{S \subset T} \mathbf{q}^{\mathbf{n}}_{\mathbf{A},\mathbf{TS}} = -f(C^n_{AT}) + f(C^{n-1}_{AT}) \qquad \forall T \in \mathcal{T}_h,$$

$$\sum_{S' \subset T} \mathbf{B_{TSS'}\, q^n_{A,TS'}} = D_A(C^n_{KT})(C^n_{AT} - \lambda^n_{AS}) \qquad \forall\, T \in \mathcal{T}_h,\ S \subset T\,,$$

$$\mathbf{q^n_{A,TS}} = -\mathbf{q^n_{A,T'S}} \qquad \forall\, S \in \mathcal{S}_h,\ T, T' \supset S\,.$$

The algorithm was implemented in the software toolbox *ug* (version 3.8, cf. [1]) and the computations were performed on a Sun Blade 1000 workstation.

4 Results and Discussion

In this section we shall present a real case study. First, we derive the model parameters by using experimental data. Then the results of a two-dimensional numerical simulation are shown.

Measurements of the diffusion coefficients D^0_E and D^0_A were performed with fluorescence correlation spectroscopy with dense collagen matrices, a method established in the group of Prof. Rädler (Department of Physics, LMU Munich); cf. [3] for details. We found $D^0_E = 1.76 \cdot 10^{-3}$ cm^2/h, $D^0_A = 2.1 \cdot 10^{-3}$ cm^2/h and $\beta = 2.0$. To determine the sorption isotherm ϕ and the sorption proportionality constant k_p, samples with different initial concentrations of enzyme were prepared. The concentrations of the free and adsorbed enzyme were measured after three hours and in the equilibrium state; cf. Figure 2. The sorption isotherm was fixed by fitting the data from the equilibrium state with the prescribed forms of the isotherm (cf. Table 1). The best fit was obtained for a Freundlich isotherm (cf. also [9]), $\phi(x) = a \cdot x^b$, with $a = 4.75$ and $b = 0.45$. The values after three hours and the resulting isotherm were then used to determine also the sorption proportionality constant $k_p = 0.021$/h. Since $b < 1$ here, a regularization was necessary for the numerical simulation. In our computations, the isotherm was replaced by the more regular function

$$\phi_\epsilon = \begin{cases} a \cdot \epsilon^b \cdot x/\epsilon & \text{for } x < \epsilon \\ a \cdot x^b & \text{otherwise} \end{cases}$$

with ϵ being a regularization parameter, chosen as 0.0001 in this study.

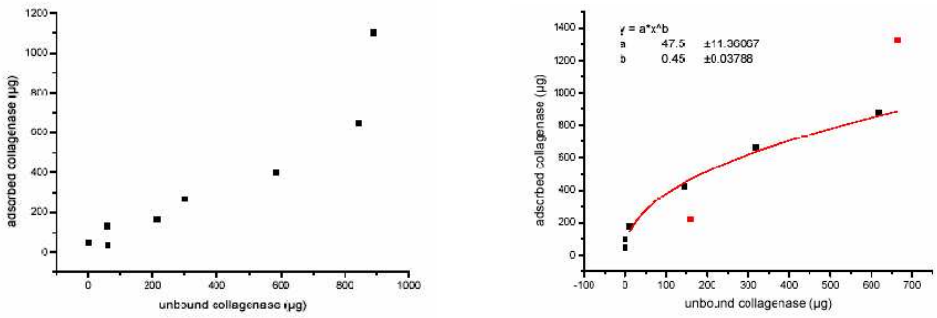


Figure 2: Adsorbtion of enzyme. Results after 3 hours (left) and 24 hours (right).

To get the Michaelis–Menten constant, typical Michaelis–Menten and Lineweaver–Burk diagrams were used; cf. [3]. We obtained $K_M = 3.915$ mg/cm^3 and a maximum rate $\mu_{max} = 1.218$ mg h/cm^3. Further parameters used in the numerical simulation are the following: dimension of the polymer matrix: 4.5 cm $\times$ 0.5 cm; initial concentration of collagen and drug in the matrix: $C^0_K = 227.0$ mg/cm^3 and $C^0_K = 2.55$ mg/cm^3, respectively; enzyme concentration at the boundary: $C^{ext}_E = 0.0067$ mg/cm^3. Finally, the dependence between the concentration of the free and immobilized active agent was supposed

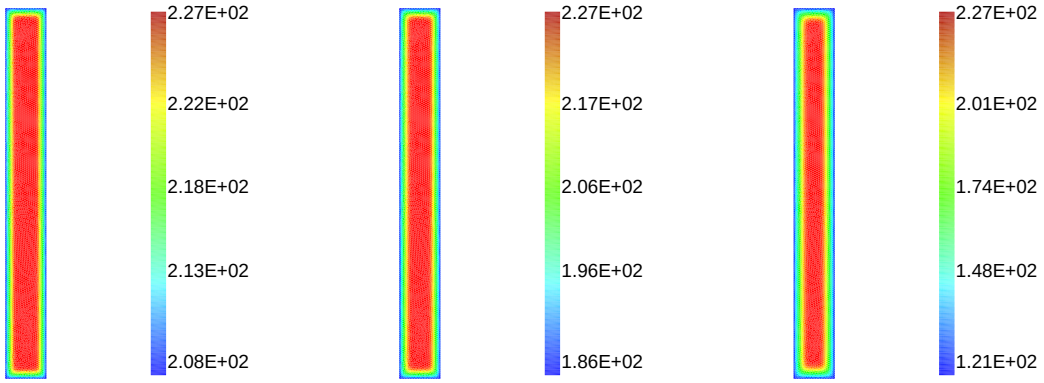


Figure 3: Concentration of collagen after three, five and ten days.

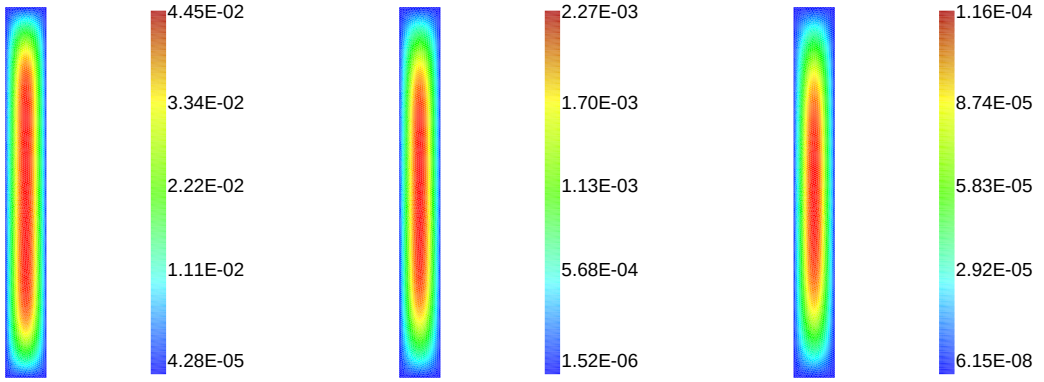


Figure 4: Concentration of the active agent after three, five and ten days.

to be linear ($f(x) = \sigma x$). We found $\sigma = 0.69$ by measuring, after some time long enough such that no release occurs anymore, the concentration of the drug in a collagen matrix which contains no enzyme such that no degradation arises.

The numerically computed concentrations of the degrading collagen after three, five and ten days are shown in Figure 3 and of the active agent in Figure 4. To validate the numerical results, additional independent experiments were done. We observed a good agreement between simulation and physical experiment.

5 Conclusions

A mathematical model for describing drug release from insoluble collagen matrices undergoing enzymatyic degradation was established, implemented and validated. Two dimensional simulations were performed. The numerical results nicely coincide with experimental data.

Acknowledgements.

This work was supported by the German Fundation Research, Grant FR 1089/4-1.

References

[1] P. Bastian et al., UG–a flexible toolbox for solving partial differential equations. Comput. Visualiz. Sci., 1 (1997), pp. 27–40.

[2] M. Bause and P. Knabner, Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods. Submitted to Adv. Water Resour., 2002.

[3] M. Bause, W. Friess, P. Knabner, I. Metzmacher and F. Radu, Modelling drug release from collagen matrices undergoing enzymatic degradation. Technical report, in preparation.

[4] F. Brezzi and M. Fortin, Mixed and Hybrid Finite Element Methods. Springer Verlag, New York, 1991.

[5] F.H. Chapelle, Ground-Water Microbiology and Geochemistry. John Wiley & Sons, New York, 2001.

[6] W. Friess, Drug Delivery Systems Based on Collagen. Habil. Thesis., Erlangen, 1999.

[7] H. Fujita, Diffusion in polymer-diluent systems. Fortschr. Hochpolym.-Forsch. 3 (1961), pp. 1-47.

[8] F. Radu et al. 2002, Modelling of drug release from collagen matrices. J. Pharm. Sci., Vol. 91, Issue 4 (2002), pp. 964-972.

[9] D.N. Rubingh and M.D. Bauer, Catalysis of Hydrolysis by Proteases at the Protein-Solution Interface. Polymer Solutions, Blends, and Interfaces, Elsevier Science Publishers B.V., 1992.

[10] A.G. Thombre and K.J. Himmelstein, Simultaneous Transport-Reaction Model fro Controlled Drug Delivery from Catalyzed Bioerodible Polymer Matrices. AIChE Journal, Vol. 31, No. 5 (1985), pp. 759-766.

[11] A.R. Tzafriri, Mathematical modeling of diffusion-mediated release from bulk degrading matrices. Journal of Controlled Release 63 (2000), pp. 69-79.

THE IMPACT OF INTRAPARTICLE CONVECTION ON THE MULTIPLICITY BEHAVIOUR OF LARGE-PORE CATALYST PARTICLES

Cristina Almeida Costa, Rosa M. Quinta-Ferreira*
Department of Chemical Engineering – University of Coimbra, Pólo II, 3030-290 Coimbra, PORTUGAL

Abstract
This paper analyses the impact of intraparticle convection on the steady-state multiplicity of isothermal large-pore catalyst particles with external mass and heat resistances, where a first order irreversible exothermal reaction is carried out. The coexistence of internal convection and diffusion results in a maximum of five steady-state solutions against the maximum of three obtained for diffusion only. The emergence of a second hysteresis loop due to convection increases the number of different possible types of bifurcation diagrams. The individual effects of diffusion and convection on the overall behaviour of the catalyst particle were determined.

Keywords: intraparticle convection, large-pore catalysts, steady-state multiplicity, bifurcation diagram

1. Introduction

The occurrence of multiple steady states in isothermal catalyst pellets due to the interactions between mass and thermal resistances in the fluid-solid interface and internal concentration gradients is a well-known feature of catalytic processes. If large-pore pellets are used instead of conventional porous catalysts, the additional mass transport by convection inside the particles must be accounted for besides diffusion, leading to changes in the internal concentration profiles (Rodrigues and Quinta-Ferreira, 1988) and consequently to a different pellet behaviour in what concerns steady-state multiplicity. This work deals with the study of the multiplicity features of an isothermal catalyst particle with simultaneous mass transport by diffusion and convection and a first order irreversible exothermal reaction (*o*-xylene oxidation to phthalic anhydride).
The mathematical technique developed by Balakotaiah and Luss (1982) for the global analysis of multiplicity features of lumped-parameter systems is used. When applied to the problem under study, this methodology allows the prediction of the maximal number of steady-state solutions for the particle temperature and the different types of bifurcation diagrams representing the particle temperature as a function of an operating variable such as concentration or temperature on the bulk phase.

2. Evaluation of the bifurcation sets and corresponding bifurcation diagrams

The dimensionless model equations describing the referred phenomena in a catalyst slab include the mass balance to the reactant inside the catalyst and the mass and thermal balances in the solid-fluid interface:

$$\frac{\partial^2 f_p}{\partial r_p^{*2}} - 2\lambda \frac{\partial f_p}{\partial r_p^*} - 4\phi^2 f_p \theta_s \exp\left[-\gamma\left(1/\theta_s - 1\right)\right] = 0 ; \quad r_p^* = 0 \text{ and } r_p^* = 1 : f_p = f_s \tag{1}$$

$$\left(1 - f_s\right) = \eta \, Da \, f_s \, \theta_s \exp\left[-\gamma\left(1/\theta_s - 1\right)\right] \tag{2}$$

$$\left(\theta_s - \theta_b\right) = \eta \beta \, Da \, f_s \, \theta_s \exp\left[-\gamma\left(1/\theta_s - 1\right)\right] \tag{3}$$

where subscripts *p* and *s* refer to particle and surface conditions, respectively. The model parameters are listed in Table 1. The numerical values indicated were obtained in previous studies as a function of the system properties evaluated at a reference temperature T_o=625 K, for R_p=0.0013 m (Quinta-Ferreira, Costa and Rodrigues, 1996).

Table 1 – Dimensionless system parameters.

Name	*Definition*	*Value*	*Name*	*Definition*	*Value*
Arrhenius number	$\gamma = E/R\,T_o$	21.8	intraparticle mass Peclet number	$\lambda = v_o\,R_p/D_e$	0,10,25,50
Damkhöler number	$Da = k(T_o)\,R_p/k_f$	8.7×10^{-3}	adiabatic temperature rise	$\beta = (-\Delta H)\,k_f\,C_b/h\,T_o$	0 – 3
Thiele modulus	$\phi = R_p\sqrt{k(T_o)/D_e}$	0.76	dimensionless bulk temperature	$\theta_b = T_b/T_o$	0.8 – 1.24

* Author to whom correspondence should be addressed.

Among these well-known parameters, the intraparticle mass Peclet number (λ) is the one accounting for the convective flow inside the particles; it represents the competition between the internal transport rates by convection and diffusion. When the convective flow is negligible in relation to diffusion, λ is set to zero and diffusion is the sole mechanism of transport accounted for, as it is commonly assumed in the classic mathematical treatment of porous particles.

The mass balance to the catalyst particle subjected to the corresponding boundary conditions (eq. (1)), has analytical solution, which allows an explicit expression for the catalyst effectiveness factor:

$$\eta(\theta_s,\phi,\lambda) = \frac{1/\alpha_1 - 1/\alpha_2}{\coth\alpha_1 - \coth\alpha_2} \tag{4}$$

where

$$\alpha_{1,2}(\theta_s,\phi,\lambda) = \lambda/2 \pm \sqrt{\lambda^2/4 + \phi^2 \exp\left[-\gamma(1/\theta_s - 1)\right]} \tag{5}$$

By combining the mass and energy balances in the solid/fluid interface, a single algebraic equation is obtained, as a function of the dimensionless pellet temperature, θ_s, and the vector **p,** containing the six model parameters (θ_b, β, Da, γ, ϕ, λ):

$$F(\theta_s,\mathbf{p}) = \theta_s - \theta_b - \frac{\beta\,\mathrm{Da}\,\eta(\theta_s,\lambda,\phi)\,\theta_s \exp\left[-\gamma(1/\theta_s - 1)\right]}{1 + \mathrm{Da}\,\eta(\theta_s,\lambda,\phi)\,\theta_s \exp\left[-\gamma(1/\theta_s - 1)\right]} = 0 \tag{6}$$

where η is given by eq.(4). The dimensionless solid temperature θ_s is bounded by θ_b and $\theta_b + \beta$.

In the forthcoming analysis, four of the six model parameters - Da, γ, ϕ, λ - are fixed (see values in Table 1) in order to reduce the problem dimension. The solution of the resulting equation $F(\theta_s,\theta_b,\beta)=0$ is a three-dimensional surface called the steady-state manifold, while the simultaneous solution of $F(\theta_s,\theta_b,\beta)=0$ and $\partial F(\theta_s,\theta_b,\beta)/\partial\theta_s=0$ defines the singular set. When this set is projected in the θ_s direction, by eliminating θ_s from these last two equations, a two-dimensional bifurcation set is obtained in the θ_b-β plane. These graphs demarcate the regions of β (linearly dependent on reactant bulk concentration) and θ_b (linearly dependent on bulk temperature), for which a different number of steady-state solutions of equation (6) exist. Figures 1 show the bifurcation sets obtained for different values of λ: 0, 10, 25 and 50, where the number of steady-state solutions (1, 3 or 5) is indicated.

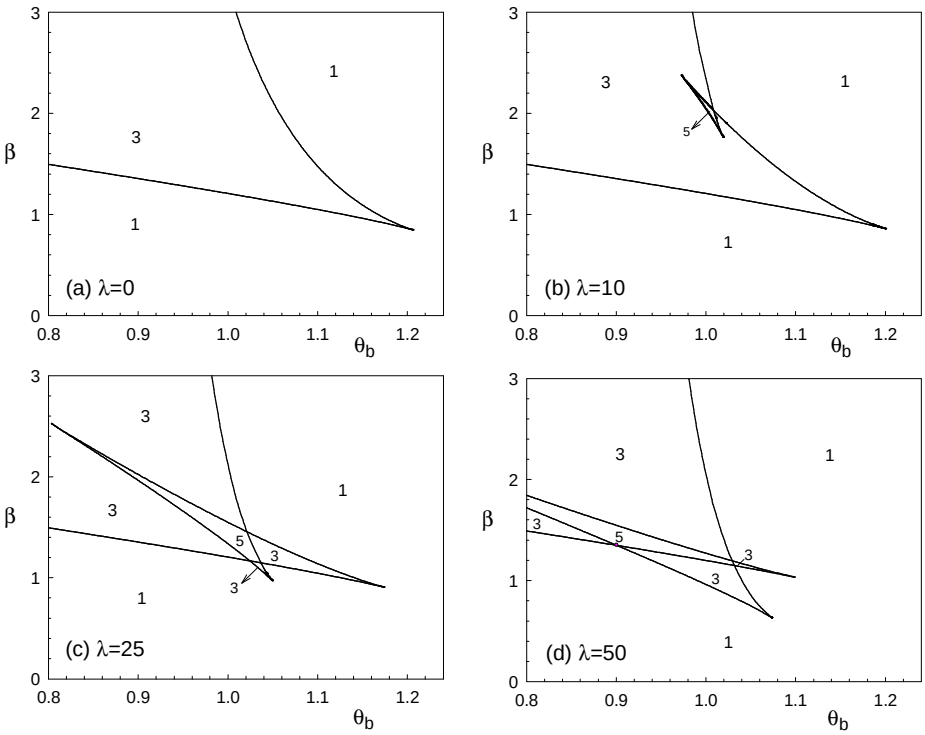


Figure 1 – Bifurcation sets for different values of λ (Da=0.0087, γ=21.8, ϕ=0.76).

For λ=0, the pellet temperature exhibits at most 3 solutions (Fig. 1 (a)) and the lines shown in the graph are the locus of extinction and ignition points. However, when intraparticle convection is taken into consideration (λ>0), the shape of the bifurcation set changes significantly and a maximum of 5 multiple solutions can be achieved (Figs. 1 (b), (c) and (d)), as a result of the joint effect of diffusion and convection.

The curves showing the dependence of the state variable θ_s on a bifurcation variable, β or θ_b, are nominated bifurcation diagrams. Each bifurcation set can be divided in a number of regions where the bifurcation diagrams have different shapes. In this paper the bifurcation diagrams θ_s *vs* θ_b are analysed. In such case, the shape of these bifurcation diagrams can change if an ignition or extinction point exists at any boundary of θ_b or either if β crosses one of the following varieties:

a) the hysteresis variety, which is the locus of all feasible values of β that satisfy the equations

$$F(\theta_s,\theta_b,\beta)=\frac{\partial F(\theta_s,\theta_b,\beta)}{\partial\theta_s}=\frac{\partial^2 F(\theta_s,\theta_b,\beta)}{\partial\theta_s^2}=0 \tag{7}$$

b) the double limit variety, which is the locus of all feasible values of β that satisfy the equations

$$F(\theta_{s1},\theta_b,\beta)=F(\theta_{s2},\theta_b,\beta)=\frac{\partial F(\theta_{s1},\theta_b,\beta)}{\partial\theta_s}=\frac{\partial F(\theta_{s2},\theta_b,\beta)}{\partial\theta_s}=0,\quad \theta_{s1}\neq\theta_{s2} \tag{8}$$

Figure 2 (a) shows again the bifurcation set obtained for λ=0 and the corresponding qualitative features of the θ_s *vs* θ_b bifurcation diagrams that can be found in each one of the intervals of β: $0\leq\beta<0.85$, $0.85<\beta<1.5$ and $1.5<\beta\leq3$ – Figs. 2 (b), (c) and (d), respectively. The transition from region **I** to **II** occurs because β crosses a hysteresis variety, which appears in the bifurcation set as a cusp point, marked with a C, for which β=0.85, while the change from region **II** to **III** is due to the existence of an extinction point at the lower boundary of θ_b (θ_b=0.8, β=1.5). Figure 2 (b) shows a bifurcation diagram characteristic of the uniqueness region **I**, where θ_s is a single value function of θ_b. In region **II** two limit points (extinction and ignition) arise and a hysteresis loop emerges as shown in bifurcation diagram of Fig. 2 (c), where the 1-3-1 multiplicity pattern is represented by a S-shaped curve. Extinction and ignition points are indicated by arrows pointing down and up, respectively, and a dashed line is used to represent the unstable steady-state solution. In region **III**, the extinction point is below the lower boundary specified for θ_b, therefore the corresponding bifurcation diagrams show just the ignition point, as depicted in Figure 2 (d).

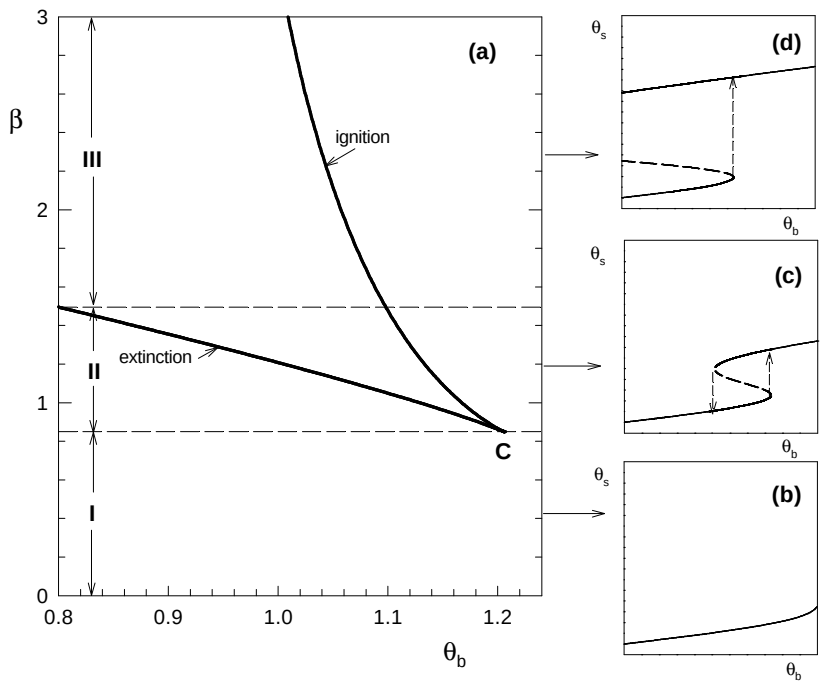


Figure 2 – Bifurcation set and bifurcation diagrams obtained for λ=0 (Da=0.0087, γ=21.8, ϕ=0.76).

For $\lambda>0$, the bifurcation set is more complex and some more intricate bifurcation diagrams can be obtained, as illustrated in Figures 3, for $\lambda=50$. The bifurcation set (Fig. 3 (a)), is now divided in nine subintervals of β numbered from **I** to **IX**, where the bifurcation diagrams θ_s *vs* θ_b have different shapes, exemplified in the nine small graphs of Figs. 3 (b) –3 (i).

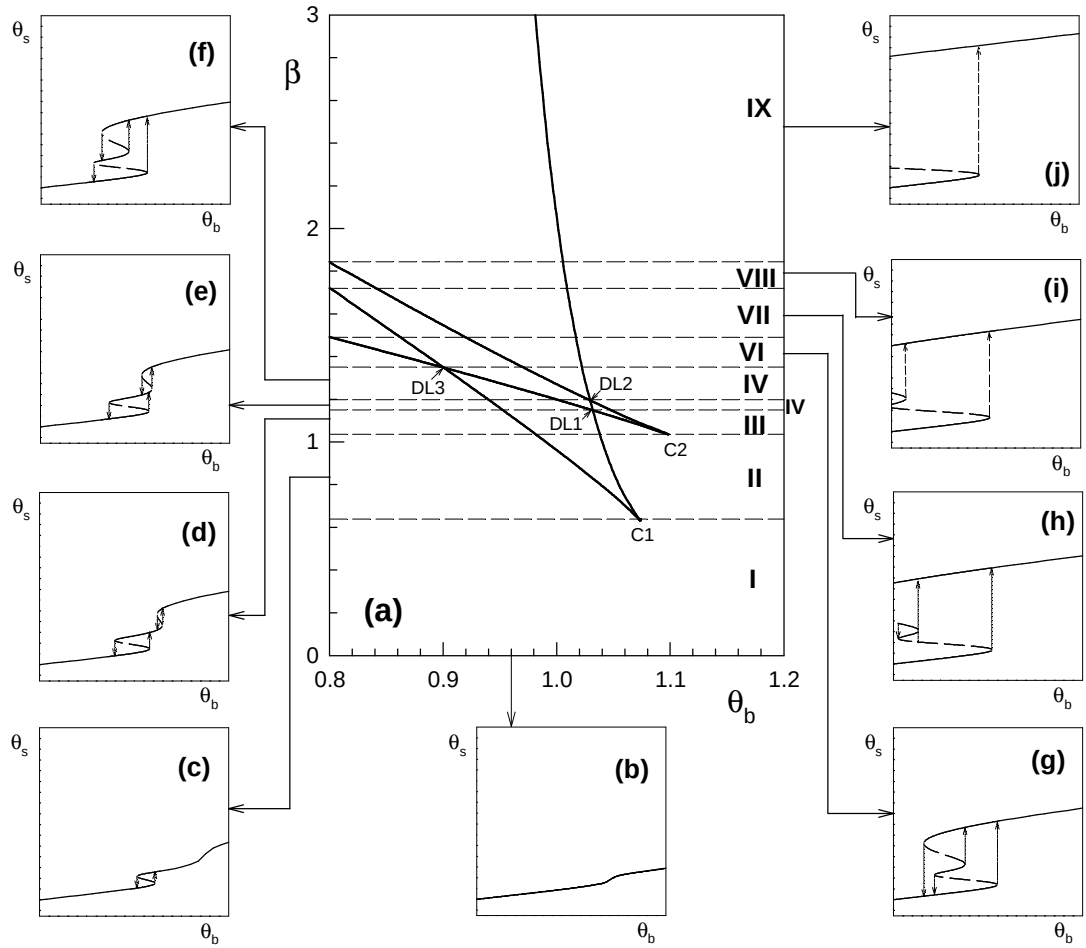


Figure 3 – Bifurcation set and bifurcation diagrams obtained for $\lambda=50$ (Da=0.0087, γ=21.8, ϕ=0.76).

The transitions between regions **I**-**II** and **II**-**III** occur when β crosses the two hysteresis varieties corresponding to the two cusp points C1 and C2 indicated on the bifurcation set, respectively. The bifurcation diagrams typical of regions **I** and **II**, shown in Figs. 3 (b) and (c), are similar to those found for $\lambda=0$; however, in region **III** a second hysteresis loop emerges due to convection, generating a 1-3-1-3-1 multiplicity pattern – Fig. 3 (d). The following transitions between regions **III**-**IV**, **IV**-**V** and **V**-**VI** happen when β traverses double-limit points DL1, DL2 and DL3, respectively. This type of variety results from the intersection of two branches of the bifurcation set (see Fig. 3 (a)), with different values of θ_s at each branch. When β crosses a double-limit variety, the number of limit points (extinction and ignition points) in the bifurcation diagrams does not change, but the relative position of two limit points changes, as it can be observed by comparing the different 1-3-5-3-1 multiplicity patterns of Figs. 3 (e), (f) and (g). In these cases the two even solutions for θ_s, represented by dashed lines are unstable. The remaining transitions between regions **VI**-**VII**, **VII**-**VIII** and **VIII**-**IX** arise from the intersection of the two extinction branches and one ignition branch of the bifurcation set with the lower boundary of the bifurcation variable θ_b, leading to the type of bifurcation diagrams shown in Figs. 3 (h), (i) and (j). The bifurcation diagram of Fig. 3 (j) is qualitatively similar to the one depicted in Fig 2 (d) obtained for $\lambda=0$.

3. The individual effect of diffusion and convection on particle multiplicity

In order to get a better understanding of the relative contribution of diffusion and convection to the global behaviour of the catalyst particle, the individual effect of each one of these mechanisms of mass transport (coupled with chemical reaction) was evaluated.

3.1 The influence of diffusion

In conventional porous catalysts pellets, the mass transport rate by convection is usually negligible when compared to diffusion. As referred before, for an isothermal particle this situation is mathematically described by the mass balance of eq. (1), taking $\lambda=0$. The corresponding effectiveness factor is given by eqs. (4) and (5) with $\lambda=0$, or alternatively by:

$$\eta(\theta_s,\phi)=\frac{\tanh\left\{\phi\exp\left[-\gamma\left(1/\theta_s-1\right)\right]\right\}}{\phi\exp\left[-\gamma\left(1/\theta_s-1\right)\right]} \tag{9}$$

Combining again equations (2) and (3) a unique equation $F(\theta_s, \mathbf{p^d})=0$ is obtained, with the same expression of equation (6), but with $\eta(\theta_s,\phi)$ given by eq. (9) and $\mathbf{p^d}=(\theta_b,\beta,Da,\gamma,\phi)$. In the following analysis, the singular set was calculated by solving equations $F(\theta_s,\mathbf{p^d})=0$ and $\partial F(\theta_s,\mathbf{p^d})/\partial\theta_s=0$ for fixed values of Da, γ and ϕ. The resulting bifurcation sets are shown in Figure 4 for different values of ϕ: $\phi=0$, 0.08, 0.24, 0.76 and 2.42 correspondingly to $D_e/D_e(T_o)=\infty$, 100, 10, 1 and 0.1. For $\phi=0$, the solution was obtained by solving the system of equations with $\eta(\theta_s,\phi)=1$.

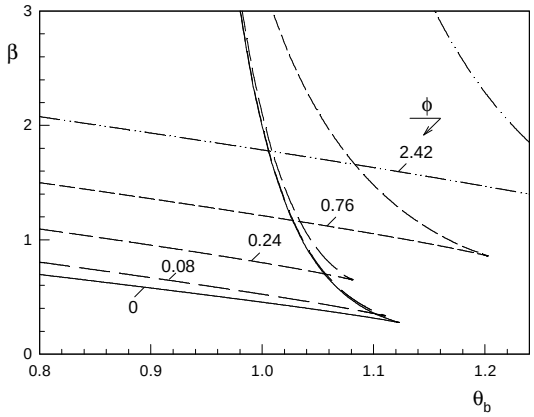


Figure 4 – Bifurcation sets for different values of ϕ (Da=0.0087, γ=21.8).

For each ϕ, three steady-state solutions exist inside the regions bounded by the extinction and ignition lines, while a unique solution exists in the complementary space. The multiplicity region moves in the β-θ_b plane as ϕ changes, in such a way that for increasing values of ϕ both extinction and ignition branches appear at higher values of β and θ_b.

3.2 The influence of convection

If a large-pore catalyst particle is considered with such a porous structure that transport rate by convection is far greater than diffusion, very large values of λ are obtained. A rearrangement of equation (1) gives:

$$\frac{1}{2\lambda}\frac{\partial^2 f_p}{\partial r_p^{*2}}-\frac{\partial f_p}{\partial r_p^*}-2Da_p\, f_p\,\theta_s\exp\left[-\gamma\left(1/\theta_s-1\right)\right]=0\,;\;\; r_p^*=0 \text{ and } r_p^*=1: f_p=f_s \tag{10}$$

where Da_p is a new dimensionless parameter, the *particle Damkhöler number*, expressing the competition between reaction rate and transport rate by intraparticle convection:

$$Da_p=\frac{\phi^2}{\lambda}=\frac{k(T_o)R_p}{v_o} \tag{11}$$

To study the multiplicity features of this catalyst, a very small effective diffusivity of $D_e=2.79\times10^{-10}$ m^2/s (10^4 times smaller than $De(T_o)$) was fixed, in order to assure mass transport by diffusion negligible when compared to

convection. Some bifurcation sets are going to be obtained for several values of Da_p, calculated for different values imposed to the intraparticle fluid velocity v_o. With fixed values for v_o and D_e, λ can be calculated (see Table 1) and thus it is no longer an independent parameter. A unique equation $F(\theta_s, \mathbf{p}^c)=0$ is then obtained by eliminating f_s from equations (4) and (5), being the parameter vector $\mathbf{p}^c=(\theta_b,\beta,Da,\gamma,Da_p)$ and $\eta(\theta_s,Da_p)$ calculated by eqs. (4)-(5) with ϕ^2 replaced by the product $Da_p \times \lambda$. The singular set was evaluated by solving equations $F(\theta_s,\mathbf{p}^c)=0$ and $\partial F(\theta_s,\mathbf{p}^c)/\partial\theta_s=0$ for fixed values of Da_p: Da_p=0, 0.0012, 0.0117, 0.0234, 0.0584 correspondingly to v_o= ∞, 1, 0.1046, 0.0523, 0.0209 and the resulting bifurcation sets are shown in Figure 5.

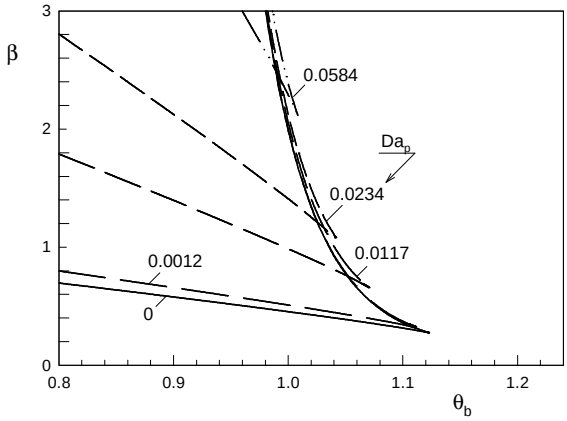


Figure 5 – Bifurcation sets for different values of Da_p (Da=0.0087, γ=21.8, $D_e=2.79\times10^{-10}$ m^2/s).

The bifurcation sets obtained for the different values of Da_p are similar to the ones of Figure 4, with three steady-state solutions inside the regions bounded by the extinction and ignition branches and a unique solution outside. For increasing values of Da_p, the multiplicity region decreases mainly due to the displacement of the extinction branch towards higher values of β.

3. The combined effect of diffusion and convection on particle multiplicity

The relative influence of diffusion and convection on the overall behaviour of the particle when both mass transfer mechanisms coexist can be assessed by comparing the results of the complete model, that includes convection and diffusion, (*dc* model) to those of the models that consider convection and diffusion separately (*c* and *d* models, respectively). To perform such comparisons, the parameters of the different models must be matched, such that λ of the *dc* complete model is the result of ϕ and Da_p used in *d* and *c* models:

$$\lambda(\mathrm{dc}) = \frac{\phi^2(\mathrm{d})}{Da_p(\mathrm{c})} \tag{12}$$

In Figure 6, the bifurcation sets obtained for the complete model and the curves resulting from the isolated effect of each one of the transport mechanisms are plotted together. In each graph, the value of λ used in the *dc* model and the values of ϕ and Da_p used in *d* and *c* models satisfy eq. (12). The sets of parameters used are indicated in Figs. 6 (a), (b) and (c). The number of steady-state solutions (1, 3 or 5) in the different regions of the graphs was not indicated in order to avoid the graphs overload, but it is easily identifiable from the previous bifurcation sets.

The influence of each transport mechanism on the overall behaviour of the catalyst particle is obvious: while the extinction branch of the complete model is mainly governed by diffusion, the location of the ignition branch is highly influenced by intraparticle convection. When the relative importance of convection over diffusion increases, the shape of the multiplicity region of the *dc* model becomes more identified with the one obtained for convection only, as shown in Fig. 6 (c) for λ=50. The regions of five steady-states appear for those operating conditions (β, θ_b) for which diffusion and convection have approximately the same relative importance.

Finally, the discussion of the individual effect of diffusion and convection on the overall behaviour of the particle is complemented with the analysis of the different types of bifurcation diagrams θ_s *vs* θ_b corresponding to Figure 6 (c). For λ=50, the *dc* model predicts nine different types of these bifurcation diagrams shown in Figures 3 (b)-(i). In Figures 7, these are compared with the bifurcation diagrams obtained through *d* and *c* models.

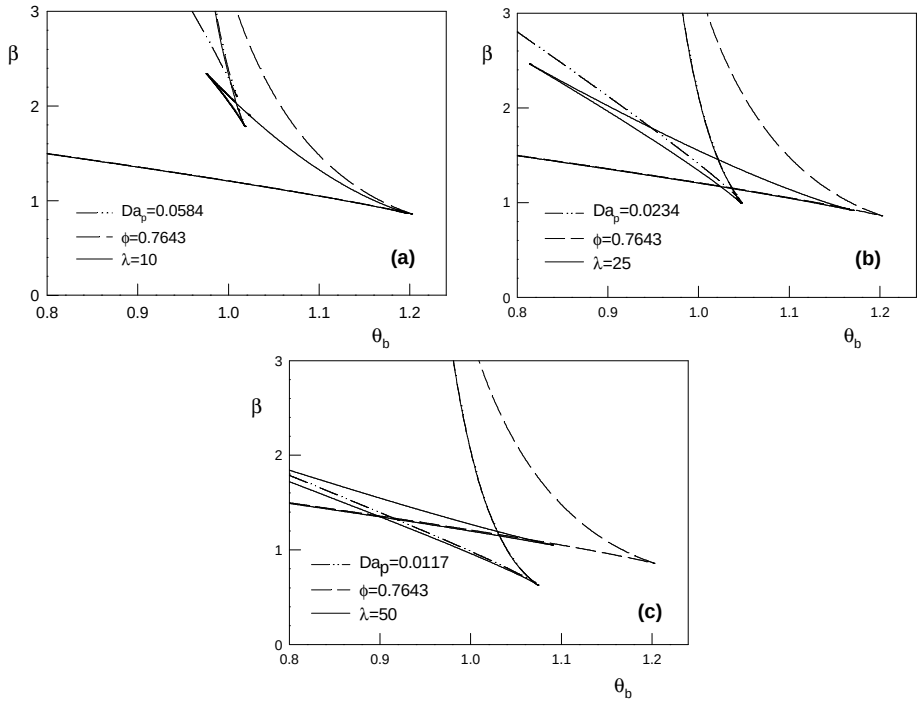


Figure 6 – Bifurcation sets obtained for *dc* model (λ=10,25,50), *d* model (ϕ=0.7643) and *c* model (Da_p=0.0584,0.0234,0.0117); (Da=0.0087, γ=21.8).

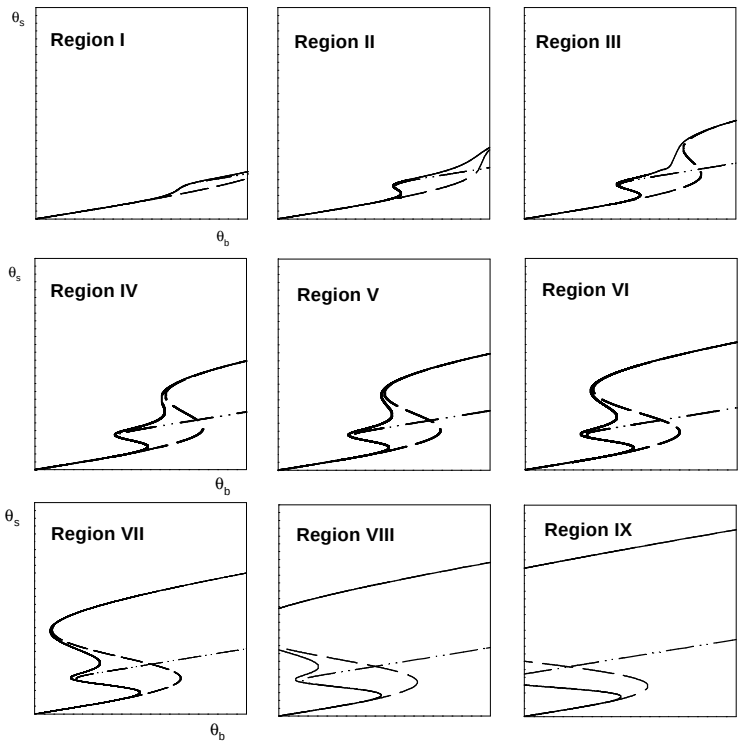


Figure 7 – Bifurcation diagrams obtained for *dc* model (λ=50), *d* model (ϕ=0.7643) and *c* model (Da_p=0.0117); (Da=0.0087, γ=21.8).

Some general comments can be done based on Figures 7. In the lower temperature branches, the particle operates in chemical regime, with chemical reaction controlling the overall process rate and therefore in each graph the three curves remain coincident for lower values of θ_s. On the other hand, in the higher temperature branch, the particle operates in difusional regime and consequently the predictions of *dc* model follow the evolution of *d* model. When multiplicity occurs, the first hysteresis loop is clearly due to convection, as shown by the overlapping of the curves from *dc* and *c* models, while the appearance of a second hysteresis loop in region **IV** is caused by diffusion.

4. Summary and Conclusions

This study showed interesting multiplicity features produced by the coexistence of internal diffusion and convection on an isothermal large-pore catalyst particle with external resistances. By using the technique developed by Balakotaiah and Luss (1982), a maximum of five steady-state solutions was calculated for the particle temperature. Moreover, nine different types bifurcation diagrams representing the dependence of particle temperature on the dimensionless bulk temperature were determined, some of them showing two hysteresis loops. The separate effects of diffusion and convection on the pellet multiplicity were evaluated and related to the overall catalyst performance when the two mechanisms act simultaneously.

Notation

C – reactant concentration, mol/m^3
D_e - reactant effective diffusivity, m^2/s
Da - Damkhöler number, dimensionless
E – activation energy, J/mol
f – dimensionless reactant concentration, C/C_o
h – film heat transfer coefficient, J/m^2 s
k – rate constant, 1/s
k_f – film mass transfer coefficient, mol/m^2 s
R – perfect gas constant, J m^3/mol K
R_p - half-thickness of the slab, m
T – temperature, K
v_o - intraparticle fluid velocity, m/s

β - adiabatic temperature rise, dimensionless
ϕ - Thiele modulus, dimensionless
γ - Arrhenius number, dimensionless
η- effectiveness factor, dimensionless
λ - intraparticle mass Peclet number, dimensionless
θ - dimensionless temperature, T/T_o
ΔH – heat of reaction, J/mol

Superscripts:
c – convection
d - diffusion

Subscripts:
b – bulk phase conditions
s – pellet surface conditions
o – reference conditions

References

Balakotaiah, V. & Luss, D. (1982). Structure of the steady-state solutions of lumped-parameter chemically reacting systems. Chemical Engineering Science, 32, 1611-1623.

Rodrigues, A. E. & Quinta-Ferreira, R. M. (1988). Convection, diffusion and reaction in a large-pore catalyst particle, AIChE Symposium Series, 84, 80-87.

Quinta-Ferreira, R. M., Costa, C. A. & Rodrigues, A. E. (1996) . Effect of intraparticle convection on the transient behaviour of fixed-bed reactors: finite differences and collocation methods for solving unidimensional models, Computers and Chemical Engineering, 20, 1201-1225.

Mass Transfer and Dispersion Around an Active Sphere Buried in a Packed Bed of Inerts

J.M.P.Q.Delgado[(1)], M.A.Alves[(2)] and J.R.F. Guedes de Carvalho[(2)]
(jdelgado@fe.up.pt; mmalves@fe.up.pt; jrguedes@fe.up.pt)

[(1)] Departamento de Engenharia Civil;
[(2)] Departamento de Engenharia Química
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, 4200-418 Porto – PORTUGAL

ABSTRACT

The dissolution of a slightly soluble sphere buried in a packed bed of sand, through which water flows is considered in the present work with due consideration given to the processes of transverse and longitudinal dispersion. Numerical solution of the equations was undertaken to obtain point values of the Sherwood number, as a function of the Peclet and Schmidt numbers over a wide range of values of the relevant parameters. A correlation is proposed that describes accurately the dependence found numerically between these dimensionless parameters.

INTRODUCTION

In several situations of practical interest a large solid mass interacts with the liquid flowing around it through the interstices of a packed bed of inerts. Examples are the leaching of buried rocks and the contamination of underground waters by compacted buried waste. The dissolution of a slightly soluble sphere buried in a packed bed of sand, through which water flows, is a useful model for such processes, and it is considered in the present work.

In a recent study on transverse dispersion in liquids, Delgado and Guedes de Carvalho [1] showed that there is a significant dependence between the transverse dispersion coefficient (D_T) and the Schmidt number (Sc), for $Sc < 550$. Since the rate of mass transfer around a buried sphere, exposed to a flowing fluid, is strongly determined by D_T [2], it may be expected that mass transfer from a buried sphere will show a significant dependence on Sc.

THEORY

In terms of analysis, we consider the situation of a slightly soluble sphere of diameter d_1 buried in a bed of inert particles of diameter d (with $d << d_1$), packed uniformly (void fraction ε) around the sphere. The packed bed is assumed to be "infinite" in extent and a uniform interstitial velocity of liquid, $\boldsymbol{u}_0$, is imposed, at a large distance from the sphere.

[(1)] Corresponding author.

Darcy's law, $\boldsymbol{u} = -K\ \mathbf{grad}\ p$, is assumed to hold, and if it is coupled with the continuity relation for an incompressible fluid, $\mathrm{div}\,\boldsymbol{u} = 0$, Laplace's equation $\nabla^2\phi = 0$ is obtained for the flow potential, $\phi = K\ p$, around the sphere.

In terms of spherical coordinates (r, θ), the potential and stream functions are, respectively [2],

$$\phi = -u_0\left[1+\frac{1}{2}\left(\frac{R}{r}\right)^3\right] r\cos\theta \tag{1}$$

$$\psi = \frac{u_0}{2}\left[1-\left(\frac{R}{r}\right)^3\right] r^2\sin^2\theta \tag{2}$$

and the velocity components are

$$u_r = \frac{\partial\phi}{\partial r} = -u_0\cos\theta\left[1-\left(\frac{R}{r}\right)^3\right] \tag{3}$$

$$u_\theta = \frac{1}{r}\frac{\partial\phi}{\partial\theta} = u_0\sin\theta\left[1+\frac{1}{2}\left(\frac{R}{r}\right)^3\right]. \tag{4}$$

The analysis of mass transfer is based on a steady state material balance on the solute crossing the borders of an elementary volume, limited by the constant potential surfaces ϕ and $\phi+\delta\phi$, and the stream surfaces ψ and $\psi+\delta\psi$. The resulting equation is [2],

$$\frac{\partial c}{\partial\phi} = \frac{\partial}{\partial\phi}\left(D_\mathrm{L}\frac{\partial c}{\partial\phi}\right) + \frac{\partial}{\partial\psi}\left(D_\mathrm{T}\omega^2\frac{\partial c}{\partial\psi}\right) \tag{5}$$

The boundary conditions, to be observed in the numerical integration of Eq. (1), are: (*i*) the solute concentration is equal to the background concentration, c_0, far away from the sphere; (*ii*) the solute concentration is equal to the equilibrium saturation concentration, $c = c^*$, on the surface of the sphere and (*iii*) the concentration field is symmetric about the flow axis.

In order to integrate Eq. (5), with the auxiliary Eqs. (1) and (2), it is convenient to define the dimensionless variables:

$$C = \frac{c-c_0}{c^*-c_0} \tag{6}$$

$$U = \frac{u}{u_0} = \frac{\left(u_r{}^2+u_\theta{}^2\right)^{1/2}}{u_0} \tag{7}$$

$$\Re = \frac{r}{R} \tag{8}$$

$$\Phi = \frac{4}{3}\frac{\phi}{u_0 d_1} \tag{9}$$

$$\Psi = \frac{\psi}{u_0 d_1^2} \tag{10}$$

Equation (5) may be re-arranged to

$$\frac{\partial C}{\partial \Phi} = \frac{\partial}{\partial \Phi}\left[\left(\frac{4}{3\mathrm{Pe}'}\frac{D_\mathrm{L}}{D'_\mathrm{m}}\right)\frac{\partial C}{\partial \Phi}\right] + \frac{\partial}{\partial \Psi}\left[\left(\frac{3}{16}\frac{\Re^2 \sin^2\theta}{\mathrm{Pe}'}\frac{D_\mathrm{T}}{D'_\mathrm{m}}\right)\frac{\partial C}{\partial \Psi}\right] \tag{11}$$

and the appropriate boundary conditions are

$$\Phi \to -\infty,\ \Psi \geq 0 \qquad C \to 0 \tag{12}$$

$$\Phi \to +\infty,\ \Psi \geq 0 \qquad C \to 0 \tag{13}$$

$$\Psi = 0 \begin{cases} -1 \leq \Phi \leq 1 & C = 1 \quad (14a) \\ |\Phi| > 1 & \dfrac{\partial C}{\partial \Psi} = 0 \quad (14b) \end{cases}$$

$$\Psi \to +\infty,\ \text{all } \Phi \qquad C \to 0 \tag{15}$$

Discretisation

Equation (11) was solved numerically using a finite-difference method similar to that adopted in [2]. A second-order central differencing scheme (CDS), in a general non-uniform grid, was adopted for the discretisation of the diffusive terms that appear on the right hand side of Eq. (11) [3]. The convection term, that appears on the left hand side of Eq. (11), was discretised with the SMART high-resolution scheme [4], which preserves boundedness even for convective dominated flows.

The discretised equation resulting from the finite-difference approximation of Eq. (11) reads:

$$\frac{C_{i+1/2,j} - C_{i-1/2,j}}{\Phi_{i+1/2} - \Phi_{i-1/2}} = \frac{\left(\frac{4}{3\mathrm{Pe}'}\frac{D_\mathrm{L}}{D'_\mathrm{m}}\right)_{i+1/2}\frac{C_{i+1,j} - C_{i,j}}{\Phi_{i+1} - \Phi_i} - \left(\frac{4}{3\mathrm{Pe}'}\frac{D_\mathrm{L}}{D'_\mathrm{m}}\right)_{i-1/2}\frac{C_{i,j} - C_{i-1,j}}{\Phi_i - \Phi_{i-1}}}{\Phi_{i+1/2} - \Phi_{i-1/2}} + \\ + \frac{\left(\frac{3}{16}\frac{\Re^2 \sin^2\theta}{\mathrm{Pe}'}\frac{D_\mathrm{T}}{D'_\mathrm{m}}\right)_{j+1/2}\frac{C_{i,j+1} - C_{i,j}}{\Psi_{j+1} - \Psi_j} - \left(\frac{3}{16}\frac{\Re^2 \sin^2\theta}{\mathrm{Pe}'}\frac{D_\mathrm{T}}{D'_\mathrm{m}}\right)_{j-1/2}\frac{C_{i,j} - C_{i,j-1}}{\Psi_j - \Psi_{j-1}}}{\Psi_{j+1/2} - \Psi_{j-1/2}} \tag{16}$$

where the values of the Φ and Ψ coefficients are easily computed using their definitions (Eqs. (9) and (10)). Please note that these coefficients only have to be computed once, since they are dependent on *a priori* known quantities, and are not influenced by the unknown concentration field.

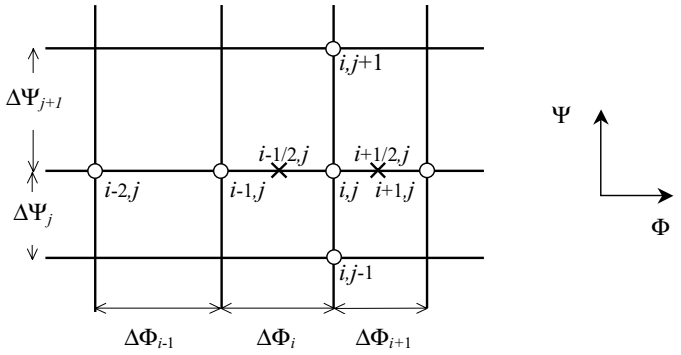


Figure 1 - Sketch of computational grid.

The $C_{i+1/2,j}$ and $C_{i-1/2,j}$ values are conveniently interpolated from the known grid node values (represented as circles in Figure 1) using the SMART high-resolution scheme to ensure numerical stability and good precision:

$$C_{i+1/2,j} = C_{i-1,j} + \widehat{C}_{i+1/2,j}\left(C_{i+1,j} - C_{i-1,j}\right) \tag{17}$$

The limiter function $\widehat{C}_{i+1/2,j}$ used in this work is expressed as [4]:

$$\widehat{C}_{i+1/2,j} = \begin{cases} 3\dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} & \text{if} \quad \dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} \in \left[0, \dfrac{1}{6}\right[\\ \dfrac{3}{8}+\dfrac{3}{4}\dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} & \text{if} \quad \dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} \in \left[\dfrac{1}{6}, \dfrac{5}{6}\right] \\ 1 & \text{if} \quad \dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} \in \left]\dfrac{5}{6}, 1\right] \\ \dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} & \text{if} \quad \dfrac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}} \notin [0,1] \end{cases} \tag{18}$$

or, in compact form:

$$\widehat{C}_{i+1/2,j} = \max\left[\frac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}},\ \min\left(3\frac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}},\ \frac{3}{8}+\frac{3}{4}\frac{C_{i,j}-C_{i-1,j}}{C_{i+1,j}-C_{i-1,j}},\ 1\right)\right] \tag{19}$$

Similarly, for the left face relative to node (i, j), one obtains:

$$C_{i-1/2,j} = C_{i-2,j} + \widehat{C}_{i-1/2,j}\left(C_{i,j} - C_{i-2,j}\right) \tag{20}$$

with

$$\widehat{C}_{i-1/2,j} = \max\left[\frac{C_{i-1,j}-C_{i-2,j}}{C_{i,j}-C_{i-2,j}},\ \min\left(3\frac{C_{i-1,j}-C_{i-2,j}}{C_{i,j}-C_{i-2,j}},\ \frac{3}{8}+\frac{3}{4}\frac{C_{i-1,j}-C_{i-2,j}}{C_{i,j}-C_{i-2,j}},\ 1\right)\right] \tag{21}$$

Substitution of Eqs. (17) and (20) into Eq. (16) leads to the final form of the discretised equation, which can be casted in compact form as:

$$C_{i,j} = (F\ C_{i-2,j} + G\ C_{i-1,j} + H\ C_{i+1,j} + I\ C_{i,j-1} + J\ C_{i,j+1})/E \tag{22}$$

The resulting system of equations (22) was solved iteratively using the successive over-relaxation (SOR) method [3], and the implementation of the boundary conditions was done in the same way as described in our previous work [2].

It should be noted that we always started our calculations with a zero concentration field on a coarse grid. A converged solution could be obtained very quickly (O(10 s) in a desktop PC with a 1.4GHz AMD® processor) and then we proceeded to a finer grid (doubling the number of grid points in each direction). Instead of restarting the calculations with a zero concentration field in this new finer grid, we simply interpolated the solution obtained in the coarse-level grid, leading to a significant decrease in the time of CPU required to attain convergence. This fully automated procedure was repeated until the finest mesh calculations were performed. The use of Richardson´s extrapolation to the limit allowed us to obtain very accurate solutions (with errors in the computed Sh′ value below 0.1%). A more elaborate multigrid technique could have been implemented to further increase the convergence rate, but we found that this simple technique was sufficient to obtain mesh-independent solutions in affordable CPU times.

RESULTS

The converged solution calculated yields values of $C_{i,j}$, from which the overall mass-transfer rate from the sphere, n, could be calculated and expressed by means of an average Sherwood number,

$$\mathrm{Sh}' = \frac{k d_1}{D'_\mathrm{m}} = \left[n / (\pi\, d_1^2)\,(c^* - c_0) \right] d_1 / D'_\mathrm{m} . \tag{23}$$

The value of n was evaluated by numerically integrating the diffusive/dispersive flux of solute perpendicular to the sphere along its surface,

$$n = -\varepsilon \sum_i D_\mathrm{T} R \sin(\theta_i) \frac{3}{2} u_0 \sin(\theta_i) \left(\frac{\partial c}{\partial \psi} \right)_{i,1} 2\pi R^2 \left[\frac{\cos(\theta_{i-1}) - \cos(\theta_{i+1})}{2} \right] \tag{24}$$

which in dimensionless discretised form reads:

$$\frac{\mathrm{Sh}'}{\varepsilon} = -\frac{3}{8} \sum_i \left(\frac{D_\mathrm{T}}{D'_\mathrm{m}} \right)_{i,1} \sin^2(\theta_i) \left(\frac{\partial C}{\partial \Psi} \right)_{i,1} \left[\frac{\cos(\theta_{i-1}) - \cos(\theta_{i+1})}{2} \right] \tag{25}$$

Values of D_T for liquid flow have been reported recently in [1], in what seems to be the only available study on the influence of Sc on D_T. Their data showed the dependence of D_T on the Schmidt number for the range $\mathrm{Sc} \le 550$, and an empirical correlation was found to describe the measured data of D_T for $\mathrm{Sc} \le 550$:

$$\frac{D_\mathrm{T}}{D'_\mathrm{m}} = 1 + \frac{\mathrm{Pe}'_\mathrm{p}}{12} - \left(\frac{\mathrm{Sc}}{1500} \right)^{4.8} \mathrm{Pe}'^{\,4.83 - 1.3 \log_{10}(\mathrm{Sc})}_\mathrm{p} \tag{26}$$

For $\mathrm{Sc} > 550$ the transverse dispersion coefficient is found to be independent of the Schmidt number, and is given by:

$$\frac{D_\mathrm{T}}{D'_\mathrm{m}} = 1 + \frac{\mathrm{Pe}'_\mathrm{p}}{12} - 8.1 \times 10^{-3}\, \mathrm{Pe}'^{\,1.268}_\mathrm{p} \tag{27}$$

As for D_L, it is fortunate that its value is not needed with accuracy, since for $\mathrm{Pe}'_\mathrm{p} > 1$, the boundary layer for mass transfer around the sphere is thin, provided that the approximate condition $d_1 / d > 10$ is observed. Indeed, for $\mathrm{Pe}' \left(= \mathrm{Pe}'_\mathrm{p} d_1 / d \right) > 10$, the boundary layer is thin and the term with D_L, in Eq. (11), may be neglected; numerical computations were undertaken in the present work that confirm the insensitivity of Sh′ to D_L, for $\mathrm{Pe}' > 10$.

For different values of Sc, Eq. (5) was solved numerically, with the point values of D_T given by the corresponding fitted curve (Eqs. (26) or (27)). From the numerical simulations, plots of Sh'/ε *vs.* Pe′ were prepared, for given values of d / d_1, in a similar fashion to what was done in [2]; in the present case, a set of plots had to be made for each value of Sc. The results of the numerical computations are shown as points in Figure 2, and an expression was sought to describe the functional dependence observed, with good accuracy. The following equation is proposed for $\mathrm{Sc} \le 550$:

$$\frac{\mathrm{Sh}'}{\varepsilon} = \left[4 + \frac{4}{5} (\mathrm{Pe}')^{2/3} + \frac{4}{\pi} \mathrm{Pe}' \right]^{1/2} \left[1 + \frac{\mathrm{Pe}'_\mathrm{p}}{9} - \left(\frac{\mathrm{Sc}}{1500} \right)^{4.8} \left(\frac{4}{3} \mathrm{Pe}'_\mathrm{p} \right)^{4.83 - 1.3 \log_{10}(\mathrm{Sc})} \right]^{1/2} \tag{28}$$

For $\mathrm{Sc} > 550$, the value of Sh' is independent of Sc, since $D_\mathrm{T} / D'_\mathrm{m}$ is independent of Sc. Substituting $\mathrm{Sc} = 550$ in Eq. (28) leads to

$$\frac{\mathrm{Sh}'}{\varepsilon} = \left[4 + \frac{4}{5} (\mathrm{Pe}')^{2/3} + \frac{4}{\pi} \mathrm{Pe}' \right]^{1/2} \left[1 + \frac{\mathrm{Pe}'_\mathrm{p}}{9} - 1.16 \times 10^{-2} \ \mathrm{Pe}'^{1.268}_\mathrm{p} \right]^{1/2} \tag{29}$$

and this may be expected to predict mass transfer coefficients for $\mathrm{Sc} \geq 550$. In the plots shown in Figure 2, the solid lines represent Eqs. (28) or (29) and it may be seen that they describe the results of the numerical computations with good accuracy.

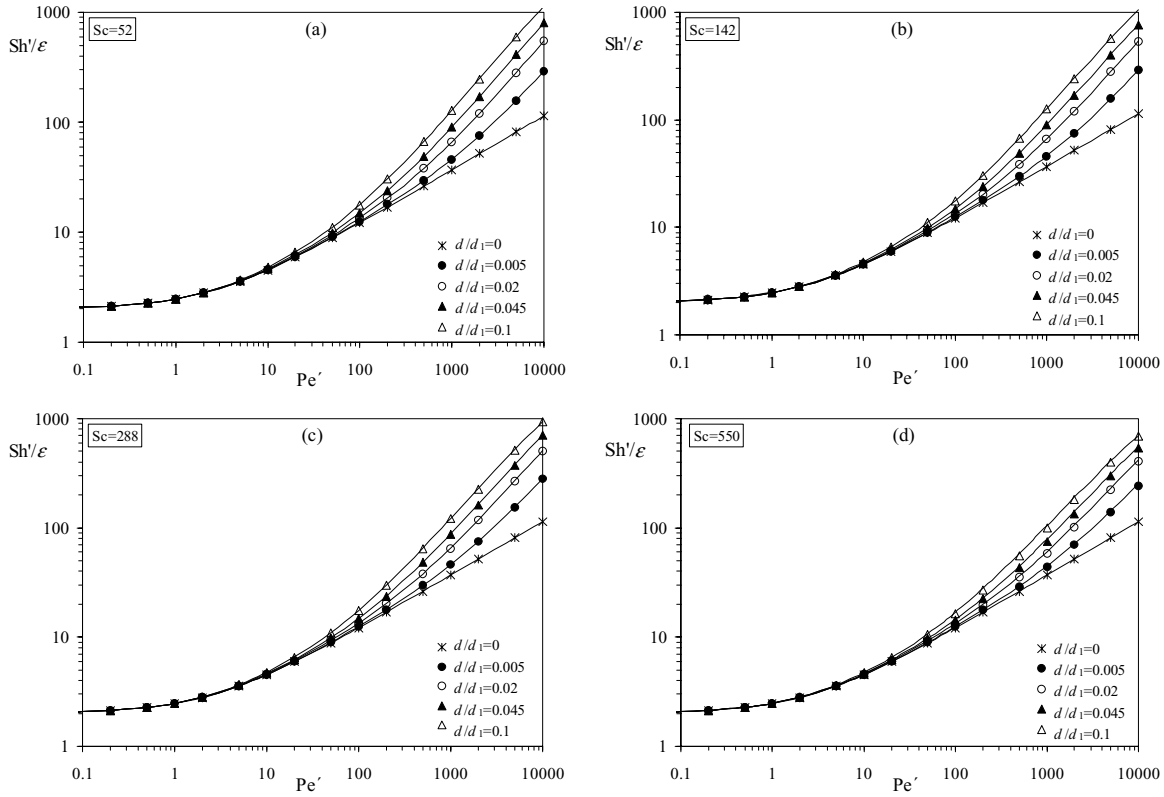


Figure 2 - Dependence of Sh'/ε on Pe' for different values of d / d_1 at (a) $\mathrm{Sc} = 52$; (b) $\mathrm{Sc} = 142$; (c) $\mathrm{Sc} = 288$; (d) $\mathrm{Sc} = 550$.

It is worth emphasizing some important features of Eqs. (28) and (29). First of all, the fact that the first term on the right hand side of both equations gives the dimensionless mass transfer coefficient for low Pe'_p, when both longitudinal and transverse dispersion are due to molecular diffusion alone ($D_\mathrm{L} = D_\mathrm{T} = D'_\mathrm{m}$). This result was obtained in [2], where that fact was emphasized by writing (for $\mathrm{Pe}'_\mathrm{p} < 0.1$)

$$\frac{\mathrm{Sh}'}{\varepsilon} = \frac{\mathrm{Sh}'_\mathrm{md}}{\varepsilon} = \left[4 + \frac{4}{5} (\mathrm{Pe}')^{2/3} + \frac{4}{\pi} \mathrm{Pe}' \right]^{1/2} \tag{30}$$

The second term (with square brackets) on the right hand side of Eqs. (28) and (29) is therefore an "enhancement factor" due to convective dispersion. It will be noticed that this

enhancement factor is independent of Sc, for high values of this parameter, and dependent on Sc for $Sc \leq 550$. This is because mass transfer rates around the sphere depend strongly on D_T, and the value of D_T / D'_m is independent of Sc only for Sc>550, as shown recently by Delgado and Guedes de Carvalho (2001) in a detailed study on dispersion in liquids.

The approximate conditions of validity of Eqs. (28) and (29) are: $Re_p < 25$, $d_1 / d > 10$ and $Pe'_p < 1300$, which are observed in a number of situations of practical interest. The limitation on Re_p ensures that Darcy's law is observed with good approximation, $d_1 / d > 10$ is an approximate condition establishing that the active particles are large compared to the inerts, and the limitation on Pe'_p is related to the dispersion data available used to obtain Eqs. (26) and (27).

CONCLUSION

The present work shows that a theory for mass transfer between a sphere buried in a packed bed of inerts and the fluid flowing past it, may be derived from first principles, that is valid for any value of Sc. The numerical solution of the partial differential equation representing this theory gives the "exact" values of Sh'/ε, which are well represented by Eqs. (28) and (29).

NOTATION

c	Solute concentration
c_0	Bulk concentration of solute
c^*	Saturation concentration of solute
C	Dimensionless solute concentration (as defined in Eq. 6)
d	Diameter of inert particles
d_1	Diameter of active sphere
D_L	Longitudinal dispersion coefficient
D'_m	Effective molecular diffusion coefficient
D_T	Transverse (radial) dispersion coefficient
K	Permeability in Darcy's law
k	Average mass transfer coefficient
n	Mass transfer rate
p	Pressure
R	Radius of the sphere
$\Re$	Dimensionless spherical radial co-ordinate ($= r / R$)
r	Spherical radial coordinate (distance to the centre of the soluble sphere)
U	Dimensionless interstitial velocity ($= u / u_0$)
u	Absolute value of interstitial velocity
$\boldsymbol{u}$	Interstitial velocity (vector)
u_0	Absolute value of interstitial velocity far from the active sphere
u_r, u_θ	Components of fluid interstitial velocity

Greek letters

ε	Bed voidage

Φ Dimensionless potential function (as defined in Eq. 9)
ϕ Potential function (defined in Eq. 1)
μ Dynamic viscosity
θ Spherical angular coordinate
ρ Density
τ Tortuosity
ω Cylindrical radial coordinate (distance to the axis)
Ψ Dimensionless stream function (as defined in Eq. 10)
ψ Stream function (defined in Eq. 2)

Dimensionless groups

Pe' Peclet number based on diameter of active sphere ($= u_0 d_1 / D'_\mathrm{m}$)
Pe'_p Peclet number based on diameter of inert particles ($= u_0 d / D'_\mathrm{m}$)
Re_p Reynolds number based on diameter of inert particles ($= \rho u d / \mu$)
Sc Schmidt number ($= \mu / \rho D_\mathrm{m}$)
Sh' Sherwood number ($= k d_1 / D'_\mathrm{m}$)
Sh'_md Sherwood number when $D_\mathrm{T} = D_\mathrm{L} = D'_\mathrm{m}$ (*i.e.* $\mathrm{Pe}'_\mathrm{p} < 0.1$)

REFERENCES

[1] Delgado, J.M.P.Q. and Guedes de Carvalho, J.R.F., "Measurement of the Coefficient of Transverse Dispersion in Packed Beds over a Range of Values of Schmidt Number (50-1000)", *Transport in Porous Media*, **44** (1), 118 (2001).

[2] Guedes de Carvalho, J.R.F. and Alves, M.A.M., "Mass transfer and dispersion around active sphere buried in a packed bed", *AIChE J.*, **45,** 2495-2502 (1999).

[3] Ferziger, J.H. and Peric, M., *Computational Methods for Fluid Dynamics*, Springer-Verlag, Berlin, p.42 (1996).

[4] Gaskell, P.H. and Lau, A.K.C., "Curvature compensated convective transport: SMART, a new boundedness preserving transport algorithm", *Int. J. Numer. Meth. Fluids*, **8**, 617-171 (1988).

A Simulation Study on the Transport Phenomena in Ultrafiltration

Licínio M. Ferreira[1]*, Paulo Brito and António Portugal*
Department of Chemical Engineering, University of Coimbra, Pólo II, Pinhal de Marrocos, 3030 Coimbra, Portugal

Maarten Blox and Piet Kerkhof
Laboratory of Separation Technology and Transport Phenomena, Faculty of Chemical Engineering and Chemistry, Eindhoven University of Techonology, P.O.Box 513, 5600 MB Eindhoven, Netherlands

Abstract

A coupled model of concentration polarization and membrane transport is used to study the crossflow ultrafiltration of PEG-3400 solutions. For the intramembrane transport, the model incorporates the binary friction model (BFM) derived by Kerkhof [4] and that is a modification of the Maxwell-Stefan-Lightfoot equation. Good agreement between model predictions and experimental data (apparent rejections and pressure drops as function of the flux) has been obtained. A value of 0.49 for the equilibrium partition coefficient K, the only adjustable parameter, was found. The model predictions also enabled us to study the effects of circulation velocity and partition coefficient on the apparent rejection and to get an insight into the concentration profiles in the polarization layer and in the membrane.

Keywords: Apparent rejection; Adaptive method; Maxwell-Stefan; Transport; Ultrafiltration.

Introduction

Modeling of mass transport phenomena present in the separation of solutes using inert membranes is important for the design and/or optimization of these new separation processes. In recent years, there has been an increased awareness on these type of processes since they can be an alternative to the conventional separations processes like distillation, centrifugation and others. They also find applications in a variety of fields, being the most prominent the food and bioprocess areas.

A number of mathematical models and algorithms for their solution have been explored for the description of the transport of components through membranes. Some of them are special cases of the generalized Maxwell-Stefan equations [1-2] and can be derived from either statistical-mechanics or thermodynamics of irreversible processes [3]. In fact, the approach based on the Maxwell-Stefan theory for the transport in both the polarization layer and the membrane give a rigorous description of the problem and the thermodynamics effects involving more than one species can be well predicted. However, this kind of mathematical formulation results, for the binary case, in two partial differential equations (PDE's) defined in two different spatial regions, corresponding to the boundary layer and the membrane and the use of commercial packages such the PDECOL and PDASAC to achieve a transient solution is difficult, since they were developed to solve straightforward PDE's. In most of the studies the steady-state behavior is considered, where the problem is described by a set of ODE's and the solution is obtained by employing numerical methods based on finite differences. With this approach the coupled equations are solved via an iterative procedure and in many situations, problems of convergence and stability of the numerical method occur.

This work focuses on the ultrafiltration of PEG solutions. For the intramembrane transport the binary friction model derived by Kerkhof [4 -5] was used. That is a modification of the Maxwell-Stefan-Lightfoot equation, and includes both interspecies (diffusive) and species-wall forces. The numerical scheme used for solution of the equations is based on the application of an adaptive method with grid refinement developed by Brito and Portugal [6].

[1] Author to whom the correspondence should be addressed

Mathematical Model

The modeling of the transport of solutes through membranes involves a couple solution of two transport models. The first model describes the transport phenomena in the concentration polarization layer on the feed side adjacent to the membrane, while the second model deals with the intramembrane (inside membrane pores) transport.

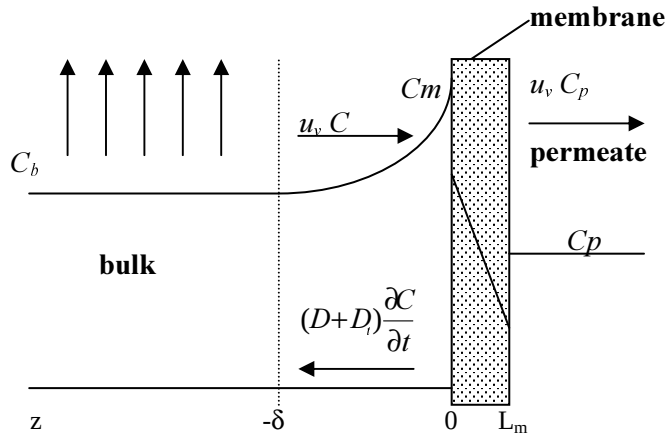
The governing equations for the unsteady-state transport of species through the membrane can be described by the continuity equations. For the polarization layer (see Figure 1), we have,

$$\frac{\partial (c)}{\partial t} = -\frac{\partial (N)}{\partial z} \tag{1}$$

where the flux per unit area of the membrane is written as $(N) = -([D] + [D_t])\frac{\partial c}{\partial t} + u_v c$, in which the Fickian molecular and turbulent diffusion matrixes are given by $[\boldsymbol{D}] = [\boldsymbol{B}]^{-1}[\boldsymbol{\Gamma}_c]$ and $[\boldsymbol{D}_t] = D_t\,[\mathrm{I}]$, respectively; for the membrane, the equations are written as

$$\varepsilon \frac{\partial (c')}{\partial t} = -\frac{1}{\tau}\frac{\partial (N_m)}{\partial z} \tag{2}$$

where N_m is the flux per square meter of membrane area.



The intramembrane transport can be subdivided into intermolecular fricction between different components, and the effective friction of each component with the wall. The detailed momentum balance for each component, and the averaging over the pore cross section, results in the binary friction model (BFM) that is an extension of the Lighfoot model [3],

$$\frac{x_i}{RT}\nabla_{T,P}\mu_i + \frac{c_i \bar{V}_i}{c_t R T}\nabla P = \sum_{j=1}^{n}\frac{x_i N_j - x_j N_i}{c_t D_{ij}} - \frac{1}{B_o}k_i \phi_i u_i \tag{3}$$

where the last term of right side of the eq. (3) enables us to quantify the friction between the components and the membrane. This term includes the fractional viscosity coefficients k_i that can be evaluated from the bulk mixture viscosity data.

Finally, considering the binary case that involves the transport of a single solute, we obtain for the boundary layer,

$$\frac{\partial c}{\partial t} = -\frac{\partial N}{\partial z} \tag{4}$$

with $N = -(D + D_t)\frac{\partial c}{\partial t} + u_v c$ and $D = D_{12}\Gamma_c$.

For the intramembrane transport, we have,

$$\varepsilon \frac{\partial c'}{\partial t} = -\frac{1}{\tau}\frac{\partial N_m}{\partial z} \tag{5}$$

From eq. (3), by converting the chemical potential gradients in concentration gradients and developing the equation in terms of molar fluxes, N_m is given by,

$$N_m = -\frac{\Gamma_c}{G}\frac{\partial c'}{\partial z} + u_v \frac{F}{G} c', \quad \text{where } F = \left(\frac{1}{D_{12}} + \frac{c_t^2 \bar{V}_1 \bar{V}_2}{B_0} k_2\right)\frac{\tau}{\varepsilon} \text{ and}$$

$$G = \left(\frac{1}{D_{12}} + \frac{c_t}{B_0}(\phi_2 k_1 \bar{V}_1 + \phi_1 k_2 \bar{V}_2)\right)\frac{\tau}{\varepsilon},$$

with boundary conditions,

z = $-\delta_{pol}$: $c = c_b$

z = *0* (interface polarization layer/membrane): $c' = K c$ and $N = N_m$

z = L_m(interface membrane/permeate): $N = u_v c_p$ ($c_p = c'/K$)

and initial conditions,

t = *0*: $c = c_b$ for $z = -\delta_{pol}$ and $c = c' = 0$ for $z > -\delta_{pol}$.

For turbulent conditions the diffusion mechanism in the polarization layer should incorporate an additional transport contribution by the turbulent eddies. The usual procedure for prediction of D_t is to proceed through the calculation of kinematics viscosity, υ_t. Defining the turbulent Schmidt number as $Sc = \upsilon_t / D_t$ and considering that for most practical design purposes Sc value is taken equal to 1, i.e., $D_t = \upsilon_t$. If the turbulent viscosity is taken to vary according to Vieth correlation,

$$\frac{\upsilon_t}{\upsilon} = \left(\frac{2}{9}\Pi\sqrt{3}\right)^3 \left(\frac{f}{2}\right)^{3/2} \left(y^+\right)^3 \tag{6}$$

The reduced distance from the membrane wall is expressed as $y^+ = \frac{y < u_t > \sqrt{f/2}}{\upsilon}$ in which the fanning friction factor, is evaluated using the Blasius equation, $f = (0.3164/4)\, Re^{-0.25}$. According to Kerkhof [4], a region limited by $y^+ \leq 5$ was considered as having the sufficient distance for the development of the composition profiles within of the boundary layer of thickness δ.

The coupled model of concentration polarization and membrane transport presented provides a consistent procedure for predicting the concentration and molar flux profiles throughout the system, the permeate concentration c_p and the apparent rejection of the solute, given by,

$$R_{app} = 1 - \frac{c_p}{c_b} \tag{7}$$

For the total pressure gradient over the membrane $\Delta P_{tot} = \Delta P_{flow} + \sigma \Delta \prod$, in which $\Delta \prod$ is the osmotic pressure difference that depend on the concentrations on both membrane sides and of the osmotic reflection coefficient σ, the following expression was used for ΔP_{flow} evaluation,

$$\Delta P_{flow} = \frac{\tau R T}{\varepsilon B_o} \int_0^{Lm} c_t \sum_{i=1}^{2} k_i N_i \bar{V}_i dz \tag{8}$$

The relationship between the pressure differences ΔP and the flux Uv for experiments with pure water enables the determination of the membrane resistance R_m.

$$R_m = \frac{\Delta P}{\eta_w J_v} \tag{9}$$

where $R_m = \frac{L_m \tau}{B_o \varepsilon} = \frac{8 L_m \tau}{r_p^2 \varepsilon}$. Hence, the ratio ε/τ can be evaluated from the R_m value and the geometrical properties of the membrane.

Numerical Procedure

The adaptive mesh algorithm was developed for one-dimensional evolutive systems of Algebraic-Differential Equations that can be resumed by the following general model:

$$F\left(\underline{u}_t, \underline{u}, \underline{u}_z, \underline{u}_{zz}\right) = 0$$
$$G\left(\underline{u}\right) = 0$$

subjected to the boundary conditions: $\underline{u}\left(z^L, t\right) = \underline{u}^L(t)$ and $\underline{u}\left(z^R, t\right) = \underline{u}^R(t)$ and the initial condition: $\underline{u}(z,0) = \underline{u}^0(z); \quad z \in \left[z^L, z^R\right]$. The algorithm can be structured in two main stages: estimation of the discretization error and identification of the adaptive sub-domains; and solution of the sub-problems generated in the first stage, by the introduction of an adaptive grid technique.

Stage I - Discretization

The error estimation is based on the comparison of the solution obtained by solving the original problem on two different grids: a fine and a coarse grid (Grids of level 2 and 1, respectively). Initially, the fine grid is constructed by the bissection of each interval of the coarse one. The nodes in level 1 grid, that do not satisfy the error criterion, are grouped together with the level 2 nodes placed between them, to define the sub-domains over which the adaptive sub-problems are generated and then solved.

Stage II - Adaptive Integration of the Sub-problem

The sub-problems are generated with increasing refinement level, by the repetition of the procedure described in Stage I, until every node in every grid verifies the tolerance condition associated with the error estimated by:

$$EU^i_{j,k+1} = Wh^i_{j,k+1} - W2h^i_{j,k+1}; \quad j = 1, \cdots, NP_{n-1}, \quad i = 1, \cdots, NPDE$$

In this case, $EU^i_{j,k+1}$ represents the approximation to the spatial error, in a node j of a grid of refinement level n; $Wh^i_{j,k+1}$ and $W2h^i_{j,k+1}$ are the approximations to the component i of the solution, obtained through integration between the times t_k and t_{k+1}, on the finer (level n) and the coarser (level n-1) grids, respectively; NP_{n-1} is the number of nodes in the grid of level n-1; and *NPDE* is the number of partial differential equations of the problem.
The sub-domains of level n+1 are obtained by joining all nodes n-1 that satisfy the condition:

$$\left|EU^i_{j,k+1}\right| > TOL_i; \quad i = 1, \cdots, NPDE$$

In each refinement procedure, the profiles of the solution are computed by interpolation of the profiles of level 2, at all the intermediary positions.
The algorithm is coupled with a strategy for the treatment of boundary conditions in the refinement sub-problems that simply defines fixed Dirichlet conditions on each internal bound. The position of each bound, for the refinement level n+1 (for n = 2, … , N_{MAX-1}, where N_{MAX} is the maximum refinement level) are coincident with the positions of the first nodes of level n-1 that verify the

specified tolerance. The constant value of the boundary conditions is given by the solution obtained in the integration over the level n-1 grid. This kind of procedure is very simple and prevents discontinuities on the overall profiles but tends to introduce significant errors in the solution, in very specific cases.

The model is divided in the overall length in relation to the concentration (normalized by the bulk conditions). The fluxes are later calculated using the concentration profiles. The spatial coordinate ($-\delta \leq z \leq L_m$) is also normalized using the overall length: $z^* = \frac{z+\delta}{L_m+\delta}$. Therefore, in the bulk position: $z = -\delta$, $z^* = 0$, and for the permeate position: $z = L_m$, $z^* = 1$. The differential equations that describe the time evolution of the concentrations are spatially discretized by finite differences approximations (either on the polarization layer as on the membrane sides) and solved simultaneously. Initially we assume a zero concentration profile on the whole domain (with the obvious exception of the bulk border). The bulk border is treated as a fixed *Dirichlet* condition with $C = C_b$. The inner border (which represents the transition between the polarization layer and the membrane) and the permeate border are treated with the introduction of two nodes (very close to one another) that represent the inner and outer conditions related to the membrane. For the inner border (positioned at $z = 0$; $z_{ib}^* = \frac{\delta}{L_m+\delta}$) the solution on both nodes is calculated by the solution of two algebraic equations: the equality of the fluxes on both positions and the equilibrium condition. The strategy used for the treatment of the permeate border (positioned at $z = L_m$; $z^* = 1$) is similar and it is based on the solution of two algebraic equation also: the equality of fluxes, which has to satisfy the border condition: $N = u^v \times C_p$; and the equilibrium condition.

Results and discussion

Comparison of model predictions with experimental data

In order to validate the model, we used data from ultrafiltration experiments of PEG-3400 performed by Box, [7] in a cross-flow ultrafiltration module containing a tubular polysulfone membrane of 1 meter length, with two separated permeate sections. Samples for analysis were obtained from the second section of the tube, where the entrance effects are absent. The operating conditions of the system under study are given in Table 1.

Regarding the physical properties of the solution, we used $\eta_{sp} = 0.1(\eta_{int})\rho_p + 0.0033(\eta_{int}\rho_p)^2$ and $\eta_{int} = 6.04 \times 10^{-5} M_n^{0.90}$ with M_n(number-averaged molecular Mass) = 3158, for the calculation of the PEG3400 viscosity. From the viscosity data, the fractional viscosity coefficients k_1 e k_2 were estimated according to the following relationships: $k_1 = k_2(1+\eta_{sp}/\phi_1)$, where k_2 may be related directly to the pure solvent (water), η_w=ct RT k_2. For the determination of the molecular Fickian diffusion coefficient in aqueous PEG-3400 solutions, the following relation was taken into account: $D = (5\omega_p + 1.37) \cdot 10^{-10}$ m^2.s^{-1}.

Table 1-Conditions used in the simulation of ultrafiltration of PEG3-400 solutions

Membrane characteristics	Solution properties	Flow conditions
$L_m = 5.10^{-7}$ m	$c_b = 2.78 \cdot 10^{-3}$ kmol/m^3	u_t = 1.08; 1.57 m/s
$R_m = 1.62 \cdot 10^{16}$. Uv+5.39x10^{12}m^{-1}	M=3600	$\delta = 86.10^{-6}$ m
$r_p = 9.00.10^{-9}$ m^2	$\bar{V}_1$ =2.83 m^3/kmol	
MWCO=50 kDa	$\bar{V}_2$ =0.018 m^3/kmol	
ε = 0.50	$\eta_w = 8.0 \cdot 10^{-4}$ Pa.s	
	T= 298 K	

The comparison of experimental apparent rejections of PEG-3400 with the values predicted by the model are presented in Fig. 1. As can be seen, the model is able to describe well the experimental results for K = 0.49. In Fig.2, it is also visible a good agreement between the pressure drop calculated for σ= 0 and the ones measured.

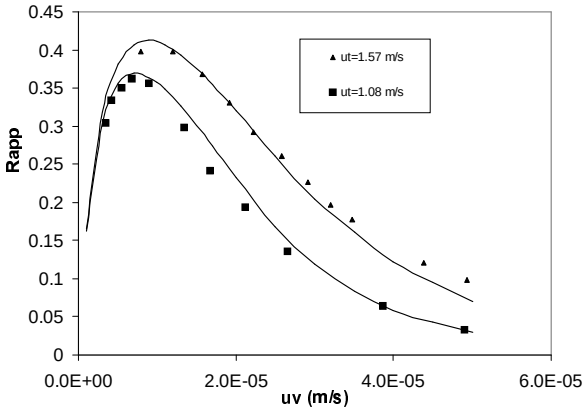


Figure 1 – Apparent rejection of PEG-3400 as function of the flux at two different circulation velocities. The predictions are the solid lines whereas the experimental data are presented by the symbols.

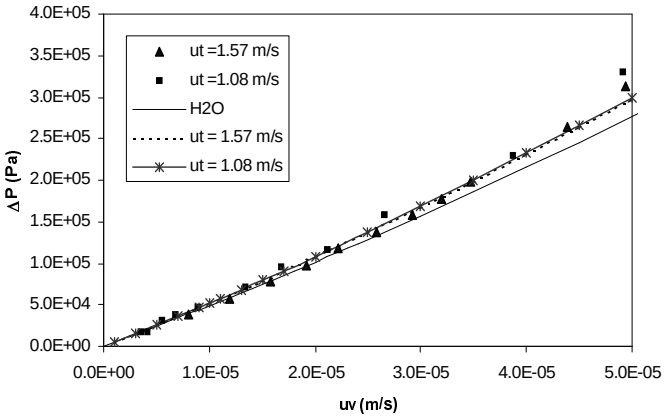


Figure 2 – Pressure drop vs membrane flux. The predictions are the solid lines whereas the experimental data are presented by the symbols.

Effect of circulation velocity u_t

The apparent rejection is affected by the circulation velocity along membrane tube as shows the Fig. 1. It can be observed that increasing u_t results in a increase of the apparent rejection. Since the turbulent contribution for the mass transport in the polarization layer increases at high circulation velocities, the amount of solute accumulated on the surface membrane is expected to decrease. This will correspond to a lower value of the membrane surface concentration and thus less solute will be transported through the membrane and hence, the rejection will increase.

Effect of equilibrium partition coefficient K

Concentrations in the membrane pore are related to the interfacial bulk concentrations through the partition coefficients. Once the interfacial region is very small and therefore the differences in the chemical potential are negligible, to assume interfacial equilibrium conditions ($K_i = c_i'/c_i$) is a valid approximation. The K_i value is determined by geometrical factors and by specific interactions of solute and pore wall. For spherical solutes in cylindrical pores and according to the exclusion theory, K_i only depends on the ratio of the molecular radius and the pore radius: $K_i = (1 - \lambda_i)^2$. Thus, for a given solute, the decrease of the partition coefficient is consistent with the use of membranes that exhibit lower pore radius and thereby, the solute concentration in the pores will tend to increase. Consequently, high values of solute rejection will be obtained as is depicted in Fig. 3.

Concentration Profiles

For a flux of 10^{-5} m/s and at circulation velocity of 1.04 m/s, the evolution of PEG-3400 concentration along the spatial coordinate for various time values are shown in Figures 4a and 4b. The conditions used in this simulation corresponds to the ultrafiltration experiment of PEG-3400 reported by Kerkhof [4]. In Fig. 4a, the behavior of the solute transport in the polarization layer can be observed together with the propagation of the bulk concentration towards the membrane driven by the

contribution of the convective and diffusive fluxes. At higher times the concentration near the membrane increases surface due to the exclusion of the solute by the membrane, therefore originating a back diffusion is generated influencing strongly the evolution of concentration inside the membrane. This phenomena is illustrated in Fig. 4b.

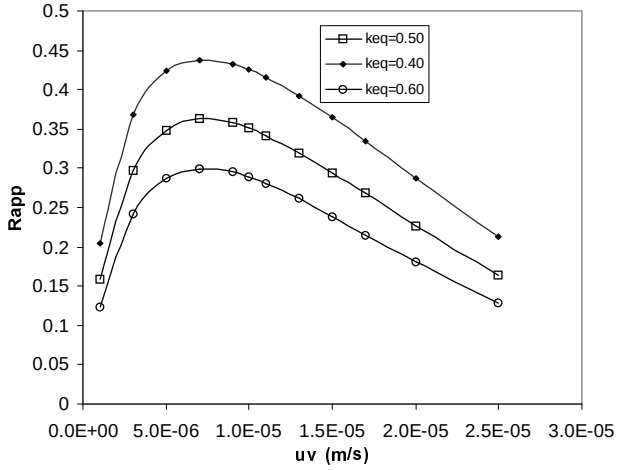


Figure 3 – Effect of equilibrium partition coefficient on the apparent rejection of PEG-3400

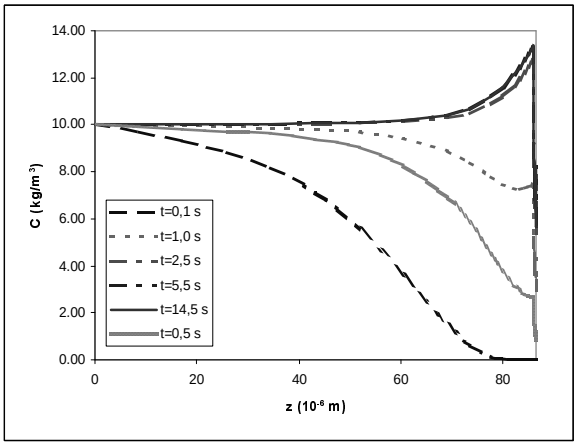


Figure 4a – Concentration profiles of PEG-3400 as a function of time in the polarization layer.

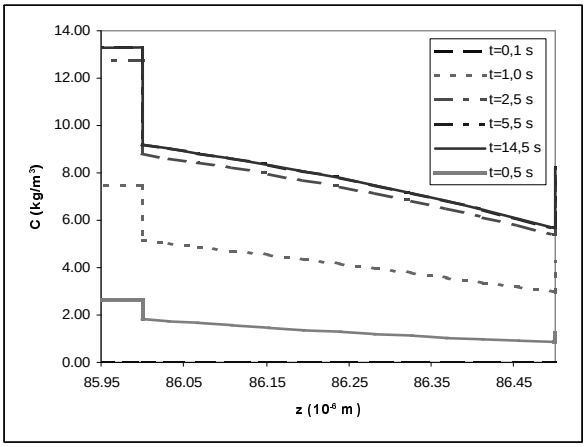


Figure 4b – Concentration profiles of PEG –3400 as function of time in the membrane.

Conclusions

A coupled model of concentration polarization and membrane transport described by the binary friction model (BFM) is used to study the crossflow ultrafiltration of PEG-3400 solutions.

A numerical procedure based on the adaptive method with grid refinement, in which the model differential equations are spatially discretized by finite differences and solved simultaneously, was able to give the solution of the system without much computational power and yield a rigorous solution of the problem. It has been shown that the solution predicts quite well the apparent rejection of PEG-3400 and the pressure drop as a function of the flux. The model is capable of predicting the influence of fundamental physico-chemical parameters and operating conditions on the apparent rejection of the solute. In fact, the influence of some of these parameters namely, the circulation velocity and the equilibrium partition coefficient was shown in this study. The predictions of the model also provides a good insight regarding the concentration and flux profiles in the polarization layer and in the membrane.

The numerical description of the ultrafiltration model used is versatile and allows in the future to extend this study to the multicomponent transport.

Nomenclature

c: molar concentration (kmol.m^{-3})
B_0: permeability parameter (m^2)
D: Fick diffusion coefficient (m^2 .s^{-1})
D_{12}: Maxwell-Stefan diffusion coeficient (m^2 .s^{-1})
D_t: turbulent diffusion coefficient (m^2 .s^{-1})
k: fractional viscosity coefficient
K: equilibrium partition coefficient
L_m: membrane thickness (m)
M: molecular mass (kg .kmol^{-1})
N: flux with respect to stationary coordinate (kmol.m^{-2}.s^{-1})
P: pressure (Pa)
r_p: pore radius (m)
R: gas constant (J.kmol^{-1}.K^{-1})
x: mole fraction
t: time (s)
T: temperature (K)
u_t: circulation velocity (m.s^{-1})
u_v: average permeate flux (m.s^{-1})
$\bar{V}$: specific molar volume (m^3.kmol^{-1})
z: spatial coordinate (m)
( ): vector
[ ]: square matrix

Greek letters

δ: thickness of polarization layer (m)
ε:porosity
ϕ: volume fraction
Γ_c: thermodynamic factor
τ: tortuosity
η:viscosity (Pa.s)
μ: chemical potential (J.kmol^{-1})
ρ: mass concentration (kg.m^{-3})
ω_p: weight fraction (kg.kg-1)

Subscripts

b: bulk
m: membrane
p: permeate
pol: polarization
t: total
w: water.

References

[1] T.R. Noordman and J.A. Wesselingh, *J. of Membrane Science* 210 (2002) 227-243

[2] W.M. Deen, *AIChE Journal* (1987) 1409-1425

[3] E.N.Lightfoot, *Transport in Living Systems*, Wiley, New York, 1974

[4] P.J.A.M. Kerkhof, *The Chem. Eng. Journal* 64 (1996) 319-343

[5] P.J.A.M. Kerkhof, M. M. Geboers and K. Ptasinski, *The Chem. Eng. Journal* 83 (2001) 107-121

[6] P.M.P. Brito and A.A.T.G.Portugal, *Adv. Comput. Meth. In Eng.,* ACOMEN'98, in Proceeding of Conference on Advanced Computational Methods in Engineering

[7] M.J.J. Blox, *MSc. Thesis,* Eindhoven University of Technology (June 2003)

PROCESS MODELLING THROUGH KNOWLEDGE INTEGRATION – COMPETITIVE AND COMPLEMENTARY MODULAR PRINCIPLES

Georgieva[1] P, J. Peres[1], R. Oliveira[2], S. Feyo de Azevedo[1*]

[1] Department of Chemical Engineering – Institute for Systems and Robotics Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias s/n 4200-465 Porto, Portugal
petia@fe.up.pt; jperes@fe.up.pt; sfeyo@fe.up.pt
[2] REQUINTE/CQFB, Departamento de Química - Centro de Química Fina e Biotecnologia, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, P-2829-516 Caparica, Portugal
rui.oliveira@dq.fct.unl.pt

Abstract: In the last decade hybrid modelling in the sense of knowledge has gained an increasing interest as a technique for identification of biochemical processes. This new approach is based on a combination of partial (traditional) first principles models with data-driven models (such as ANN). There are two main schemes of such a modular integration – competitive and complementary.

The aim of this paper is to report our experience applying both hybrid modelling approaches to relevant case studies: the competitive modular principle applied to a *Sacharomyces cerevisae* yeast (a biological process) and the complementary modular principle to a fed-batch evaporative sugar crystallization (a chemical process). Due to their specific nonlinear nature we were challenged to model the process kinetics sufficiently well by first principles models only.

1. Introduction

Modelling through knowledge integration aims at exploring all available sources of a priori knowledge/information about the process that should be optimally combined and incorporated in the process model. There are different modelling techniques based on the nature of the information available. Most generally, the models can be classified as first principles (or deterministic) models, fuzzy (based on heuristic knowledge) models, statistical models and more recently, black-box (usually ANN) models.

The modular principle in knowledge integration consists of division of the process in several modules according to the kind of knowledge available in the different process parts. There are two main modular architectures – complementary and competitive. In the competitive structure different modules concur for the right to represent the same part of the process (parameters, outputs, etc.). In the complementary structure different kinds of knowledge complement themselves. Usually, for the known physical constraints (e.g. mass and energy balances) the most reliable models are still the first principles models while for the less known parts the data driven modelling is more efficient.

Two benchmark problems are considered in this paper: for the first case study, *Sacharomyces cerevisae* yeast, the competitive modular principle of modelling is adopted; for the fed-batch evaporative sugar crystallization the complementary modular principle of modelling is implemented.

2. Competitive modular modelling

The application of Artificial Neural Networks (ANNs) for modelling the reaction kinetics in biological systems has been exemplified in many works (e.g. Schubert et al. (1994), Montague and Morris (1994)). Conventional BP networks and RBF networks are the most employed architectures. One important issue related to the nature of the cell system is the fact that cells may process substrates through different metabolic pathways. This is the case of diauxic growth on two carbon sources. Or the case of aerobic/anaerobic growth depending on the presence or absence of dissolved oxygen in the medium. For example, *S. cerevisae* can grow through three

[*] To whom the correspondence should be addressed

different metabolic pathways for exploiting energy and basic material sources and is able to switch between a respiratory metabolic state and a reductive metabolic state (Sonnleitner and Kappäli (1986)).
BPs and RBFs networks have some limitations for approximating discontinuous input-output systems. BPs tends to exhibit erratic behaviour around discontinuities (Haykin, 1994). RBFs are voted for local mappings and suffer from generalisation problems especially for resolution of fine details. There are strong reasons to believe that modular networks architectures may be advantageous for modelling reaction kinetics in biological systems. A modular network architecture consists of two or more (small) network modules mediated by a so-called gating network which decides how to combine their outputs to form the final output of the system. The learning of such networks is based on the principle of divide-and-conquer, i.e., the network modules compete to learn the training patterns. This type of architecture performs task decomposition in the sense that it learns to partition a task into two or more functionally independent tasks and allocates distinct networks to learn each task (Jacobs et al. (1991). Microorganisms reaction kinetics are ruled by a rather complex network of metabolic reactions that can be viewed as being composed by a set of interconnected modules representing different pathways: glycolysis, TCA cycle, etc. Hence a modular network structure is hypothetically highly compatible with the internal structure of the system 'cell reaction kinetics'. A second relevant point in favour of modular networks is that they fit better discontinuous input-output systems (Haykin, 1994). These features indicate that this type of networks could be advantageous to model the reaction kinetics.
Three types of networks are compared: the ME, BP and RBF networks. The *S. cerevisae* yeast serves as an example to illustrate the application of the networks. The main objective of this study is to verify if modular network architectures, which are supposed to be able to perform task decomposition, are able to discriminate between reaction pathways in complex biological reaction schemes.

2.1 Methods

The Mixture of Experts (ME) network developed by Jacobs and Jordan (1991) was adopted in this work. The ME architecture consists of a set of *k* expert networks and one gating network (Fig. 1). The task of each expert *i* is to approximate a function $f_i : \mathbf{x} \rightarrow \mathbf{y}$ over a region of the input space. The task of the gating network is to assign an expert network to each input vector **x**. The final output **y** is a linear combination of the expert networks.

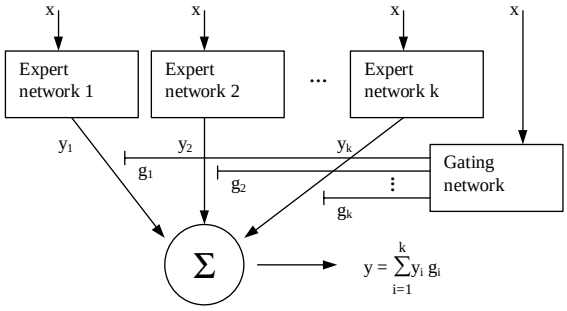


Fig. 1. Block diagram of a 'mixture of experts' network;
the outputs of the expert networks are mediated by a gating network.

The interesting property of this network is that it is able to learn to partition a task into two or more functionally independent tasks and to allocate distinct networks to learn each task. The training of the ME network may be performed using a maximum likelihood parameter estimator. For the class of nonlinear regression problems (which is our case) the objective is to map a set of training patterns {**x**,**d**}.
The goal of the learning algorithm is to model the probability distribution of {**x**,**d**}. The output vector of each expert can be interpreted as a parameter of a conditional target distribution. In the case of a Gaussian distribution, the probability of a desired target **d** of dimension *q*, given the input **x** of dimension *p* and given the expert *i* is

$$P(\mathbf{d} \mid \mathbf{x}, i) = \frac{1}{(2\pi)^{q/2}} \exp\left(-\frac{1}{2} \| \mathbf{d} - \mathbf{y}_i \|^2\right) \quad (1)$$

The expert outputs $\mathbf{y}_i$ corresponds in this case to the conditional mean of the desired response **d** given the input vector **x** and that the ith expert network is used, $\mathbf{y}_i = E[P(\mathbf{d}\,|\,\mathbf{x},i)]$. The outputs of the gating networks g_i are interpreted as the conditional probability $P(i|\mathbf{x})$ of picking the expert i given de input **x**. The probability of the desired target given the input **x** is thus

$$P(\mathbf{d}\,|\,\mathbf{x}) = \sum_{i=1}^{k} P(i\,|\,\mathbf{x}) P(\mathbf{d}\,|\,\mathbf{x},i) = \sum_{i=1}^{k} g_i \frac{1}{(2\pi)^{q/2}} \exp(-\frac{1}{2}\|\,\mathbf{d}-\mathbf{y}_i\,\|^2) \tag{2}$$

The learning algorithm for this architecture, and in the light of the probabilistic interpretation made so far, can be viewed as a maximum likelihood parameter estimation problem. The criterion for estimating the synaptic weights $\mathbf{w}_i$ of each expert i and of the synaptic weights **a** in the gating network is to maximise the density function of Eq. (2). Usually the natural logarithm of $P(\mathbf{d}|\mathbf{x})$ is preferable to use (notice that $P(\mathbf{d}|\mathbf{x})$ is a monotonic increasing function of its arguments). Over a set of p training patterns and after some manipulation the maximum likelihood function $l(\mathbf{x},\mathbf{w},\mathbf{a})$ is

$$l(\mathbf{x},\mathbf{w},\mathbf{a}) = \sum_{t=1}^{p} \ln \sum_{i=1}^{k} g_i(\mathbf{x}_t,\mathbf{a}_i) \exp(-\frac{1}{2}\|\,\mathbf{d}_t - \mathbf{y}_i(\mathbf{x}_t,\mathbf{w}_i)\,\|^2) \tag{3}$$

being $\mathbf{w}=[\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_k]^T$ and $\mathbf{a}=[\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_k]^T$ the vector of weights of the expert networks and gating network respectively. The expert modules may be linear, $\mathbf{y}_i = \mathbf{w}_i\mathbf{x}$, or nonlinear functions, for instance, a small BP network. The gating network outputs have a probabilistic interpretation and must obey to two constrains: all g_i must be positive and they must sum to one for each **x**. The gating network may be defined by a set of k 'softmax' processing units (Jacobs and Jordan (1991)):

$$g_i = \exp(\mathbf{u}_i) / \sum_{j=1}^{k} \exp(\mathbf{u}_j), i = 1,...,k \tag{4}$$

being $\mathbf{u}_i$ a linear combination of input vector **x** and connection weights $\mathbf{a}_i$, $\mathbf{u}_i = \mathbf{a}_i\mathbf{x}$. The softmax functions provide normally a 'soft' partition of the input space.

The learning algorithm must update the synaptic weights $\mathbf{w}_i$ of all expert networks and weights $\mathbf{a}_i$ in the gating network in order to maximise function (3). Jacobs and Jordan (1991) applied gradient ascent weights updating algorithm where the weights $\mathbf{w}_i$ and $\mathbf{a}_i$ are updated simultaneously. Jordan and Jacobs (1994) applied the Expectation Maximisation (EM) algorithm for training the network, which proved to converge much faster then the gradient ascent algorithm.

2.2 Results and discussion

Case Study 1: Model of the Specific Growth Rate by Blackman

In this simple example the objective is to approximate the Blackman model for the specific growth rate (μ) as a function of substrate concentration (S):

$$\mu(S) = \begin{cases} \dfrac{\mu^*}{K_M} S & S \le K_M \\ \mu^* & S > K_M \end{cases} \tag{5}$$

being μ^* and K_M two kinetic parameters. Eq. (5) has a discontinuity for $S=K_M$; the objective of this study is to assess the behaviour of the networks when dealing with such discontinuous models. Eq. (5) was used to generate data, with K_M=0.2 g/L and μ^*=0.17 h^{-1}, and for glucose concentrations ranging between (0 ,1) [g/l] with intervals of 0.002 g/l.

A ME network with 2 linear experts and a softmax gating network was trained on this data with the gradient ascent method. The total number of parameters was 8, which is the minimum number possible. The training algorithm converged very easily and rapidly, yielding a final mean square error of 9.7×10^{-8}. The results are shown in Fig. 2. The ME network was able to partition the input space at the discontinuity, as expected, and each of the experts were assigned to one or the other partition. One can notice a small curvature around the discontinuity because the 'softmax' functions produce a soft partition of the input space. A BP network with one

hidden layer, sigmoid activation functions and with 7 and 10 parameters produced a mean square error of 5.5×10^{-7} and 5.7×10^{-8} respectively. The performance of the BP net with the same number of parameters is quite similar to that of the ME network. The RBF network with 8 and 31 parameters produced an error of 3.6×10^{-6} and 3.1×10^{-7} respectively. For describing this fine detail around the discontinuity, the RBF net requires much more parameters than the other two networks.

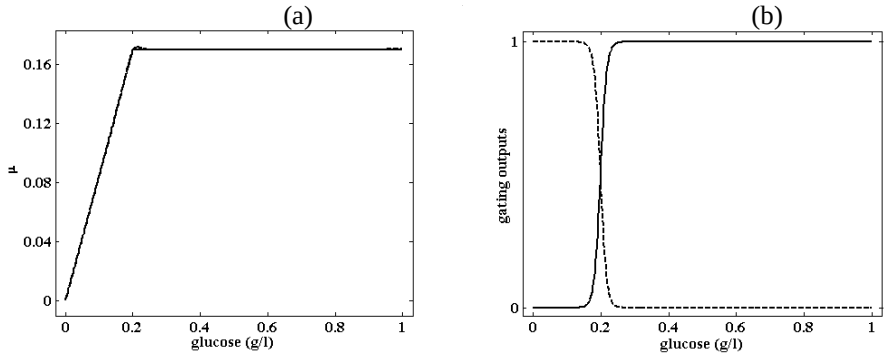


Fig. 2. Approximation results of the ME network to the Blackman model. (a) specific growth rate, (b) gating network outputs g_1 (-, solid line) and g_2 (--, dash line).

Case Study 2: S. cerevisae cultivation process

The *S. cerevisae* cells can metabolise glucose via two pathways under aerobic conditions: oxidative and/or reductively, with ethanol being the end product of the reductive pathway. The cells are able to use ethanol as a second substrate (the phenomenon of diauxic growth), but ethanol can be metabolised oxidative only. The 3 metabolic pathways may be stated by the following macroscopic reactions:

$S + NH_3 + O_2 \rightarrow X + CO_2 + H_2O$ ($\square_{os}$) (R1) - oxidative glucose uptake
$S + NH_3 \rightarrow X + E + CO_2 + H_2O$ ($\square_{rs}$) (R2) - reductive glucose uptake
$E + NH_3 + O_2 \rightarrow X + CO_2 + H_2O$ ($\square_{oe}$) (R3) - oxidative ethanol uptake

where S is glucose, X is biomass and E is ethanol. μ_{os}, μ_{rs} and μ_{oe} are three specific growth rates associated with each pathway. Sonnleitner and Käppeli (1986) proposed a kinetic model, assuming this reaction mechanism, based on the bottleneck concept. The key concept in the bottleneck model is that there is a maximum rate for oxidative glucose and ethanol uptakes, which are governed by the yeast' maximum respiratory capacity. The cells cannot grow simultaneously though pathways 2 and 3. Growth switches between pathways 2 and 3 depending on the available respiratory capacity (which depends on the concentration of dissolved oxygen) and on the actual glucose uptake rate (which is dependent on the glucose concentration in the medium). The total growth rate is the sum of three growth rates related to 3 pathways. The main goal in this case study is to model the specific growth rate and to verify if the ME network is able to detect the switch between pathway 2 and 3. Three batches were simulated with constant feed rates of 0.05, 0.5 and 2.5 l/h. Data of total growth rate as a function of glucose concentration and ethanol concentration (we assumed that oxygen was never a limiting substrate) was collected with sampling intervals of 0.1 h. The total number of points used for training was 78. This data was used to train and compare the 3 networks. The results obtained with the ME network with 3 linear experts (9 parameters) are plotted in Fig. 3. The gating network employed was a gaussian network and the training algorithm was the EM algorithm. The mean square error obtained was 1.6×10^{-5}. The interesting point to be noticed in this example is that the ME was able to discriminate between the 3 possible combinations of reactions. A BP network with 9 parameters produced a mean-square error of the same order of magnitude (1.38×10^{-5}), indicating that there is no apparent advantage of using a ME network in this example. The results produced by a RBF network with 9 parameters are worst as it was in the previous case study. The mean square error obtained was 1.7×10^{-3}.

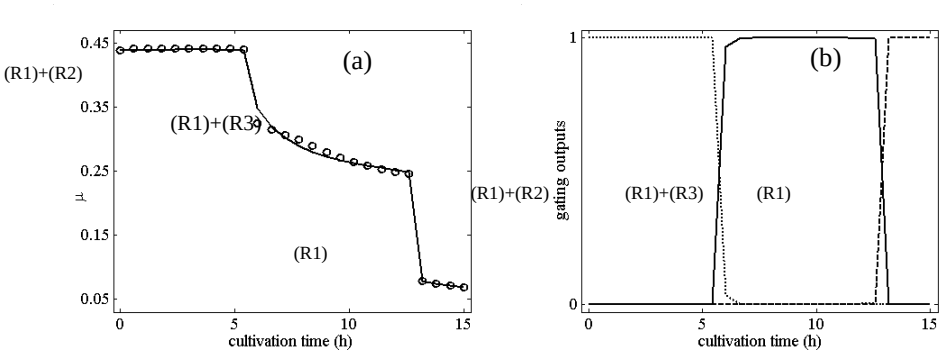


Fig. 3. Results for one batch. (a) specific growth rate estimates with a ME with 3 experts (9 parameters): measured values (o, dots), estimated values (-, solid line). (b) Gaussian gating network outputs: g_1 (…, dot line), g_2 (-, solid line) and g_3 (--, dash line).

3. Complementary modular modelling

3.1 Sugar crystallization

In sugar production the purpose is to grow sucrose crystals with a required standard of quality, essentially measured by the purity, by the shape and by the crystal size distribution (CSD). The crystallization process occurs through mechanisms of nucleation, growth and agglomeration, which are known to be affected by several operating conditions. Agglomeration, in particular, is an undesired phenomenon, to a large extent not yet understood, which has significant effect in the CSD, i.e. in the final product quality. The search for efficient process model is thus linked both to the scientific interest of understanding fundamental mechanisms of the crystallization and to the relevant practical interest of daily production requirements, i.e. mainly optimisation and control purposes.

The difficulty in crystallization modelling is essentially on the accurate description of the CSD and their related quantities – mass averaged crystal size (MA) and coefficient of variation (CV). The experience with models neglecting agglomeration and/or nucleation mechanisms shows that the CSD predictions do not correspond to the experimentally obtained AM and CV at the end of the process run. Therefore, accurate modelling can only be achieved by incorporating agglomeration and nucleation mechanisms.

3.2 Partial mechanistic model

The mechanistic model considered below is investigated by several authors (Feyo de Azevedo et al. 1993,1994) and proved to give a relevant interpretation to the physical nature of the process considered.

Mass balance. The mass of water (M_w), impurities (M_i), dissolved sucrose (M_s) and crystals (M_c) are included in the following set of conservation mass balance equations

$$\frac{dM_w}{dt} = F_f \rho_f (1 - B_f) + F_w \rho_w - J_{vap}, \quad \frac{dM_i}{dt} = F_f \rho_f B_f (1 - Pur_f), \tag{6.1}$$

$$\frac{dM_s}{dt} = F_f \rho_f B_f Pur_f - J_{cris}, \qquad \frac{dM_c}{dt} = J_{cris} \tag{6.2}$$

Energy balance. The second part of the model is the energy balance

$$\frac{dT_m}{dt} = aJ_{cris} + bF_f + cJ_{vap} + d \tag{7}$$

where *a*, *b*, *c*, *d* incorporate the enthalpy terms and specific heat capacities derived as functions of physical and thermodynamic properties. Details with respect to the evaporation rate (J_{vap}) and the thermal conditions can be found elsewhere (Georgieva et al, 2003a, 2003b).

Population balance (in volume coordinates). The kinetics mechanisms of nucleation, crystal growth and particle agglomeration are defined by the population balance. There are different mathematical representations of it depending on the crystallisation phenomena taken into account. Most of the crystalliser models reported in the literature neglect the agglomeration effect. For the process in hand this assumption appears to be irrelevant since agglomeration is registered in the process run. The population balance is expressed by the leading moments of CSD in volume coordinates since agglomeration must obey mass conservation low,

$$\frac{dm_0}{dt} = B - \frac{1}{2}\beta m_0^2 \quad , \quad \frac{dm_1}{dt} = \overline{G}_v m_0 \, , \tag{8.1}$$

$$\frac{dm_2}{dt} = 2\overline{G}_v m_1 + \beta m_1^2 \, , \quad \frac{dm_3}{dt} = 3\overline{G}_v m_2 + 3\beta m_1 m_2 \, . \tag{8.2}$$

The main process nonlinearities are included in the crystallisation rate

$$J_{cris} = \rho_c \frac{dm_1}{dt} \, . \tag{9}$$

The kinetic parameters considered are the nucleation rate (B), the agglomeration kernel (β) and the linear growth rate (G) from which the volume growth rate can be determined

$$\overline{G}_v = 3k_v^{1/3}\left(\frac{V_c}{m_0}\right)^{2/3} G \tag{10}$$

3.3 Complementary hybrid structure

The complementary structure is a combination between an ANN and first principle equations in a serial hybrid structure, where the known physical constrains (the mass, energy and population balances) are modelled by their analytical expressions (eqs. 6-10) and the kinetic parameters are approximated by an ANN, (see Fig.4). The process variables S (supersaturation), Pur_{sol} (purity of the solution) T_m (temperature) and υ_c (volume fraction of crystals) are known to determine the process kinetics (resp. the kinetic parameters) therefore they are considered as network inputs. Direct measurements are available only for T_m, the other variables are computed through software sensors (Feyo de Azevedo et al., 1993). Each kinetic parameter can be approximated either by training of individual neural networks or all of them simultaneously as outputs of one common ANN. The latter structure was preferred in this study as less computationally involved.

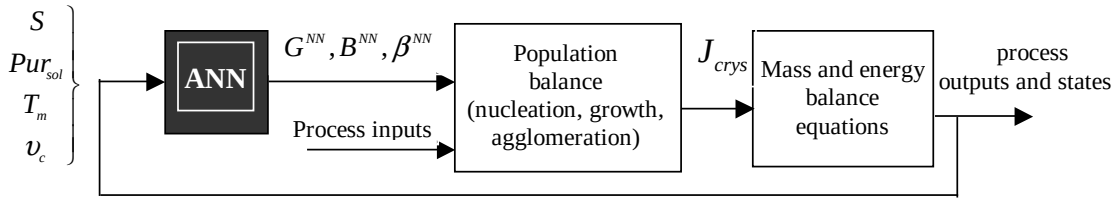


Fig. 4 Complementary hybrid (analytical and data-driven) modular structure

The supervised mode of network training requires target values for the kinetic parameters. As measurements of these process variables are not available the sensitivity approach of hybrid network training is performed (Psichogios and Ungar, 1992). The network outputs are propagated through a partial mechanistic model, to get an output for which measurements or reliable estimations are available (see Fig.5). Note that the partial mechanistic model involved in the hybrid network training is viewed as a fixed parameterised part of the network. In the particular case the mass of crystals (M_c^{hyb}) is considered as the hybrid network output, which is compared with software sensor estimations (M_c^{obs}) for the same variable. The residual between them is termed as the observation error ($e_{obs=}M_c^{hyb} - M_c^{obs}$). The (training) error signal for updating the network weights is set

as the observation error multiplied by the partial derivatives of the hybrid model output with respect to the network outputs (see Georgieva et al, 2003b for more details)

$$e_{tr} = e_{obs}\left[\partial M_c^{hyb}/\partial G \quad \partial M_c^{hyb}/\partial B \quad \partial M_c^{hyb}/\partial \beta\right]^T . \tag{11}$$

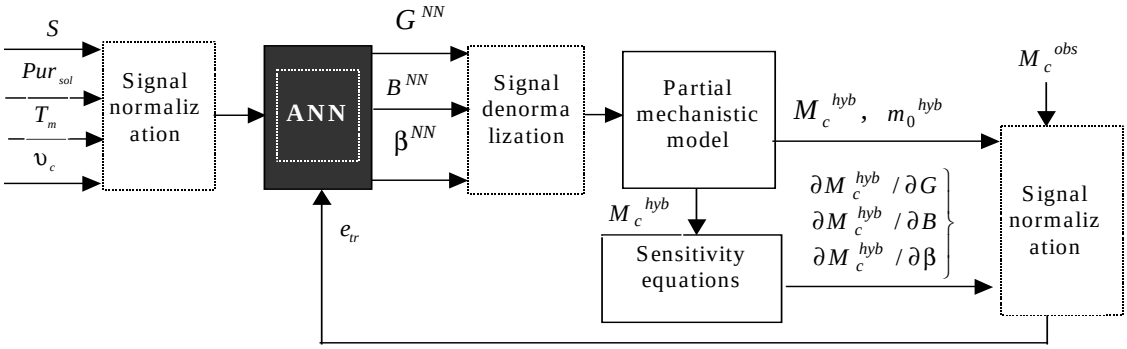


Fig.5 Hybrid ANN training procedure (sensitivity approach)

The main drawback of the hybrid modelling structure is that it suffers of a relatively long computational time, as for every training step a solution of the set of ordinary differential equations is required: the partial mechanistic model to get the mass of crystals and the sensitivity equations to get the partial derivative in eq. (11). In Fig. 6 the main CSD parameters, namely AM and CV, at the end of 10 batches are compared with corresponding experimental data obtained by off-line laboratory (sieve) analysis of mass-size distribution. The hybrid model predictions closely match the real data, which serves as a test for evaluating the model reliability. The model is now investigated as being an essential part of nonlinear model based predictive control algorithm.

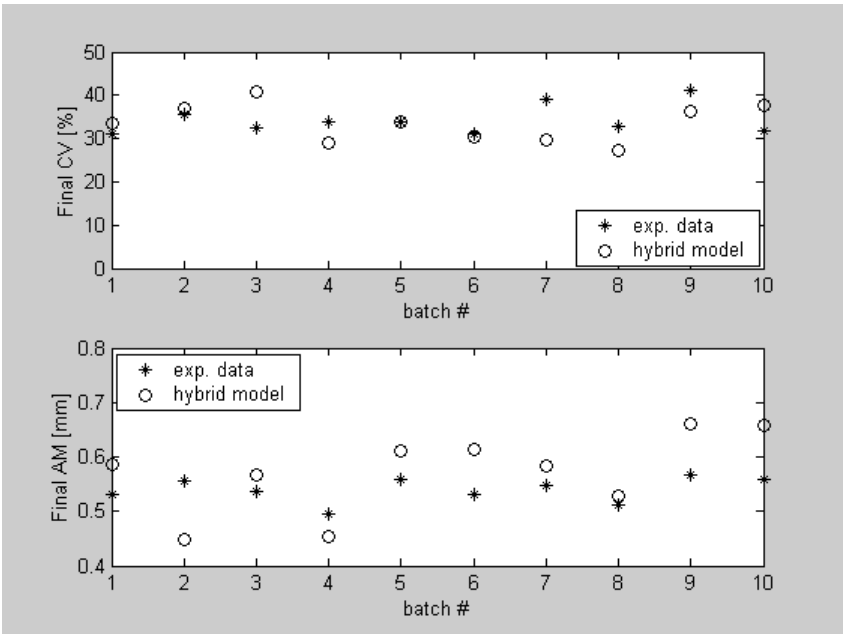


Fig. 6 Final CSD (CV and AM) – experimental data and hybrid model predictions

4. Conclusions

The present work illustrated the application of two modelling alternatives: competitive and complementary architectures.

The competitive approach was formulated in the framework of modular networks for modelling reaction kinetics in biological processes. The study was restricted to the very simple ME architecture, with linear expert modules. The main results showed that the ME was able to perform task decomposition, in the sense that it could decompose the input space in three partitions that in reality correspond to 3 different growth pathways. In terms of modelling errors, it was shown that the ME did not represent an advantage in relation to the BP network, at least for the 2 simple case studies presented. Additional studies with more complex multidimensional problems, with the ASM2 wastewater treatment model (Gujer et al., 1995), are in progress. Nonlinear expert networks were tested. The results obtained so far show that the expert networks are able to discriminate and to develop expertise in describing all metabolic pathways involved.

The complementary modular principle applied to sugar crystallization modelling consists of a serial combination of a partial mechanistic model reflecting the mass, energy and population balances and the poorly known kinetic parameters (nucleation rate, growth rate, agglomeration kernel), are replaced by a feedforward ANN. This knowledge-based hybrid model demonstrates good agreement with the experimental data available.
A reliable description of the kinetic parameters is of special importance not only for the academic understanding of the crystallisation phenomena but also for the purposes of optimising the manipulated input time profiles, with the objective to obtain sugar crystals with desired quality characteristics.

Modelling based on the competitive or complementary modular principles of integrating the process knowledge offers a reasonable compromise between the extensive efforts to get a fully parameterised structure, as are the mechanistic models and the poor generalisation of the complete data-based modelling approaches.

References

1. Feyo de Azevedo, S., Chorão, J., Gonçalves, M.J., & Bento, L. (1993). On-line Monitoring of White Sugar Crystallization through Software Sensors - Part I. *Int. Sugar JNL.*, 95, 483-488.
2. Feyo de Azevedo, S., Chorão, J., Gonçalves, M.J., & Bento, L. (1994). On-line Monitoring of White Sugar Crystallization through Software Sensors - Part II. *Int. Sugar JNL.*, 96, 18-26.
3. *Georgieva P.*, Feyo de Azevedo, M. J. Goncalves, P. Ho (2003a). Modelling of sugar crystallization through knowledge integration. *Eng. Life Sci.*, WILEY-VCH 3 (3) 146-153.
4. *Georgieva* P., M. J. Meireles, S. Feyo de Azevedo (2003b). Knowledge-based hybrid modelling of fed-batch sugar crystallization when accounting for nucleation, growth and agglomeration phenomena. *Chemical Engineering Science* (in press).
5. Gujer,W., Henze, M., Mino, T., Matsuo, T., Wentzel, M. C., and Marais, G. V. R. (1995). The activated sludge model Nº 2: biological phosphorus removal. Water Sci. and Technol., London, England, 31(2),1-12.
6. Haykin S. (1994), *Neural Networks: A comprehensive foundation*, Prentice Hall, UK.
7. Jacobs R.A., M.I. Jordan, and A.G. Barto, (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science* 15, 219-250.
8. Jacobs, R.A., Jordan, M.I. (1991). A competitive modular connectionist architecture, Advances in Neural Information Processing Systems 3, R.P. Lippman, J.E. Moody and D.J. Touretzky Eds, pp. 767-773. San Mateo, CA Morgan Kaufmann.
9. Jordan, M.I. and Jacobs, R.A., (1994). Hierarchical mixtures of experts and the EM algorithm, Neural computation, **6**, pp. 181-214.
10. Montague, G., Morris, J., (1994). Neural network contribution in biotechnology, Trends Biotechnol., 12, pp. 312-324.
11. Psichogios D.C., L. H. Ungar (1992). A hybrid neural network - first principles approach to process modelling, *AIChE J.*, 38(10) 1499-1511.
12. Schubert, J., Simutis, R., Doors, M., Havlik, I. and Lübbert, A., (1994). Hybrid Modelling of Yeast Production Processes, Chem. Eng. Technol., 17, pp. 10-20.
13. Sonnleitner, B. and Käppeli, O., (1986). Growth of Saccharomyces cerevisiae is controlled by its Limited Respiratory Capacity: Formulation and Verification of a Hypothesis, Biotech. Bioeng., 28, pp. 927-937.

DEAD CORE IN POROUS CATALYSTS: MODELLING AND SIMULATION OF A CASE PROBLEM USING MATHEMATICA

Miguel Angelo Granato*, Luiz Carlos de Queiroz
m_granato@uol.com.br, queiroz@dequi.faenquil.br
* corresponding author

Chemical Engineering Department
FACULDADE DE ENGENHARIA QUÍMICA DE LORENA – FAENQUIL
Lorena – São Paulo – Brazil

Abstract

In this work, an approach of the concept of dead core in a porous catalytic particle is made, and a mathematical model for analysis of the dead core for a single, irreversible and isothermal steady state chemical reaction is presented. The main factors that influence the existence of the dead core are defined, the distribution of reactant concentration and the position of dead core for zeroth and first order reactions, in catalysts with classical geometry of an infinite slab are calculated. The software *Mathematica*, which generates the solution of the differential equations, implements the calculation and the corresponding graphs that confirm the required conditions to the existence of the dead core. The results agreed with those published in the literature.

Notation

u = dimensionless concentration
X = dimensionless coordinate
f = Thiele modulus
a = magnitude of dead core ($0 < a < 1$)
α = geometric factor ($\alpha = 1$, slab; $\alpha = 2$, cylinder; $\alpha = 3$, sphere)
n = reaction order

Introduction

For some cases in heterogeneous catalysis, the catalyst has the shape of a porous grain and reactant diffusion into the grain occurs.

If the reaction rate is low, when compared with the diffusion rate, the size of the grain does not represent any problem for the concentration in inner points be almost the same from those at the surface.

Otherwise, if reaction occurs much faster than diffusion, equilibrium can be reached even before that all reactants have spread inside the whole catalyst particle. In this case, a region within the catalyst particle will appear, where reaction will never take place. This region is called Dead Core.

Depending on the dimensions of the grain, the catalyst is not entirely active, and then the reaction yield is low.

Mathematical Model of Dead Core

A mathematical model of the reaction-diffusion phenomenon for analysis of the dead core in porous catalysts for a single, irreversible and isothermal steady state chemical reaction, was developed. For an isothermal particle of any geometry, diffusion and a chemical reaction of nth order are described by an ordinary second order differential equation , [1] and [2]:

$$X^{1-a}\,\frac{d}{dX}\left(X^{a-1}\,\frac{du}{dX}\right) = f^2 u^n \tag{1}$$

Assuming the existence of the dead core, the problem can be posed in the form of Equation (1), with the following boundary conditions:

$$X = 1 \quad \Rightarrow \quad u = 1 \tag{2}$$

$$X = a \quad \Rightarrow \quad \frac{du}{dX} = 0 \tag{3}$$

and the condition:

$$X = a \quad \Rightarrow \quad u = 0 \tag{4}$$

where "*a*" represents the dead core position, with 0 < a < 1.

Case 1: Slab, Zeroth Order Reaction:

The resulting equation for a zeroth order reaction is an ordinary second order linear differential equation as follows:

$$\frac{d^2u}{dX^2} = f^2 . \tag{5}$$

The analytical solution of (2) is:

$$u(X) = 1 - \frac{f^2}{2} + f^2 a - f^2 aX + \frac{f^2}{2} X^2 . \tag{6}$$

The position of the dead core will be given when $u(a)=0$

$$1 - \frac{f^2}{2} + f^2 a - f^2 a^2 + \frac{f^2}{2} a^2 = 0 , \tag{7}$$

if $f > \sqrt{2}$, then the dead cores exists and its position is:

$$a = 1 - \frac{\sqrt{2}}{f} . \tag{8}$$

The concentration profile is

$$u(X) = \begin{cases} 0 \\ \frac{f^2}{2}\left[X - 1 + \frac{\sqrt{2}}{f} \right]^2 \end{cases} . \tag{9}$$

Case 2: Slab, First Order Reaction

Applying a =1 and n = 1 to equation (1) gives:

$$\frac{d^2u}{dX^2} = f^2 u . \tag{10}$$

Which has the solution:

$$u(X) = \frac{e^{f - fX}\left(e^{2fa} + e^{2fX}\right)}{e^{2f} + e^{2fa}} . \tag{11}$$

Dead core position will be given when $u(a) = 0$

$$Sech\ (f - fa) = 0 . \tag{12}$$

However, on determination of the dead core position, $u(a) \to 0$ for $a \to \pm\infty$ and, as $0 < a < 1$, there is no occurrence of dead core in this case.

Mathematica Solution

The following procedure models the dead core, [3].

Case 1 Solution:

1. Clear all previous inputs and assign the corresponding values for geometric factor and reaction order:

```
In[1]:= Clear [x, a, α, n];
In[2]:= α = 1;
In[3]:= n = 0;
```

2. Define a function for equation (1):

```
In[4]:= eqn = x^(1-α) ∂x (x^(α-1) ∂x u[x]) - ϕ^2 (u[x])^n
Out[4]= -ϕ^2 + u''[x]
```

3. Solve equation (1) with the assigned boundary conditions:

```
In[5]:= DSolve[{eqn == 0, u[1] == 1, u'[a] == 0}, u[x], x]
Out[5]= {{u[x] → 1/2 (2 - ϕ^2 + 2 a ϕ^2 - 2 a x ϕ^2 + x^2 ϕ^2)}}
```

4. Simplify solution:

```
In[6]:= Apart[%, x = a]
Out[6]= {{u[a] → 1/2 (2 - ϕ^2 + 2 a ϕ^2 - a^2 ϕ^2)}}
```

5. Solve for "*a*":

```
In[7]:= Solve[1/2 (2 - ϕ^2 + 2 a ϕ^2 - a^2 ϕ^2) == 0, a]
Out[7]= {{a → (-√2 + ϕ)/ϕ}, {a → (√2 + ϕ)/ϕ}}
```

6. Select root $0 < a < 1$:

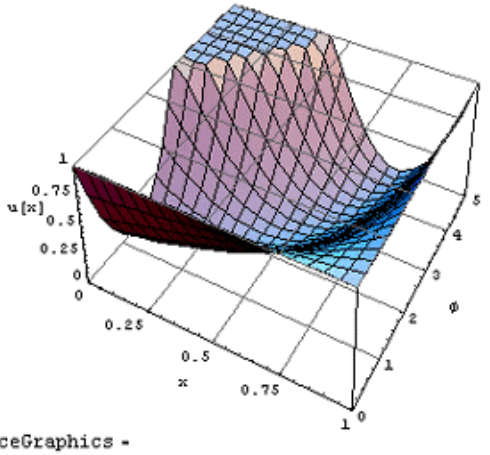
```
In[8]:= Apart[(-√2 + ϕ)/ϕ]
Out[8]= 1 - √2/ϕ
```

7. Simplify concentration expression assigning the value of "*a*", for which the concentration is zero:

```
In[21]:= FullSimplify[u[x] -> 1/2 (2 - ϕ^2 + 2 a ϕ^2 - 2 a x ϕ^2 + x^2 ϕ^2), a = 1 - √2/ϕ]
Out[21]= u[x] → 1/2 (2 + (-1 + x) ϕ (2 √2 + (-1 + x) ϕ))
```

One observes that concentration is a function of position "X" and of Thiele Modulus "f". Three-dimensional graphs of the function $u = u$ (X, f) will be generated, as shown in Figures 1 and 2.

```
In[10]:= Plot3D[1/2 (2 + (-1 + x) ϕ (2 √2 + (-1 + x) ϕ)), {x, 0, 1}, {ϕ, 0, 5},
  PlotRange → {{0, 1}, {0, 5}, {0, 1}}, AxesLabel → {"x", "ϕ", "u[x]"},
  PlotPoints → 20, FaceGrids → {{0, 0, 1}, {0, 0, -1}}, AspectRatio → 1]
```

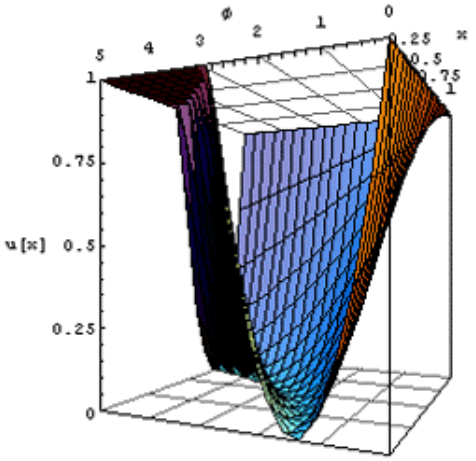



```
Out[10]= - SurfaceGraphics -
```

Figure 1 – Sample 3D Graph: concentration vs. Thiele modulus vs. position.

Figure 2 shows the option **`ViewPoint`** which allows a different graph perspective. **`PlotRange`** zooms in the image for a better definition of a particular region.

```
In[11]:= Show[%, PlotRange → {{0, 1}, {0, 5}, {0, 1}},
          ViewPoint -> {-2, -.9, 0}, AspectRatio → 1.2]
```



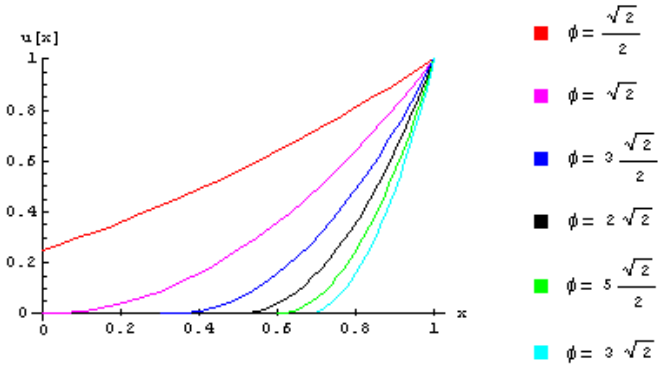

```
Out[11]= - SurfaceGraphics -
```

Figure 2 – Zoom in for detailed view.

The following procedure generates the graph that shows the positions of dead core and reactant concentration profiles for several values of Thiele modulus:

```
In[22]:= group = Table[{1/2 (2 + (-1 + x) ϕ (2 √2 + (-1 + x) ϕ))}, {ϕ, 0, 3 √2, √2/2}];

In[23]:= p1 = Plot[group[[2]], {x, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[1, 0, 0]];

In[24]:= p2 = Plot[group[[3]], {x, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[1, 0, 1]];

In[25]:= p3 = Plot[group[[4]], {x, .3, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[0, 0, 1]];

In[26]:= p4 = Plot[group[[5]], {x, .5, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[0, 0, 0]];

In[27]:= p5 = Plot[group[[6]], {x, .6, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[1, 1, 1]];

In[28]:= p6 = Plot[group[[7]], {x, .7, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[0, 1, 1]];
```

```
In[29]:= Show[p1, p2, p3, p4, p5, p6]
```




```
Out[29]= - Graphics -
```

Figure 3 –Positions of dead core for several values of ϕ.

The purple line in Figure 3 represents a value of $f = \sqrt{2}$. For such a value, the concentration is zero exactly at the catalytic particle center (a = 0). As the Thiele modulus increases, one can see that the dead core occurs at positions closer from the surface and $a \to 1$ for $f \to \infty$. Then, the dead core tends to occupy the whole particle when $f \to \infty$.

Case 2 Solution:

An analog procedure for solving Case 2 equation was adopted and it is shown below:

1. Clear all previous inputs and assign the corresponding values for geometric factor and reaction order:

```
In[1]:= Clear [x, a, α, n]

In[2]:= α = 1;

In[3]:= n = 1;
```

2. Define a function for equation (1):

```
In[4]:= eqn = x^(1-α) ∂x (x^(α-1) u'[x]) - ϕ^2 u[x]^n

Out[4]= -ϕ^2 u[x] + u''[x]
```

3. Solve equation (1) with the boundary conditions:

```
In[5]:= DSolve[{eqn == 0, u[1] == 1, u'[a] == 0}, u[x], x]

Out[5]= {{u[x] → (E^(ϕ-xϕ) (E^(2aϕ) + E^(2xϕ)))/(E^(2ϕ) + E^(2aϕ))}}
```

4. Attempt to solve for "*a*":

```
In[6]:= Solve[(E^(ϕ-xϕ) (E^(2aϕ) + E^(2xϕ)))/(E^(2ϕ) + E^(2aϕ)) == 0, x = a]

Solve::ifun : Inverse functions are being used by Solve, so some solutions may not be found.

Out[6]= {{a → -∞/ϕ}}
```

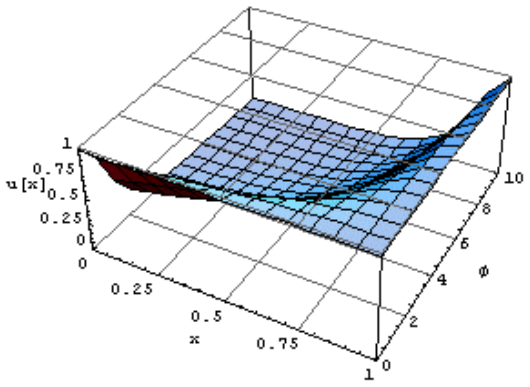
As seen in the previous analytical solution, there is no dead core.

5. Simplify the solution of Equation (1):

```
In[7]:= FullSimplify[(E^(ϕ-xϕ) (E^(2aϕ) + E^(2xϕ)))/(E^(2ϕ) + E^(2aϕ))]

Out[7]= Cosh[(a - x) ϕ] Sech[ϕ - aϕ]

In[8]:= Apart[Cosh[(a - x) ϕ] Sech[ϕ - aϕ], x = a]

Out[8]= Sech[ϕ - aϕ]
```

The following three-dimensional graphs have been generated, plotting concentration versus Thiele Modulus versus position.

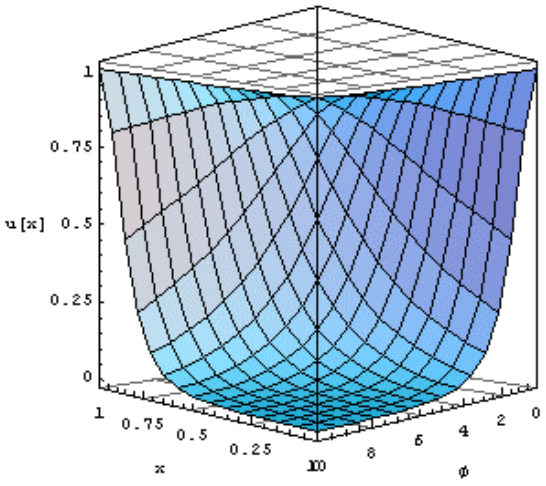
```
In[9]:= s1 = Plot3D[%, {a, 0, 1}, {ϕ, 0, 10},
          AxesLabel → {"x", "ϕ", "u[x]"},
          FaceGrids → {{0, 0, 1}, {0, 0, -1}}]
```




```
Out[9]= - SurfaceGraphics -
```

Figure 4 – Sample 3D graph for first order reaction in slab geometry.

```
In[10]:= Show[%, ViewPoint -> {-2, 2, 0}]
```



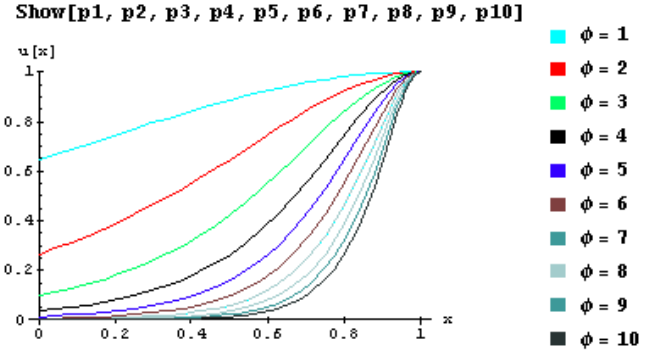

```
Out[10]= - SurfaceGraphics -
```

Figure 5 – Different view of concentration profiles vs. Thiele Modulus vs. position.

To generate plots of reactant concentration profiles for several values of f, a similar procedure to the Case 1 was adopted.

```
In[11]:= group = Table[(Sech[ϕ - a ϕ]), {ϕ, 0, 10, 1}] ;
In[12]:= p1 = Plot[group[[2]], {a, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[0.5]}];
In[13]:= p2 = Plot[group[[3]], {a, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[2]}];
In[14]:= p3 = Plot[group[[4]], {a, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[.4]}];
In[15]:= p4 = Plot[group[[5]], {a, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → RGBColor[0, 0, 0]];
In[16]:= p5 = Plot[group[[6]], {a, 0, 1}, PlotRange → {0, 1},
          AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[.7]}];
```

```
In[17]:= p6 = Plot[group[[7]], {a, 0, 1}, PlotRange → {0, 1},
         AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[1, .5, .5]}];
In[18]:= p7 = Plot[group[[8]], {a, 0, 1}, PlotRange → {0, 1},
         AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[1.2, .8, .8]}];
In[19]:= p8 = Plot[group[[9]], {a, 0, 1}, PlotRange → {0, 1},
         AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[2.5, .2, .8]}];
In[20]:= p9 = Plot[group[[10]], {a, 0, 1}, PlotRange → {0, 1},
         AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[1.5, .6, .6]}];
In[21]:= p10 = Plot[group[[11]], {a, 0, 1}, PlotRange → {0, 1},
         AxesLabel → {"x", "u[x]"}, PlotStyle → {Hue[5.5, .2, .2]}];

In[22]:= Show[p1, p2, p3, p4, p5, p6, p7, p8, p9, p10]
```

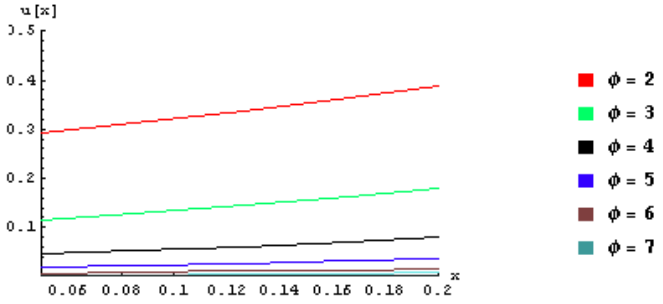



```
Out[22]= - Graphics -
```

Figure 6 – Reactant concentration profiles for a first order reaction in a slab.

The plots show that reactant concentration never reaches zero when the reaction order is 1 for a slab and there is no occurrence of dead core in this case, that agrees with results from [4]. This is confirmed by viewing a zoomed-in section of the previous graph.

```
In[23]:= Show[%, PlotRange → {{.05, .2}, {0, .5}}, AxesOrigin → {0.05, 0}]
```




```
Out[23]= - Graphics -
```

Figure 7 – Zoomed-in concentration profiles.

Conclusions

The analytical solutions and the results generated by *Mathematica* agree with those published in the literature. The use of *Mathematica* as a computational tool to solve the proposed problem provides an application of the software in catalysis.

Acknowledgements

Dr. Cláudio Umberto Granato – Hospital S. João – Porto – Portugal. – financial support.
CAPES.

References

[1] - *The Mathematical Theory of Diffusion and Reaction of Permeable Catalysts,* Vol. 1, R. Aris, Ed. Clarendon Press, Oxford, **1975**.
[2] - F. Garcia-Ochoa, A. Romero *AIChE. J*., **1988,** *34*, 1916-1918.
[3] - *Mathematica: A System for Doing Mathematics by Computer* S. Wolfram,. 2nd. ed. Addison Wesley, **1991**.
[4] - Chemical Reactor Analysis and Design, G. F. Froment, K. B. Bischoff, 2nd ed. John Wiley & Sons, **1990.**

Multiple Nonlinear Regression Analysis for the Baker's Yeast Fermentation Parameters Estimation

Celina P. Leão[1], Filomena O. Soares[2], and Edite M. G. P. Fernandes[1]
[1] Department of Systems and Production Engineering, University of Minho,
Campus de Gualtar, 4710-057 Braga, Portugal
{cpl,emgpf}@dps.uminho.pt
[2] Department of Industrial Electronics, University of Minho,
Campus de Azurém, 4800-058 Guimarães, Portugal
filomena.soares@dei.uminho.pt

Abstract

The baker's yeast used in the bread making and beer industries as a microorganism, has an important industrial role. A system of differential-algebraic equations is used to predict the behavior of the concentration of the state variables over a 20 hours time period, in a well-mixed reactor. Kinetic and yield coefficients parameters in the model, usually obtained from the literature, can be considered unknown or be estimated by fitting the model to the experimental data through an optimization procedure. A multiple nonlinear regression can be performed by simultaneously fitting all the differential equations to the experimental data collected for the state variables. This technique will enable us to obtain the value of the parameters of the model, which minimize the overall sum of the squared residuals. In order to reduce the dimension of the optimization procedure, only the most significant model parameters were used. A runs statistical test on the deviations between the experimental and predicted values of the state values to ensure random distribution was also performed to accompany the regression analysis.

1. Introduction

The baker's yeast, essentially, composed by living cells of *Saccharomyces cerevisiae*, used in the bread making and beer industries as a microorganism, has an important industrial role. Baker's yeast production is a fed-batch fermentation that uses as substrate feed (carbon) a glucose solution. The simulation procedure represents then a necessary tool to understand clearly the baker's yeast fermentation process. A mathematical model was developed [1] in order to predict the behavior of the concentration of the state variables over a 20 hours time period, in a well-mixed reactor. Considering the kinetics and the gas transfer rates relations, which define algebraic equations, as well as the volume rate equation during the fed-batch process, we end up with a system of differential-algebraic equations [2]. Kinetic [3] and yield coefficients [4] parameters in the model can be obtained from the literature. However, these parameters can be considered unknown and can be estimated fitting the model to the experimental data through an optimization procedure.

Experiments for this fermentation process were carried out in a fed-batch fermentation and data, for all dependent variables, were collected. A multiple nonlinear regression can be performed by simultaneously fitting all differential equations to the data. This technique will enable us to obtain the value of the parameters of the model, which minimize the overall sum of the squared residuals. In

order to reduce the dimension of the optimization procedure, a heuristic sensitivity analysis was performed to identify the most significant model parameters, *i.e.*, the parameters that give the most significant differences between experimental and simulated data [5]. The maximum uptake rate for glucose and oxygen and the yield coefficients were seen to be the most significant model parameters.

For the nonlinear regression analysis, a Marquardt algorithm [6] for multiresponse data, which uses an interpolation technique to combine the Gauss-Newton and Steepest Descent methods, was coded in Matlab. For the solution of the differential system of equations, a stiff Matlab solver was integrated in the developed code.

A runs statistical test on the deviations between the experimental and predicted values of the state values to ensure random distribution was also performed to accompany the regression analysis.

The paper is organized as follows. Section 2 gives a brief description of the baker's yeast fermentation process and presents the model equations. Section 3 presents a multiple nonlinear regression analysis and Section 4 a randomness test of the regression results. Finally, some conclusions are shown in Section 5.

2. Baker's yeast fermentation – equations model

In the baker's yeast fermentation process three metabolic pathways can be distinguish: (1) respirative growth on glucose, (2) fermentative growth on glucose and (3) respirative growth on ethanol. Respirative pathways occur in presence of oxygen and the fermentative one in its absence (with production of ethanol) [1]. The metabolic pathways of fermentative growth on glucose and oxidative growth on ethanol are competitive. This competition is governed by the respiratory capacity of the cells. If the instantaneous oxygen uptake capacity exceeds the oxygen need for total respiratory glucose uptake, then, all sugar uptakes follow the respiratory pathway (1) with the remaining oxygen being spent on ethanol respiratory uptake (3). Otherwise, if the instantaneous oxygen uptake capacity is not enough, part of glucose uptake follows the respiratory pathway (1) while the remaining follows the fermentative pathway (2).

The mechanistic model for the fed-batch fermentation is obtained from mass balances for all the components [1]. It is assumed that the yield coefficients, $\left(Y^{O}_{X/S},\ Y^{r}_{X/S},\ Y^{r}_{X/E},\ Y^{OE}_{X/E},\ Y^{O}_{X/O},\ Y^{OE}_{X/O},\ Y^{O}_{X/C},\ Y^{r}_{X/C},\ Y^{OE}_{X/C}\right)$, are constant and the dynamics of the gas phase can be neglected. The kinetics equations for baker's yeast growth, $\left(\mu^{O}_{S},\ \mu^{r}_{S},\ \mu^{O}_{E}\right)$, are considered as Monod equations [2]. Then the set of differential-algebraic equations is:

· mass balance for the biomass, $$\frac{dX}{dt} = \left(\mu_S^O + \mu_S^r + \mu_E^O - D\right)X \quad (1)$$

· mass balance for the sugar, $$\frac{dS}{dt} = \left(-\frac{\mu_S^O}{Y_{X/S}^O} - \frac{\mu_S^r}{Y_{X/S}^r}\right)X + \left(S_f - S\right)D \quad (2)$$

where S_f is the substrate concentration in the feed,

· mass balance for the ethanol, $$\frac{dE}{dt} = \left(\frac{\mu_S^r}{Y_{X/E}^r} - \frac{\mu_E^O}{Y_{X/E}^{OE}}\right)X - DE \quad (3)$$

· mass balance for the oxygen, $$\frac{dO}{dt} = \left(-\frac{\mu_S^O}{Y_{X/O}^O} - \frac{\mu_E^O}{Y_{X/O}^{OE}}\right)X - DO + OTR \quad (4)$$

· mass balance for the carbon dioxide, $$\frac{dC}{dt} = \left(\frac{\mu_S^O}{Y_{X/C}^O} + \frac{\mu_S^r}{Y_{X/C}^r} + \frac{\mu_E^O}{Y_{X/C}^{OE}}\right)X - DC - CTR \quad (5)$$

· accumulation of the working volume during the fed-batch process, $$\frac{dV}{dt} = DV \quad (6)$$

where D is dilution rate (ratio feed rate/volume) defined by $D = F/V$.

The gas transfer rates are given by

$$OTR = K_L^O a\left(O^* - O\right), \quad CTR = K_L^C a\left(C - C^*\right) \quad (7\text{-}8)$$

where $K_L^i a$ are overall mass transfer coefficients for oxygen and carbon dioxide and O^* and C^* are the corresponding equilibrium concentrations.

The kinetics equations for the respirative regime are:

$$\mu_S^O = Y_{X/S}^O . q_S, \quad \mu_S^r = 0, \quad \mu_E^O = min\left(\mu_{E_1}^O, \mu_{E_2}^O\right) \quad (9a\text{-}11a)$$

or, for the respiro-fermentative regime:

$$\mu_S^O = Y_{X/S}^O . \frac{q_O}{a}, \quad \mu_S^r = Y_{X/S}^r \left(q_S - \frac{q_O}{a}\right), \quad \mu_E^O = 0. \quad (9b\text{-}11b)$$

Equations (12-15) must be added to the kinetics equations, (9a-11a) or (9b-11b), for the estimation of the specific growth rate on ethanol, defined as:

$$\mu_{E_1}^O = \mu_E^{max} \frac{E}{E + K_E} \frac{K_i}{S + K_i} \quad (12)$$

$$\mu_{E_2}^0 = \frac{Y_{X/O}^{OE}}{Y_{X/E}^{OE}} \left(q_O - a q_S\right) \quad (13)$$

$$q_S = q_S^{\max} \frac{S}{S + K_S} \tag{14}$$

$$q_O = q_O^{\max} \frac{O}{O + K_O} \tag{15}$$

where $\mu_E^{\max}$ is the maximal specific growth rate, K_i is the inhibition parameter, K_E is the saturation parameter, a is the stoichiometric coefficient of the oxygen in the respiratory pathway of glucose, $q_S^{\max}$ is the maximal specific glucose uptake rate, K_S and K_O are saturation parameters and $q_O^{\max}$ is the maximal specific oxygen uptake rate. The kinetic coefficients $\left(q_S^{\max},\ q_O^{\max},\ \mu_E^{\max},\ K_E,\ K_i,\ K_S,\ K_O\right)$ were considered as constants. The set of differential-algebraic equations, (1-8, 12-15, 9a-11a) or (1-8, 12-15, 9b-11b), defines the model for the baker's yeast fermentation process.

3. Multiple nonlinear regression analysis

The most significant parameters must be computed by fitting the baker's yeast model to experimental data. Experiments for this fermentation process were carried out in a fed-batch fermentation and data for all the state variables $\left(X,\ S,\ E,\ O,\ C\right)$ and the overall mass balance $\left(V\right)$, defining the six dependent variables, were collected.

A multiple nonlinear regression is then performed by simultaneously fitting all six equations (1-6) to the data in order to obtain values for the parameters of the model. To reduce the size of the differential system to be solved all the algebraic equations (7-15) were replaced in the dynamic model (1-6). From now on and for simplicity, we consider the dynamic model (1-6) written in the form

$$\frac{dY_j}{dt} = f_j\left(t, Y, p\right),\quad j = 1,\ ...,\ 6 \tag{16}$$

where t is the independent variable, $Y = \left(X,\ S,\ E,\ O,\ C,\ V\right)^{\mathrm{T}}$ is the vector of the dependent variables and $p = \left(q_S^{\max},\ q_O^{\max},\ Y_{X/S}^{O},\ Y_{X/S}^{r},\ Y_{X/E}^{r},\ Y_{X/O}^{O}\right)^{\mathrm{T}}$ is the vector containing the parameters to be estimated. A heuristic sensitivity analysis [5] identified these six parameters as the most important, from the sixteen previously defined, reducing the dimension of the optimization procedure. We denote the i^{th} element of the vector p by p_i and the j^{th} element of vector Y by Y_j. Given initial conditions and estimated values for the parameters, the differential equations (16) can be numerically integrated to give

$$Y_j = F_j\left(t, p\right),\quad j = 1,\ ...,\ 6. \tag{17}$$

In a multiple regression method we minimize the overall sum of squared residuals given by

$$\Phi = \sum_{j=1}^{6} r_j^{\mathrm{T}} r_j = \sum_{j=1}^{6} \left(Y_j^e - Y_j\right)^{\mathrm{T}} \left(Y_j^e - Y_j\right) \tag{18}$$

where r_j is the residual vector of Y_j and Y_j^e is the vector of m experimental observations of the dependent variable Y_j. For a nonlinear model, the Gauss-Newton iterative method computes a sequence of estimates to the vector $p_{new} = p_{current} + \Delta p$, where Δp is the solution to the system

$$\sum_{j=1}^{6} \mathrm{J}_j^{\mathrm{T}}\, \mathrm{J}_j\, \Delta p = \sum_{j=1}^{6} \mathrm{J}_j^{\mathrm{T}} \left(Y_j^e - Y_j\right) \tag{20}$$

and J_j is the Jacobian matrix of the partial derivatives of Y_j with respect to p, evaluated at all m points where experimental data are available,

$$\mathrm{J}_j = \begin{pmatrix} \dfrac{\partial Y_{j,1}}{\partial p_1} & \cdots & \dfrac{\partial Y_{j,1}}{\partial p_6} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial Y_{j,m}}{\partial p_1} & \cdots & \dfrac{\partial Y_{j,m}}{\partial p_6} \end{pmatrix}. \tag{21}$$

This process is iteratively repeated until Δp becomes small.

It is possible that, for some iterations, the matrices $\mathrm{J}_j^{\mathrm{T}}\, \mathrm{J}_j$ are not positive definite so that the correction vector Δp, given by (20), is not downhill for Φ at $p_{current}$. A Marquardt algorithm, which uses an interpolation technique to combine the Gauss-Newton and steepest descent methods, is used instead, where the coefficient matrix in (20) is substituted by

$$\sum_{j=1}^{6} \left(\mathrm{J}_j^{\mathrm{T}}\, \mathrm{J}_j + \lambda \mathrm{I}\right), \text{ for some } \lambda > 0. \tag{22}$$

Further, to improve efficiency, $\mathrm{J}_j^{\mathrm{T}}\, \mathrm{J}_j$ is scaled so that its diagonal elements become equal to unity, which is equivalent to solving the following set of equations for Δp:

$$\sum_{j=1}^{6} \left(\mathrm{J}_j^{\mathrm{T}}\, \mathrm{J}_j + \lambda \mathrm{D}_j\right) \Delta p = \sum_{j=1}^{6} \mathrm{J}_j^{\mathrm{T}} \left(Y_j^e - Y_j\right), \ \lambda > 0, \tag{23}$$

where D_j is a diagonal matrix containing the diagonal elements of $\mathrm{J}_j^{\mathrm{T}}\, \mathrm{J}_j$ [6]. When the model consists of algebraic equations, the elements of each J_j are easily obtained by differentiating the model.

However, the baker's yeast model consists of differential equations, and its elements must be obtained through the variational equations

$$\frac{d}{dt}\left(\frac{\partial Y_j}{\partial p_i}\right)=\frac{\partial}{\partial p_i}\left(\frac{dY_j}{dt}\right)=\frac{\partial f_j(t,\ Y,\ p)}{\partial p_i},\ j=\ 1,\ ...,\ 6;\ i=\ 1,\ ...,\ 6. \tag{24}$$

This set of ordinary differential equations can be then integrated simultaneously with the model equations (16) to give $Y_j, j=1, \ldots, 6$, and $\frac{\partial Y_j}{\partial p_i}$, $j=1, \ldots, 6$, $i=1, \ldots, 6$, which are required to construct the vectors and matrices in (23).

The final results, for a $\Phi < 1.5$, are shown in Table 1.

Table 1. Comparison between multiple nonlinear regression analysis final results and literature.

	q_S^{max} (h^{-1})	q_O^{max} (h^{-1})	$Y_{X/S}^O$	$Y_{X/S}^r$	$Y_{X/E}^r$	$Y_{X/O}^O$
Model	5.3	0.203	0.45	0.09	1.48	1.71
Literature [3, 4]	3.5	0.256	0.49	0.05	0.10	1.20

4. Randomness statistical test

A test will be performed to investigate the goodness of fit of the model. In a least squares regression, the assumption is made that the model being fitted is the correct one and that the observations deviate from the model in a random fashion. The residuals, r_j, between the experimental values Y_j^e and the computed values Y_j can be either positive or negative. However, if they are random, the sign of the residuals should change in a random fashion. The randomness of the distribution of the residuals (or lack of it) can be visually detected by plotting the residuals versus the independent variable *t*. Figure 1 illustrates the residuals for the six dependent variables. This randomness can also be measured and tested by the *runs statistical test* [7]. This test is based upon the number of runs, R, which represents the number of series of identical signs in the residuals sequence. If the number of positive signs is p_s and the number of negative signs is n_s, then the null hypothesis, H_0, that the model is the correct one and that the residuals are randomly distributed should not be rejected, if the sample R value falls between the lower and the upper critical values taken from [7], at a 0.05 level of significance. Table 2 contains the sample values of R for each variable. Comparing these sample values with the critical values, the conclusions are as shown in the last row of Table 2.

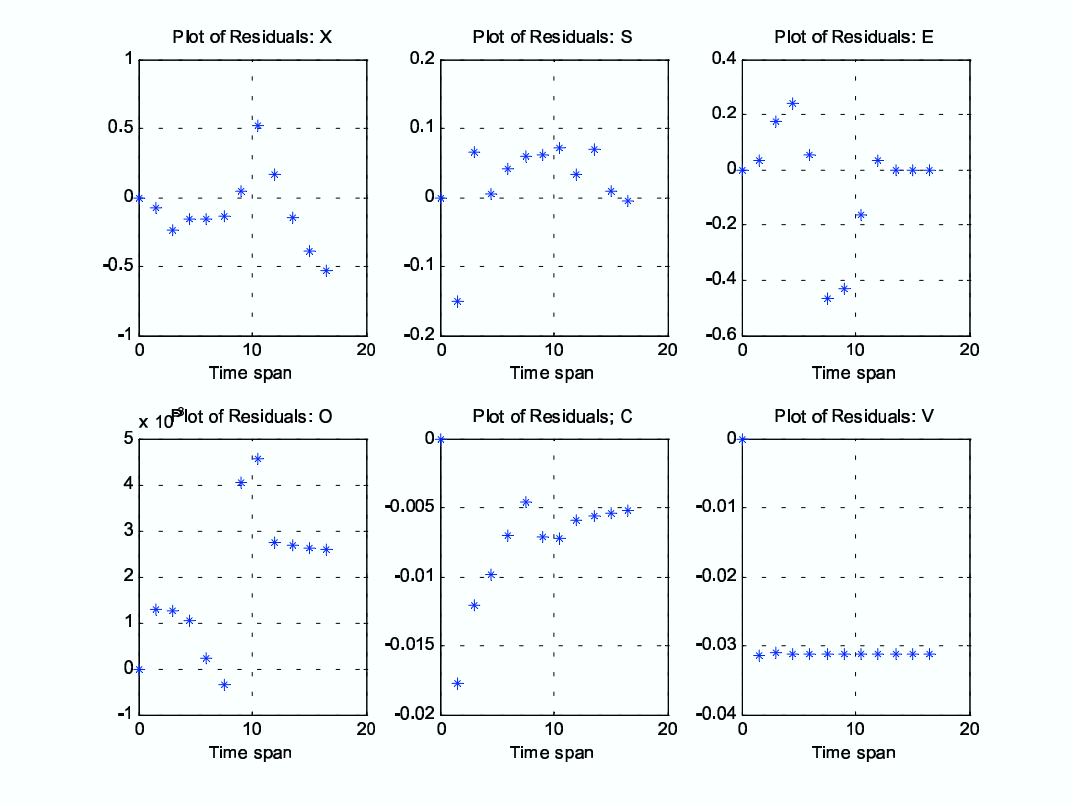


Figure 1. Residuals versus *t*, for the dependent variables.

Table 2. Sample values of *R* and conclusions (* no critical values for these cases).

Runs test for variable	1	2	3	4	5	6
p_s	4	10	7	11	1	1
n_s	8	2	5	1	11	11
Sample value of *R*	4	4	4	3	2	2
H_0 not rejected	yes	*	yes	*	*	*

5. Conclusions

The method used in this paper enables us to fit a baker's yeast model, consisting of multiple dependent variables, to multi-response experimental data in order to obtain the best values for the most significant parameters, which minimize the overall sum of squared residuals between the data and the model.

Two of the significant parameters obtained through the multiple nonlinear regression analysis, q_S^{max} and $Y_{X/E}^r$, comparative to the literature values are very different. However, they allow obtaining good predictions values.

The statistical analysis of the regression results show that the model seems to represent adequately the data and the residuals are randomly distributed for two of the variables being fitted, the biomass and ethanol. For the sugar, nevertheless no critical value for this case, the model follows the experimental trend. The model seems to give a low prediction of oxygen and a high prediction of the carbon dioxide for all values of t, and no significant differences in the volume profile.

References

[1] Soares, F. O., *Monitorização e controlo de fermentadores – Aplicação ao fermento de padeiro*, PhD. Thesis, Universidade do Porto, 1997 *(in portuguese)*.

[2] Leão, C. P., Soares, F. O., *Two Different Strategies for Baker's Yeast Fermentation Process Simulation*, in Nikos E. Mastorakis (Edi.), Recent Advances in Simulation, Computational Methods and Soft Computing, WSEAS Press, 2002, pp- 11-16.

[3] Sonnleitner, B., Käppeli, O. *Growth of Saccharomyces cerevisiae is controlled by its limites respiratory capacity: formulation and verification of a hypothesis,* Biotech. Bioeng., 28, 1986, pp. 927-937.

[4] Pomerleau, Y., *Modélisation et Contrôle d'un Procédé Fed-Batch de Culture des Levures à Pain (Saccharomyces cerevisiae),* PhD Thesis, Université de Montreal, 1990.

[5] Leão, C. P. and Soares, F. O., Heuristic Sensivity Analysis for Baker's Yeast Kinetics Parameters, accepted in Journal Investigação Operacional, 2003.

[6] Wolfe, M. A., *Numerical Methods for Unconstrained Optimization- An introduction*, Van Nostrand Reinhold Company, 1978.

[7] D.H. Sanders, R.J. Eng and A.F. Murph, *Statistics. A Fresh Approach,* Third Edition, McGraw-Hill Book Co., 1985.

COMPUTER MODELING AND SIMULATION IN CHEMICAL PROCESSES POLLUTION PREVENTION

Teresa M. Mata[a] and Carlos A. V. Costa[b]

Laboratory of Processes, Environment and Energy Engineering
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, 4200-465 Porto, PORTUGAL
[a] tmata@fe.up.pt
[b] ccosta@fe.up.pt

Abstract

In this work a strategy is described for designing chemical processes with improved economic and environmental performance, using computer modeling and simulation. This strategy is applied during process development stages of process synthesis and conceptual design, where the flowsheet of a chemical process is developed and evaluated. The several steps of this strategy include process synthesis, modelling and simulation, preliminary assessment, generation of alternative design options, detailed assessment, feasibility analysis and finally screening of alternatives to arrive to the best design option. This strategy takes into account the process flowsheet, the open and fugitive emissions and the potential environmental impacts (PEI's) of a chemical process. It places a high emphasis on pollution prevention and waste minimisation, focusing on those chemical components and process streams that have the largest contribution to the PEI's and the largest economic potential, revealing where attention should be focused when designing a chemical process.

Introduction

The chemical industry provides a vast array of products and materials that are essential for modern societies. However gaseous, liquid and solid wastes are inevitably generated during the manufacture of any product. In the past three decades industry have been in the position of responding to legislation imposed as a consequence of a perceived environmental crisis. The mode of operation was essentially unplanned and always reactive rather than proactive. There has been little operational guidance about how to do better.

Apart from creating potential environmental problems, wastes represent losses from the production process of valuable raw materials and energy, requiring significant investment in pollution control practices. The waste generated by chemical industries is often associated to inefficient processes. Thus reducing waste by improving efficiency will maximise profits, while reducing the environmental impacts. To address this challenge, rather than using the traditional end-of-pipe approaches chemical engineers need to assess, improve and integrate the environmental performance of processes with the objective of avoiding waste generation.

During process synthesis and conceptual design important decisions are made that will determine the economic viability, safety and environmental impact of the final design. In these preliminary stages of process development an optimised structure of a chemical process is determined and a number of suitable process alternatives or possible structures are identified and then evaluated to get the best solution.

While detailed environmental impact assessments have been performed for decades, their implementation has generally been restricted to evaluations of final designs. A better approach would be to evaluate environmental performance in the design development process. At the earliest stages of process design, only the most elementary data on raw materials, products and by-products may be available resulting in a large number of design alternatives that need to be considered. Although there are trade-offs between different environmental impacts decisions must be made. Supporting these

decisions require environmental assessment tools that chemical engineers will need to master. Environmental assessment tools are required not only to quickly assess the potential environmental impacts and toxicity potential of products and processes but also to identify key compounds of concern or emission points in a chemical process. For example, the waste reduction algorithm (WAR) from U.S. EPA (Young *et al.*, 2000) is an environmental assessment tool, which can be used to evaluate the potential environmental impacts of alternative design options.

The original version of the WAR algorithm, developed by Hilaly and Sikdar (1994), introduced the concept of a pollution balance, which was strictly mass based. Cabezas *et al.* (1999) introduced the generalised WAR algorithm with a PEI balance, which assigned environmental impact values to different pollutants, as an improvement upon the original WAR algorithm. Young and Cabezas (1999) extended the PEI balance to include the consumption of energy by the process into the environmental evaluation.

Chemical engineering practice has traditionally relied on experience-based and heuristic or rule-of-thumb type methods to evaluate some feasible process design (Douglas 1988). Mathematical algorithms are used to find the optimal solution from these manually determined feasible process design options. The fault in this process is that it is virtually impossible to manually define all of the feasible process system options comprising more than a few operating units (Bumble 2000). Chemical process simulation techniques have emerged as tools for providing process design and developing clean technology for pollution prevention and waste minimisation. Most state of the art process simulators are powerful tools for the analysis of pollution prevention alternatives in a wide range of industrial processes.

Steady state process simulators make it possible to run the plant as a model on a computer and test out operation scenarios (e.g. higher flowrates, different feedstocks, modified operating conditions, etc.) before they are tried on the actual plant. Examples of commercially available process simulators that can be used to model chemical processes are ASPEN PLUS™ by Aspen Technology Inc., CHEMCAD™ by ChemStations, Inc., HYSYS™ by Hyprotech Ltd. and PRO/II® by Simulation Sciences Inc., etc. With the ever-increasing capabilities in computer power and accurate models for describing process units, process simulators make it possible to do rigorous analyses and exploring different design alternatives. In addition to the classical experimental approaches (e.g. bench scale, mini-plant, pilot plant, market development plant), the use of modelling and simulation tools is becoming increasing popular and powerful.

In this work a strategy is described for designing chemical processes with improved economic and environmental performance, using computer modelling and simulation. This strategy allows the identification and evaluation of different process design alternatives, resulting on the creation of more energy-efficient, mass-efficient and environmental benign industrial processes. This strategy has been tested through example processes, which results can be viewed in Smith *et al.* (2001a, 2001b) and Mata *et al.* (2001, 2003). By applying this strategy one can easily and quickly evaluate and identify chemical process design options with superior economic and environmental performance. Also it incorporates the assessment of the potential environmental impacts of a chemical process, which are usually ignored in traditional process design, where attention is only paid to bring the process into compliance with discharge standards. This strategy allows the identification of the tradeoffs between process economics and potential environmental impacts, revealing where attention should be focused when designing a chemical process.

Strategy for Chemical Process Design

This strategy consists in the selection and design of cost-effective alternatives for chemical processes with significant environmental and economic improvements. Figure 1 shows the several steps included in the strategy for designing chemical processes with good economic and environmental performance, using computer modeling and simulation. The several steps of this strategy include process synthesis, modelling and simulation, preliminary assessment, generation of alternative design options, detailed assessment, feasibility analysis and finally screening of alternatives to arrive to the best design option.

After basic research and development, process synthesis is the earliest stage in the developing

process of a chemical process design. Then it proceeds to conceptual design, preliminary design, detailed design and finally to construction and start up. The strategy described in Figure 1 is applied during the earliest stages, i.e. process synthesis and conceptual design. In these stages a conceptual flowsheet of the chemical process is developed and evaluated.

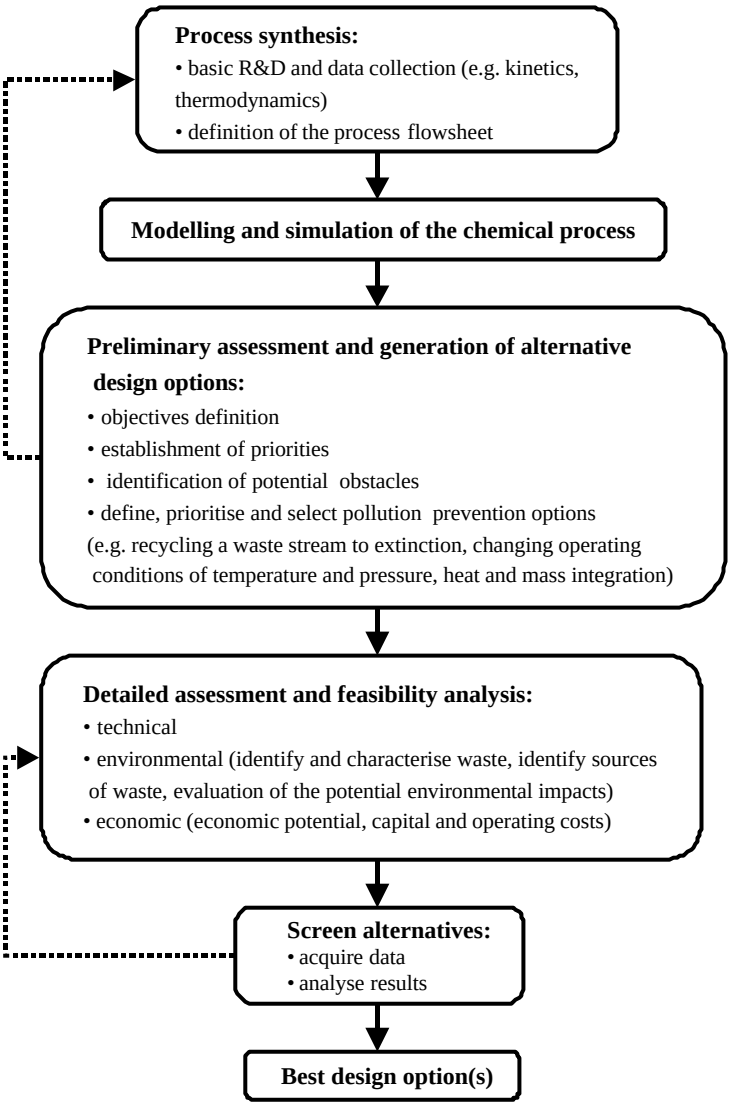


Figure 1. Strategy for designing chemical processes with good economic and environmental performance, using computer modeling and simulation

Process Synthesis. Normally starting with a market need or a business opportunity basic research and development is performed and an input-output diagram may be sketched out. This overall transformation of raw materials into desired products is then divided into several processing steps that provide intermediate transformations (e.g. reaction, separation, mixing, heating and cooling). One can break down the process into its basic functional elements such as the reaction and separation sections. Then identify recycle streams and unit operations to reach desired temperature and pressure conditions. These basic elements lead to a generic process block flow diagram. After preliminary equipment specifications the process flow diagram is made.

The process of selection and evaluation of the individual transformation steps and their interconnections to form a complete structure that achieves the required overall transformation is usually called process synthesis. The outcome of process synthesis is normally expressed in terms of process flowsheets. A "flowsheet" is the diagrammatic representation of the process steps and their interconnections, i.e. it is composed by pieces of equipment (e.g. reactors, heat exchangers, compressors and distillation columns) interconnected by streams.

Generally process synthesis starts at the reactor, if one is required, since it is the place where raw materials are converted into the desired products. Following the reactor and according to the normal sequence in the process flowsheet, the separation and recycle systems are designed. Then follows the design of the process heating and cooling duties, which are dictated by the reactor, separation and recycle systems together. Finally those heating and cooling duties, which cannot be satisfied by heat recover dictate the need for external utilities (steam, cooling water, fuel, etc.). This hierarchy can be represented symbolically by the layers of the "onion diagram" as described by Smith (1995). The "onion diagram" diagram emphasises the sequential or hierarchical nature of process design. When a process do not require a reactor (e.g. in some refinery processes) the design starts with the separation system and moves outward to the heat exchanger network and utilities.

Modelling and Simulation of Chemical Processes. After the structure of the process is determined models are needed as partial substitutes for their prototypes to assist in designing, understanding and predicting the behaviour of the prototype. They must represent significant characteristics of their prototype. Simulation is the use of the model to predict plant's performance and its economics.

The flowsheet generated can be further refined using process simulators. They use more rigorous models of process units, impossible to be performed without a computer. They also provide a way to integrate all the relevant aspects in the process synthesis, therefore reducing the development and implementation time.

Preliminary Assessment and Generation of Alternative Design Options. When the design is specified, methods for generating alternatives are used. Pollution prevention and waste minimisation options must be analysed in this step. Pollution prevention consists of eliminating or minimising waste generation at source, i.e. reducing waste or pollutants before they are created, prior to recycling, treatment or disposal. Pollution prevention via source reduction of a chemical process involves replacing or modifying conventional chemical production processes. There are some basic strategies for reducing process wastes at their source. The flowrate of a purge stream can be reduced by decreasing the purge fraction, by using a higher purity feedstock, or by adding a separation device to the purge or recycle stream that will remove the inert impurity. Reaction by-product production can be reduced by using a different reaction path, by improving catalyst selectivity, or by recycling by-product back to the reactor so that they accumulate to equilibrium levels. Waste minimisation via alternative reactor operating conditions and parameters are other possible examples.

Waste minimisation can be achieved through for example, changes in design and operating conditions that alter the flowrate and composition of pollutant-laden streams, by promoting substitution, recycling and reuse, by applying strategies to minimise, moderate and simplify or by addressing the fundamental chemistry of processes. Other measures such as process modifications (temperature, pressure changes, etc.), unit replacement, feedstock substitution and reactor separation network design can be manipulated to achieve cost-effective waste minimisation.

In process synthesis there are a very large number of ways that one might consider to accomplish the same goal, i.e., there are a very large number of possible alternative processes for converting raw materials into the desired products (Douglas 1988). The analysis of the alternatives usually starts with basic engineering analysis such as mass and energy balances. Predictions are made of the expected performance of the system. Inputs and outputs of the process, flow rates, compositions, pressure, temperature and physical properties of material streams, energy consumption and sizing of the equipment units are listed and analysed.

Chemical process simulators simplify the process of evaluating the different design alternatives without the need of making to much process assumptions and considering the entire process structure. A process simulator has the capability to input and modify the configuration of the process flowsheet and to perform design calculations considering the complete process flowsheet, before they are tried on the actual plant. This way it is possible to model and predict the behaviou of the process flowsheet and to study different operation scenarios (e.g. higher flowrates, different feedstocks, modified operating conditions, various levels of energy integration, etc.) in combination with evaluations of the process economics and potential environmental impacts.

Detailed Assessment and Feasibility Analysis. If a process design appears to be profitable, more rigorous design calculations can be used to develop a final design for the best alternative or the best few alternatives. Usually more rigorous design and costing procedures are used for the most expensive equipment items. However to improve the accuracy of the approximate-material and energy-balance calculations, it is also important to add detail in terms of small and inexpensive equipment items that are necessary for the process operations but do not have a major impact on the total plant cost (e.g., pumps, flash drums, etc.) (Douglas 1988).

A feasibility analysis is then performed. As the mechanical and instrumentation details are considered and the piping and instrumentation diagram is created, estimations of equipment size and costing and the economic and environmental merits of the process are analysed. For example, the economic performance can be readily quantified, by estimating capital investment and operating costs using simple correlations that approximate the actual costs (Biegler *et al.* 1997, Peters and Timmerhaus 1991). Other criteria such as safety, environmental constraints, flexibility, easy control and operation are not readily quantifiable and yet often requires the judgement of the designer (Smith 1995). Properly done, it requires a balance of reliability, safety and economics, while having an acceptable impact on the environment and society. The initial choice of the process is not expected to be optimal. However it is usually possible to improve the process by a different choice of process flows and conditions, e.g., by parameter optimisation.

Open and fugitive emissions of chemicals escape to the atmosphere posing a large risk to public, employee and environmental health. While open emissions are usually controlled or remediated, fugitive emissions are still escaping from processes and are becoming a relatively large source of environmental impacts. Emissions from equipment leaks occur in the form of gases or liquids that escape to the atmosphere through many types of connection points (e.g. flanges, fittings, etc.) or through the moving parts of valves, pumps, compressors, pressure relief devices and certain types of process equipment. Valves are usually the single largest source of fugitive emissions (Goyal 1999). Point sources of fugitive emissions, such as a single piece of equipment are usually small. However, cumulative emissions throughout a plant can be very large, based on the large number of equipment pieces that can leak such as valves, pumps, flanges, compressors, etc.

In order to determine fugitive emissions losses, the U.S. EPA conducted emission test programs in petroleum refineries, which resulted in a set of average emission factors for process equipment (U.S. EPA, 1980, 1996). These average factors are listed in AP-42 (1995) and total losses are estimated by combining the losses for all the pieces of equipment based on their average factors. The *Protocol for Equipment Leak Emission Estimates* (U. S. EPA, 1995) describes the testing procedures, such as screening or bagging (or both) involved in the development of emission factors. According to Sydney (1989) there are several variables that can affect the emission factors such as the fluid phase, pressure, temperature, unit type, equipment size, type of valve, flange, compressor, pump, etc. The use of emission factor methods is based on the assumption that the leak frequency and the equipment emission rates are similar to those of the average process in EPA's studies (Schaich, 1991). Therefore, these methods are most valid for estimating emissions from a process or population of equipment and for a large period of time.

For the environmental analysis, environmental assessment tools can be used such as the waste reduction algorithm (WAR) (Young *et al.*, 2000). WAR has made available a method to simply evaluate processes with a library of approximately 1600 chemicals. The WAR algorithm applies a balance equation around a process to evaluate the potential environmental impacts. It considers eight impact categories including human toxicity potential by ingestion (HTPI), human toxicity potential by exposure (HTPE) through dermal and inhalation routes, aquatic toxicity potential (ATP), terrestrial toxicity potential (TTP), photo-oxidation chemical (smog) potential (POCP), acidification potential (AP), ozone depletion potential (ODP), and global warming potential (GWP). Potential environmental impact scores are available in the WAR database for chemicals based on a representative measurement. For a more complete description of the WAR algorithm see Cabezas *et al.* (1999) and Young *et al.* (2000).

Screen Alternatives. Finally after detailed assessment and feasibility analysis, data is acquired and results are analysed to arrive to the best design option or the best few options.

Conclusions

The described strategy aims to link traditional process design with environmental assessment, i.e. to evaluate the potential environmental impacts of a chemical process into the environment, while recognising the diversity of value judgements regarding the environmental issues. This strategy focuses attention on those chemical components and process streams that have the largest contribution to the potential environmental impacts and the largest economic potential. It takes into account the process's fugitive emissions and their potential environmental impacts when designing a chemical process, which is normally ignored in traditional process design.

With this strategy one can easily incorporate environmental assessment in chemical process design, while devising alternative designs with superior environmental and economic performance. Since the environmental protection is an important aspect of the performance of chemical processes, this strategy has many advantages for the modern industry. It makes possible for a company to anticipate compliance with environmental regulations, representing a new procedure and tool capable of exploring different alternative designs and of identifying design features leading to potential environmental problems and process costs. It can be applied to the design of new processes or to the retrofit existing ones. It allows that process design alternatives with superior environmental and economic performance are identified, which is often associated with materials and energy efficiency.

References

AP-42: *Compilation of Air Pollutant Emission Factors. Volume I: Stationary Point and Area Sources*. U.S. Environmental Protection Agency: North Carolina, 1995.

Biegler L. T., Grossmann I. E., Westerberg A. W.: *Systematic Methods of Chemical Process Design*. Prentice Hall, Inc., New Jersey, 1997.

Bumble S.: Computer Simulated Plant Design for Waste Minimization/Pollution Prevention. Lewis Publishers, CRC Press LLC, NewYork, 2000.

Cabezas H., Bare J. C., Mallick S. K.: Pollution Prevention with chemical process simulators: the generalized waste reduction (WAR) algorithm – full version. *Computers and Chemical Engineering*, 1999, 23, 623-34.

Douglas, J. M.: *Conceptual Design of Chemical Processes*, McGraw-Hill, New York, 1988.

Goyal O. P.: Reduce HC losses plant-wide: part 1. *Hydrocarbon Processing*, August, 1999, 97-104.

Hilaly A. K., Sikdar S. K.: Pollution balance: a new method for minimizing waste production in manufacturing processes. *Journal of Air & Waste Management Association*, 1994, 44, 1303-8.

Mata T. M., Smith R. L., Young D. M., Costa C. A. V. Simulation of Ecological Conscious Chemical Processes: Fugitive Emissions versus Operating Conditions. *Proceedings of CHEMPOR´01: 8th International Chemical Engineering Conference*, ed. Ribeiro, F. R.; Cruz Pinto, J. J. C., Aveiro, Portugal, 12-14 September 2001, 907-913.

Mata T. M., Smith R. L., Young D. M., Costa C. A. V.: Evaluating the Environmental Friendliness, Economics and Energy Efficiency of Chemical Processes: Heat Integration. *Clean Technologies and Environmental Policy* (in press).

Peters M. S., Timmerhaus K. D.: *Plant Design and Economics for Chemical Engineers*, 4th ed. McGraw-Hill International Editions, Chemical and Petroleum Engineering Series, 1991.

Schaich J. R.: Estimate Fugitive Emissions from Process Equipment. *Chem. Eng. Prog.*, August 1991, 31-35.

Smith R. L., Mata T. M., Young D. M., Cabezas H., Costa C. A. V.: Designing Environmentally Friendly Chemical Processes with Fugitive and Open Emissions. Process Integration. *Proceedings of PRES'01: 4th Conference on Process Integration, Modelling and Optimisation for Energy Saving and Pollution Reduction*, Florence, Italy, 20-23 May 2001, 129-134.

Smith R. L., Mata T. M., Young D. M., Cabezas, H., Costa C. A. V.: Designing Efficient, Economic and Environmentally Friendly Chemical Processes. In: European Symposium on Computer Aided Process

Engineering –11, Series: *Computer-Aided Chemical Engineering*, 9, ed. Jorgensen, S.; Gani, R., Kolding, Danmark, 27-30 May 2001, 1165-1170.

Smith R.: *Chemical Process Design*. McGraw-Hill, Inc., New York, 1995.

Sydney L.: Fugitive Emissions. *Chem. Eng. Prog.*, June, 1989, 42-47.

U. S. EPA: *Assessment of Atmospheric Emissions from Petroleum Refining. Volume I*; U.S. Environmental Protection Agency, Washington, DC, 1980; EPA-600/2-80-075a.

U. S. EPA: *Protocol for Equipment Leak Emission Estimates*; U.S. Environmental Protection Agency: North Carolina, 1995; EPA-453/R-95-017.

U. S. EPA: *Preferred and Alternative Methods for Estimating Fugitive Emissions from Equipment Leaks*; U.S. Environmental Protection Agency: North Carolina, 1996; EPA-454/R-97-004b.

Young D. M., Cabezas H.: Designing Sustainable Processes with Simulation: The Waste Reduction (WAR) Algorithm. *Computers and Chemical Engineering*, 23, 1999, 1477-91.

Young D. M., Scharp R., Cabezas H.: The waste reduction (WAR) algorithm: environmental impacts, energy consumption and engineering economics. *Waste Management*, 20, 2000, 605-15.

Global Simulation and Optimization of a Chemical Plant

Filipe J.M. Neves, Dulce C.M. Silva, Joana I.L.C. Tourais and Nuno M.C. Oliveira

GEPSI — PSE Group, Department of Chemical Engineering, University of Coimbra

Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal

Tel. +351-239-798700. E-mail: {fneves, dulce, joana, nuno}@eq.uc.pt

Abstract

This paper addresses the main phases and challenges met during the global simulation and optimization of a continuous process for the production of aniline. A fixed process topology is considered, based on the process implemented by Quimigal S.A., using the liquid phase hydrogenation of nitrobenzene. Simulation and optimization studies addressed separately the reaction and purification stages of the process. This illustrates the application of a systematic mathematical approach for the simulation and optimization of a chemical process, using detailed mechanistic models of the units.

1 Introduction

The process under analysis, for production of aniline by liquid phase hydrogenation of nitrobenzene, can be divided in two major stages: reaction and purification. The reaction occurs in slurry (three phase) reactors, while the purification stage consists in a complex configuration of liquid-liquid separators and distillation columns.

Different steps were involved in the construction of process models for the individual units, and for simulation/optimization studies. These included the selection and validation of property and parameter estimation methods, the choice of the degree of model complexity for the individual units, and the numerical methods for their solution. In both cases, the overall strategy was to start with the simplest approaches available, adding complexity only when required for reliability and accuracy of the results (Levenspiel, 2002).

2 Knowing the process

Industrial data was extensively gathered during the initial phase of this work, by collecting samples and monitoring flows of all process streams. A data reconciliation methodology was defined, to produce a consistent stationary view of the process. These values were later used as an essential simulation target and as a base case for process benchmarking. The reconciliation of process values was accomplished in 2 steps, due to the diversity of the data available, and the uncertainties present. In the first part, the overall mass flowrates were reconciliated using a simple least squares formulation, assuming the remaining variables fixed. This was done in GAMS language, by solving a quadratic problem of the form:

$$\begin{aligned} \min_{F,\epsilon} \quad & \sum_{i=1}^{\mathrm{nc}} \epsilon_{\mathrm{BP},i}^2 p_i + \epsilon_{\mathrm{BT}}^2 \\ \text{s.t.} \quad & \sum_{j=1}^{\mathrm{ne}} w_{i,j} F_j = \sum_{k=1}^{\mathrm{ns}} w_{i,k} F_k - \epsilon_{\mathrm{BP},i} \qquad i \in C \\ & \sum_{j=1}^{\mathrm{ne}} F_j = \sum_{k=1}^{\mathrm{ns}} F_k - \epsilon_{\mathrm{BT}} \\ & F_{u,l} \leq F_j \leq F_{u,u}, \qquad j \in E, S \end{aligned} \tag{1}$$

In these equations, F_j are the flowrates to be determined, ϵ_i are errors associated with each balance equation, and p_i are the weights attributed to each of these error terms. This formulation was applied to simultaneously to all process units, with additional linear mass balance equations to express the relations between the different process streams. This step was followed by simultaneous reconciliation of flowrates and composition, using a NLP formulation, and the previous determined values of the flowrates as reference values. The mathematical formulation of this second phase can be expressed as:

$$\begin{aligned}
\min_{F^{\mathrm{C}},w^{\mathrm{C}},\gamma} \quad & \sum_u \gamma_{\mathrm{F},u}^2 p_{\mathrm{F},u} + \sum_i \sum_u p_{\mathrm{w},i,u} \gamma_{\mathrm{w},i,u}^2 \\
\text{s.t.} \quad & \sum_{j=1}^{\mathrm{ne}} w_{i,j}^{\mathrm{C}} F_j^{\mathrm{C}} = \sum_{k=1}^{\mathrm{ns}} w_{i,k}^{\mathrm{C}} F_k^{\mathrm{C}} \\
& \sum_{j=1}^{\mathrm{ne}} F_j^{\mathrm{C}} = \sum_{k=1}^{\mathrm{ns}} F_k^{\mathrm{C}} \\
& F_u^{\mathrm{C}} = F_u + \gamma_{\mathrm{F},u} \\
& w_{i,u}^{\mathrm{C}} = w_{i,u} + \gamma_{\mathrm{w},i,u} \\
& F_{u,l}^{\mathrm{C}} \leq F_u^{\mathrm{C}} \leq F_{u,u}^{\mathrm{C}} \\
& w_{i,u,l}^{\mathrm{C}} \leq w_{i,u}^{\mathrm{C}} \leq w_{i,u,u}^{\mathrm{C}}
\end{aligned} \tag{2}$$

The GAMS language was used to solve this nonlinear problem involving around 900 variables and 740 equations. During this preliminary phase, laboratory experiments were also performed, to characterize the complexity of the equilibria phenomena that need to be considered in the separation phase, including the identification of azeotropic mixtures and the selection and validation of equilibria estimation methods (UNIFAC for V/L, and NRTL for L/L equilibria).

3 Simulation of the reaction phase

Distinct physical and chemical processes are known to occur inside the multiphase reactors used, including gas-liquid and liquid-solid mass transfer, diffusion, adsorption and reaction on the catalyst, as well as desorption of the products. These reactions are often described through elaborate schemes, with several intermediate chemical species and alternative pathways to the desired products and byproducts. This can lead to systems with complex behavior, where certain sets of variables exert a major influence on overall system performance. Given the complexity of the phenomena that occur in the three-phase hydrogenation reactors, and the limited amount of measurements available relative to these systems, two detailed mechanistic models were built for them, combined published kinetic information (Turek et al., 1986) with mass transfer models for this type of systems (Chaudhari and Ramachandran, 1980). The following main hypothesis were used in the development of these models:

- Perfectly agitated liquid phase, with catalyst particles and hydrogen bubbles uniformly dispersed in the reacting mixture.
- Efficient removal of the reaction heat, allowing a constant liquid phase temperature.
- Constant volume of the reactant mixture.
- Monodisperse catalyst particles (of identical diameter), with active centers uniformly distributed and equally accessible throughout the solid volume.

Both models were implemented computationally, using the *Mathematica* language (Wolfram, 1999). The first model neglects intraparticle diffusion processes. A simplified description of the catalyst particles is used, in a pseudo-homogeneous approach, by considering the temperature and concentrations constant in the solid phase.

In the second case the concentration and temperature profiles within the particles are explicitly considered. The balance to a reactant (e.g. hydrogen) in the solid phase is expressed by an equation of the form

$$D_{\mathrm{eff},\mathrm{H_2}}\left(\frac{\partial^2 C_{\mathrm{H_2,S}}}{\partial r^2}+\frac{2}{r}\frac{\partial C_{\mathrm{H_2,S}}}{\partial r}\right)=-\alpha_{\mathrm{H_2}}kf(C_{\mathrm{NB,S}},C_{\mathrm{H_2,S}})\frac{m_p}{V_p}, \tag{3}$$

valid in the domain $r\in[0,r_p]$. In this expression $C_{\mathrm{H_2,S}}(r)$ and $C_{\mathrm{NB,S}}(r)$ are functions of the radial position in the particle, resulting in a distributed parameter model. The corresponding boundary conditions are

$$\left.\frac{\partial C_{\mathrm{H_2,S}}}{\partial r}\right|_{r=0}=0,\qquad \left.D_{\mathrm{eff},\mathrm{H_2}}\frac{\partial C_{\mathrm{H_2,S}}}{\partial r}\right|_{r=r_p}=K_{\mathrm{LS,H_2}}\big(C_{\mathrm{H_2,L}}-C_{\mathrm{H_2,S}}(r_p)\big),$$

with similar partial balances considered for the remaining components and conservation of energy in the solid phase, originating a system of algebraic-differential equations.

The system of differential equations of the second model was solved using finite differences, with centered formulas, resulting in a classical scheme with convergence of second order relatively to the spacement in the space grid. Since both models were written in the *Mathematica* language, a generic discretization package was written, using the symbolic manipulation capabilities of this system, including the automatic treatment of the boundary conditions and their singularities.

The algebraic equations corresponding to both models were also solved in the *Mathematica* system. While the simplest model (composed of 11 variables and 11 algebraic equations) was easily solved, the more complex one exhibited serious convergence problems during the solution of the discretized model, using variations of Newton's method. This was due to the presence of very steep intraparticular profiles that caused convergence to solutions of the system without physical meaning (e.g., negative concentrations), if the initial guess was not extremely (i.e., pathologically) close to the final solution, even after proper scaling of the variables and equations.

Since traditional methods for the solution of systems of algebraic equations did not seem to provide efficient solutions for this problem, an alternative strategy was implemented by directly imposing known solution bounds during the determination of the search direction of a Newton-type method. This is done through the solution of a linear program of the form

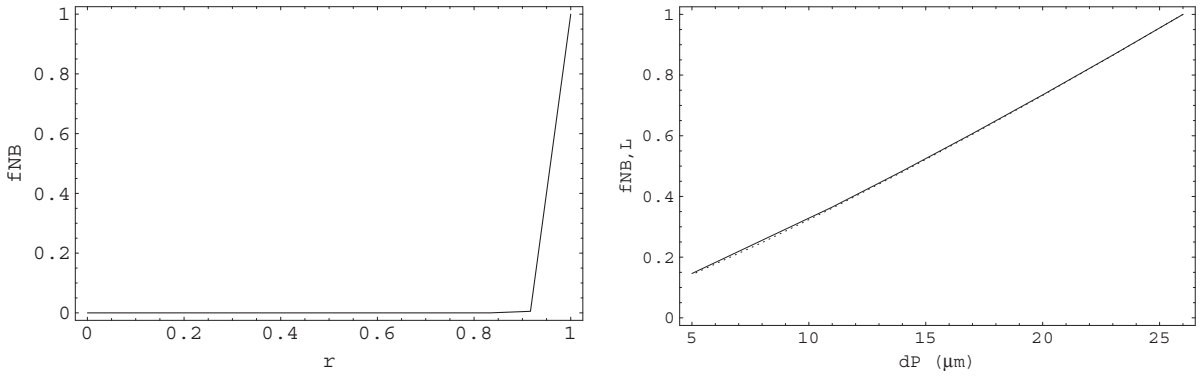
$$\begin{aligned}\min_{\Delta x_n,\epsilon}\quad & \|\epsilon\|_1\\ \text{s.t.}\quad & f(x_n)+J(x_n)\Delta x_n=\epsilon\\ & x_l-x_n\le\Delta x_n\le x_u-x_n\end{aligned} \tag{4}$$

instead of solving the linear system $J(x_n)\Delta x_n=-f(x_n)$. The former approach reduces to the solution of the linear system when an entirely feasible solution can be found. Bullard and Biegler (1991) propose a similar approach for the simulation of constrained systems. However, in their approach infeasible intermediate iterates can be generated, if associated with a sufficient decrease of the merit function used. In the case of (4), by guaranteeing that the iterate candidates always remain inside the feasible region, and considering explicitly its bounds, model failures can be avoided, and less effort can be required during the step search phase.

Simulation results exhibit good agreement between the models, and also with available industrial data. Figure 1 illustrates some of the results obtained. As can be concluded by observation of Figure 1(a), reaction takes place in a thin layer close to the particle surface, with a thickness of only 5-7% of the particle radius, resulting in an effectiveness factor of 10^{-4}. This is mainly due to the depletion of the reagent MNB in the catalyst particles, caused by an extremely high mass transfer resistance for this component in the solid-liquid film. As a consequence, the concentration and temperature profiles within the pellets are nearly flat, after the external layer, and this allows good agreement between the results obtained with both models (Neves et al., 2002).

4 Reaction phase — optimization studies

Optimization of the reactors units can be easily performed by a sensitivity analysis of the simulation results relatively to each of the main process variables and design parameters available. In the present configuration, the



(a) (b)

Figure 1: (a) intra-particular profile of MNB concentration, (b) dependence of the residual concentration of nitrobenzene with particle diameter.).

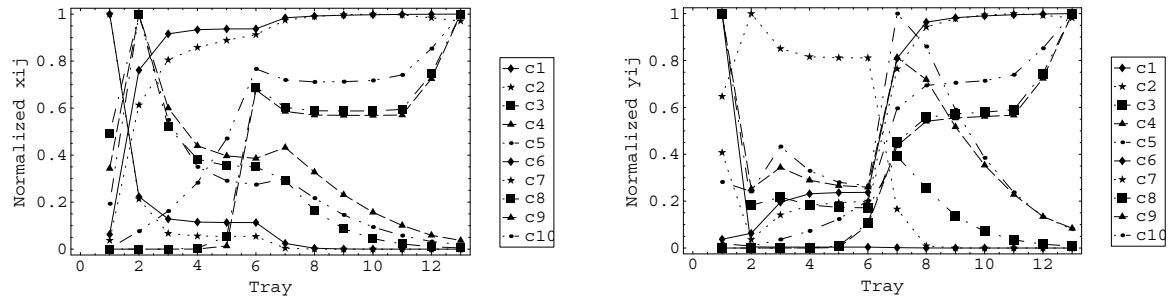
most important optimization variable was found to be the diameter of the catalyst particles. Figure 1(b) illustrates the variation in the concentration of mono-nitrobenzene in the liquid phase with the catalyst dimensions. As can be observed, a decrease in the particle diameter from 25 μm to 5 μm allows an 80% reduction in the residual concentration of MNB in the effluent, without further process changes.

The models were also used to diagnose the variability of the operating data from identical industrial reactors, where significant differences in the catalyst consumption and effluent concentration of MNB were observed in practice (Neves et al., 2002). The analysis of the catalyst present in the reactors with better performance showed that a significant portion of the catalyst in use (approximately 2/3) had diameters of the order of 1-3 μm, although the fresh catalyst that is added to the reactors has a mean diameter of 20 μm. These values are in agreement with Turek et al. (1986), where significant degradation of the catalyst was reported to occur in a similar (although laboratory) reactor, due to the effects of intense agitation. The analysis of the specific area (BET) and porous volume indicated that the catalyst in use had suffered a significant decrease in its values relative to the fresh catalyst, without noticeable loss of activity. These results, together with the model predictions, suggest that since the reaction occurs essentially at the solid surface, the mechanical degradation of the catalyst actually improves the reactor performance. Simultaneously, in the industrial reactors where higher catalyst consumptions and higher MNB effluent concentrations were observed, the analysis showed a closer proximity between the properties of the used and fresh catalysts. Their lower performance is therefore attributed mostly to problems in the separation system, unable to adequately retain the catalyst particles of smaller dimensions that lead to greater conversion.

5 Simulation of the purification phase

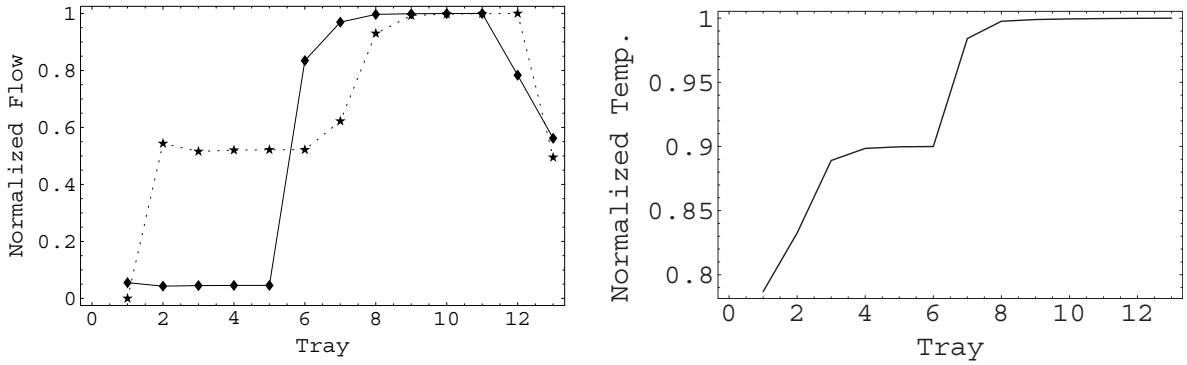
The models developed for the various separation units are composed of systems of algebraic nonlinear equations, describing the various equilibrium stages that are assumed to occur in this phase of the process. These are large-scale and highly nonlinear, mostly due to the nonideal models necessary to accurately describe the relations between the compositions of the various phases, and physical variables such as pressure and temperature. Instead of relying on purely algebraic handling, these models are usually more conveniently solved by a combination of shortcut and self initialized equation-tearing methods (Seader and Henley, 1998).

Various equation-tearing methods can be used with these models — plate-to-plate, matricial or relaxation. The (matricial) rigorous iterative method of Wang-Henke was chosen to refine the results provided by the short-cut method of Fenske-Underwood-Gilliand-Kirkbridge. In fact, the simplicity of implementation of the Wang-Henke method, together with good convergence properties, was decisive to exclude the relaxation (time consuming) and plate-to-plate (difficult to converge). With the examples tested, the Wang-Henke method produced accurate estimates of the separation profiles (validated by the split fractions obtained in the reconciliation data exercise) and,



(a) (b)

Figure 2: Normalized profiles of the liquid (a) and vapor (b) phase concentrations of one of the distillation columns.



(a) (b)

Figure 3: (a) Liquid (....) and vapor (__) phase flowrates and (b) temperature profile for one the distillation columns.

in 4 units, was able to converge reasonably fast to the solution. On other hand, for 2 of the 6 distillation columns involved in this process, the Wang-Henke method presented severe difficulties in handling wide boiling mixtures — the method failed, because negative values of compositions and flowrates were calculated at a given iteration, causing the "blow up" of terms in some of the equations. This was a characteristic reported by Friday and Smith (1964), who suggested some empirical modifications to the base algorithm, whenever in presence of such cases.

To improve the convergence properties of Wang-Henke method's for wide boiling mixtures ($\Delta T_B > 50°$C), a strategy based on dumping of the loop variables was implemented. This consisted, essentially, in constraining the admissible changes on the values of V_j and T_j, between two consecutive iterations, by a factor of, e.g., 10% of the full correction. It has also been observed that, in some cases, perturbing the initial values of the K_{ij} coefficients in the first iteration was also beneficial. The solutions obtained by the Wang-Henke method were compared with the solution of these same models in the commercial simulator ASPEN PLUS 11.1. Figures 2 and 3 show these profiles for one of the distillation columns simulated.

The next step consisted in trying to converge, simultaneously, all of the separation units, taking in consideration the connections between them. The presence of several recycle streams increased the difficulty of this task. According to Barton (2000), a sequential-modular approach corresponds usually to the best choice. For convergence of outer loops (resulted from the tearing of recycle streams), Newton, Quasi-Newton, successive substitutions and Weigstein methods, among others, are valid options. The solution adopted is represented schematically in Figure 4. It implements a successive substitutions strategy, performed in 2 steps. During the first step, the rigorous models concerning each unit are solved by the most appropriated scheme. This requires the estimation of all of the unknown input streams for each unit, in the first iteration. For distillation columns the Wang-Henke method is used, as discussed, and for the liquid-liquid separators, a classic Newton method is able to converge easily.

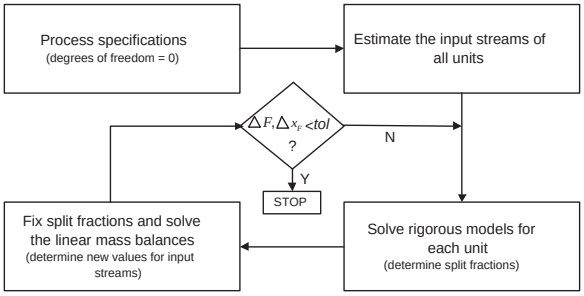


Figure 4: Schematic representation of the strategy for overall simulation of the purification stage.

The second step involves the solution of a linear system of equations corresponding to the partial mass balances around each unit, using the split fractions obtained in the first step. The solution of these equations provides updated values for the flowrates and compositions of the input streams, and a new loop takes place until the differences between the updated values of two consecutive iterations satisfy a pre-specified tolerance. The great advantage of this strategy consists in the easy implementation and the reduced calculation effort at each iteration. Although no guarantees can be made in general, the speed of convergence observed with the present case was very reasonable: only 5–6 iterations were needed to achieve the solution of this flowsheet.

When the flowsheet was solved, the results obtained for the composition and flowrates of every stream were compared with the results obtained during the data reconciliation exercise. If some deviations were observed, the specifications of the problem were adjusted, and the flowsheet solved again until the simulation of the purification stage matched the industrial reality.

6 Purification phase — optimization studies

Once the performance of the purification phase was considered to be conveniently reproduced in the simulation, the next step consisted in finding optimization opportunities. Distillation columns are in general responsible by the consumption of a great share of the energy resources available on most processes. For equipment with fixed physical specifications it is possible to adapt operational features like the localization of the feed plate and reflux ratio in order to order to decrease the consumption of utilities for a given separation. For this purpose, the strategy shown in Figure 5 was developed.

The sequential-modular strategy is not suitable to solve optimization problems, and therefore an equation-oriented (EO) strategy had to be employed for this purpose. However, trying to solve directly the *MESH* (Mass-Equilibrium-Summation-Heat) for a distillation column is not a easy task. The simultaneous solution of this highly nonlinear set of equations requires extremely good initial values, bounds and scaling factors, to avoid the failure of existing implementations of optimization algorithms (e.g., SQP or GRG). For this reason, extreme care was taken in building a suitable initialization phase.

Another key to the successful application of an optimization algorithm to this problem is the introduction of additional slack variables in the MESH equations. These usually allow a faster solution start, avoiding problems caused by infeasibilities during the early solution stages. After the first feasible point is determined, maintaining bounds for the maximum magnitude of these variables corresponds to the definition of a trust region, that constraints the maximum amount of deviation from a feasible physical configuration at any point during the iteration. This procedure is also generally beneficial to converge rate of these problems, given its highly nonlinear nature.

With the capability of solving a column model with fixed NT, NF and RR, by a EO strategy, the next step was to define how to optimize the consumption of utilities in a given column. This corresponds to the minimization of reflux ratio (RR), subject to additional operational constraints, such as the the degree of separation of some components, internal flows, etc. Typical results for the process considered indicate that savings of 10 000 euro/year per column are possible, through the solution of these optimization problems.

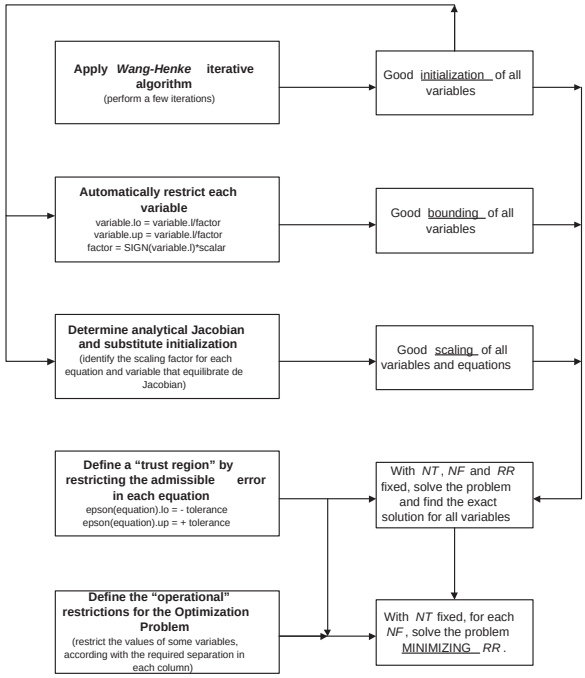


Figure 5: Schematic representation of the strategy for individual optimization of the distillation columns.

The GAMS environment was used to solve these optimization problems. Among the solvers tested, *CONOPT3* presented usually the best performance, converging fast and without difficulties to the optimal solution. *MINOS* exhibited some difficulties if few iterations of the Wang-Henke method were performed in the pre-processing phase. This means that the initial point required by this solver needs to be better than the one required by *CONOPT3*. The weakest performance belonged to *SNOPT*, a SQP implementation, where much slower convergence was always observed. However, the systematic interpretation of these results is still currently being considered.

7 Future work

An additional aspect of optimization of these separation systems, not considered in the present work, is the optimal design problem. Here, more degrees of freedom are available, and different strategies to deal with discrete variables (such as the total number of equilibrium stages, or the location of the feed streams) are available. This problem has been addressed over the past decade as a mixed integer nonlinear programming (MINLP) problem (Barttfeld et al., 2003). But tools for solving MINLPs are not widespread, especially in connection with detailed simulation models. The other alternative, presently available, is the introduction of a differentiable distribution function (DDF) (Lang and Biegler, 2002). In this formulation, all streams around a column, except the top and bottom products, are directed to all of the column trays using the DDF. However, due to their nature, these functions can introduce numerical ill-condition in the problem to be solved. Developments of this technique are needed to allow a wider applicability of optimization to solution problems of this type.

Acknowledgements

The authors would like to acknowledge support in different forms from Quimigal S.A., and financial support from Agência de Inovação (AdI), through the consortium research projects AP2000 and INOVA.

References

P.I. Barton (2000). “The Equation Oriented Strategy for Process Flowsheeting”, *Internal Report*, Massachusetts Institute of Technology.

M. Barttfeld, P.A. Aguirre, I.E. Grossman (2003). “Alternative representations and formulations for the economic optimization of multicomponent distillation columns”, *Comp. and Chem. Eng.*, 27, 363–383.

A. Brooke, D. Kendrick, A. Meeraus (1992). “Gams: A User’s Guide”, The Scientific Press, San Francisco.

Bullard, L.G., L.T. Biegler (1991). “Iterative Linear Programming Strategies for Constrained Simulation”, *Comp. and Chem. Eng.*, 15(4), 239–254.

R.V. Chaudhari, P.A. Ramachandran (1980). “Three Phase Slurry Reactors”, *AIChE J.*, 26(2), 177–201.

J.R. Friday, B.D. Smith (1964). “An analysis of the equilibrium stage separation problems – formulation and convergence”, *AIChE J.*, 698–706.

Y-D Lang, L.T. Biegler (2002). “A distributed stream method for tray optimization”, *AIChE J.*, 48(3), 582–595.

O. Levenspiel (2002). “Modelling in Chemical Engineering”, *Chem. Eng. Sci.*, 57, 4691–4696.

F.J.M. Neves, C.M.G. Baptista, P.A.P. Araújo, N.M.C. Oliveira (2002). “Mechanistic models of a slurry hyrogenation reactor as a tool for process diagnosis and optimization”, *17th International Symposium on Chemical Reaction Engineering, ISCRE’17*, Hong-Kong.

J.D. Seader, E.J. Henley (1998). *Separation Process Principles*, Wiley & Sons, Inc., New York, NY.

F. Turek, R. Geike, R. Lange (1986). “Liquid-phase Hydrogenation of Nitrobenzene in a Slurry Reactor”, *Chem. Eng. Process.*, 20(4), 213–219.

S. Wolfram (1999). *The Mathematica Book*, 4th edition, Cambridge University Press, Cambridge.

MICRO–SCALE ANALYSIS OF CRYSTAL DISSOLUTION AND PRECIPITATION IN POROUS MEDIA

I. S. POP AND C. J. VAN DUIJN

Abstract. A micro–scale model for precipitation and dissolution processes in porous media is discussed. Weak solutions are shown to exist in case of general domains. Next we consider the case of thin strips, where we look for dissolution and precipitation fronts. These are located at a free boundary, which is continuous and monotone. Letting the ratio between the thickness and the length of the strip go to 0 we end up with the upscaled transport–reaction model proposed in [9]. This paper summarizes the results obtained in [3].

1. Introduction

Mathematical models for reactive flow in porous media are of great importance for understanding soil chemistry processes. In general such models are coupled systems of partial and ordinary differential equations, involving different kinds of nonlinearities describing reaction, adsorption, precipitation or dissolution rates.

A significant amount of mathematical literature is devoted to the macroscopic (core–scale) models. In this sense we mention [1], [2], [4], [5], or [10] for questions concerning existence and uniqueness of a solution and of travelling waves.

Upscaled models can be derived form microscopic ones by homogenization techniques. An extended overview in this sense can be found in [6]. To make the upscaling procedure mathematically rigorous, not only the upscaled model, but also the microscopic one has to be analyzed. In this respect, rigorous homogenization results are obtained in [7] (for linear reaction rates and isotherms, see also [11]) and extended to certain types of nonlinearities in [8].

In this paper we consider the microscopic (pore) scale situation, which is strongly related to the upscaled model introduced in [9]. Two ions are dissolved into a fluid occupying the void space of a porous medium. The ions can precipitate in form of a crystalline solid, which is attached to surface of the porous matrix (the grains). The reversed process is also possible. Here we make a simplifying assumption: the flow geometry, as well as the fluid density and viscosity are not affected by the chemical processes.

Modelling aspects are detailed in [3]. Here we restrict ourselves in studying the resulting dimensionless problem. Let $\Omega \subset \mathbb{R}^d$ $(d > 1)$ be a bounded, simply connected domain in $\mathbb{R}^d$ (the pore space). Its boundary $\partial\Omega$ is assumed Lipschitz and consisting of three disjoint parts: an internal (grain) boundary Γ_G and an external boundary where Dirichlet (Γ_D) or Neumann (Γ_N) conditions are prescribed. Both Γ_G and $\Gamma_D \cup \Gamma_N$ have positive measure. Further, $\vec{\nu}$ denotes the outer normal to $\partial\Omega$ and $T > 0$ is a maximal value of time. With $X^T := (0,T) \times X$, the model under consideration reads:
Ion transport (in the pore space):

$$
\left\{
\begin{array}{rcll}
\partial_t u + \nabla \cdot (\vec{q}u - D\nabla u) & = & 0, & \text{in } \Omega^T, \\
-D\vec{\nu} \cdot \nabla u & = & \varepsilon n \partial_t v, & \text{on } \Gamma_G^T, \\
u & = & u_D, & \text{on } \Gamma_D^T, \\
\vec{\nu} \cdot \nabla u & = & 0, & \text{on } \Gamma_N^T, \\
u & = & u_I, & \text{in } \Omega, \text{ for } t = 0,
\end{array}
\right. \tag{1.1}
$$

Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (*email:* {C.J.v.Duijn, I.Pop}@tue.nl).

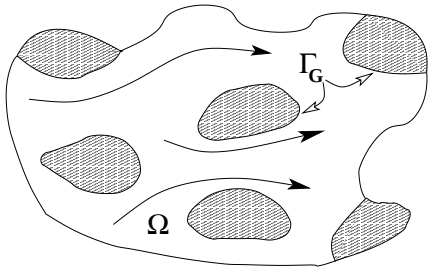


FIGURE 1. Flow domain with grains.

Precipitation/dissolution (on the grains):

$$\left\{\begin{array}{rcll} \partial_t v & = & k(r(u,c)-w), & \text{on } \Gamma_G^T, \\ w & \in & H(v), & \text{on } \Gamma_G^T, \\ v & = & v_I, & \text{on } \Gamma_G, \text{ for } t=0. \end{array}\right. \tag{1.2}$$

Here $\vec{q}$ denotes the fluid velocity, which is obtained by solving a Stokes system in the pore space. We assume no flow at the grain surface. c stands for the electric charge inside the fluid, which, assuming both solutes have the same diffusion coefficient D, is a conserved quantity. It can be seen as the solution of a convection–diffusion problem (like (1.1_1)), with no–flow conditions on grains. By u and v we denote the cation concentration (relative to the water volume), respectively the precipitate concentration (relative to the grain surface). The third unknown w is introduced for describing a multi-valued nature of the dissolution rate in (1.2_2), where H stands for the Heaviside graph.

The model studied in [3] is completed by equations for the flow and the charge. Here we restrict ourselves to the description of the chemical processes, which is the challenging part of the model. Specifically, we investigate (1.1)–(1.2), a parabolic advection–diffusion problem that is coupled to an ordinary differential equation on a lower dimensional manifold (the grain surface). Moreover, the dissolution rate in (1.2_2) is multi-valued. Following [9], the anion concentration is eliminated from the model, since it can be obtained straightforwardly if the cation concentration and the total charge are known.

For the precipitation rate $r(u,c)$ in (1.2) we assume:

(A_r) (i) $r : \mathbb{R}^2 \to [0,\infty)$, $r \geq 0$ and locally Lipschitz in $\mathbb{R}^2$;

(ii) $r(u,c) = 0$ for all $u \leq 0$;

(iii) for each $c \in \mathbb{R}$ there exists a unique $u_* = u_*(c) \geq 0$, with $u_*(c) = 0$ for $c \leq 0$ and u_* is strictly increasing for $c \geq 0$, such that

$$r(u,c) = \left\{\begin{array}{l} 0, \qquad \text{for} \quad u \leq u_*, \\ \text{strictly increasing for } u > u_* \text{ with } r(\infty,c) = \infty; \end{array}\right.$$

(iv) for each $u > 0$, $r(u,c)$ strictly decreases with respect to c whenever $r > 0$.

With $[x]_+$ denoting the positive cut of x, a typical example is

$$r(u,c) = K\left([u]_+\right)^m \left(\left[\frac{mu-c}{n}\right]_+\right)^n, \tag{1.3}$$

for some $K > 0$, where m and n are natural numbers (the valences of the two ions). Thus $u_*(c) = \frac{[c]_+}{m}$ (see also Figure 2).

Boundary and initial data are assumed essentially bounded and non-negative. Boundary data are traces of H^1–functions.

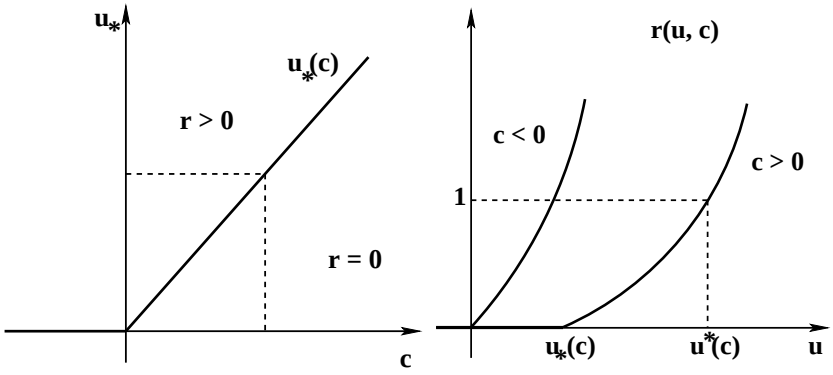


FIGURE 2. Typical examples for u_* (left) and r (right).

The parameter ε in (1.1_2) expresses the ratio of two length scales: the characteristic pore scale length and the problem related scale. When upscaling to a macroscopic model one takes $\varepsilon \searrow 0$. However, this limit case is considered here only for thin strips.

The results given below are obtained in two cases. For general general domains we prove existence of a weak solution. If the flow domain is a two–dimensional strip, a dissolution front occurs after a waiting time, and its location is a free boundary. Letting ε - the ratio between the width and the length of the strip - go to 0, we end up with the upscaled model proposed in [9] (see also [2]), for which we also obtain uniqueness.

2. General domains

Below we use function spaces and notions that are commonly encountered in books for functional analysis and partial differential equations (see, e. g., [12]). The main difficulty in the analysis is due to the multi-valued function describing the precipitation and dissolution. With

$$\mathcal{U} := \{u \in u_D + L^2(0,T; H^1_{0,\Gamma_D}(\Omega))/\partial_t u \in L^2(0,T; H^{-1}(\Omega))\},$$
$$\mathcal{V} := \{v \in H^1(0,T; L^2(\Gamma_G))\},$$

we look for weak solutions of (1.1)–(1.2), which are defined as below.

Definition 2.1. Find $(u,v,w) \in \mathcal{U} \times \mathcal{V} \times L^\infty(\Gamma_G^T)$ such that $(u(0),v(0)) = (u_I, v_I)$, and

$$(\partial_t u, \varphi)_{\Omega^T} + D(\nabla u, \nabla\varphi)_{\Omega^T} - (\vec{q}u, \nabla\varphi)_{\Omega^T} = -\varepsilon n(\partial_t v, \varphi)_{\Gamma_G^T}, \tag{2.1}$$

$$\begin{aligned} (\partial_t v, \theta)_{\Gamma_G^T} &= k(r(u,c) - w, \theta)_{\Gamma_G^T}, \\ w &\in H(v), \end{aligned} \tag{2.2}$$

hold for all $(\varphi,\theta) \in L^2(0,T; H^1_{0,\Gamma_D}(\Omega)) \times L^2(\Gamma_G^T)$.

By definition w is between 0 and 1. Here u and v stand for concentrations, so we expect similar properties.

Lemma 2.1. *If (u,v,w) is a weak solution of (1.1)–(1.2), then u, v and w are positive and bounded. Specifically, two constants M_u and M_v depending on the boundary and initial data can be found such that*

$$0 \le u \le M_u, \qquad \text{and} \qquad 0 \le v \le M_v, \tag{2.3}$$

almost everywhere. Here M_v may also depend on T and on the precipitation rate r.

Remark 2.1. Assuming Γ_D of 0–measure we obtain the following mass balance

$$\int_\Omega u(t,x)dx + \varepsilon n \int_{\Gamma_G} v(t,s)ds = \int_\Omega u_I(x)dx + \varepsilon n \int_{\Gamma_G} v_I(s)ds.$$

2.1. **Existence of a solution.** Above we have stated the definition of a weak solution, and seen that such solutions are essentially bounded. Nevertheless, it is not clear yet that such solutions exist. To prove this we use a regularization argument.

With $\delta > 0$ being an arbitrary small parameter, we consider the following regularization of the Heaviside graph:

$$H_\delta(v) := \begin{cases} 0, & \text{if } v < 0, \\ v/\delta, & \text{if } v \in (0, \delta), \\ 1, & \text{if } v > \delta. \end{cases} \tag{2.4}$$

Now a regular perturbation of (2.1)–(2.2) can be defined.

Definition 2.2. Find $(u, v) \in \mathcal{U} \times \mathcal{V}$ such that $(u(0), v(0)) = (u_I, v_I)$ and the following hold

$$(\partial_t u, \varphi)_{\Omega^T} + D(\nabla u, \nabla \varphi)_{\Omega^T} - (\vec{q} u, \nabla \varphi)_{\Omega^T} = -\varepsilon n (\partial_t v, \varphi)_{\Gamma_G^T}, \tag{2.5}$$

$$(\partial_t v, \theta)_{\Gamma_G^T} = k(r(u, c) - H_\delta(v), \theta)_{\Gamma_G^T}, \tag{2.6}$$

for all $(\varphi, \theta) \in L^2(0, T; H^1_{0,\Gamma_D}(\Omega)) \times L^2(\Gamma_G^T)$.

To show existence and uniqueness of a solution for the problem above we proceed by iteration. To this end we consider the following closed and convex sets

$$\begin{aligned} \mathcal{K}_U &:= \{u \in u_D + L^2(0, T; H^1_{0,\Gamma_D}(\Omega))/\ 0 \le u \le M_u \text{ a. e. in } \Omega^T\}, \\ \mathcal{K}_V &:= \{v \in \mathcal{V}/\ 0 \le v \le M_v \text{ a. e. in } \Gamma_G^T\}. \end{aligned} \tag{2.7}$$

Given an $u \in \mathcal{K}_U$, equation (2.6) has a Lipschitz–continuous right hand side. For the initial data v_I, it has a unique solution $v \in \mathcal{V}$, which is also bounded by 0 and M_v. Analogous, given $v \in \mathcal{K}_V$, equation (2.5) with initial data u_I has a unique solution $u \in \mathcal{U}$, which is uniformly bounded by 0 and M_u.

Thus for any $u \in \mathcal{K}_U$ we have constructed a unique element $\mathcal{T} u \in \mathcal{K}_U$. In other words, we have defined an operator

$$\mathcal{T} : \mathcal{K}_U \to \mathcal{K}_U. \tag{2.8}$$

A solution of (2.5)–(2.6) is a fixed point of $\mathcal{T}$. If a fixed point exists and it also belongs to $H^1(0, T; H^{-1}(\Omega))$, it also solves (2.5)–(2.6).

In proving existence of such a fixed point we make use of a–priori estimates that are uniform w. r. t. δ. Once these are obtained we can show that, for small times, $\mathcal{T}$ is a contraction in $\mathcal{K}_U$ with the usual norm associated to $L^2(0, T; H^1_{0,\Gamma_D}(\Omega))$. This upper time limit does not depend on the data, so the fixed point can be extended for all $t \in (0, T]$. This is summarized by

Theorem 2.2. *For any $\delta > 0$, the regularized problem stated in Definition 2.2 has a solution (u_δ, v_δ), which is a fixed point of $\mathcal{T}$. A constant $C > 0$ not depending on δ exists s. t. for any $t \in (0, T]$ we have*

$$\begin{aligned} \|u_\delta(t)\|^2_\Omega + \|\nabla u_\delta\|^2_{\Omega^T} + \|\partial_t u_\delta\|^2_{L^2(0,T;H^{-1}(\Omega))} &\le C, \\ \|v_\delta(t)\|^2_{\Gamma_G} + \|\partial_t v_\delta\|^2_{\Gamma_G^T} &\le C/\varepsilon. \end{aligned} \tag{2.9}$$

Remark 2.2. In a porous medium, Γ_G denotes the total surface of the porous skeleton, while $meas(\Omega)$ the total void volume. Since we have interpreted ε as the ratio between the pore scale and the characteristic length, a natural assumption is that

$$\varepsilon meas(\Gamma_G) \approx meas(\Omega).$$

When upscaling to a macroscopic model, the total internal surface goes to infinity as $\varepsilon \searrow 0$. The assumption above allows us to control this growth and is usually made in homogenization of periodic structures (see, e. g., [6]). In this setting, the a–priori estimates are independent not only on δ but also on ε, offering us a useful result for the homogenization procedure.

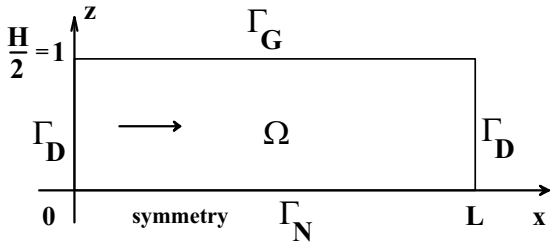


FIGURE 3. Particular domain: strip.

Theorem 2.2 provides a sequence $\{(u_\delta, v_\delta, w_\delta := H_\delta(v_\delta))\}_{\delta>0} \subset \mathcal{U} \times \mathcal{V} \times L^\infty(\Gamma_G^T)$, and the corresponding a-priori estimates. Then, by compactness arguments, a solution of the original problem exists.

Theorem 2.3. *The problem stated in Definition 2.1 has a solution (u, v, w). This solution is uniformly bounded (as shown in (2.3)) and satisfies the a–priori estimates in (2.9).*

3. Flow in a strip

In this section we investigate the formation of dissolution and precipitation fronts. These fronts are located at a free boundary separating regions of Γ_G where no crystals are present ($v = 0$) from those including some precipitate ($v > 0$). To this aim we restrict ourselves to a particular geometry in two spatial dimensions, $\Omega = (0, L) \times (0, H/2)$, with $L > 0$ possibly much larger than $H > 0$. We assume symmetry at $z = 0$, and take $\Gamma_G = (0, L) \times \{H/2\}$, $\Gamma_D = \{\{0\} \times (0, H/2)\} \cup \{\{L\} \times (0, H/2)\}$, and $\Gamma_N = (0, L) \times \{0\}$ (see also Figure 3). In agreement with previous interpretations we have $\varepsilon = H/L$.

In this case the flow has a parabolic profile and can be written explicitly,

$$\vec{q}(x, z) = (q(z), 0), \quad \text{with} \quad q(z) = C_q(H^2/4 - z^2), \tag{3.1}$$

where C_q is a given maximal velocity. Following [9] and [2], we assume here a homogeneous total charge c. This situation occurs if the charge is constant ($c_0 \in \mathbb{R}$) both initially and at Γ_D, and this value is compatible with the boundary data for the solute. Then the charge remains constant everywhere. By (A_r), a unique pair of positive reals (u_*, u^*) exists such that

$$r(u_*, c_0) = 0 \qquad \text{and} \qquad r(u^*, c_0) = 1. \tag{3.2}$$

Since now the charge is assumed constant, in this section we skip the second argument of r.

We first look for dissolution fronts on Γ_G. In doing so we assume that initially crystals are present everywhere on Γ_G, and the system is in equilibrium. This situation is perturbed by injecting fluid containing less solute, but having the same charge c_0. Specifically we take

$$v_I(x) \equiv v_0 > 0,\ u_I(x, z) = u_D(t, L, z) = u^*,\ u_D(t, 0, z) = u_*, \tag{3.3}$$

for all $x \in (0, L)$, $z \in (0, H/2)$, and $t > 0$.

Under the assumptions above some additional properties of u and v can be given.

Lemma 3.1. *If initial and boundary data are taken as mentioned in (3.3), then both u and v are decreasing in time and increasing in the x–direction. Moreover, $v \in C(\Gamma_G^T)$, while $u(t)$ is continuous up to the boundary of Ω for almost every $t > 0$.*

As follows from above, a dissolution front moves in the flow direction and separates regions on Γ_G where all the crystals have been dissolved from those where precipitate is still present. Denoting by $s(t)$ the position of the dissolution front at time $t > 0$, we expect that $v(t, x) = 0$ for all $x \le s(t)$, while $v(t, x) > 0$ for all $x > s(t)$. This situation is displayed intuitively in Figure 4, showing the evolution in time for both the precipitate and the free boundary.

The free boundary is defined rigorously in

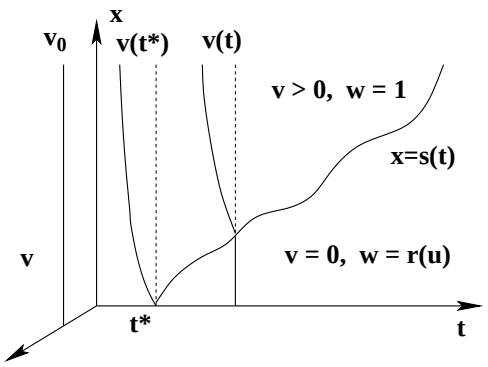


FIGURE 4. Evolution of the precipitate v and of the free boundary s.

Definition 3.1. For any t we define $s : [0,T] \to [0,L]$ as

$$s(t) = \sup \left\{ x \in [0,L] \,/ \int_0^x v(t,y)dy = 0 \right\}. \tag{3.4}$$

Remark 3.1. Due to the regularity of v, s is well defined for all t. Moreover, because v is positive, we get $v(t,x) = 0$ for a. e. $x < s(t)$.

Viewing s as the position of the dissolution front is justified by the following theorem, which also shows that a waiting time t^* has to pass until the dissolution front starts to move.

Theorem 3.2. *For the free boundary s we have*

(i) $v(t,x) = 0$, $w(t,x) = r(u(t,x,H/2), c_0)$ *for a. e.* $x < s(t)$;
(ii) $v(t,x) > 0$, $w(t,x) = 1$ *for a. e.* $x > s(t)$;
(iii) $s(t) = 0$ *for all* $t < t^*$, *where*

$$t^* = \frac{v_0}{k(r(u^*) - r(u_*))};$$

(iv) s *is continuous and strictly increasing for* $t > t^*$.

Remark 3.2. Theorem 3.2 holds for the initial and boundary data given in (3.3). The results can be extended to more general data, assuming these are compatible and satisfy

$$v_I(x) \geq 0,\ u_I(x,z) \leq u^*,\ u_* \leq u_D(t,0,z) < u_D(t,L,z) \leq u^*, \tag{3.5}$$

for all $x \in (0,L)$, $z \in (0,H/2)$, and $t > 0$. In such situations precipitation cannot occur even locally, or for a short time, since v is decreasing in time. In particular, assuming that v_I is "hat–shaped", while u fulfills (3.3), then two dissolution fronts will appear and move toward each other until crystals are completely dissolved. The support of v is shrinking in time.

Remark 3.3. Similar results can be obtained for precipitation fronts. Specifically, if the initial and boundary data are such that

$$v_I(x) = 0,\ u_I(x,z) = u_D(t,L,z) = \underline{u} = u^*,\ u_D(t,0,z) = \bar{u} > u^*,$$

then a precipitation front will move in the flow direction. It separates regions on Γ_G where precipitate is present from those not containing crystals.

3.1. **Thin strips.** Now we turn our attention to thin strips. The small parameter $\varepsilon = H/L$ plays an essential role, thus below all quantities depending on it are indexed. We maintain the setting above, and study the limit as $\varepsilon \searrow 0$.

As before, the flow takes place only in the x–direction, while the velocity q^ε is z–dependent,

$$\vec{q}^{\varepsilon}(x,z) = (q^\varepsilon(z), 0), \quad \text{with} \quad q^\varepsilon(z) = C_q \left(1 - \frac{z^2}{\varepsilon^2}\right), \tag{3.6}$$

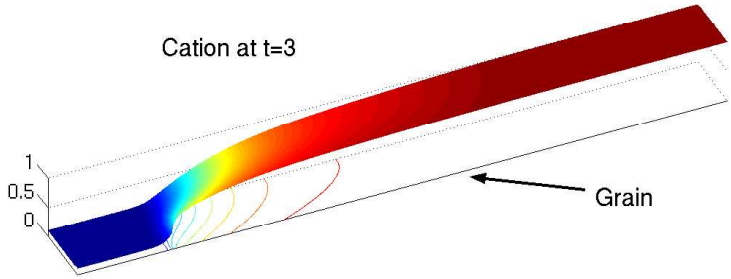


FIGURE 5. Cation at $t = 3$.

where $C_q = \frac{3}{2}Q$ and $Q = \frac{1}{\varepsilon}\int_0^\varepsilon q^\varepsilon(z)dz$ is the averaged velocity.

All the properties shown in Sections 2 and 3 are valid here too. For any $\varepsilon > 0$, the problem posed in Definition 2.1 (for the domain Ω_ε, and the corresponding boundaries) has a solution $(u^\varepsilon, v^\varepsilon, w^\varepsilon) \in \mathcal{U} \times \mathcal{V} \times L^\infty(\Gamma^T_{G,\varepsilon})$. Further, u^ε and v^ε are uniformly bounded, continuous, decreasing in time and increasing in x, and we can define a continuous and strictly increasing free boundary s_ε as in (3.4). With

$$U^\varepsilon(t,x) := \frac{1}{\varepsilon}\int_0^\varepsilon u^\varepsilon(t,x,\xi)\,d\xi, \qquad V^\varepsilon(t,x) := v^\varepsilon(t,x), \qquad W^\varepsilon(t,x) := w^\varepsilon(t,x), \tag{3.7}$$

and letting $\varepsilon \searrow 0$, we expect that $(U^\varepsilon, V^\varepsilon, W^\varepsilon)$ approaches the solution of the one-dimensional upscaled model proposed in [9] and [2]:

$$\begin{cases} \partial_t\,(U + nV) + Q\partial_x U &=& D\partial^2_{x^2}U, \\ \partial_t V &=& k(r(U) - W), \\ W &\in& H(V), \end{cases} \tag{3.8}$$

in $Q^T = (0,T)\times(0,L)$, satisfying

$$\begin{cases} U(t,0) = u_*, \quad U(t,L) = u^*, & t \in (0,T], \\ U(0,x) = u^*, \quad V(0,x) = v_0(x), & x \in (0,L). \end{cases} \tag{3.9}$$

As before, ε–independent a–priori estimates for $(U^\varepsilon, V^\varepsilon, W^\varepsilon)$ and compactness arguments give

Theorem 3.3. *The upscaled model has a unique solution (U,V,W), which is the limit of $\{(U^\varepsilon, V^\varepsilon, W^\varepsilon)\}_{\varepsilon>0}$.*

Remark 3.4. The uniqueness result is a consequence of Gronwall's lemma. This also implies weak convergence for the entire sequence $\{(U^\varepsilon, V^\varepsilon, W^\varepsilon)\}_{\varepsilon>0}$, and not only for a subsequence.

Remark 3.5. For each $\varepsilon > 0$ a free boundary s^ε exists in the sense of Definition 3.1. Similarly, a free boundary S can be defined for the upscaled model, featuring the same properties as s^ε. As $\varepsilon \searrow 0$ we also obtain that $s^\varepsilon(t) \to S(t)$ for all t.

4. Numerical example

Here we present some numerical results obtained for the particular geometry considered in Section 3. We take $\Omega_\varepsilon = (0,1)\times(0,\varepsilon)$, where $\varepsilon = 1/50$. The initial and external boundary conditions are as given as in (3.3), with $v_0 = 1.0$, $u_* = 0.1$ and $u^* = 1.0$. We also take $D = k = 1.0$, while q is given in (3.6) with $Q = 9.0$. The precipitation rate r (in 1.3) is obtained for $m = n = 1$, $K = 10/9$ and $c_0 = 0.1$, namely $r(u, c_0) = K[u]_+[u - c_0]_+$.

Computations are done by finite differences with explicit time stepping. The results are obtained for a constant time step $\tau = 0.0001$ and a uniform grid of mesh–size $h = 0.05$.

Figure 5 shows the cation at $t = 3$. Here the strip width is enlarged 5 times, and the picture is flipped over the symmetry axis. Flow takes place from left to right.

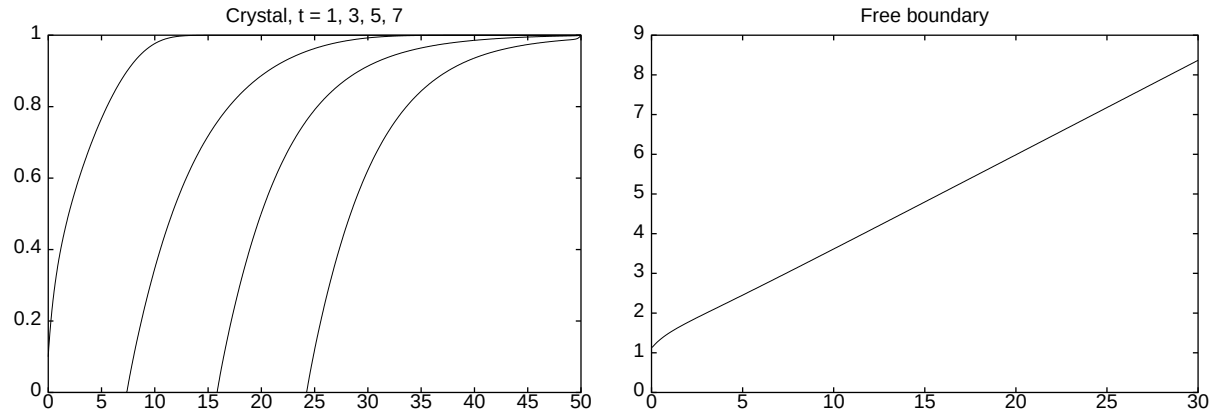


FIGURE 6. Evolution of the free boundary.

Numerical results for the precipitate at different moments are presented in the left picture of Figure 6. The horizontal axis stands for the grain boundary, and the dissolution front moves from left to right. The picture on the right displays the evolution of the free boundary. Time is represented on the vertical axis, while the free boundary location can be measured on the horizontal axis. In our computation, up to $\tilde{t}^* = 1.1424$ crystal is present everywhere on the grain. According to Theorem 3.2, the waiting time here should be $t^* = 1/0.9 = 1.11\ldots$. After a short time the dissolution front moves to the right with a constant velocity, which we estimate numerically to $\tilde{a} = 4.202$. This is a reasonable approximation of the travelling wave velocity determined in Proposition 1.2 of [2], $a = Q\,\frac{u^*-u_*}{(u^*-u_*)+v_0} = 4.263\ldots$. Refining both the time step and the spatial mesh gives a better approximations for the waiting time and the front speed, therefore we conclude that numerical results are in good agreement with the theoretical ones.

Acknowledges. We would like to thank Prof. A. Mikelić (Université Lyon), Dr. M. Peletier (CWI Amsterdam) and Dr. G. Prokert (TU Eindhoven) for useful discussions and suggestions.

REFERENCES

[1] C. J. VAN DUIJN AND P. KNABNER, *Solute transport in porous media with equilibrium and nonequilibrium multiple-site adsorption: travelling waves*, J. Reine Angew. Math. 415 (1991), pp. 1–49.

[2] C. J. VAN DUIJN AND P. KNABNER, *Travelling wave behaviour of crystal dissolution in porous media flow*, European J. Appl. Math., 8 (1997), pp. 49–72.

[3] C. J. VAN DUIJN AND I. S. POP, *Crystal dissolution and precipitation in porous media: pore scale analysis*, RANA Preprint (2003), Eindhoven University of Technology.

[4] A. FRIEDMAN AND P. KNABNER, *A transport model with micro- and macro- structure*, J. Differential Equations, 98 (1992), pp. 328–354.

[5] D. HILHORST AND M. A. PELETIER, *Convergence to travelling waves in a reaction-diffusion system arising in contaminant transport*, J. Differential Equations 163 (2000), pp. 89–112.

[6] U. HORNUNG, *Homogenization and Porous Media*, Interdisciplinary Applied Mathematics, Vol. 6, Springer Verlag, Berlin, 1997.

[7] U. HORNUNG AND W. JÄGER, *Diffusion, convection, adsorption, and reaction of chemicals in porous media*, J. Differ. Equations 92 (2001), pp. 199–225.

[8] U. HORNUNG, W. JÄGER AND A. MIKELIĆ, *Reactive transport through an array of cells with semipermeable membranes*, RAIRO Modél. Math. Anal. Numér. 28 (1994), pp. 59–94.

[9] P. KNABNER, C. J. VAN DUIJN AND S. HENGST, *An analysis of crystal dissolution fronts in flows through porous media. Part 1: Compatible boundary conditions*, Adv. Water Res., 18 (1995), pp. 171–185.

[10] P. KNABNER AND F. OTTO, *Solute transport in porous media with equilibrium and nonequilibrium multiple-site adsorption: uniqueness of weak solutions*, Nonlinear Anal. 42 (2000), pp. 381–403.

[11] M. NEUSS-RADU, *Some extensions of two-scale convergence*, C. R. Acad. Sci. Paris Sr. I Math., 322 (1996), pp. 899–904.

[12] E. ZEIDLER, *Nonlinear Functional Analysis and its applications*, Vol II/A (*Linear Monotone Operators*), Springer-Verlag, Berlin, 1990.

SEQUENTIAL METHOD FOR KINETIC MODELS DISCRIMINATION

Paula Portugal, Hélio Jorge, Rosa M. Quinta-Ferreira
Chemical Engineering Department, Pólo II, 3030-290 Coimbra, PORTUGAL

On the Chemical Reaction Engineering area is mandatory to access to valuable kinetic expressions for modelling and simulation of reactional systems. Bos *et al.*[1] published a survey indicating the need for improved methods to determine reaction kinetics. Based on the work presented by Donati *et al.* [2,3,4] that uses a sequential method for kinetic models discrimination, it was made an effort to develop a toolkit for optimal kinetic model development. The economic advantage of the present work arises from the fact that it would result in time saving and effectiveness in experimentation, parameters estimation, and searching for the optimum model from the usually wide range of theoretical models proposed for a reactional system.

Since kinetic models are usually non-linear in parameters, we used commercial routines for non-linear regression: NL2SOL and GREG. Statistical criteria were also used to reject models that fit worse experimental results such as F Hypothesis Test and Model Probability Estimation, with Bayes Theorem. Commonly just one model must be selected (the optimum one) so, discrimination process is iterative and after a statistical discrimination step it continues by designing a new set of experiments providing data for a new discrimination cycle. The used Maximum Divergence Criteria [5][6], maximises the mean difference between different model previsions.

The developed computational code (in MATHLAB) was tested for two examples referred by Donati *et al.*[4], where a simulator model was pre-defined and algorithm robustness was measured by the convergence for the defined model, which happened for all studied cases. In despite of encouraging results, additional efforts have to be made in order to meet our goal, namely seek for best non-linear fit codes, and other examples (even more complex) for testing the developed code.

Keywords: Kinetic models, sequential methods, Parameters estimation, Design of Experiments, Models Discrimination

1. Sequential Methods

Engineering problems involve frequently process models construction. Since models are built by considering the contribution of different competitive physique-chemical phenomena, they can be presented theoretically with many different configurations. The problem is, then, centred in finding out the best model. It must be, by one hand, the one which fits closely the experimental results and, by the other hand, the one which contains physically acceptable values for parameters. Such as pre-exponential factor and activation energy in kinetical models. The quality of the accepted (chosen) kinetic model influences greatly the global process model behaviour, that's why Bos *et al.*[1] published a survey indicating the need for improved methods to determine reaction kinetics in industrial processes. For this purpose, statistical discrimination methods have already been used, namely, by Donati *et al.* [2,3,4]. These are usually called Sequential Methods, because they involve three basic sequential steps:

1. Parameters Estimation for all theoretical candidate models, using optimisation methods for available experimental data fit. Usually the Method of Least Squares.
2. Models quality comparative statistical analysis (with experimental data), using statistical known parameters and tests. This step ends by worse models rejection.
3. Design of experiments, providing better experimental data for a new calculation cycle.

After step 3, starts a new calculation cycle in step 1 (see figure 1). The iterations continue

till the convergence point, where just one model is left. At the beginning of the discrimination process, if there is available any information that one group of models can be better than the others, there is no need of considering the last ones as candidates for the sequential methods discrimination.

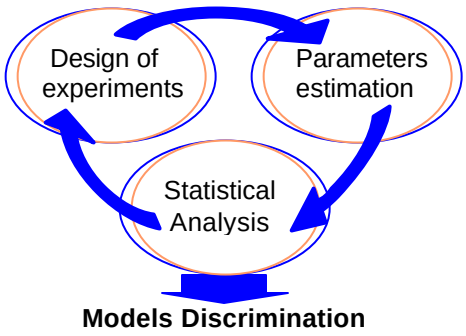


Figure 1 – Sequential Methods for Models Discrimination

1.1 Parameters Estimation

The common method for parameters estimation from experimental data fit is the Method of Least Squares (MLS):

$$\min_{\underline{\beta}} \quad S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (1)$$

$$\text{with} \quad \hat{y}_i = f(\underline{X}_i, \underline{\beta}) \qquad i = 1, 2, \ldots n$$

f can be either linear or non-linear. In kinetic models, f is usually non-linear, because it often appears, for instance, exponential terms in denominator (ex. Hougen- Watson Models).

The MLS gives good results when the error of the measured value(y_i), E, has a normal distribution – $E \sim N(0, \sigma_E^2)$. To assure this error behaviour, the experimental conditions must be well controlled and all the used apparatus should be in perfect calibrated conditions.

Sá[7] e Gouveia[8] made a previous comparative study of two different routines for non-linear fit: GREG and NL2SOL. They concluded that these routines are generally not appropriated for the required purpose. GREG does not produce well fitted results, and his poorly results became even worse as the number of estimated parameters rise (>4). On the other side, despite better fitting results of NL2SOL, it demands the consideration of parameters constraints in the optimisation problem formulation, such as, fixed partial orders for kinetic Hougen-Watson models. According to the authors this probably is due to the huge different size between the partial orders and the other parameters, such as, pre-exponential factors and activation energy.

Since we used MATHLAB to construct our computational code, the results for data fit here presented were obtained with the non-linear regression toolbox from MATHLAB, which uses the Gauss-Newton Method for optimisation of the MLS objective Function.

1.2 Statistical Analysis

The statistical parameters considered for comparison between candidate models were the Mean Relative Error, E_r:

$$E_r(\%) = \frac{1}{n}\sum_{i=1}^{n} E_{r,i} \times 100 = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \qquad (2)$$

that should be as less as possible, and the Determination Index[9], R^2:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2} \qquad (3)$$

where $0<R^2<1$, and it is the ratio between the model previsions variability and the experimental data variability. Models that have R^2 closer to 1 have better chances of being considered good models. It is a general rule that an acceptable model should not have a R^2 value less than 0,99. It is important remember that if a model is statistically considered as not good, we should simply accept it as not good, or eventually change it and test it again. For the contrary, if it is statistically considered as a good model, that does not prove that it is correct, just that there is no statistical evidence, with the available data, that allow us to reject it.

Two statistical criteria were used to discriminate models:

- F Hypothesis Test (F Test of Variances)[9]
- Model Probability Estimation (Bayes Theorem)[5,6]

1.2.1 F test

F test considers two basic hypothesis:

- H0 : The model fits well the experimental results
- H1: The model does not fit well the experimental results.

If $\Pr(F \le F_0) = \alpha$ the model is accepted, where F is a statistical function with $F_{n_p, n-n_p}$ distribution, calculated by:

$$F = \frac{\sigma_m^2}{\sigma_\varepsilon^2} = \frac{\dfrac{\sum y_i^2 - \sum (y_i - \hat{y}_i)^2}{n_p}}{\dfrac{\sum (y_i - \hat{y}_i)^2}{n - n_p}} \qquad (4)$$

and α is the significant level, usually it assumes the values 0,10 , 0,05 or 0,01. F_0 is the cumulative distribution $F_{(1-\alpha)\cdot 100\%, n_p, n-n_p}$. In other words: if $F \ge F_0$, H_0 is accepted, if not, H_1 is accepted. Donati and Ferraris[2,3,4] used F test, but they refer that some researchers use ψ coefficients in order to get better confidence in the method for linear models discrimination. In this case the acceptance rule is $F_m \ge \psi F_0$. Donati and Ferraris suggest, also, an empirical value of 4 for the ψ parameter, but for non-linear fit the F test is less adjustable, so the same authors suggest a value of 10 for these cases.

1.2.2 Models Probability

The model probability is the probability of being the best model among the other candidates. This parameter is actualised from one calculation cycle to another. It can be seen as a model behaviour indicator as actualised experimental data is provided from iteration to iteration. If there is no information that can make us think that one model is better than the others, the initial model probability is the same for all of them. Anyway it should be true that:

$$\sum_{i=1}^{m} \Pr(M_{i,0}) = 1 \qquad (5)$$

Where m is the number of candidate models.

In the n^{th} iteration the actualised i model probability is calculated by using the Bayes Theorem with normalised values to respect the constraint in eq. 5:

$$\Pr(M_{i,n}) = \frac{\Pr(M_{i,n-1})\, L(M_i | y_n)}{\sum_{i=1}^{m} \Pr(M_{i,n-1})\, L(M_i | y_n)} \qquad (6)$$

where $L(M_i | y_n)$ is the likelihood function, which is:

$$L(M_i | y_n) = \frac{1}{\sqrt{2\pi(\sigma_E^2 + \sigma_i^2)}} \exp\left[-\frac{1}{2(\sigma_E^2 + \sigma_i^2)} (y_n - \hat{y}_{i,n})^2 \right] \qquad (7)$$

1.2 Design of Experiments

If the design of experiments is made to improve confidence in parameters estimation we say that the sequential method is constructed for optimal estimation. In the present work the goal is to find out a method for optimal model discrimination, that is why we do not use a traditional design of experiments, but rather the Maximum Divergence Criteria[5,6]. This method selects the best experimental conditions that maximise the difference between the mean previsions of all candidate models. The new set of experimental conditions, $\underline{X}_n$, is then given by:

$$\max_{\underline{X}_n} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (\hat{y}_i - \hat{y}_j)^2 \qquad (8)$$

In the present work a direct search method was used to find out the optimum. A discrete grid of feasible operating conditions is predefined and the maximum objective function (Divergence) value is inspected by comparison of a set of two values at a time. Note that we could also made a continuous search by using any optimisation method for multivariable functions in a predefined range of experimental feasible values.

2. Developed Code for Models Discrimination

As it was already said the programming language used was MATHLAB. Figure 2 is a program flowsheet that shows the needed data, and the information flow. The program was tested without experimental real data. In fact we choose one of the candidate models to simulate experimental results for all the cycles. These data was generated by random extraction from a normal population with a

pre-defined variance (called experimental variance).
After some attempts to achieve convergence in the data fit optimisation process, we noted that it greatly depends on the initial estimated values, so the program was improved to overcome this problem by changing the initial estimated value.

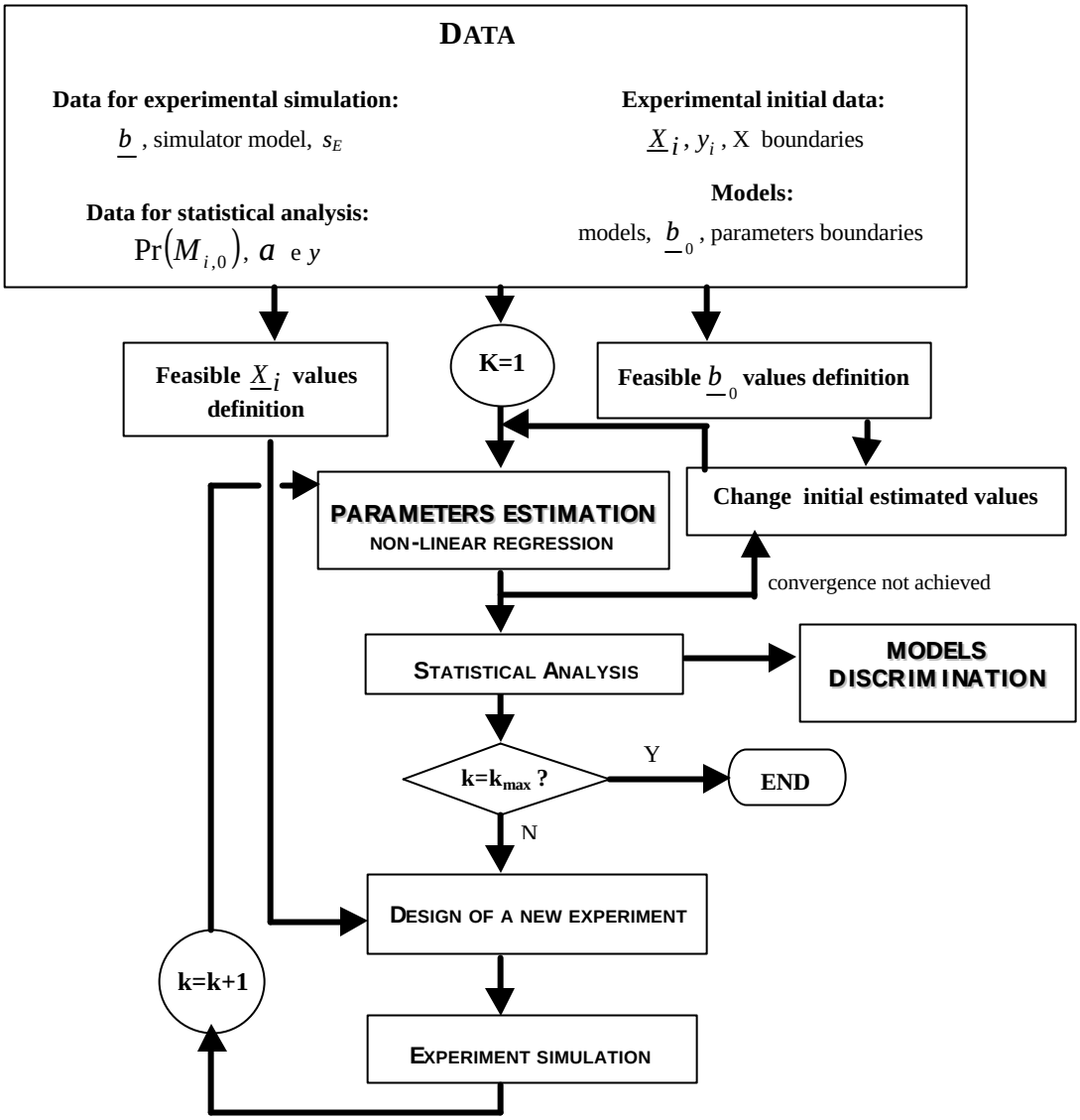


Figure 2 – Program Flowsheet for Models Discrimination.

3. Results

Two examples referred by Donati and Ferraris[4] were used to test the program, but just one is here presented.
The four models candidates to describe the kinetic behaviour of a reaction A → B are:

Model 1: $y^{(1)} = \exp\left[-b_{11}\, x_1 \exp\left(\frac{-b_{12}}{x_2}\right)\right]$

Model 2: $y^{(2)} = \frac{1}{1 + b_{21}\, x_1 \exp\left(-b_{22}/x_2\right)}$

Model 3: $y^{(3)} = \frac{1}{\left[1 + 2b_{31}\, x_1 \exp\left(-b_{32}/x_2\right)\right]^{1/2}}$

Model 4: $y^{(4)} = \frac{1}{\left[1 + 3b_{41}\, x_1 \exp\left(-b_{42}/x_2\right)\right]^{1/3}}$

Where y is the concentration of A, x_1 the reaction time, x_2 the temperature and b_{i1} e b_{i2} the model i parameters.
Nine tests were made to study different effects in the calculated results.

Test 1 - Model 2 was chosen as simulator with $b_{11} = 50$ and $b_{12} = 3500$ and $s_E = 0{,}05$. The x_1 and x_2 boundaries are:

$$0 \le x_1 \le 150 \qquad 450 \le x_1 \le 600$$

The F test significance level (α) was 0,05 and the first starting four experiences were planed by Donati and Ferraris[4] with a two level factorial design of experiments (table 1):

Table 1 – Simulated experimental data for the first test 1 iteration.

Exp	x_1	x_2	Y
1	25,0	575,0	0,4854
2	25,0	475,0	0,7231
3	125,0	575,0	0,1059
4	125,0	475,0	0,3523

The initial model probabilities were assumed to be the same for all of them, which means $\Pr(M_{i,0}) = 1/m = 0{,}25$.
Just 5 iterations were sufficient to conclude about best model 2 behaviour, as can be seen in table 2 and figure 3. In fact they can be ordered from the best to the worse as $2 \rightarrow 3 \rightarrow 1 \rightarrow 4$. The F test may be a wrong indicator for models discrimination because F_0 decreases as interations progress, which means that F test becomes less demanding in model selection. For instance, it is easier to reach $10xF_0$ in the 5th iteration than in the 1st one, which means that as iterations progress all the models can easier satisfy the F test and be considered good models.

Tests 2 and 3 - As in test 1, model 2 was chosen as simulator with the same initial conditions. The objective was to study the influence of the program random error effect generator in the model discrimination. The obtained results indicated that this factor has a neglected effect in the models selection. Just as in test 1 the model 2 continues to be the best one and model 3 is the closest one. It was also observed that there was no significant effect in the new planed experiments, which leads us to conclude that the Maximum Divergence Criteria is an applicable method for design of experiments.

Tests 4, 5 and 6 - Model 3 was chosen as simulator for the same conditions of tests 1 to 3. The results presented in the figures 4 to 6 and tables 3 to 5, are apparently discordant

Table 2 – Results from the Test 1 (model 2 as simulator)

		Model					Model		
		1	**2**	**3**	**4**		**1**	**2**	**3**	**4**
		R^2				**E_r**				
Iteration n.	1	0,983	0,996	0,982	0,958		23,9	11,2	31,5	46,2
	2	0,975	0,996	0,977	0,944		30,6	9,7	33,3	50,7
	3	0,000	0,997	0,983	0,957		100,0	8,3	29,8	46,7
	4	0,980	0,998	0,989	0,973		29,7	7,6	25,6	41,1
	5	0,978	0,997	0,983	0,960		33,9	10,8	34,7	54,2
		F				**F_0**		**$10´F_0$**		
Iteration n.	1	59,3	*273,4*	53,1	22,9		19,00		190,0	
	2	58,8	*416,0*	63,5	25,5		9,55		95,5	
	3	0,0	*700,6*	*116,0*	44,9		6,94		69,4	
	4	*122,4*	*1007,1*	*233,5*	*89,1*		5,78		57,8	
	5	*130,8*	*954,7*	*178,5*	*71,9*		5,14		51,4	

Underlined values – fitted values where $F \ge 10\ F_0$

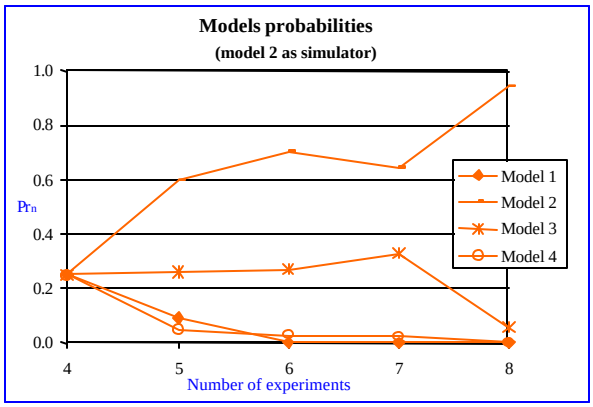


Figure 3 – Models probabilities for test 1.

between the statistical parameters (R^2 E_r(%) Ftest) and the model probabilities, and also discordant about the best model election. In test 4, model 4 was elected the best one (figure 4), in test 5 it was chosen model 2 (figure 5) and in test 6 the model 3 (figure 6). This may be due to similar 2, 3 and 4 models behaviour when model 3 is selected to be the simulator.

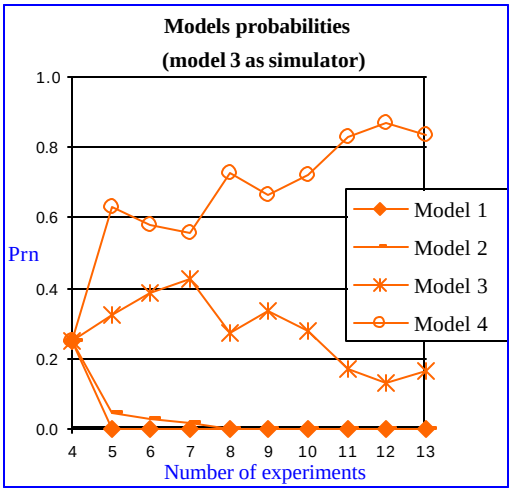


Figure 4 – Models probabilities for test 4

Table 3 – Results from the 3 last iterations for test 4 (model 3 as simulator).

I	Model 1	Model 2	Model 3	Model 4
R^2				
8	0	0,9821	0,9950	0,9947
9	0	0,9832	0,9952	0,9952
10	0	0,9847	0,9953	0,9953
E_r(%)				
8	100	15,319	7,397	7,311
9	100	14,274	7,218	6,886
10	100	13,225	8,072	6,782
F				
8	0	246,9	893,6	845,0
9	0	291,8	1026,1	1028,8
10	0	354,7	1159,5	1154,3

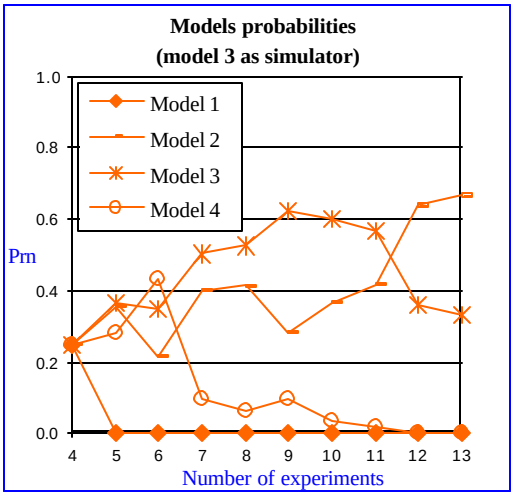


Figure 5 – Models probabilities for test 5

Table 4 – Results from the 3 last iterations for test 5 (model 3 as simulator).

i	Model 1	Model 2	Model 3	Model 4
R^2				
8	0	0,9920	0,9934	0,9890
9	0	0,9927	0,9930	0,9879
10	0	0,9935	0,9937	0,9888
E_r(%)				
8	100	12,872	13,806	17,047
9	100	12,260	13,365	17,059
10	100	11,408	12,550	16,365
F				
8	0	557,6	677,9	406,1
9	0	679,1	708,1	409,5
10	0	841,4	862,5	483,6

Table 5 – Results from the 3 last iterations for test 6 (model 3 as simulator).

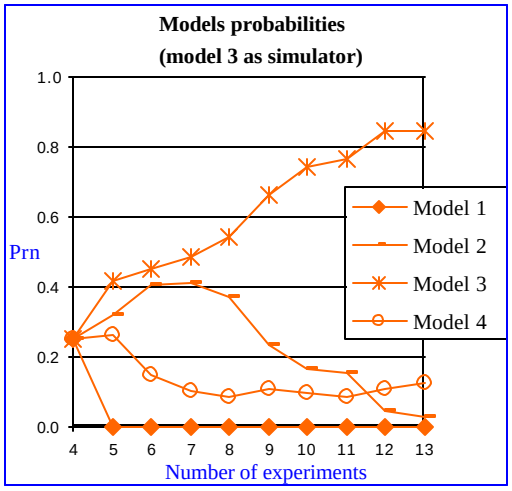


Figure 6 – Models probabilities for test 6

i	Model 1	Model 2	Model 3	Model 4
R^2				
8	0	0,9882	0,9976	0,9956
9	0	0,9881	0,9977	0,9959
10	0	0,9887	0,9977	0,9962
E_r(%)				
8	100	14,018	6,340	7,165
9	100	13,174	6,236	6,674
10	100	12,285	6,138	6,272
F				
8	0	375,5	1863,9	1007,7
9	0	416,1	2127,0	1220,2
10	0	482,3	2374,5	1448,0

Alternated elections between these 3 models probably occur because the simulated error variance is, in this case (similar models behaviour), large enough that, with some simulated experiments the model 2 slightly "benefits", and with other simulated experiments the model 4 or 3 slightly "benefits". The previous presented explanation is even more plausible as we observe the results from tests 7, 8 and 9, where experimental error variance was reduced.

Tests 7 , 8 and 9 – As for tests 4 to 6, model 3 was chosen as simulator, but the variance of simulated experimental errors was changed to half of his value (σ_E = 0,025). Figure 7 and the correspondent table 6 are the results of test 9, here presented as an example from the set 7 to 9 tests, because for the others tests we obtained similar results. The model 3 is in this case clearly better than the others comparatively to the simulated cases (test 4 to 6) where the variance was 0,05. It proves that the experimental error variance has a determinant effect in models selection.

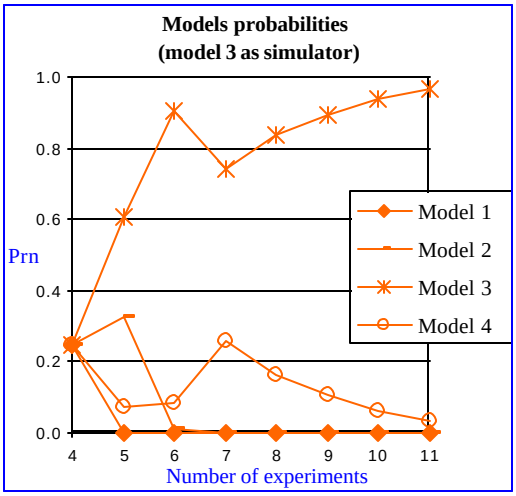


Figure 6 – Models probabilities for test 9

Table 6 – Results from the 3 last iterations for test 9 (model 3 as simulator).

i	Model 1	Model 2	Model 3	Model 4
R^2				
6	0	0,9816	0,9968	0,9961
7	0	0,9832	0,9972	0,9964
8	0	0,9851	0,9975	0,9967
E_r(%)				
6	100	17,831	7,574	6,618
7	100	16,380	6,861	6,237
8	100	14,972	6,310	5,912
F				
6	0	186,72	1083,96	884,35
7	0	234,22	1420,14	1107,92
8	0	297,07	1802,13	1344,44

4. Conclusions

The present work shows that it is possible to construct a sequential method for kinetic models discrimination. The starting steps for the toolkit development are here presented. Improvements pass for search for better non-linear fit routines. It is important to keep in mind that may be necessary a set of different fit routines to satisfy particular kinetic models demands. As it was said before, a statistically acceptable model does not make it good for process simulation, because it can contain physically impossible parameters values. So it is important to introduce other discriminating criteria, or optimisation problem constraints, or even more sophisticated design of experiments in order to assure that the chosen model is really the right acceptable one to incorporate the global reactor model. As final statement it should be remembered that it was here proven that experimental errors (even simulated) play an important role in the discrimination process. So it is important, as expected, that researchers provide validated experimental data for optimal model discrimination.

Nomenclature:

$\underline{X}$	Array of independent variables of a model
$L(M_i \mid y_n)$	*Likelihood* function of model i after the y_n observation
ε	Residual value of a model estimated y value
σ^2	Variance
E_r	Mean relative error
F	F function = Variances ratio
J	Jacobian matrix with dimension $(n-1)\times n_p$ = model differential equations in order to parameters in the n-1 experiments
j_n	Jacobian matrix with dimension $1\times n_p$ in the point $\underline{X}_n$
m	Number of candidate models to be discriminated
MLS	Least Squares Method
n	Number of experiments (system observations)
n_p	Number of model parameters
$Pr(M_i)$	Model i probability
R^2	Determination Index
y	Dependent variable of a model

Greek

α	Significant level of F statistical test
$\underline{\beta}$	Array of model parameters
σ_E^2	Experimental error variance
σ_i^2	Variance of the mean estimated values by model i
σ_ε^2	Residual values variance
σ_m^2	Model fit variance
ψ	Safety coefficient for non-linear application of F test

References:

[1] Bos, A.; Lefferts, L.; Marin, G.; Steijns, M.; " Kinetic Research on Heterogeneously Catalysed Processes: A Questionnaire on the State-of-the-art in Industry"; Appl. Catal. A : General, 160 (1997) 185-190

[2] Ferraris, G. B., Donati, G., "The Use of a Stepwise Non-linear Procedure For Hougen and Watson Kinetic Model Building"; Ing. Chim. Ital.; vol 10; 7-8 (1974) 121-126

[3] Ferraris, G., Donati, G.; "Analysis of the Kinetic Models for the Reaction of Synthesis of Methanol"; Ing. Chim. Ital.; vol. 7; 4 (1971) 53-64

[4] Donati, G.; Ferraris, G.; "Experience With an Algorithm for Model Fitting and Discrimination"; Ing. Chim. Ital.; Vol. 8, 7-8 (1972) 183-192

[5] Burke, A. et al.; "Choosing the Right Model: Case Studies on the Use of Statistical Model Discrimination Experiments"; Ca. J. Chem. Eng., vol. 75 (1997) 422-436

[6] Burke, A.; "Discriminating Between the Terminal and Penultimate Models Using Designed experiments: An Overview"; Ind. Eng. Chem. Res.; 36 (1997) 1016-1035

[7] Sá, P.; Montagem e Operação de Reactores Catalíticos Laboratoriais", Relatório de Seminário I/II; Depto Eng. Química da FCTUC (1995)

[8] Gouveia, V.; "Estudos Cinéticos de Sistemas Heterogéneos"; Relatório de Seminário I/II; Depto Eng. Química da FCTUC (1996)

[9] Montgomery, D. ; "Design and Analysis of Experiments", 2nd. ed. ; John Wiley & Sons, USA, 1983

Modelling and numerical simulation of variably saturated flow and coupled reactive, biogeochemical transport

Alexander Prechtel, Florin Radu and Peter Knabner
Institute for Applied Mathematics, University of Erlangen-Nuremberg
Martensstr. 3, D-91058 Erlangen, Germany
Phone: +49 9131 8527013, Fax: +49 9131 8527670
E-mail: prechtel@am.uni-erlangen.de

Abstract

The simulation of the fate of contaminants in the subsurface is a demanding task concerning modelling and accurate numerical solution of the problems. Reliability and correctness of modelling results are particularly important when site remediation strategies like natural attenuation are being considered.

We present a flexible, modular simulation tool that owns a robust and efficient mathematical, algorithmic kernel. The model components include the Richards equation for (un-)saturated fluid flow and the advection-diffusion-reaction-equations for transport of multiple species with (nonlinear) terms for sorption processes, biodegradation, and chemical reaction kinetics. Advanced numerical methods like a hybrid mixed finite element discretization guarantee local conservation of mass and asymptotic exact approximations of the solution. The systems of equations are treated with Newton's method, and coupled systems are solved implicitely fully coupled. We present an example of a biodegradation problem that encourages the use of complex transport models.

1 Introduction

Numerical models have proven their value along with laboratory and field experiments in processing the available site information and predicting the the migration and extent of contaminant plumes in many case studies. They support the decision for or against a certain remediation strategy not only by attempting a prognosis, but also by identifying and accessing the interplay of the complex processes and evaluating hypothetical worst case scenarios or action variants. Therefore it is essential to have a comprehensive flexible software tool at hand that relies on a robust and accurate mathematical kernel.

The processes that determine the fate of organic contaminants are highly complex, nonlinear and case specific. Thus we need a modular conceptual basis to facilitate the combination of existing and incorporation of new model components. It is beyond the scope of this paper to depict every process that has been integrated in the mathematical model with its defining equations in detail. We will restrict ourselves to a presentation of the general framework, designate the advanced mathematical solution strategies and exemplify a problem of interest for site remediation studies. The tool named RICHY [1] currently solves the sets of partial differential equations corresponding to the following problem classes:

- Heat conduction,
- variably saturated flow,
- solute transport with (non-)linear equilibrium and (multiple site) kinetic sorption including carrier facilitation (see [2] for an application),
- biodegradation with multiplicative Monod kinetics, inhibition and dynamic biomass,
- reactive multicomponent transport with geochemical kinetics,
- surfactant transport interacting with fluid dynamics, and
- multiphase flow.

These modules may be combined to perform complex simulations and take the interactions of different processes into consideration. This means not only the unidirectional combination of e.g. water flow and solute transport, where the first problem may be solved without knowing the solution of the second one (but not vice versa). In particular, this includes the simultaneous solution of mutually coupled problems in every time step without splitting the equation systems. An example is coupled fluid flow and surfactant transport, where the effects on hydraulic conductivity (clogging) and surface tension alter the fluid flow (see [3]).
RICHY is implemented in the language C, making use of OpenGL and Tcl/TK, and is thus portable to Unix, Linux and Windows-platforms.

2 Some Model Equations

2.1 Fluid Flow

The description of the flow regime in the saturated and the vadose zone is based on the conservation of mass and Darcy's law. We establish the well known Richards equation for fluid flow in its pressure formulation:

$$\partial_t \Theta(p) + \nabla \cdot \vec{q} = 0 \qquad \vec{q} = -\frac{k}{\mu} k_r(p) \nabla(p + \rho g z) \tag{1}$$

Here t denotes the time, Θ the volumetric water content, p is the pressure head, $\vec{q}$ the Darcy flux, k is the intrinsic permeability of the porous medium, μ the viscosity, k_r is the relative hydraulic conductivity, ρ the density of the fluid, g the acceleration due to gravity and z is the elevation head.
This model is augmented with two coefficient functions: As indicated, the water content Θ is a function of the pressure head p – the water retention curve – and the relative hydraulic conductivity k_r depends as well on p in a nonlinear form. For these functional relationships different parametrizations exist in the literature, and can be incorporated in the model. We added e.g. the van Genuchten - Mualem model, but also a form-free ansatz based on spline interpolation that may be the result of an inverse modelling procedure for parameter identification, for which also efficient tools are integrated into RICHY [4]. The coupled simulation of flow and carrier facilitated transport in the vadose zone has been demonstrated in [2].

2.2 Mass Transport – General Formulation

A general model for the transport of solutes dissolved in groundwater that includes advection, dispersion, diffusion and general reaction terms reads [5]:

$$\partial_t(\Theta c_i) - \nabla \cdot (\vec{D}\nabla c_i - \vec{q}c_i) = \sum_{j=1}^{N_R} \nu_{ij} R_j. \tag{2}$$

c_i denotes the solute concentration of species X_i ($i \in \{1, \ldots, N_S\}$, with the total number of species N_S), $\vec{D}$ is the diffusion-dispersion tensor (which we assume to be the same for all mobile species), N_R is the total number of reactions, ν_{ij} a stoichiometric coefficient, and R_j the reaction rate expression of the j-th general reaction. These reaction terms may account for sorption phenomena, decay or other biogeochemical reactions and potentially depend on other parameters or concentrations: $R_j = R_j(c_1, \ldots, c_{N_S}, x, t, T, \ldots)$.
For immobile species the transport terms are omitted and we have

$$\rho_b \partial_t(c_i) = \sum_{j=1}^{N_R} \nu_{ij} R_j. \tag{3}$$

We do not consider effects of the species concentrations on the water flow here. The reaction terms may couple the transport equations of the species to each other and require their simultaneous solution.

2.3 Biodegradation

Biodegradation models exist on different levels of complexity. Besides decay of 0th order with $R_j = -\text{const}$ we may think of elementary, irreversible first-order decay, resulting in linear reaction networks $X_1 \xrightarrow{k_1} X_2 \xrightarrow{k_1} \cdots \xrightarrow{k_{n-1}} X_n$ with terms

$$R_j = R_j(c_{i-1}, c_i) = k_{i-1}c_{i-1} - k_i c_i \tag{4}$$

in the equation for substance X_i. A more complex but also widely used model to quantify biodegradation rates in the subsurface is the so-called multiplicative Monod-model [6], derived from enzyme kinetics. As biodegradation of organic contaminants is often a redox process, catalyzed by microorganisms, we have to take into account the availability and dynamics of electron acceptors (such as oxygen or nitrate) and electron donors (an organic substrate), and the biomass, which we assume to be immobile.
A general Monod model combines growth terms of the type $\frac{c_i}{K_i + c_i}$ (K_i denotes the Monod half saturation concentration) with inhibition terms $h(c_i)$. Widdowson et al. [7] proposed e.g.

$$h(c_i) = \frac{K_{I_i}}{K_{I_i} + c_i}. \tag{5}$$

with an inhibition concentration K_{I_i}. Thus a substance may exclusively enhance or inhibit the degradation reaction, or even both in different concentration ranges. Thus the reaction rate for such a general microbial reaction $r \in \{1, \ldots, N_R\}$ with arbitrary electron donors, acceptors, inhibitors, and biomass (concentration c_{X_r}) reads

$$
\begin{aligned}
R_r &= R_r(c_1, \ldots, c_{N_S}, c_{X_r}) \\
&= -\mu_{\max_r} c_{X_r} \prod_{i \in I_r \subset \{1,\ldots,N_S\}} \left(\frac{c_i}{K_{M_i} + c_i} \right) \prod_{j \in J_r \subset \{1,\ldots,N_S\}} h(c_j).
\end{aligned} \tag{6}
$$

The index set I_r contains the indices of the species that are transformed (and thus necessary) in that reaction, the set J_r consists of the species, that inhibit the degradation reaction. $\mu_{\max_r}$ is the maximum specific growth rate of the biomass in this reaction.
These biodegradation processes occur in both vadose and saturated zone, thus a coupling to the Richards equation is important. Current work deals with the application of this model to quantify the potential of contaminated sites for natural attenuation.

2.4 Geochemical Reaction Kinetics

To account for additional geochemical reactions, we incorporate the following general rate expression of the rth elementary kinetic reaction, which can be formulated with the help of rate constants for forward and backward reaction k_f and k_b under the common assumption of mass action kinetics [5]:

$$
R_r = \left(k^f \prod_i c_i^{\nu_{ir}} - k^b \prod_j c_j^{-\nu_{jr}} \right). \tag{7}
$$

Index i refers to the reactants (educts), Index j to the product species, which have stoichiometric coefficients $\nu_{jr} < 0$. The implemented prototype of such a general multicomponent model may account for an arbitrary number of such kinetic reactions with several species, which may also take part at the same time in arbitrary biogeochemical degradation reactions of the type (6).
Formulation (7) results in the law of mass action (describing thermodynamic equilibrium) for $R = 0$. Note that rate constants may vary by several orders of magnitude.

3 Discretization Techniques and Numerical Methods

Advanced mathematical strategies are essential to guarantuee the efficiency and accuray of the calculations. Many conventional methods lead to qualitatively wrong results in demanding situations because of numerical diffusion, splitting errors or other deficiencies of the algorithms. Bause and Knabner [8], e.g., show in the case of a biodegradation problem, that improvements in accuracy by a higher order finite element scheme and adaptive time stepping lead to substantially different predictions of the contaminant migration.
The presented model equations are discretized by the fully implicit backward Euler method in time and by finite elements in space. To cope with stiff reaction problems, a two-step method in time, namely the backward differential formula of second order (BDF-2) has been incorporated. The Richards equation and the coupled water/surfactant problem are discretized by hybrid mixed finite element methods to ensure the local conservation of mass and the continuity of the flux (also for heterogeneous media), a crucial quality for subsequent transport processes depending on that fluid flow. The standard conforming finite element method with mass lumping

is used for the discretization of the other transport modules in 1D, in 2D/3D we also rely on the hybrid mixed FEM.
A damped version of Newton's Method (Armijo's rule) solves the local and the global nonlinear equations that result from discretizing the partial differential equations. When two-step methods in time are used, the convergence of the Newton iteration is improved by an initial iterate that is generated by interpolation of previous time steps. The global system of linear equations is solved by a direct sparse matrix solver in 1D, and the multigrid method in 2D/3D. Based on the equivalence of nonconforming and mixed finite elements the multigrid method is in this case built from grid transfer operators, derived for the Crouzeix-Raviart element [9]. The linear problem on the base level of the grid hierarchy is solved by LU decomposition. On fine grid levels smoothers like Gauss-Seidel or ILU are used.
Coupled multicomponent problems (2) and (3) are solved implicitely fully coupled. This means that in the Jacobian, all terms $\partial R_j/\partial c_i$ are calculated. This strategy is more memory demanding than a decoupling of the entries in the Jacobian by neglecting off-diagonal terms, since on each element we store a $(N_S \times N_S)$-submatrix. On the other hand, convergence of the Newton iteration is improved by a better approximation of the problem (compare [10]).
RICHY supports adaptive strategies to control the sizes of time steps and grid spacing. These techniques ensure the efficient utilization of the available resources of a computer, that otherwise would be restrictive for complex multicomponent scenarios. Using error indicators for the finite element discretization of the model equations, the grid representing the underlying domain of the simulation may automatically be refined and coarsened, corresponding to the form of the solution. Additional indicators for the error of the time discretization allow for an adaptive time stepsize control. This automative adaption of discretization parameters is currently implemented for (un-)saturated fluid flow [11] and will be applied also to the remaining model components.

4 Numerical Example

The study of the fate of organic contaminants in the subsurface involves a variety of complex processes. In practice the models of these complex processes are often facilitated. This may be appropriate in some specific situations, but is often not justified and mostly due to limited availability of field data. However we want to show an example of reactive solute transport, where the facilitations entail misleading interpretations of the contaminant migration. Thus we want to emphasize the importance of complex models for complex processes.
Firstly we present a comparison of the popular approach of first-order decay to Monod kinetics to account for biodegradation processes. The degradation of benzene depends on the availability of electron acceptors like oxygen and the activity of biomass. Only under certain quasi-stationary conditions and in specific concentration ranges, first-order models are an appropriate simplification of the Monod model [12]. Unfortunately the first order model is widely misapplied and accepted also where it may not be tolerable [12, 13]. The parameters we use are based on experimental findings of Schirmer et al. (see [14] for the reaction parameters). They are derived from the same laboratory experiments, and give reasonable results for those batch experiments.
We simulate the continuous infiltration of benzene in the centre of the top of our domain (see Fig. 1). Oxygen is available in the whole domain (10 m x 10 m) at the beginning of the simulation and also infiltrates from the top. To make the effect of active versus inactive biomass more evident, we allowed biomass not to grow until a depth of 1 m. At $t = 50$ d all three models

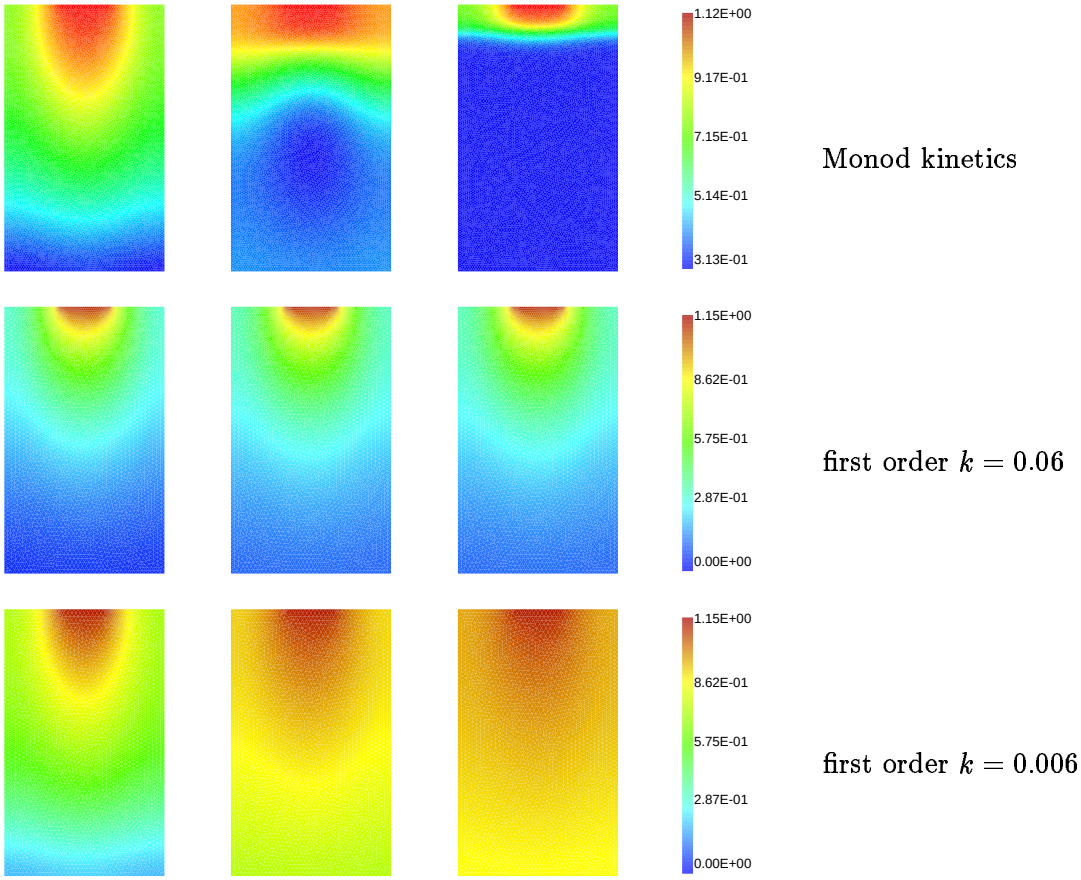


Figure 1: Simulation of benzene degradation with Monod kinetics (top), first order decay with $k = 0.06$ 1/d (middle), and first order decay with $k = 0.006$ 1/d (bottom) for $t = 10$ d (left column), $t = 20$ d (middle column) and $t = 50$ d (right column). Note the different ranges.

have reached a quasi-stationary state, which would be the basis for a long-term prediction.
We see (left column in Fig. 1) that in the beginning at $t = 10$ d, the three models differ not substantially. Monod model and $k = 0.006$ 1/d give very similar results. However, as biomass activity will now increase, also the degradation will increase – a fact that is reflected only by the Monod model. We see that at $t = 50$ d the benzene propagation is limited to the upper layers, before in a sharp reactive zone at 1 m the main part of the degradation occurs. In the stationary state (right column) about 0.3 mg/l will still infiltrate in the deepest layer. The first order model ($k = 0.006$ 1/d) however predicts, that the contaminant reaches the deepest layer in high concentrations almost without substantial degradation (0.85 mg/l) in the long term. A contrary evolution shows the first order model with $k = 0.06$ 1/d. This model will predict an almost complete degradation in the deepest layer.
This example shows that the application of an inadequate model may lead to a fatal misinterpretation of the risk potential of a contaminated site, because the conditions to apply a first order model are not met here during the first, decisive period of the simulation. Here we meet highly dynamic, non stationary conditions with biomass growth, and donor and acceptor variations.

Note that a realistic field study will even show more heterogeneities, e. g. in the material properties or boundary conditions, so that the discrepancies should even increase.

5 Conclusions and Future Work

We presented a simulation tool capable of treating complex flow and transport scenarios, in particular applications where steady state assumptions and other simplifications are not appropriate. This is the case, e.g., for highly mobile components in the subsurface where variations of the fluid flow have an immediate impact on travel times and transport behaviour of the contaminants, or for highly nonlinear reactive processes. The application of the model allows us to study complex transport phenomena for a better understanding of the interactions of the underlying processes that have been identified in experimental studies. In current site remediation studies the tool is applied for estimating the potential of natural attenuation at several sites. While we want to encourage the development and application of complex models, computer resources are still a limitation for 3D case studies. Therefore reduction strategies of coupled multicomponent systems are a vital field of current research. Future work includes the extension of the model to geochemical equilibrium reactions and the transfer of the adaptive time step and grid size routines to the transport problems.

Acknowledgements Parts of this work have been supported by the Bavarian State Ministry for Regional Development and Environmental Affairs (Verbundvorhaben "Nachhaltige Altlastenbewältigung unter Einbeziehung des natürlichen Reinigungsvermögens", BayStMLU).

References

[1] Institute of Applied Mathematics, University of Erlangen (2003), RICHY's Manual, http://www.am.uni-erlangen.de/software/RichyDocumentation/

[2] A. Prechtel, P. Knabner, E. Schneid and K. U. Totsche (2002), *Simulation of carrier facilitated transport of phenanthrene in a layered soil profile*, J. Cont. Hydrol. 56, 209-225.

[3] P. Knabner, S. Bitterlich, R. Iza-Teran, A. Prechtel and E. Schneid (2003), *Influence of Surfactants on Spreading of Contaminants and Soil Remediation*, in: W. Jäger and H.J. Krebs (eds.), Mathematics - Key Technology for the Future, Springer, Berlin, 152-161.

[4] S. Bitterlich and P. Knabner (2002), *An efficient method for solving an inverse problem for the Richards equation*, J. Comput. Appl. Math. 147, 153-173.

[5] J. C. Friedly and J. Rubin (1992), *Solute transport with multiple equilibrium-controlled or kinetically controlled reactions*, Water Resour. Res. 28, 1935-1953.

[6] R. C. Borden and P. B. Bedient (1986), *Transport of dissolved hydrocarbons influenced by oxygen-limited biodegradation, 1. Theoretical development*, Water Resour. Res. 22, 1973-1982.

[7] M. A. Widdowson, F. J. Molz and L. D. Benefield (1988), *A numerical transport model for oxygen- and nitrate-based respiration linked to substrate and nutrient availability in porous media*, Water Resour. Res. 24, 1553-1565.

[8] M. Bause and P. Knabner (2003), *Numerical simulation of contaminant biodegradation by higher order methods and adaptive time stepping*, submitted to Computing and Visualization in Science, 2003.

[9] D. Braess and R. Verfürth (1990), *Multigrid methods for nonconforming finite element methods*, SIAM J. Numer. Anal. 27, 979-986.

[10] B. A. Robinson, H. S. Viswanathan and A. J. Valocchi (2000), *Efficient numerical techniques for modeling multicomponent ground-water transport based upon simultaneous solution of strongly coupled subsets of chemical components*, Adv. Water Resour. 23, 307-324.

[11] E. Schneid (2000), *Hybrid-Gemischte Finite-Elemente-Diskretisierung der Richards-Gleichung*, PhD Thesis, Institute of Applied Mathematics, University of Erlangen, Germany.

[12] B. A. Bekins, E. Warren and E. M. Godsy (1998), *A comparison of zero-order, first-order, and Monod biotransformation models*, Ground Water 36, 261-268.

[13] M. Alexander and K. M. Scow (1989), *Kinetics of biodegradation in soil*, in: B. L. Sawhney and K. Brown: Reactions and Movement of Organic Chemicals in Soils, Soil Science Society of America, Madison, 243-269.

[14] M. Schirmer, J. W. Molson, E. O. Frind and J. F. Barker (2000), *Biodegradation modelling of a dissolved gasoline plume applying independent laboratory and field parameters*, J. Contam. Hydrol. 46, 339-374.

Modelling of heterogeneous reactions: simultaneous mass transfer and chemical reaction

Paulo A. Quadros, Nuno M.C. Oliveira and Cristina M.S.G. Baptista

Gepsi-PSE Group, Department of Chemical Engineering, University of Coimbra,

Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
Tel.: +351-239-798700. E-mail: {quadros3, nuno, cristina}@eq.uc.pt

Abstract

This work considers the solution of various fundamental models for simultaneous mass transfer and chemical reaction in heterogeneous fluid media. Several operational regimes are possible for these systems, depending on their relative Hatta numbers. When mass transfer phenomena assume a similar importance to chemical reaction, the predictions from these models can differ significantly. However, this type of regime corresponds to the conditions of specific industrial interest for some heterogeneous reactions. An accurate solution of the combined mass transfer and reaction models is therefore important to understand the behaviour of these processes.

The nitration of liquid benzene with nitric acid, using sulphuric acid as a catalyst, is used to illustrate the application of these models.

1. Introduction

Fluid phase heterogeneous reactions play an important role in the chemical industry. In these systems the chemical compounds are located in distinct physical phases and mass transfer occurs between them by diffusion and/or convection, simultaneously with chemical reaction. A modelling approach often used reduces the complexity of these problems by identifying the rate determining steps and using the corresponding asymptotic solutions (Doraiswamy and Sharma, 1984). However, in some cases, the different phenomena that take place cannot be considered independently and an accurate description of the system behaviour requires more complex mathematical models and sophisticated numerical solution techniques.

An example of complex interaction between reaction and mass transfer occurs in the nitration of liquid benzene with nitric acid, where sulphuric acid is used as a catalyst. The reaction takes place in the aqueous phase and involves the mass transfer of the organic compounds into and from the reacting phase. In the more interesting conditions for industrial operation, reaction and mass transfer assume similar importance. This leads to complex dependence of conversion and amount of secondary products formed on the input flow rates, ratio of organic/aqueous reactants, temperature and degree of mixing in the reactors. In this work a continuous stirred pilot plant reactor was used to conduct benzene nitrations under industrial operating conditions and the experimental data will be used to validate the mathematical models developed.

2. Modelling the benzene nitration process

This liquid-liquid reaction involves an organic phase dispersed in an aqueous one. The organic reactant, benzene (B), is transferred into the aqueous phase where it reacts with the nitronium ion, formed from the nitric acid (N), in the presence of sulphuric acid (S) acting as catalyst (Hughes, *et al.*, 1950 and Olah, *et al.,* 1989). The product, mononitrobenzene (MNB), is then transferred to the organic phase. The strength of the sulphuric acid used is extremely important to define the operating mode. This may range from a purely kinetic regime, limited by the reaction rate, to a fast reaction regime, controlled by the mass transfer between the two liquid phases (Cox and Strachan, 1972).

The modelling procedure for the process involves the simultaneous mass transfer and chemical reaction steps which depend on important parameters related to the assumed mechanisms.

2.1 – Mass transfer with chemical reaction

Several studies involving gas-liquid and liquid-liquid reactions accompanied by chemical reactions have been undertaken and important contributions are reported in Danckwerts (1970), Doraiswamy and Sharma (1984) and Westertertp, *et al.* (1990).

To quantify the reaction regime the Hatta number *(Ha)* or reaction-diffusion modulus is used. For a first or pseudo-first order reaction *Ha* can be calculated by (Westertertp, *et al.,* 1990)

$$Ha = \frac{\sqrt{kD}}{k_L}, \tag{1}$$

where k is the first or pseudo-first order reaction rate constant, D is the diffusion coefficient of the specie and k_L is the mass transfer coefficient. When *Ha* is less than 0.3, the process is controlled by the reaction rate, corresponding to the kinetic regime. If *Ha* is greater than 2, the reaction is very fast, occurring predominantly near the liquid-liquid interface, and the diffusion resistances to mass transfer dominate the global process rate. In the intermediate regime both phenomena prevail and it is not possible to dissociate their influences. In order to characterise the regime in the reactor the Hatta number is one of the first parameters to be calculated. Parameters like rate constant and diffusion and mass transfer coefficients can be difficult to obtain, especially when the heterogeneous reactions are catalysed, as is the case of the aromatic nitrations. Therefore, there is a degree of uncertainty associated to the value of the Hatta number. The studies reported in literature consider well defined regimes: the slow, the fast or instantaneous, and avoid the intermediate regime since it is difficult to work in a region where mass transfer and chemical reaction compete (Zaldivar, *et al.,* 1995 and 1996, Roizard and Wild, 2002).

The diffusion coefficient of aromatic compounds in mixed acid can be obtained by equations 2 and 3 according to Perkins and Geankoplis (1969) and Cox and Strachan (1972) and modified latter by Chapman and Strachan (1976),

$$D = \frac{7{,}4\times10^{-18}\left(\phi M\right)^{0,5} T}{V_b^{0,6}\,\mu_{aq}^{0,8}} \tag{2}$$

$$\phi M = 2{,}6x_W M_W + 2{,}0x_S M_S + 1{,}05x_N M_N \tag{3}$$

where x and M are the mole fraction and the molecular weight, respectively.

The mass transfer coefficient in the continuous phase can be obtained by expression 4 suggested by Calderbank and Moo-Young (1961) and used in recent works on liquid dispersions (van-Woezik and Westerterp, 2000):

$$k_L = 0{,}13\left[\left(\frac{P}{V_c}\right)\left(\frac{\mu_c}{\rho_c^2}\right)\right]^{1/4}\left[\frac{\mu_c}{\rho_c D}\right]^{-2/3} \tag{4}$$

Here P is the power dissipated by the agitator

$$P = P_o \rho_{mixture}\, n^3 D_i^5 \tag{5}$$

and P_o is the power number, which for a two paddle impeller stirrer and the range of Reynolds number used in this work has the value of 0,63 (Azbel and Cheremisinoff, 1983).

3. Experimental results

Several experiments were conducted in the pilot plant described by Quadros and Baptista (2003), and their main operating conditions are summarised in Table 1. These experiments were conducted under realistic industrial operating conditions, with a Hatta number ranging from 0.4 to 1.5, which corresponds to the intermediate (competing) regime. The corresponding output concentrations were measured. These results are presented in Figure 1.

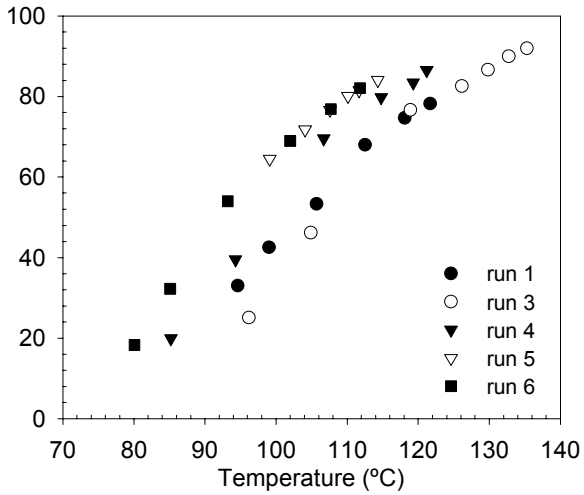


Figure 1 – MNB production as a function of the reaction temperature for runs 1, 3, 4, 5 and 6.

It is desired to correlate these process data, in order to assess the relative importance of the steps involved, and to develop a mechanistic model that can be used to diagnose and optimise the corresponding industrial process. This requires the use of kinetic data for this particular reaction (Quadros *et al*., 2003), combined with a model to describe the mass transfer step.

Table 1 – Operation conditions used in the experimental runs.

Run	F_B/F_N	$T_{Mixed\ acid}$ (ºC)	$T_{Nitration}$ (ºC)	Stirring speed (rpm)	ε	Q_{aq}/Q_{org}	% HNO_3 (w/w)
1	0,96	99,7	94,6 - 121,7	398 - 895	0,132 - 0,169	8,43	5,64
2	0,98	89,5	85,7 - 113,2	395 - 911	0,130 - 0,169	8,38	5,61
3	1,07	102,6	96,2 - 135,3	394 - 1700	0,130 - 0,170	8,11	4,99
4	1,07	88,9	85,2 - 121,2	396 - 1342	0,142 - 0,172	8,09	5,10
5	1,09	81,3	99,1 - 114,3	858 - 1381	0,156 - 0,171	8,09	4,97
6	1,10	84,8	81,1 - 111,4	398 - 885	0,131 - 0,175	7,86	5,29
7	1,15	90,2	86,2 - 117,2	394 - 870	0,132 - 0,180	7,75	5,06

4. Mechanistic models for mass transfer

Among the mechanistic models available to model mass transfer, the film model is probably the simplest one. It considers a stagnant film near the interface between the two phases where the resistance to mass transfer is concentrated, and assumes a steady state transfer process. In contrast, the penetration models of Higbie and Danckwerts make use of non-stationary conditions to describe the diffusional transport process, requiring more complex solution methods.

Each of these three approaches is considered to be a one parameter model, since it depends on a fundamental parameter. They originate the same asymptotic solutions when the reaction rate is fast and the rate determining step is the mass transfer mechanism or, at the other extreme, when mass transfer is fast and the reaction rate is slow. However, when intermediate conditions are used and neither mass transfer nor chemical reaction prevails, the solutions can differ significantly according to the model used (Westerterp *et al.*, 1990).

4.1 – The film model

The usual approach to model heterogeneous reactions uses the film model to quantify the phenomena involved in these reactions. Despite being known as the simplest representation of a very complex phenomenon, recent papers report that its use can lead to reasonable predictions of the corresponding mass transfer rates (Zaldivar, *et al.,* 1995 and 1996, Roizard and Wild, 2002 and D'Angelo, *et al.*¸ 2003). This model considers the mass transfer process as stationary and divides the fluid where the reaction occurs into two different zones: a stagnant film of thickness δ (the fundamental parameter) and a well mixed bulk. Figure 2 represents a simplified scheme of the pilot reactor using the film model, and Figure 3 illustrates the expected concentration profiles across the two phases and interface.

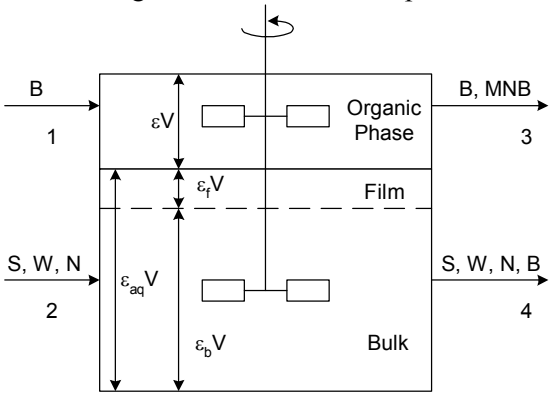


Figure 2 – Schematic representation of the film model in the pilot reactor.

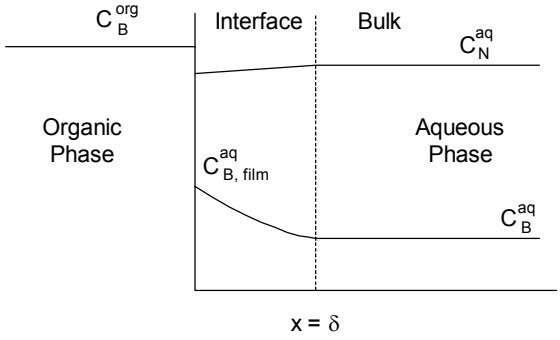


Figure 3– Representation of concentration profiles across the two phases and interface.

As the mass transfer is assumed to proceed via a stationary process the film model leads to a set of algebraic equations that describe the simultaneous mass transfer and reaction steps.

Considering the basic assumptions of the film model, it is possible to express the mass transfer and the chemical reaction at stationary state:

$$D_B \frac{d^2 C_B}{dx^2} - kC_B = 0 \tag{6}$$

Boundary Conditions: $C_B = C_{B,\,film}\,(0)$ for $x = 0$
$C_B = C_{B,\,bulk}$ for $x = \delta$

Additionally, the film thickness δ can be defined by the ratio between the diffusion and the mass transfer coefficients. The solution of this differential equation for a pseudo-first order reaction leads to

$$\frac{C_B}{C_{B,\,film}(0)} = \frac{1}{\sinh(Ha)}\left\{\sinh\left[Ha - x\sqrt{\frac{k}{D_B}}\right] + \frac{C_{B,\,bulk}}{C_{B,\,film}(0)}\sinh\left[x\sqrt{\frac{k}{D_B}}\right]\right\} \qquad \text{for } 0 \le x \le \delta \tag{7}$$

The molar flux of benzene across the interface between the organic phase and the aqueous film can be obtained by differentiation of this equation at $x = 0$. Applying the same strategy for $x = \delta$ gives the molar flux from the stagnant film to the bulk phase.

A mass balance to the organic phase can be written as:

$$0 = Q_1 C_{1B} - Q_3 C_{3B} - k_L\left(C_{B,\,film}(0) - \frac{C_{B,\,bulk}}{\cosh(Ha)}\right)\frac{Ha}{\tanh(Ha)}aV \tag{8}$$

The mass balance to the bulk of the aqueous phase results is in this case:

$$0 = -\frac{\varepsilon_b}{\varepsilon_{aq}}Q_4 C_{B,\,bulk} + \sqrt{Dk}\left(\frac{C_{B,\,film}(0)}{\sin(Ha)} - \frac{C_{B,\,bulk}}{\tanh(Ha)}\right)aV - kC_{B,\,bulk}\varepsilon_b V \tag{9}$$

4.2 – The penetration model – Higbie and Danckwerts models

The penetration models consider that the interface is covered by small stagnant fluid elements that remain there for a specific contact time and are replaced by new fluid elements when they move into the well-mixed bulk. In these models the mass transfer process is assumed as non stationary and, like the film model, they require the use of one fundamental parameter. In the Higbie model this is the specific contact time θ of the fluid element with the interface, which is assumed to be constant for all stagnant fluid elements. The Danckwerts model uses as main parameter the probability (*s)* of replacement of a element of fluid at the interface. At any time, each element, independently of its age, has this probability of being removed from the interface. Both models assume that the mass transfer into or from the stagnant elements takes place by diffusion (Westerterp, *et al.,* 1990) and the combined reaction mass transfer model includes partial and ordinary differential equations, even for steady-state processes.

Figure 4 presents a schematic representation of the concentration profiles in the stagnant elements as a function of contact time with the interface.

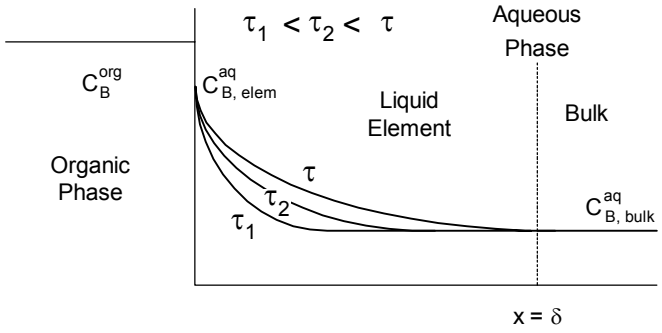


Figure 4 – Concentration profiles in the stagnant element as function of contact time – penetration model.

The Higbie penetration model assumes that after a contact time θ, the liquid element at the interface mixes with the bulk, producing a homogeneous aqueous phase concentration, before being replaced by a new fresh element from the bulk. The following equations describe the mass balances to the organic phase and to the interface and bulk of the aqueous phase:

Mass balance to the benzene in the organic phase:

$$\varepsilon V \frac{dC_{3B}}{dt} = Q_1 C_{1B} - Q_3 C_{3B} - J_B a V \tag{10}$$

Mass balance to the benzene in the liquid element at the interface, for $0 < t < \theta$:

$$\frac{\partial C_{B,elem}}{\partial t} = D_B \frac{\partial^2 C_{B,elem}}{\partial x^2} - k C_{B,elem} \tag{11}$$

Initial Conditions: $C_{B,elem}(0,x) = C_{4B}$

Boundary Conditions: $C_{B,elem}(t,0) = C_B^{i2}$

$C_{B,elem}(t,\delta) = C_{B,bulk}$

Mass balance to the benzene in the bulk of the aqueous phase for the time $0 < t < \theta$:

$$\frac{dC_{B,bulk}}{dt} = -k C_{B,bulk} \tag{12}$$

Initial Conditions: $C_{B,bulk}(0) = C_{4B}$

At the time $t = \theta$, the liquid element mixes with the bulk, and an average concentration of aqueous phase is attained instantly. The corresponding average concentration can be calculated by:

$$\int_0^{\delta} \left(C_{B,elem}(\theta,x) - C_{4B}(\theta) \right) dx.aV + C_{B,bulk}(\theta) \varepsilon_{bulk} V = C_{4B} \varepsilon_{aq} V \tag{13}$$

Results

Figure 5 compares the experimental values for the output MNB concentration with the corresponding predictions, using the film model for the mass transfer with simultaneous chemical reaction. As can be observed, good agreement is obtained between these sets of values, with estimation errors smaller than 15%. This can be considered to be very satisfactory, given the number of estimated parameter involved (mass transfer and diffusion coefficients, rate constants, solubilities, power input and effective interfacial area, among others). Work in progress to compare these experimental results with the predictions of the penetration models, for the same system.

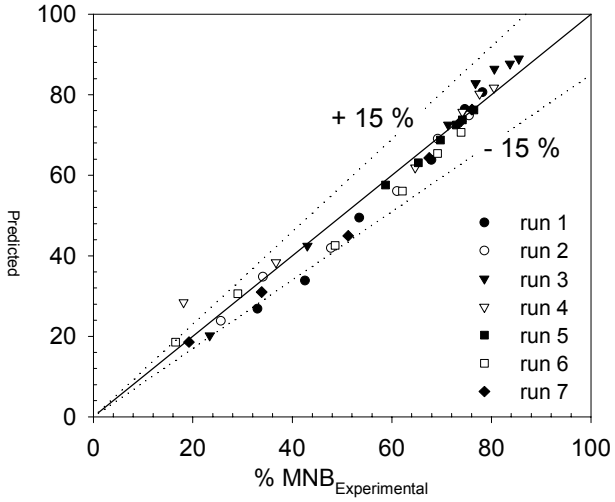


Figure 5 – Experimental versus predicted values of MNB concentrations in the organic phase outlet.

Conclusions

The operating conditions of industrial interest for some heterogeneous reaction systems correspond often to the intermediate regime between kinetic and mass transfer controlling steps. The simulation of these units requires an accurate solution of process models combining mass transfer and kinetic information. Although producing identical asymptotic solutions, the choice of the mass transfer model for this intermediate region can have a significant influence in the prediction capabilities of the combined models.

Acknowledgements

Financial support from Fundação para a Ciência e Tecnologia (FCT), for the Ph.D. grant SFRH/BD/1266/2000 and from Quimigal S.A., Portugal is gratefully acknowledged.

References

Azbel, D. S. and Cheremisinoff, N. P. (1983). Fluid mechanics and unit operations. Ann Arbor Science Publishers. Michigan.

Calderbank, P. H. and Moo-Young, M. B. (1961). The continuous phase heat and mass-transfer properties of dispersions. *Chem. Eng. Sci.*, 16, 39-54.

Chapman, J. W. and Strachan, A. N. (1976). Two phase nitration of clorobenzene in 79,8 % sulphuric acid. *Industrial and laboratorial nitrations,* Alright, L. F. and Hanson, C. (eds.), ACS Symposium series 22, 219-224.

Cox P. R. and Strachan, A. N. (1972). Two-phase nitration of toluene. Part II. *Chem. Eng. J.,* 4, 253-261.

Danckwerts, P. V. (1970). Gas-liquid reactions. McGraw-Hill Book Company.

D'Angelo, F. A., Brunet, L., Cognet, P. and Cabassud, M. (2003). Modelling and constraint optimisation of an aromatic nitration in liquid-liquid medium. Chem. Eng. J., 91, 75-84.

Doraiswamy, L.K. and Sharma, M.M. (1984). Heterogeneous reactions: Analysis, examples and reactor design, Vol. 2, John Wiley & Sons, New York.

Hughes, E. D., Ingold, C. K. and Reed, R. I. (1950). Kinetics and mechanisms of aromatic nitration. Part II. Nitration by the nitronium ion, NO_2^+, derived from nitric acid. *J. Chem. Soc.*, 2400-2415.

Olah, G. A., Malhotra, R. and Narang, S. C. (1989). Nitration – methods and mechanisms. *VCH publishers*, New York.

Perkins, L. R. and Geankoplis, C. J. (1969). Molecular diffusion in a ternary liquid system with the diffusing component dilute. *Chem. Eng. Sci.*, 24, 1035-1042.

Quadros, P. A. and Baptista, C. M. S. G. (2003). Effective interfacial area in agitated liquid-liquid continuous reactors. *Accepted for publication in the Chemical Engineering Science.*

Quadros, P. A., Oliveira, N.M.C. and Baptista, C. M. S. G. (2003). Industrial adiabatic benzene nitration with mixed acid at a pilot plant scale. *Submitted for publication.*

Roizard, C. and Wild, G. (2002). Mass transfer with chemical reaction: the slow reaction regime revisited. Chem. Eng. Sci., 57, 3479-3484.

van-Woezik, B.A.A. & Westerterp, K. R. (2000). Measurement of interfacial areas with the chemical method for a system with alternating dispersed phases. *Chem. Eng. Processing*, 39(4), 299-314.

Westertertp, K. R., Van Swaaij, W. P. M. and Beenackers, A. A. C. M. (1990). Chemical reactor design and operation. John Wiley & Sons.

Zaldivar, J. M., Molga, E., Alós, M. A., Hernández, H. and Westerterp, K. R. (1995). Aromatic nitrations by mixed acid. Slow liquid-liquid reaction regime. Chem. Eng. Processing, 34, 543-559.

Zaldivar, J. M., Molga, E., Alós, M. A., Hernández, H. and Westerterp, K. R. (1996). Aromatic nitrations by mixed acid. Fast liquid-liquid reaction regime. Chem. Eng. Processing, 35(2), 91-105.

Moving Finite Elements Method: Application to Moving Boundary Systems

Rui Robalo[1], Carlos Sereno[1], Alírio Rodrigues[2]

[1] Department of Mathematics, University of Beira Interior
6200 Covilhã, Portugal
[2] Laboratory of Separation and Reaction Engineering, Faculty of Engineering, University of Porto
4200-465 Porto, Portugal

ABSTRACT

In this work the formulation of the Moving Finite Elements Method (MFEM) proposed by Sereno was expanded for two phase systems, with only one space dimension, between fixed boundaries and with an internal moving interface, constituted by a system of parabolic Partial Differential Equations (PDE´s) subject to linear boundary conditions in the external and fixed boundaries of the system and assuming non linear conditions exists at the interface. Our computer code in FORTRAN language resulting from numerical algorithm implementation was tested in the simulations of heterogeneous solid-fluid reaction and of a system of solid $\rightarrow$ fluid phase changes.

1. INTRODUTION

The classical example of a problem with moving boundaries, often called Stefan problem, is the fusion of a solid or the freezing of a liquid. Problems of this type appear in many areas, [4,5]. The numerical simulation of mathematical models of dynamic two-phase systems described by PDE's is a difficult problem particularly when a moving interface is involved and/or the solution develops steep moving fronts. Many suggestions have been made on how to overcome the difficulties to numerically solve this type of problems, [4,5,7]. A possible approach is the MFEM, proposed by the group of K. Miller of the University of Berkeley, [8,10]. Our aim in this work is to apply the formulation of MFEM proposed by Sereno [12] to the simulation of dynamic two-phase systems with moving boundaries.

In the development of the numerical algorithm Sereno [12,14] uses the MFEM with the following characteristics: i) the grid of finite elements associated to each one of the dependent variables is independent from the others ones; ii) each dependent variable is approximated by a Lagrange interpolating polynomial of any degree in each one of the finite elements; iii) the position of interior nodes in each finite element are optimized as in the orthogonal collocation method; iv) the numerical approximation of each one of the dependent variables is smoothed in a neighborhood of the separation nodes through cubic Hermite polynomials.

The implicit time-dependent ODE system resulting from the spatial discretization of the mathematical model is solved by the LSODI package developed at the Lawrence Livermore National Laboratory [6].

2. METHOD DEVELOPMENT

Let us consider the general mathematical model of two-phase system [4] with only one space dimension, between fixed boundaries and with an internal moving interface, constituted by a system of parabolic PDE´s whose *m*-th equation is

$$\frac{\partial y_m}{\partial t} = f_m\left(x,t,\vec{y},\frac{\partial \vec{y}}{\partial x}\right)\frac{\partial^2 y_m}{\partial x^2} + g_m\left(x,t,\vec{y},\frac{\partial \vec{y}}{\partial x}\right), \tag{1}$$

where $\vec{y} = \left[\vec{y}^{\,\mathrm{I}}, \vec{y}^{\,\mathrm{II}}\right]^T = \left[y_1^{\mathrm{I}},...,y_{n_{\mathrm{I}}}^{\mathrm{I}}, y_1^{\mathrm{II}},...,y_{n_{\mathrm{II}}}^{\mathrm{II}}\right]^T = \left[y_1,...,y_n\right]^T$, $y_m = y_m(x,t)$, $t \geq 0$, under the initial condition $y_m(x,0) = q_m(x)$. The space variable satisfy

$$a \leq x \leq X_s(t) \text{ if } m \leq n_{\mathrm{I}} \text{ and } X_s(t) \leq x \leq b \text{ if } n_{\mathrm{I}} < m \leq n, \tag{2}$$

and $X_s(t)$ is interface position at the instant *t*. The equation (1) is subject to linear boundary condition in the external and fixed boundary. Assuming non linear condition exists at the interface, the movement is defined by

$$\frac{dX_s}{dt} = w\left(X_s,t,\vec{y}\big|_{X_s},\frac{\partial \vec{y}}{\partial x}\bigg|_{X_s},\frac{\partial \vec{y}^{\,\mathrm{I}}}{\partial x}\bigg|_{X_s^-},\frac{\partial \vec{y}^{\,\mathrm{II}}}{\partial x}\bigg|_{X_s^+}\right). \tag{3}$$

This class of systems of PDE´s is an extension of those considered by Sereno [12, 14] where all differential equations have the same fixed spatial domain.

The MFEM is a discretization process in two stages: first spatial discretization using finite elements, in which we focus our attention, and secondly the time integration of the resulting ODE´s system.

2.1 SPATIAL DISCRETIZATION

The grid connected to the *m*-th PDE is obtained by partitioning the spatial domain of y_m in $N_m + 1$ finite elements by $N_m - 1$ interior separation nodes,

$$\mathsf{P}_m: \quad X_{m,1}(t) < X_{m,2}(t) < ... < X_{m,N_m}(t) < X_{m,N_m+1}(t), \quad t \geq 0. \tag{4}$$

One of the nodes $X_{m,1}(t)$, $X_{m,N_m+1}(t)$ is independent of time variable and the other is defined by the interface position along time. In the *j*-th finite element of grid P_m

$$I_{m,j} = \left\{x \in \Re : X_{m,j}(t) \leq x \leq X_{m,j+1}(t)\right\}, \; t \geq 0, \tag{5}$$

we approximate y_m by a $p_{m,j}-1$ polynomial obtained by Lagrange interpolation through $p_{m,j}$ interpolation points whose relative positions are optimized as in the orthogonal collocation method, [12,13]. Locally, in $I_{m,j}$, we define the polynomial approximation $Y_{m,j}(x,t)$ as

$$Y_{m,j}(x,t)=\sum_{i=1}^{p_{m,j}} Y_{m,j}^{i}(t)\,l_{m,j}^{i}(x) \tag{6}$$

where $l_{m,j}^{i}(x)$ is the *i*-th Lagrange basis function, $Y_{m,j}^{i}(t)=Y_{m,j}\left(R_{m,j}^{i},t\right)$ and $R_{m,j}^{i}$ is the *i*-th interpolation point. The first and second spatial derivatives of $Y_{m,j}(x,t)$ are also polynomials and can be defined using the same interpolation points.

Globally the approximation Y_m to y_m in the spatial domain is the continuous piecewise polynomial function defined by

$$Y_m(x,t)=\sum_{k=1}^{\tilde{N}_m} Y_m(\xi_{m,k})\,\Phi_{m,k}(x) \tag{7}$$

where $\Xi_m=\left\{\xi_{m,1},\xi_{m,2},\ldots,\xi_{m,\tilde{N}_m}\right\}$ is the ordered set of all nodes, spatial nodes and interior interpolation points, associated to P_m, $Y_m(\xi_{m,k})=Y_{m,j}^{i}(t)$, for *i* and *j* such that $\xi_{m,k}=R_{m,j}^{i}$ and $\Phi_{m,k}$ is the *k*-th global interpolation basis function defined as

$$\Phi_{m,k}(\xi_{m,l})=\begin{cases}1 & \text{if } l=k\\ 0 & \text{if } l\neq k\end{cases}. \tag{8}$$

This approximation (7), at each instant, depends on the nodal amplitudes $Y_{m,j}^{i}$ and on the nodal position $X_{m,j+1}$,

$$\begin{aligned} &Y_{m,j}^{i}, \qquad j=1,2,\ldots,N_m \quad \text{and} \quad i=1,\ldots,p_{m,j}-\tau_j\\ &X_{m,j+1}, \qquad j=1,2,\ldots,N_m-1 \end{aligned} \tag{8}$$

where τ_j is zero if $j=N_m$ and is one otherwise. Therefore, the interface position X_s is not included in list of effective dependence parameters of the method.

2.2 ODE SYSTEM

To determine the semi-discrete variables $Y_{m,j}^{i}$ and $X_{m,j+1}$ we must integrate, in time, the system of ODE´s generated by minimization of the following objective function

$$F=\sum_{m=1}^{n}\left[\left\|R_m\right\|_{L_2}^{2}+\sum_{j=1}^{N_m}\left(\varepsilon_{m,j}\frac{d}{dt}\left(X_{m,j+1}-X_{m,j}\right)-S_{m,j}\right)^{2}\right], \tag{9}$$

with respect to time derivatives of nodal amplitudes and nodes positions, where $\mathbf{R}_m$ is the PDE residual associated with the *m*-th equation

$$\mathbf{R}_m = \frac{\partial Y_m}{\partial t} - f_m\left(x,t,\vec{\boldsymbol{Y}},\frac{\partial \vec{\boldsymbol{Y}}}{\partial x}\right)\frac{\partial^2 Y_m}{\partial x^2} - g_m\left(x,t,\vec{\boldsymbol{Y}},\frac{\partial \vec{\boldsymbol{Y}}}{\partial x}\right), \tag{10}$$

$\varepsilon_{m,j}$ and $S_{m,j}$ are the internodal viscosity and spring penalty functions, respectively, as defined in [10]. Observe that we introduce Miller's penalty functions into the minimization process to avoid the singularities associated with the method, due to parallelism and element folding. The penalty functions depend on of positive constants supplied by the user for each element of each grid and we use the previous work of Sereno [12] to choose these constants.

2.3 INTEGRALS CALCULATION

To get the explicit form of general equations of method is necessary, first, define the partial derivatives of Y_m in order to time variable and in order to effective parameters listed in (8) and, secondly, calculate the integrals that results from minimization of $\mathbf{F}$, in which we focus our attention. One problem to solve is that to define the approximations of spatial derivatives of Y_m (a continuous piecewise polynomial). For this aim we used a smoothing process based on cubic Hermite polynomials in a neighbourhood of the separation nodes

$$\mathsf{V}_{\delta_j}\left(X_{m,j+1}\right) = \left\{x:\, X_{m,j+1} - \frac{\delta_j}{2} \le x \le X_{m,j+1} + \frac{\delta_j}{2}\right\},\ j = 1,\ldots,N_m - 1, \tag{11}$$

where $\delta_j > 0$. By this process, all the integrals are well defined as a limit when $\delta_j \to 0$ replacing in integrate function of those Y_m by the smooth numerical approximation $\widetilde{Y}_m$ defined as

$$\widetilde{Y}_m(x) = \begin{cases} Y_m(x), & \text{se } x \notin \mathsf{V}_{\delta_j}\left(X_{m,j+1}\right) \\ H_j(x), & \text{se } x \in \mathsf{V}_{\delta_j}\left(X_{m,j+1}\right) \end{cases}, \tag{12}$$

where H_j is the cubic Hermite polynomial satisfying

$$H_j\left(X_{m,j+1} \mp \frac{\delta_j}{2}\right) = Y_m\left(X_{m,j+1} \mp \frac{\delta_j}{2}\right);\ H_j'\left(X_{m,j+1} \mp \frac{\delta_j}{2}\right) = \frac{\partial Y_m}{\partial x}\left(X_{m,j+1} \mp \frac{\delta_j}{2}\right). \tag{13}$$

We divide the common spatial domain in a unique fine partition constituted by all separation nodes of the first n_{I} grids and after repeat the process for the last ones. After that we use numerical Radau or Lobatto quadratures to compute the integrals of the smooth approximation.

To solve the implicit time-dependent ODE system resulting from the spatial discretization by finite elements of the mathematical model, we use the package LSODI

developed at the Lawrence Livermore National Laboratory [6] and all routines JACOBI, DFOPR, RADAU and INTRP from [15].

3. NUMERICAL EXAMPLES

We present two numerical examples to demonstrate the working and performance of our MFEM. All the numerical results presented are obtained on a Power Mac 8600 at 200 MHz. The minimum permissible cell width is 10^{-5}. The ODE solver tolerances, for nodal amplitudes and for nodal position are $TOL_1 = 10^{-3}$ and $TOL_2 = 10^{-5}$, respectively. We use Lobatto quadrature with 3 interior quadrature points to compute the integrals appearing in each one of the equations of the ODE systems.

3.1 HETEROGENEOUS SOLID-FLUID REACTION

The computer code resulting from numerical algorithm implementation was tested initially in the simulation of heterogeneous solid-fluid reaction

$$\mathrm{C}_{(\mathrm{s})} + \vartheta \mathrm{A}_{(\mathrm{f})} \rightarrow \text{reaction products.} \tag{14}$$

In this model reactant A diffuses through the porous layer of reaction products, reacts at interface of the solid unreacted core C, producing porous reaction products. Assuming that the reaction is isothermal and instantaneous, which implies a zero concentration of reactant A in the solid-reaction products interface, and a plan geometry system where phase I is constituted by products of reaction and phase II by solid C, the shrinking core model is described in dimensionless form by

$$\frac{\partial C}{\partial \theta} = \frac{1}{\gamma}\frac{\partial^2 C}{\partial x^2}, \quad 0 \le x \le X_s(\theta), \quad \theta \ge 0 \tag{15}$$

with boundary conditions

$$\begin{aligned} &\frac{\partial C}{\partial x} = Bi_m (C-1), \quad x = 0, \quad \theta \ge 0 \\ &C = 0, \quad x = X_s(\theta), \quad \theta \ge 0 \end{aligned} \tag{16}$$

and initial conditions $C(x,0) = 0, \quad 0 \le x \le X_s(\theta)$ and $X_s(0) = X_{s0}$. The solid-reaction products interface movement is defined by

$$\frac{dX_s}{d\theta} = -\frac{\partial C}{\partial x}, \quad x = X_s(\theta), \quad \theta \ge 0\,. \tag{17}$$

When $Bi_m \rightarrow \infty$ and $X_{s0} \rightarrow 0$, (15)-(17) tends to the model described in [5] which have analytic solution

$$C(x,\theta) = 1 - \frac{erf(\eta)}{erf(\eta_s)}, \quad \eta = \sqrt{\frac{\gamma}{4\theta}}\,x\,, \tag{18}$$

where $erf(\eta) = (2/\sqrt{\pi})\int_0^{\eta} e^{-\xi^2} d\xi$ and η_s, which is related with interface position at instant θ by $\eta = \eta_s \Rightarrow x = X_s(\theta)$, is the solution of $1 = (\sqrt{\pi}/\gamma)\eta_s e^{(\eta_s)^2} erf(\eta_s)$. We compute the analytic solution and determine the new interface position at each instant for which we obtain numerical results using Lobatto quadrature with a multiple of 10 interior quadrature points to estimate the integral appearing in *erf* function and the Newton-Raphson method to solved the last equation, with relative errors bound by 10^{-8}.

We computed the solution for $\gamma = 2$ and $Bi_m = 1000$, with 4 finite elements and a polynomial approximation of degree 5 in each element, on a time interval from $\theta = 0$ to $\theta = \theta_M$ such that $X_s(\theta_M)$ is closeness but lower than 1. Nodes are initially concentrated near the left fixed boundary of the system. Figure 1 presents the concentration profiles in phase I for various values of θ, figure 2 shows the evolution of interface and figure 3 presents the trajectories of separation nodes. It was observed that the numerical solution is in agreement with the analytical solution of the model.

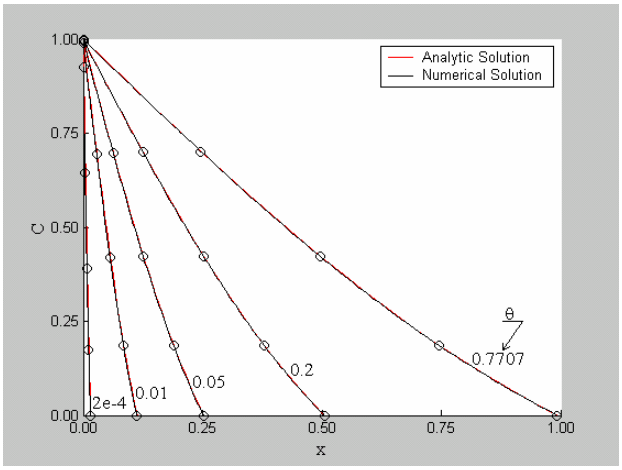


Figure 1: Concentration profiles in phase I for various values of θ

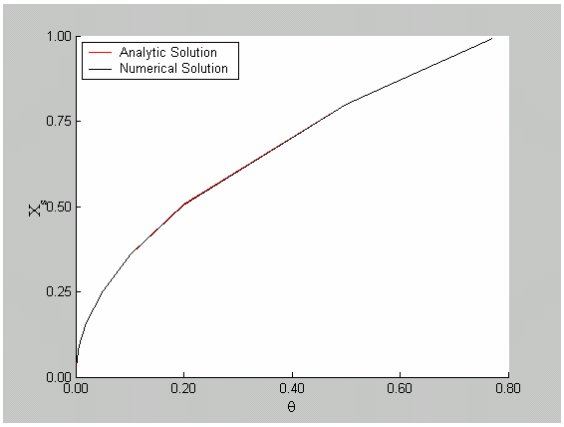


Figure 2: Evolution of interface

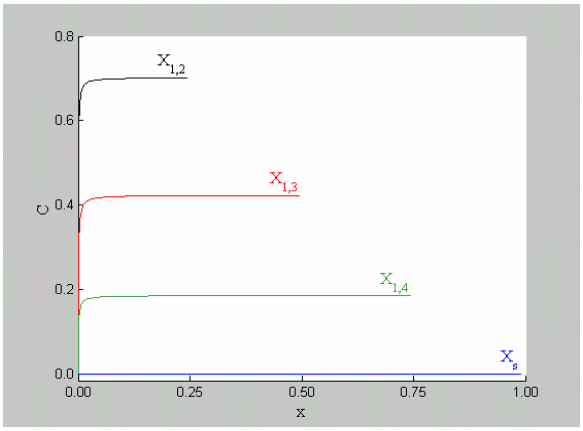


Figure 3: Trajectories of separation nodes

3.2 SOLID $\rightarrow$ FLUID PHASE CHANGES

The second problem that we studied is a model of a system of solid $\rightarrow$ fluid phase changes, described in [4]. Initially the system is at fusion temperature having heat flow in two phases and we change the temperature of environment adjacent to fluid and to solid phases to values great and lower than fusion temperature, respectively. Assuming that system has plan geometry, the phase I is constituted by fluid and phase II is the solid phase, the equations that defined this model are, in dimensionless form,

$$\frac{\partial U_1}{\partial \theta} = \frac{\partial^2 U_1}{\partial \zeta^2}, \quad 0 \le \zeta \le \zeta_s(\theta), \qquad \frac{\partial U_2}{\partial \theta} = \frac{\alpha_2}{\alpha_1}\frac{\partial^2 U_2}{\partial \zeta^2}, \quad \zeta_s(\theta) \le \zeta \le 1 \tag{19}$$

where interface movement is defined by

$$\frac{d\zeta_s}{d\theta} = \gamma\left(\kappa\frac{\partial U_2}{\partial \zeta} - \frac{\partial U_1}{\partial \zeta}\right) \tag{20}$$

with boundary conditions

$$\begin{aligned} &\frac{\partial U_1}{\partial \zeta} = Bi_{h1}(U_1 - U_{1b}), \quad \zeta = 0 \text{ e } \theta \ge 0 \\ &U_1 = U_2 = 0, \quad \zeta = \zeta_s(\theta) \text{ e } \theta \ge 0 \\ &\frac{\partial U_2}{\partial \zeta} = -Bi_{h2}(U_2 - U_{2b}), \quad \zeta = 1 \text{ e } \theta \ge 0 \end{aligned} \tag{20}$$

and initial conditions $U_1(\zeta,0) = U_2(\zeta,0) = 0$ and $\zeta_s(0) = 2\times10^{-4}$. We used the following values of model parameters: $\alpha_1 = 1.42\times10^{-3}$, $\alpha_2 = 1.15854\times10^{-2}$, $\gamma = 1.25433\ (^{o}\mathrm{C})^{-1}$, $\kappa = 4.01085$, $Bi_1 = Bi_2 = 1000$ and $U_{b1} = -U_{b2} = 10\ ^{o}\mathrm{C}$.

The numerical solution was computed using 4 finite elements in each system of finite elements and a polynomial approximation of degree 5 in each element, on a time interval from $\theta = 0$ to $\theta = \theta_u$ such that $|d\zeta_s/d\theta|$ is minimum. Nodes are initially concentrated near the fixed boundaries of the systems.

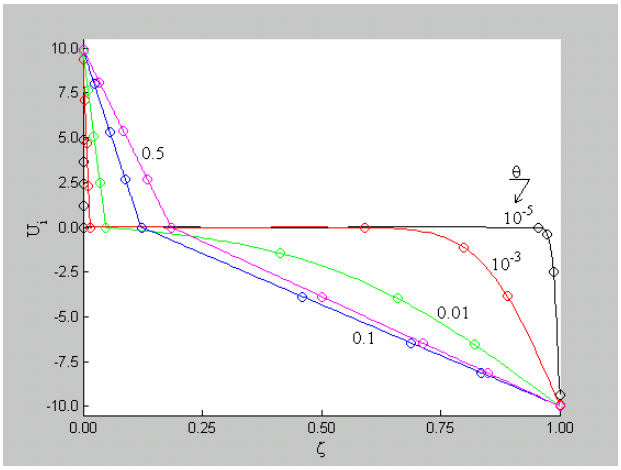


Figure 4: Temperature profiles in both phases for different values of θ

Figure 4 presents the temperature profiles in booth phases for different values of θ, figure 5 shows the evolution of interface and figure 6 presents the trajectories of separation nodes and the interface.

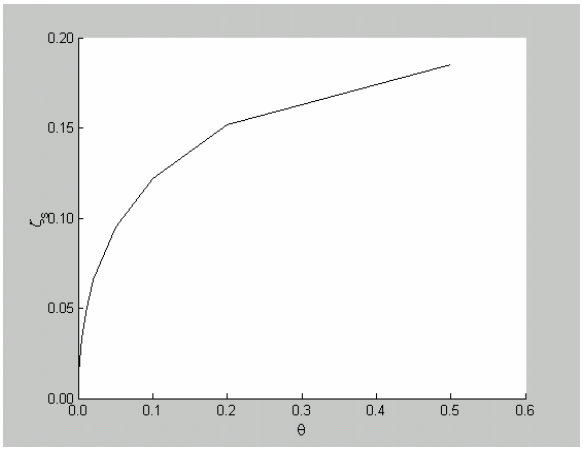


Fig 5: Evolution of interface

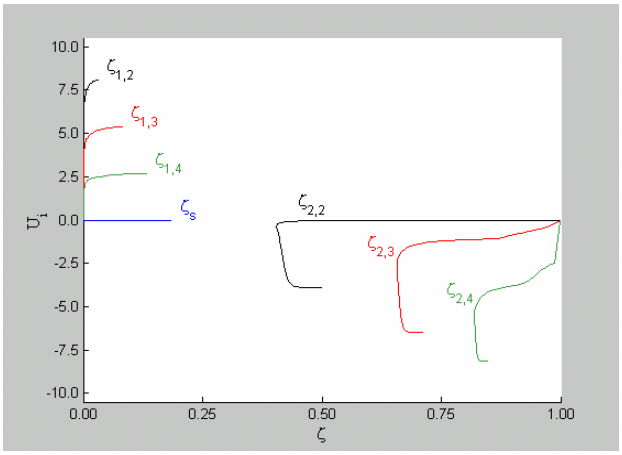


Fig. 6: Trajectories of separation nodes and interface

5. CONCLUSIONS

In this work the formulation of the MFEM proposed by Sereno was expanded for two phase systems, considering the interface just one more moving node. The computer code resulting from numerical algorithm implementation was tested in the simulations of two phase systems It was observed that the MFEM has capacity to produce quite good solutions. In particular, the results obtained in the instantaneous reaction are in agreement with the analytical solution of the present model.

6. REFERENCES

1 Baines, M. J., "*Moving* Finite *Elements*", Clarendon Press, Oxford, 1994.

2 Coimbra, M. C., Sereno, C. A. e Rodrigues, A., "Modelling multicomponent adsorption process by a moving finite elements method", J. Comput. Appl. Math., **115**, 169-179, 2000

3 Coimbra, M. C., Sereno, C. A. e Rodrigues, A., "Applications of a Moving Finite Element Method", Chemical Engng. J., **84**, 23-29, (2001)

4 Crank J., "Free and Moving Boundary Problems", Clarendon Press, Oxford, 1996.

5 Finlayson, B. A., "Numerical Methods for Problems with Moving Fronts", Ravenna Park Publishing, Inc., 1992.

6 Hindmarsh, A. C., "LSODE and LSODI, Two New Initial Value Ordinary Differencial Equation Solvers", *ACM-SIGNUM Newsletter*, Vol. **15**, No. 4, pp. 10-11, 1980.

7 Lapidus, L. and Pinder, G. F., "Numerical Solution of Partial Differential Equations in Science and Engineering", Wiley-Interscience, New York, 1982.

8 Miller, K., "Moving Finite Elements. II", *SIAM J. Numer. Anal.*, Vol. **18**, No. 6, pp. 1033-1057, 1981.

9 Miller, K., "A Geometrical-Mechanical Interpretation of Gradient-Weighted Moving Finite Elements", *SIAM J. Numer. Anal.*, Vol. **34**, No. 1, pp. 67-90, 1997.

10 Miller, K. and Miller, R. N., "Moving Finite Elements. I", *SIAM J. Numer. Anal.*, Vol. **18**, No. 6, pp. 1019-1032, 1981.

11 Rubinstein, L. I., "The Stefan Problem", Translations of Mathematical Monographs, Vol. **27**, American Mathematical Society, 1971.

12 Sereno, C. A., "Método dos Elementos Finitos Móveis: Aplicações em Engenharia Química", Ph. D. Thesis, University of Porto, 1989.

13 Sereno, C. A., Rodrigues, A. e Villadsen, J., "The Moving Finite Element Method with Polynomial Approximation of Any Degree", *Computers Chem. Engng.*, Vol. **15**, No. 1, pg. 25-33, 1991.

14 Sereno, C. A., Rodrigues, A. e Villadsen, J., "Solution of Partial Differential Equations Systems by the Moving Finite Element Method", *Computers Chem. Engng.*, Vol. **16**, No. 6, pg. 583-592, 1992.

15 Villadsen, J. and Michelsen, M. L., "Solution of Differencial Equations Models by Polynomial Approximation", Prentice-Hall, Englewood Cliffs, N. J., 1978.

Model-based Optimization of Discontinuous Chemical Polymerization Systems

Dulce C.M. Silva and Nuno M.C. Oliveira
GEPSI — PSE Group, Department of Chemical Engineering, University of Coimbra
Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
Tel. +351-239-798700. E-mail: {dulce,nuno}@eq.uc.pt

Abstract

This work describes the development of mechanistic models for the suspension polymerization of vinyl chloride (VCM), and their subsequent use for optimization and control of an industrial batch reactor. The methodology used for dynamic optimization of this system uses a feasible path approach with roots on nonlinear Model Predictive Control (MPC) theory. The approach is sufficiently flexible to accommodate general objectives and common constraints. This allows a tighter integration between the control and optimization layers, while making these problems addressable by software packages that start to be commercially available. The results obtained with both models clearly illustrate the advantages and possible improvements in the operation of a typical discontinuous processes.

1 Introduction

Pressure to reduce costs and improve competitiveness in the process industries has led to renewed interest in the development of rigorous process models. These models are frequently based on first principles, and include a detailed description of the various physico-chemical phenomena that take place in the system. When coupled with modern solution and optimization algorithms they constitute valuable tools to diagnose abnormal behavior, improve product quality, and minimize the environmental impact while simultaneously improving the productivity and safety aspects.

Within the chemical industries, polymerization systems, often operated in discontinuous (batch) mode, present interesting challenges in their control and online optimization, due to their non-stationary nature and highly nonlinear behavior. In many cases, these operations are still carried according to recipes based on heuristics and past experience. The use of detailed mechanistic models, experimentally validated, to understand and systematically optimize these processes can therefore have a clear and significant impact in their operation.

This work describes the development of mechanistic models for the suspension polymerization of vinyl chloride (VCM), and their subsequent use for optimization and control of an industrial batch reactor. This is an heterogeneous system, involving four distinct phases (monomer, polymer, aqueous and gas phases). Based on the kinetic information of Xie et al. (1991a) and Kiparissides et al. (1997), two mechanistic models were built, allowing a comparison of the optimization results and their sensitivity to be established. The methodology used for dynamic optimization of this system uses a feasible path (sequential) approach with roots on nonlinear Model Predictive Control (MPC) theory; the algorithms described can be applied either *off*- or *on*-line. This provides a well integrated methodology for optimal supervision of batch polymerization processes.

2 Problem Formulation and Solution Strategy

A dynamic optimization strategy is used to solve problems formulated in a general manner as

$$\begin{aligned} \min_{u(t)\in\mathcal{H}_{ik}} \quad & \Psi(t,u(t),x(t)) = G(x(t_{\mathrm{F}})) + \int_{t_0}^{t_{\mathrm{F}}} F(x(t),u(t),t)d\tau \\ \text{s.t.} \quad & \dot{x} = f_p(x,u,d;\theta) \\ & y = g_p(x;\theta) \\ & u_l \le u \le u_u \\ & x_l \le x \le x_u \\ & y_l \le y \le y_u \end{aligned} \tag{1}$$

where f_p and g_p are usually assumed differentiable and continuous, except perhaps at a finite number of switching points. The state variables are denoted by $x \in \mathbf{R}^{n_s}$, $u \in \mathbf{R}^{n_i}$ are the input variables and $y \in \mathbf{R}^{n_o}$ is the output vector. G and F are assumed to be general twice differentiable nonlinear functions. This formulation is sufficiently flexible to accommodate different objectives and constraints of various nature such as:

- Direct minimization of the operation time:
$$\Psi(\bullet) = t_{\mathrm{F}}. \tag{2}$$

- Treatment of soft constraints, especially related to final product properties, that can be formulated as
$$\Psi(\bullet) = (y_{\mathrm{F}} - y_{\mathrm{sp}})^{\mathrm{T}} Q (y_{\mathrm{F}} - y_{\mathrm{sp}}), \tag{3}$$
where y_{sp} represents the desired final values, y_{F} is the value of a set of output variables at the end of the run, and Q is a weighting matrix. Polymers with improved final properties can also be sought, by direct minimization or imposing restrictions on the variance of the chain length distribution,
$$\Psi(\bullet) = \int_0^{\infty} f(r)(r-\bar{r})^2\, dr$$
where r is the chain length and $f(r)$ is the polymer weight fraction with chain length r.

- Objectives related to tracking an arbitrary trajectory, for a set of properties expressed in terms of the input, state and output variables, similarly to the nonlinear MPC strategy (Oliveira and Biegler, 1995),
$$\Psi(\bullet) = \int_{t_k}^{t_k+t_{oh}} (y-y_{\mathrm{sp}})^{\mathrm{T}} Q_y(t)(y-y_{\mathrm{sp}}) + (u-u_r)^{\mathrm{T}} Q_u(t)(u-u_r)\, d\tau. \tag{4}$$

- Optimal initial conditions for the operation (e.g., amounts and composition of initiators) can be determined.

As mentioned, a feasible path approach is used to solve the dynamic optimization problem (1). The problem is first discretized using stepwise constant input profiles. To simplify the notation, augmented vectors U, X and Y are defined, containing all values of the corresponding variables inside an operating horizon. An exact linearization of the model around a nominal trajectory can be written as

$$\hat{Y} \approx \bar{Y} + \left.\frac{\partial Y}{\partial U}\right|_{U=\bar{U}} \Delta U = \bar{Y} + \mathcal{S}_m \Delta U,$$

where $\mathcal{S}_m$ represents the *dynamic matrix* of the model, containing the first order information for the system relative to the input variables. This matrix can be efficiently computed from the original differential model through the use of appropriate sensitivity equations.

When the objective has the form of (4) or (3) the algorithms described in Oliveira and Biegler (1995); Santos et al. (1995) can be directly used. However, some modifications are required in this formulation to treat minimum-time problems. In these problems, the final time is usually defined by a certain output variable, which reaches a predefined value y_F at the end of the operation. We assume that this happens during the nth discretization interval, from t_n to t_{n+1}, inside a larger horizon defined as a maximum bound on t_F. Given the previous assumptions about the model, it is possible to write t_F as an implicit function of the initial condition and input variables during this interval

$$t_F = h(x_n, u_n). \tag{5}$$

The first order information for t_F can then be obtained by writing a Taylor series in this interval:

$$t_F = \bar{t_F} + \left.\frac{\partial t_F}{\partial x_n}\right|_{x=\bar{x}} \cdot (x_n - \bar{x}_n) + \left.\frac{\partial t_F}{\partial u_n}\right|_{u=\bar{u}} \cdot (u_n - \bar{u}_n).$$

The derivatives $\left.\frac{\partial t_F}{\partial x_n}\right|_{x=\bar{x}}$ and $\left.\frac{\partial t_F}{\partial u_n}\right|_{u=\bar{u}}$ are, in some cases, difficult to obtain directly, by integration of the sensitivity coefficients, since (5) is usually not available in explicit form. However, since these coefficients are only needed in the last time interval, they can also be approximated by finite differences, without a great penalty. Applying the previous concepts, the linearization of t_F with respect to the input variables can be written as

$$t_F = \bar{t_F} + S^* \Delta U.$$

This allows formulation of the optimization problem as the successive quadratic programming (SQP) iteration of

$$\begin{aligned} \min_{\Delta U} \; & J_2 = \bar{t_F} + S^* \Delta U + \Delta U^{\mathrm{T}} H \Delta U \\ \text{s.t.} \quad & U_{ld} \le \Delta U \le U_{ud} \\ & Y_{ld} \le S_m \Delta U \le Y_{ud} \\ & S_{m_{n,j}} \Delta U = \Delta y_F, \end{aligned}$$

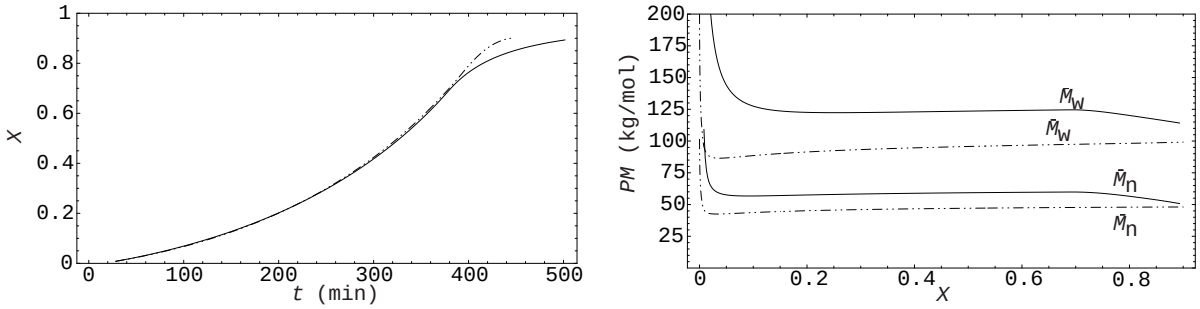
where $\Delta y_F = y_{\mathrm{sp}} - y(t_F)$, and H represents an approximation of the Hessian of the Lagrangian of (1). This formulation is closely similar to the one used in the nonlinear Newton control law, making the algorithms developed for its solution applicable for minimum-time problems as well. A more complete description of the solution strategy can be found in Silva and Oliveira (2002).

3 Suspension Polymerization of vinyl chloride

The suspension polymerization of VCM is a heterogeneous reaction involving four phases: polymer rich phase, monomer rich phase, aqueous and gas phase. Here, the kinetic information provided by Xie et al. (1991a,b) and Kiparissides et al. (1997) was used to build two detailed mechanistic process models, in order to compare the optimization results and their sensitivities. These kinetic models contain all of the important elementary reactions for two-phase polymerization, namely:

- The distribution of monomer by the different phases, as a function of the conversion and the reactor operation conditions.
- The conversion and reaction rate.
- The pressure inside the reactor.
- The characteristics of the polymer formed[1].

[1] Xie's model gives the accumulated molecular weight averages, and the molecular weight distribution. Kiparissides's model predicts the molecular weight averages, the short and long-chain branching, and the number of terminal double bounds.



(a) (b)

Figure 1: (a) Conversion profile for isothermal operation and (b) conversion dependence of number and weight average molecular weight($- \cdot - \cdot -$ Kiparissides's model, $-$ Xie's model).

Predictions from the two models are compared in Figure 1, using a constant polymerization temperature of 55°C, with an equimolar mixture of initiators (A and B). As can be observed, the conversion profiles remain close until a conversion of 70% is reached. Their divergence after this point can be attributed to the fact that in Kiparissides's model the initiation efficiency after the critical conversion is not diffusionally controlled as in Xie's model. In Xie's model the values of the propagation and termination rates at high conversion are lower than the corresponding rates in Kiparissides's model. The profiles of the average molecular weights given by both models can also be observed in Figure 1(b) . For similar operating conditions, the models predict polymers with slightly different properties at the end of the operation. This can be due to the value of the kinetic parameters used in each model, especially the chain transfer to monomer that controls the molecular weight of the polymer.

4 Optimization Results

The main decision variables available for optimization of this system are the reaction temperature and the initiator quantities. These variables can be changed in small steps during the operation and, in the case of the initiator, the amounts of each species to be added at the beginning of the operation can also be independently specified. By optimizing directly the reaction temperature, our results are relevant to polymerization reactors of different sizes. Therefore, this framework can also be helpful in identifying physical limitations of existing process equipment and to confirm retrofit decisions.

In the following two cases, optimal profiles are calculated to originate a polymer with desired properties (polidispersivity and molecular weight averages) in *minimum time*. The desired values for these properties were taken as identical to the polymer obtained using constant temperature profiles of Figure 1. In the last case, an optimal profile is calculated to manufacture products with *improved final properties*, not possible when constant temperature profiles are used.

Case I - Operation with optimal temperature profile

Figure 2 shows the optimal temperature profile that minimizes the batch time, subject to upper and lower bounds of 5°C relatively to the nominal temperature. Xie's model shows smaller deviations relatively to the nominal trajectory, due to higher sensitivity to this variable. Figure 3 compares the conversion and molecular weights obtained with Xie's model, in the optimal and base cases. As can be observed, the molecular weights are essentially identical (due to imposed constraints), although the final conversion is obtained much faster. In fact, in the suspension polymerization of VCM and in the absence of limitations in the cooling capacity, the cycle time can be reduced *between 9% and 23%*, compared with traditional isothermal operation.

The on-line implementation of the temperature profile obtained in the Kiparissides's model is considered, using a nonlinear MPC controller based on a formulation similar to the previous optimization strategy. This is illustrated

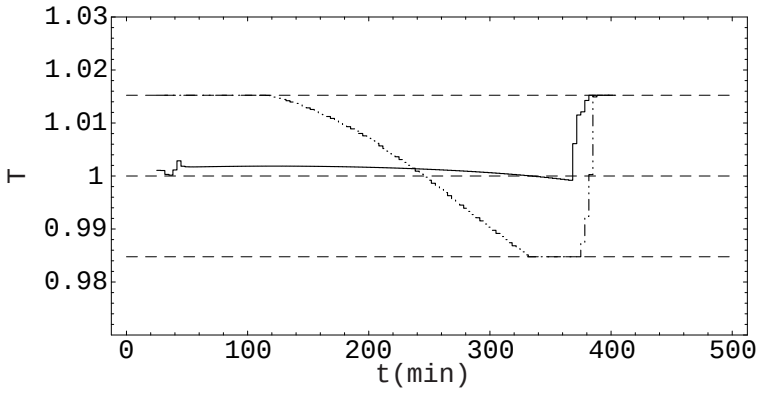


Figure 2: Normalized optimal reactor temperature policy ($- \cdot\cdot -$ Kiparissides's model, $-$ Xie's model).

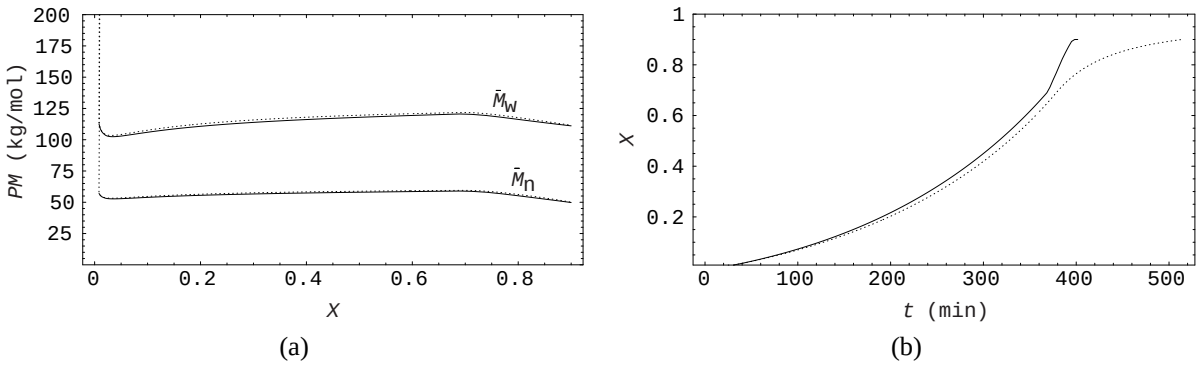


Figure 3: Comparison between the optimal and base profiles using Xie's model: (a) Average molecular weights and (b) conversion profiles ($\cdots$ base case, $-$ optimal).

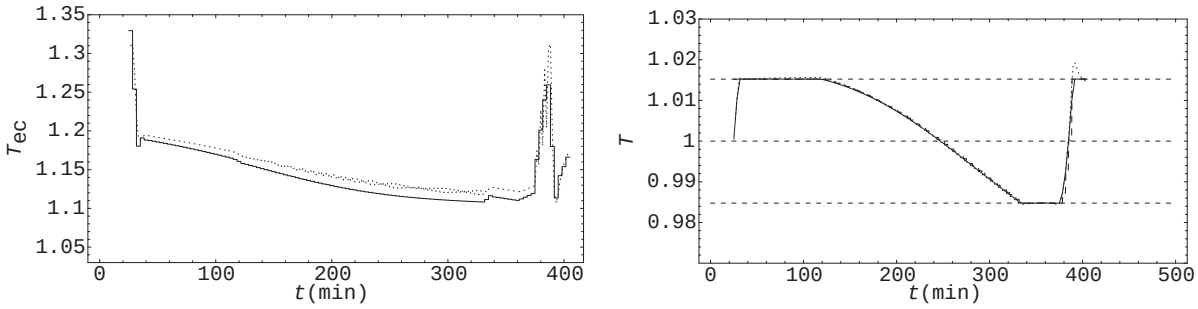
in Figure 4 and these results are also compared to a linear controller PI, representative of the current industrial practice. For the predictive controller the tuning parameters used were $Q_1 = I$, $Q_2 = 10^{-3}I$, with a sampling time of 200 seconds. The discrete PI controller used a sampling time of 100 seconds, $k_c = 20$ and $\tau_I = 1500$ seconds. As can be observed, the MPC controller is able of better tracking the optimal trajectory, while exhibiting smaller amplitude changes in the input profiles.

Case II - Operation with optimal initiator concentrations

In this section, the impact of changes in the initiator concentrations during isothermal operation is studied, considering both their addition at the beginning or in a continuous manner during the entire operation.

Case IIa - Operation with optimal initiator amounts added at the beginning of the operation

In this case, the initiator amounts to be added at the beginning of the operation are optimized. The results obtained in this situation are described in Table 1. As can be observed, both models predict a similar composition of this mixture, slightly different from the nominal case (50/50%). The optimal total amount is also slightly higher than the nominal case; if the increased amount of initiator to be used is considered problematic in terms of residuals trapped inside the polymer particles at the end of the operation, it can also be limited by including a corresponding



(a) (b)

Figure 4: (a) Normalized inlet temperature jacket profile for (b) the optimal profile of the reactor temperature ($\cdots$ PI; $-$ MPC).

Table 1: Optimization of the initial quantities of the initiators.

	Isothermal case	Xie's model	Kiparissides's model
Total amount (mol)	18,0	19,4	22,1
Initiator A (%)	50	53	53
Initiator B (%)	50	47	47
Reduction in the cycle time (%)	–	7,2	14,4

hard constraint in formulation (1). The cycle time reduction obtained with Kiparissides's model is higher, because the initiation efficiency is not considered to be diffusionally controlled for higher conversions, as in Xie's model.

Case IIb - Operation with optimal initiator amounts added during the operation

Figure 5 shows the optimal profiles of initiator amounts to be added during the operation, in order to minimize the operation time. This feed is considered an equimolar mixture of initiators A and B. Due to operation constraints, bounds of $0^{\mathrm{mol}}/_{\mathrm{min}}$ and $6^{\mathrm{mol}}/_{\mathrm{min}}$ were imposed.

As can be observed, the profiles are quite different. In Kiparissides's model the initiator is mainly added in the first stage of reaction, and then the quantity added begins to decrease. In Xie's model, the initiator should be added almost entirely enduring the last stage of operation. The optimal amount of initiators added during the

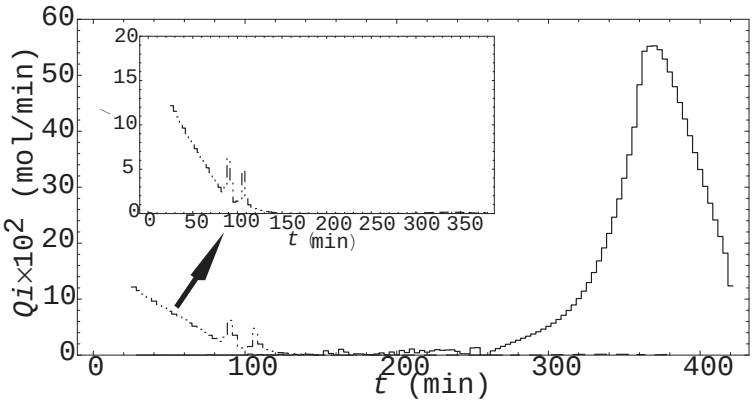


Figure 5: Optimal feed rate of initiators policy ($-\cdot\cdot-$ Kiparissides's model, $-$ Xie's model).

Table 2: Total and residual amounts of the initiators used in Xie's model.

	Isothermal case	Case IIa	Case IIb
Total amount added at the beginning (mol)	18,0	19,4	18,0
Total amount added during the operation (mol)	–	–	36,3
Total residual amount (mol)	8,0	9,8	44,1
Reduction in the cycle time (%)	–	7,2	18,7

Table 3: Total and residual amounts of the initiators used in Kiparissides's model.

	Isothermal case	Case IIa	Case IIb
Total amount added at the beginning (mol)	18,0	22,1	18,0
Total amount added during the operation (mol)	–	–	5,2
Total residual amount (mol)	7,4	10,1	11,0
Reduction in the cycle time (%)	–	14,4	14,5

operation is much higher in case of Xie's model. Xie's model predicts a total reduction of cycle time of 18,7 %. In Kiparissides's model, the reduction is smaller (14,5%) and similar to the one obtained in the case IIa.

Comparison of cases IIa and IIb

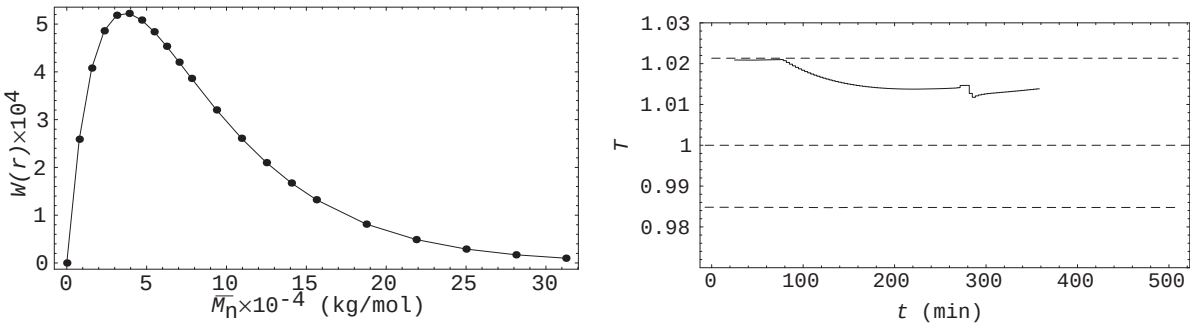
The constraints imposed on the polymer require that the final products in both of these cases have similar properties, in terms of their polidispersivity and molecular weight averages. However the total amounts of initiators used (and residual) can be different; these are described in Tables 2 and 3. These tables show that the residual amounts of initiator increase with the amount of initiator added during the operation. Kiparissides's model predicts that the addition of initiators at the beginning or during the operation has a similar similar effect, in terms of reduction in the operation time possible. Their addition at the beginning leads to smaller residual amounts and its implementation can be considered more practical. In contrast with these findings, Xie's model predicts that higher reductions in the cycle time are possible when the initiators are added continuously during the operation, at the cost of an increased residual amount of initiators in the final product.

Case III - Manufacturing of innovative products

In this section we consider the application of the previous optimization strategy to manufacture innovative products, with improved final properties. An optimal temperature profile is calculated in order to originate a polymer with desired molecular weight distribution, e.g. with a smaller polidispersivity (narrower molecular weight distribution). A temperature constraint of $50°\text{C} < T < 62°\text{C}$ is enforced. Figure 6(b) shows the optimal profile obtained. As can be observed from Figure 6(a) the distribution obtained closely matches the desired one.

5 Conclusions

The kinetic information, available in the literature, for the suspension polymerization of vinyl chloride was incorporated in two detailed mechanistic models. Their prediction was used for optimization and control of an industrial batch reactor. Optimal trajectories for this process, leading to products with specific properties in minimum time was considered. The capability of manufacturing innovative products with innovative products with



(a) (b)

Figure 6: (a) Final PVC polymer molecular weight distribution and (b) the normalized optimal profile of the temperature polymerization (• obtained; – desired).

improved properties was also considered. The results obtained with a batch suspension polymerization system clearly illustrates the advantage and possible improvements in the operation of typical discontinuous processes. This framework can also be helpful in identifying physical limitations of existing process equipment and to confirm retrofit decisions.

An important additional advantage of the application of this strategy to the study of polymerization systems, when combined with experimental tests, is that it can provide an efficient screening methodology for alternative model structures. In the present case, two mechanistic model structures that produce essentially similar results for isothermal polymerization (Figure 1a) show very different sensitivities to the main operating variables such as reaction temperature and amount of initiators used.

Acknowledgments

The authors gratefully acknowledge the financial support provided by FCT under the scholarship PRAXIS/BD/9456/96 and the research project 3/3.1/CEG/2577/95.

References

C. Kiparissides, G. Daskalakis, D.S. Achilias, E. Sidiropoulou (1997). “Dynamic Simulation of Industrial Poly(vinyl chloride) Batch Suspension Polymerization Reactors”, *Ind. Eng. Chem. Res.*, 36(4), 1253–1267.

N.M.C. Oliveira, L.T. Biegler (1995). “An Extension of Newton-type Algorithms for Nonlinear Process Control”, *Automatica*, 31(2), 281–286.

L.O. Santos, N.M.C. Oliveira, L.T. Biegler (1995). “Reliable and Efficient Optimization Strategies for Nonlinear Model Control”, *Proc. Dynamics and Control of Chemical Reactors, Distillation Columns and Batch Processes (DYCORD+’95)*, 33–38, Elsevier Science, England.

D.C.M. Silva, N.M.C. Oliveira (2002). “Optimization And Nonlinear Model Predictive Control Of Batch Polymerization Systems”, *Comp. and Chem. Eng.*, 26, 649–658.

T.Y. Xie, A.E. Hamielec, P.E. Wood, D.R. Woods (1991). “Experimental Investigation of Vinyl Chloride Polymerization at High Conversion: Mechanism, Kinetics, and Modeling”, *Polymer*, 32(3), 537–555.

T.Y. Xie, A.E. Hamielec, P.E. Wood, D.R. Woods (1991). “Experimental Investigation of Vinyl Chloride Polymerization at High Conversion: Molecular Weight Development”, *Polymer*, 32(6), 1098–1111.

Convex Programming Tools for Disjunctive Programs

João Soares,
Departamento de Matemática,
Universidade de Coimbra,
Portugal

Abstract

A Disjunctive Program (DP) is a mathematical program whose feasible region is the convex hull of the union of convex sets. The objective function is also convex. Disjunctive Programming models are very frequent as a modeling framework for setup costs or constraints, and models with constraints that are better expressed as disjunctions. Some Process Sinthesis Design models arising from Chemical Engineering are mixed integer convex programming models which are particular instances of a Disjunctive Program. In this talk we will address questions that are raised when conceptualizing a Branch-and-cut algorithm for mixed-integer convex programming.

1 Introduction

The process synthesis network problem problem in Chemical Engineering is the problem of simultaneously determining the optimal structure and operating parameters for a chemical synthesis problem. This problem can be modeled as a mixed 0-1 convex program where the continuous variables represent process parameters such as flowrates and the 0-1 variables represent the potential existence of a process unit. The nonlinear elements come from the intrinsic nonlinear input-output performance equations of some process units, see [3] where this model is proposed and [8, 4] for related models.

Other models of Network Design in communication and transportation networks are discrete by nature. The 0-1 variables represent the potential existence of multiplexers, concentrators, or interface message processors in computer communication networks, junctions in pipeline networks, interchanges in highway networks, and so on. Discrete variables may also represent the discrete quantity of physical arc units of certain characteristics between two junctions of the network. In a simple model, described in [5], the nonlinear element comes from modelling *delay* at some link (i, j) as proportional to the fraction of the rate of messages crossing the link (i, j) to the available capacity of the same link.

We propose a cutting-plane algorithm for solving the following mathematical program that we will refer to as a mixed zero-one convex program,

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & G(x) \leq 0 \\ & x_i \in \{0,1\}, i = 1, \ldots, p, \end{array} \tag{1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a closed convex function and $G\colon \mathbb{R}^n \to \mathbb{R}^m$ is a vector function of closed convex functions. The variables x_i, for $i = 1, \ldots, p$, are zero-one constrained and the variables x_i, for $i = p+1, \ldots, n$, are simply nonnegative. We will further assume that both f and G are continuous in an open set containing the continuous relaxation, i.e., when $\{0, 1\}$ is replaced by $[0, 1]$.

Our work extends to the nonlinear setting the lift-and-project approach of Balas, Ceria and Cornuéjols [1, 2], which is seen as one of the most important practical contributions to the solution of mixed zero-one linear programs by general-purpose cutting-plane based algorithms since the work of Gomory in the sixties. As proposed by Stubbs and Mehrotra [7], we solve the cut generation problem in its dual form. Some of the distinctive features of our algorithm are the following: our algorithm guarantees the existence of a cut whenever an optimal solution was not yet found; we solve the cut generation problem using standard nonlinear programming algorithms; and, we fully extend the lifting procedure to the nonlinear setting.

The article is structured in the following way. In Section 2 we describe the basic cutting-plane algorithm specialized to solve Program (1). In Section 3 we explain how the cut generation problem can be solved in a smaller-dimension space, taking advantage of the fact that some variables are already integral. In the talk, the algorithm will be illustrated on a small example. A practical implementation of this method is part of an ongoing research project.

2 The basic cutting plane algorithm

Our approach requires that we use the following equivalent formulation of program (1),

$$\begin{array}{ll} \min & x_{n+1} \\ \text{s.t.} & f(x) \leq x_{n+1} \\ & G(x) \leq 0 \\ & x_i \in \{0, 1\}, i = 1, \ldots, p. \end{array} \tag{2}$$

Since f is convex then this formulation is still a mixed zero-one convex program. Moreover, the feasible region K can be replaced by $P = \text{conv}\,(K)$ without loss of generality, where we note that P is closed. As a matter of notation, we will still use the same $f(x)$ and $G(x)$ eventhough we are refering to these functions as functions of the first n components of the vector x that now lies in $\mathbb{R}^{n+1}$.

A specialization of the basic cutting-plane algorithm is presented in Figure 1. The algorithm requires performing three basic steps in each iteration. In the first step, the *relaxation step*, we seek an optimal solution $\bar{x}$ of the following convex program

$$\begin{array}{ll} \min & x_{n+1} \\ \text{s.t.} & x \in \bar{P}, \end{array} \tag{3}$$

whose feasible region $\bar{P}$ is defined by

$$\bar{P} \equiv \left\{ x \in \mathbb{R}^{n+1} \colon \begin{array}{ll} f(x) \leq x_{n+1}, & a^i x \leq b_i, i = 1, \ldots, m_1, \\ G(x) \leq 0, & x_i \in [0, 1], i = 1, \ldots, p, \end{array} \right\}, \tag{4}$$

where m_1 is the number of cuts generated so far. In the second step, the *optimality check step*, we try to reduce as much as possible the number of fractional components of $\bar{x}$ while keeping the same value of the component $\bar{x}_{n+1}$. In the third step, the *separation step*, we

use the last index j tried in the second step to define the following disjunctive programming relaxation $\bar{P}_j$ of P,

$$\bar{P}_j \equiv \text{conv } \left(\left(\bar{P} \cap \{x: x_j = 0\}\right) \cup \left(\bar{P} \cap \{x: x_j = 1\}\right)\right). \tag{5}$$

The following proposition shows that a nonoptimal $\bar{x} \notin \bar{P}_j$, from where we are able to guarantee the existence of a separating hyperplane.

Proposition 1 *In each iteration of the algorithm BCP4MINLP, Step 2 is performed at most p times. Moreover, if j is the last index tried in Step 2 then either $\bar{x}$ is optimal or $\bar{x} \notin \bar{P}_j$.*

Proof: We recall that in Step 2 of the algorithm BCP4MINLP, the integer-constrained variables are sequentially fixed at one of their bounds, zero or one, until an index j is found such that

$$\min_{i=0,1} \left(\begin{array}{ll} \min & x_{n+1} \\ \text{s.t.} & x \in \bar{P}, \\ & x_{F'} = \bar{x}_{F'}, \quad x_j = i \end{array} \right) > \bar{x}_{n+1}, \tag{6}$$

where F' identifies the variables that are fixed in the process. Since F' can have at most p elements then Step 2 is performed at most p times until either (6) holds or all the integer constrained variables are fixed in which case we would have found an optimal solution.

Now, we prove the second part of this proposition. Let j be the last index tried in Step 2 so that (6) holds. Assume, by contradiction, that $\bar{x} \in \bar{P}_j$. Then, the point $\bar{x}$ can be represented by one of the following three possible ways:

a. $\bar{x} = \delta z + (1-\delta)y$, where $\delta \in (0,1)$, $z \in \bar{P} \cap \{x: x_j = 0\}$ and $y \in \bar{P} \cap \{x: x_j = 1\}$;

b. $\bar{x} = z + dy$, where $z \in \bar{P} \cap \{x: x_j = 0\}$ and dy is a direction of the set $\bar{P} \cap \{x: x_j = 1\}$ if this set is nonempty or the zero vector otherwise.

c. $\bar{x} = dz + y$, where $y \in \bar{P} \cap \{x: x_j = 1\}$ and dz is a direction of the set $\bar{P} \cap \{x: x_j = 0\}$ if this set is nonempty or the zero vector otherwise.

If $\bar{x}$ can be decomposed as in a. then, since $\bar{x}_k \in \{0,1\}$, for every $k \in F'$, we must have $z_k = y_k = \bar{x}_k$, for every $k \in F'$. Thus,

$$\bar{x} \in \text{conv } \left(\left(\bar{P} \cap \{x: x_{F'} = \bar{x}_{F'}, x_j = 0\}\right) \cup \left(\bar{P} \cap \{x: x_{F'} = \bar{x}_{F'}, x_j = 1\}\right)\right)$$

which contradicts (6). If $\bar{x}$ can be decomposed as in b. then, since $dy_i = 0$, for every $i \in \{1, \ldots, p\}$, $z_k = \bar{x}_k$, for every $k \in F' \cup \{j\}$. Thus, $\bar{x} \in \bar{P} \cap \{x: x_{F'} = \bar{x}_{F'}, x_j = 0\}$ which contradicts (6) once again. If $\bar{x}$ can be decomposed as in c. an analogous argument as in b. applies. □

3 The cut generation problem

We explain how the cut generation solution procedure should be implemented to take advantage of the fact that many variables have been fixed during the second step of the algorithm BCP4MINLP. Our cut generation problem uses the following duality result

$$\begin{array}{lll} \sup & \alpha\bar{x} - \beta & \\ s.t. & (\alpha, \beta) \in \text{polar } (\bar{P}_j), & = \\ & \|\alpha_F\|_* \leq 1 & \end{array} \quad \begin{array}{ll} \inf & \|x - \bar{x}\| \\ s.t. & x \in \bar{P}_j, \\ & x_{F'} = \bar{x}_{F'}. \end{array} \tag{7}$$

Data: Functions f and G. The scalars n and p.

Initialization: Set $k = 0$ and define P^0 as

$$P^0 \equiv \left\{ x \in \mathbb{R}^{n+1}\colon \begin{array}{ll} f(x) \le x_{n+1}, & \\ G(x) \le 0, & x_i \in [0,1], i = 1, \ldots, p, \end{array} \right\},$$

Iteration-k:

Step 1: (Relaxation) Let $\bar{x}$ be the optimal solution of

$$\begin{array}{ll} \min & x_{n+1} \\ \text{s.t.} & x \in P^k \end{array},$$

Step 2: (Optimality check) Define $F \equiv \{j \in \{1, \ldots, p\}\colon 0 < \bar{x}_j < 1\}$ and $F' = \{1, \ldots, p\} \setminus F$. If F is empty then stop: $\bar{x}$ is an optimal solution and $\bar{x}_{n+1}$ is the optimal value. Otherwise, let $j \in F$.

Step 2.1: Find an optimal solution $\hat{x}$ of

$$\min \left(\begin{array}{lll} \min & x_{n+1} & \\ \text{s.t.} & x \in P^k & \\ & x_{F'} = \bar{x}_{F'} & , \\ & x_j = 0 & \end{array} \quad \begin{array}{lll} \min & x_{n+1} \\ \text{s.t.} & x \in P^k \\ & x_{F'} = \bar{x}_{F'} \\ & x_j = 1 \end{array} \right)$$

Step 2.2: If $\hat{x}_{n+1} = \bar{x}_{n+1}$ then let $\bar{x} = \hat{x}$ and restart Step 2; Otherwise set $x^k = \bar{x}$ and continue to Step 3.

Step 3: (Separation) Let j be the last index tried in Step 2. Find a separating hyperplane "$a^{k+1}x \le b_{k+1}$" between P^k_j and x^k. Define $P^{k+1} = P^k \cap \left\{x\colon a^{k+1}x \le b_{k+1}\right\}$ and set $k := k + 1$.

Figure 1: The basic cutting-plane algorithm for mixed zero-one convex programming (BCP4MINLP)

Let us assume without loss of generality that the relaxation $\bar{P}$ is defined by

$$\bar{P} \equiv \{x \in \mathbb{R}^n: \quad G(x) \leq 0, \quad x \geq 0, \quad x_i \leq 1, i = 1, \ldots p\}. \tag{8}$$

and $\bar{P}_j$ is defined by (5). Let F be an index set that may or may not be related to the Step 2 of the cutting-plane algorithm, and $F' = \{1, \ldots, n\} \backslash F$ be its complement. If $\hat{x} = (\hat{x}_F, \bar{x}_{F'}) \in \bar{P}_j$ is a known optimal primal solution in (7) then the subgradient $\hat{\xi} = (\hat{\xi}_F, \hat{\xi}_{F'})$ of the function

$$f(x) = \begin{cases} \|x_F - \bar{x}_F\| & \text{if } x_F = \bar{x}_F, \\ +\infty & \text{otherwise,} \end{cases} \tag{9}$$

at the point $\hat{x}$ that satisfies $\hat{\xi}(x - \hat{x}) \geq 0$, for every $x \in \bar{P}_j$, defines an optimal dual solution $(\hat{\alpha}, \hat{\beta}) \in \text{polar } (\bar{P}_j)$.

However, the main purpose of (7) is to define the cut generation problem using a smaller number of variables. This means that after solving the primal problem

$$\begin{array}{ll} \min & \|x_F - \bar{x}_F\| \\ \text{s.t.} & x_F \in \left\{x_F: (x_F, \bar{x}_{F'}) \in \bar{P}_j\right\} \end{array} \tag{10}$$

we have at hand an optimal primal solution $\hat{x}_F$ and a subgradient $\hat{\xi}_F$ of the function $\|\cdot - \bar{x}_F\|$ at $\hat{x}_F$, such that $\hat{\xi}_F(x_F - \hat{x}_F) \geq 0$, for every $x_F \in \left\{x_F: (x_F, \bar{x}_{F'}) \in \bar{P}_j\right\}$. Thus, a natural question is whether we can extend $\hat{\xi}_F$ so that $\hat{\xi} = (\hat{\xi}_F, \hat{\xi}_{F'})$ is a subgradient of the function f defined by (9) at $\hat{x} = (\hat{x}_F, \bar{x}_{F'})$ that satisfies $\hat{\xi}(x - \hat{x}) \geq 0$, for every $x \in \bar{P}_j$. Another natural question is whether we can apply a similar mechanism even when $\hat{x}_F$ is not optimal.

Our answers to these questions require that $\bar{x}_{F'} = 0$. This can be done without loss of generality because when $\bar{x}_k$, for some $k \in F'$, is nonzero then as long as it coincides with one of its bounds on Program (2) a variable transformation allows for the requirement to hold. In this setting, x_F is feasible for Program (10) if and only if x_F belongs to

$$\text{conv } \left(\left\{x_F: (x_F, 0) \in \bar{P}, x_j = 0\right\} \cup \left\{x_F: (x_F, 0) \in \bar{P}, x_j = 1\right\}\right). \tag{11}$$

Note that the two individual sets that define this convex hull are the feasible regions in (6) and consequently at the end of Step 2 we already know whether those sets are empty or nonempty. This feature is important because it determines which is the best solution procedure to use on Program (10). If both sets are nonempty then the program can be handled using the solution procedures described on Sections 5.4 and 5.5 of [6]. If one of them is empty then Program (10) is a standard convex program, and may therefore be solved by a standard nonlinear programming algorithm. If the two sets are empty then there is no feasible solution x to Program (2) such that $x_{F'} = \bar{x}_{F'}$. In this case the following inequality

$$\sum_{k \in F': \bar{x}_k = 0} x_k + \sum_{k \in F': \bar{x}_k = 1} (1 - x_k) \geq 1,$$

separates $\bar{x}$ from the convex hull of the feasible region of Program (2).

Now, we explain how the lifting procedure works under two distinct situations, depending on the fact that one or none of the sets in (11) is empty. We start by assuming that none of them is empty. Let $\hat{x}_F$ be a feasible solution for Program (10) and $\hat{\xi}_F$ be a subgradient of the function $\|\cdot - \bar{x}_F\|$ at $\hat{x}_F$ such that for a given scalar β satisfying $\hat{\xi}_F \bar{x}_F < \beta$ the following holds

$$\min_{i=0,1} \left(\begin{array}{ll} \min & \hat{\xi}_F z_F \\ \text{s.t.} & (z_F, 0) \in \bar{P}, \\ & z_j = i \end{array} \right) \geq \beta. \tag{12}$$

We remark that $\hat{x}_F$ need not be optimal for Program (10), though if it were optimal then the existence of a subgradient and a scalar satisfying (12) would be guaranteed. Under a constraint qualification, the optimal solution $\hat{z}_F^i$ of each one of the problems in (12) also solves a linear program defined by a suitable matrix $A^i \in \partial G(\hat{z}_F^i, 0)$ so that

$$\min_{i=0,1} \left(\begin{array}{ll} \min & \hat{\xi}_F z_F \\ s.t. & \left\{ \begin{array}{l} G(\hat{z}_F^i, 0) + A^i(z_F - \hat{z}_F^i, 0) \leq 0, \\ z_k \geq 0, k \in F, \qquad z_j = i, \\ z_k \leq 1, k \in F \cap \{1, \ldots, p\}, \end{array} \right. \end{array} \right) \geq \beta \tag{13}$$

The feasible region of each one of these linear programs defines an outer-approximation of each one of the sets in (11). Our lifting procedure applies to these linear programs, so that by the outer-approximation argument it also applies to our original nonlinear sets. Proposition 2 below describes the lifting mechanism in generic terms.

Proposition 2 *Let F be an index set and $F' = \{1, \ldots, n\} \setminus F$. For a given arbitrary vector α_F, let $\hat{z}_F$ be an optimal solution of the following linear program:*

$$\begin{array}{ll} \min & \alpha_F z_F \\ \text{s.t.} & A_F z_F \leq b, \\ & l_F \leq z_F \leq u_F, \end{array} \tag{14}$$

where l_F and u_F are the, possibly infinite, lower and upper bounds, $A_F \in \mathbb{R}^{m \times |F|}$ and $b \in \mathbb{R}^m$. Then, for any extended matrix $A = [A_F, A_{F'}] \in \mathbb{R}^{m \times n}$ there is a closed-form extended vector $\alpha = (\alpha_F, \alpha_{F'})$ such that the vector $\hat{z} = (\hat{z}_F, 0)$ is an optimal solution of the following linear program:

$$\begin{array}{ll} \min & \alpha z \\ \text{s.t.} & Az \leq b, \\ & l_F \leq z_F \leq u_F, \\ & z_{F'} \geq 0. \end{array} \tag{15}$$

Proof: Let $\hat{v} \leq 0$ be the optimal dual multipliers associated with the matrix constraints in Program (14) and define

$$\alpha_k \equiv \max \left(0, \sum_{l=1}^{m} \hat{v}_l a_{lk} \right),$$

for every $k \in F'$. Now, consider Program (15) and use the same dual variables to price the new primal variables z_k, for every $k \in F'$. Since the reduced costs are $\rho_k = \alpha_k - \sum_{l=1}^{m} \hat{v}_l a_{lk} \geq 0$, for every $k \in F'$, we conclude that $\hat{z} = (\hat{z}_F, 0)$ is optimal for Program (15). □

This proposition shows by construction how to define extended vectors $\hat{\xi}^i = (\hat{\xi}_F, \hat{\xi}^i_{F'})$ such that $\hat{z}^i = (\hat{z}^i_F, 0)$, for $i = 0, 1$ are still optimal in the following linear programs

$$\min_{i=0,1} \left(\begin{array}{ll} \min & \hat{\xi}^i z \\ s.t. & \left\{ \begin{array}{l} G(\hat{z}^i) + A^i(z - \hat{z}^i) \leq 0, \\ z \geq 0, \qquad z_j = i, \\ z_k \leq 1, k \in F \cap \{1, \ldots, p\}, \end{array} \right. \end{array} \right) \geq \beta \tag{16}$$

whose feasible regions are larger than the set $\bar{P} \cap \{x : x_j = i\}$, respectively. Since $z_{F'} \geq 0$, for every $z \in \bar{P}_j$, then

$$\hat{\xi} = (\hat{\xi}_F, \max_{i=0,1}(\hat{\xi}^i_{F'}))$$

is a subgradient of f at $\hat{x}$ such that $\hat{\xi}x \geq \beta$, for every $x \in \bar{P}_j$. Moreover, since $\hat{\xi}\bar{x} < \beta$ then we have found a separating hyperplane.

Now, we assume that one of the sets in (11) is empty. Thus, Program (10) is solved as a standard convex program because its feasible region is defined by the nonempty set only. However, the fact that one of the sets in (11) is empty does not imply that the same has to occur in (5), when the variables $x_{F'}$ are no longer fixed. Proposition 3 below describes in generic terms how to define the extended vector $\hat{\xi}^i = (\hat{\xi}_F, \hat{\xi}^i_{F'})$ so that $\hat{\xi}^i z \geq \beta$, for every $z \in \bar{P} \cup \{x\colon x_j = i\}$, when the set $\bar{P} \cap \{x\colon x_{F'} = 0, x_j = i\}$ is empty.

Proposition 3 *Let F be an index set and $F' = \{1, \ldots, n\} \setminus F$. For a given arbitrary vector α_F, let $\hat{z}_F$ be the optimal value of the following linear program:*

$$\begin{array}{ll} \min & \alpha_F z_F \\ \text{s.t.} & A_F z_F \leq b + \hat{t}e, \\ & l_F \leq z_F \leq u_F, \end{array} \tag{17}$$

where l_F and u_F are the, possibly infinite, lower and upper bounds, $A_F \in \mathbb{R}^{m \times |F|}$, $b, e \in \mathbb{R}^m$ where e is a vector of "all-ones", and $\hat{t} \equiv \min\{t\colon A_F z_F \leq b + te, l_F \leq z_F \leq u_F\} > 0$. Then, for any extended matrix $A = [A_F, A_{F'}] \in \mathbb{R}^{m \times n}$ and scalar β there is a closed-form extended vector $\alpha = (\alpha_F, \alpha_{F'})$ such that $\alpha z \geq \beta$, for every z such that $Az \leq b, l_F \leq z_F \leq u_F, z_{F'} \geq 0$.

Proof: First, consider the linear program that defines $\hat{t}$. Let $(\hat{t}, \tilde{z})$ be an optimal solution and $\hat{w}$ be the optimal dual multipliers associated with the matrix constraints. Then,

$$\hat{t} = \hat{w}b + \hat{\gamma}_F \tilde{z}_F, \tag{18}$$

where $\hat{\gamma}_k = 0 - \sum_{l=1}^m \hat{w}_l a_{lk}$ is the reduced cost associated with the variable z_k, for each $k \in F$.

Now, consider Program (17) and let $\hat{v}$ be the optimal dual multipliers associated with the matrix constraints. Then,

$$\begin{array}{rl} & \alpha_F \hat{z}_F = \hat{v}\left(b + \hat{t}e\right) + \hat{\rho}_F \hat{z}_F \\ \iff & \alpha_F \hat{z}_F - \hat{t}\hat{v}e = \hat{v}b + \hat{\rho}_F \hat{z}_F, \end{array} \tag{19}$$

where $\hat{\rho}_k = \alpha_k - \sum_{l=1}^m \hat{v}_l a_{lk}$ is the reduced cost associated with the variable z_k, for each $k \in F$.

If $\beta \leq \alpha_F \hat{z}_F - \hat{t}\hat{v}e$ then define $\alpha_k = \max(0, \sum_{l=1}^m \hat{v}_l a_{lk})$, for every $k \in F'$. For every z such that $Az \leq b, l_F \leq z_F \leq u_F, z_{F'} \geq 0$ we have that

$$\begin{array}{rcll} \alpha z & = & \alpha_F z_F + \alpha_{F'} z_{F'} & \\ & \geq & \displaystyle\sum_{k \in F} \left(\hat{\rho}_k + \sum_{l=1}^m \hat{v}_l a_{lk} \right) z_k + \sum_{k \in F'} \left(\sum_{l=1}^m \hat{v}_l a_{lk} \right) z_k & (20) \\ & = & \hat{\rho}_F z_F + \hat{v} A z & \\ & \geq & \hat{\rho}_F z_F + \hat{v} b & (21) \\ & \geq & \hat{\rho}_F \hat{z}_F + \hat{v} b & (22) \\ & = & \alpha_F \hat{z}_F - \hat{t}\hat{v}e & (23) \\ & \geq & \beta, & (24) \end{array}$$

where the inequality (20) follows from the definition of $\hat{\rho}_F$, the definition of $\alpha_{F'}$ and the fact that $z_{F'} \geq 0$; the inequality (21) follows from the fact that $\hat{v} \leq 0$ and $Az \leq b$; the inequality

(22) follows the fact that $\hat{\rho}_k(z_k - \hat{z}_k) \geq 0$, for every $k \in F$, which is consequence of the values of the reduced costs at optimality; the inequality (23) follows from (19); and finally the inequality (24) holds by hypothesis.

If $\beta > \alpha_F \hat{z}_F - \hat{t}\hat{v}e$ then a similar formula works but we need to increase $\hat{v}$ by a suitable positive multiplier of $\hat{w}$. Observe that $\hat{z}_F$ is feasible for the linear program that defines $\hat{t}$ and so, from (18), we have that $\hat{t} \leq \hat{w}b + \hat{\gamma}_F \hat{z}_F$, or equivalently,

$$\beta - \left(\alpha_F \hat{z}_F - \hat{t}\hat{v}e\right) \leq \delta \hat{w} b + \delta \hat{\gamma}_F \hat{z}_F, \tag{25}$$

where $\delta = \left(\beta - \left(\alpha_F \hat{z}_F - \hat{t}\hat{v}e\right)\right)/\hat{t} > 0$. Now, define $\alpha_k = \max\left(0, \sum_{l=1}^{m} \left(\hat{v} + \delta\hat{w}\right)_l a_{lk}\right)$, for every $k \in F'$. For every z such that $Az \leq b, l_F \leq z_F \leq u_F, z_{F'} \geq 0$ we have that

$$\begin{aligned}
\alpha z &= \alpha_F z_F + \alpha_{F'} z_{F'} \\
&\geq \sum_{k \in F} \left(\hat{\rho}_k + \sum_{l=1}^{m} \hat{v}_l a_{lk} \right) z_k + \sum_{k \in F'} \left(\sum_{l=1}^{m} \left(\hat{v} + \delta\hat{w}\right)_l a_{lk} \right) z_k && (26) \\
&= \hat{\rho}_F z_F + \left(\hat{v} + \delta\hat{w}\right) Az + \delta \hat{\gamma}_F z_F \\
&\geq \hat{\rho}_F z_F + \hat{v}b + \delta\hat{w}b + \delta\hat{\gamma}_F z_F && (27) \\
&\geq \hat{\rho}_F \hat{z}_F + \hat{v}b + \delta\hat{w}b + \delta\hat{\gamma}_F \hat{z}_F && (28) \\
&\geq \hat{\rho}_F \hat{z}_F + \hat{v}b + \left(\beta - \left(\alpha_F \hat{z}_F - \hat{t}\hat{v}e\right)\right) && (29) \\
&= \beta && (30)
\end{aligned}$$

where the inequality (26) follows from the definition of $\hat{\rho}_F$, the definition of $\alpha_{F'}$ and the fact that $z_{F'} \geq 0$; the inequality (27) follows from the fact that $\hat{v} + \delta\hat{w} \leq 0$ and $Az \leq b$; the inequality (28) follows the fact that $\hat{\rho}_k(z_k - \hat{z}_k) \geq 0$ and $\hat{\gamma}_k(z_k - \hat{z}_k) \geq 0$, for every $k \in F$, which is consequence of the values of the reduced costs at optimality; the inequality (29) follows from (25); and finally the equality (30) is a consequence of (19). □

This result can be easily generalized to a situation in which a distinct t variable occurs for each constraint. This is in fact the usual procedure with most phase-one implementations of the Simplex algorithm for linear programs.

When solving a nonlinear program whose feasible region $\bar{P} \cap \{z : z_{F'} = 0, z_j = i\}$ is empty, most standard nonlinear programming algorithms are not ready to provide some point $(\hat{t}, \tilde{z}_F^i)$ that solves the following program

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \left\{ \begin{array}{l} G(z_F, 0) \leq te, \\ z_k \geq 0, k \in F \qquad z_j = i, \\ z_k \leq 1, k \in F \cap \{1, \ldots, p\}, \end{array} \right. \end{array} \tag{31}$$

and in this way proving infeasibility. In fact, it may occur that what seems to be an infeasible problem is just a numerical difficulty of meeting the constraints to a desired accuracy. The solution of the program (31) provides a verification of infeasibility and, as saw in the proof of Proposition 3, the dual variables that may be required for the lifting procedure. Since the Slater condition holds, the optimal solution $(\hat{t}, \hat{z}_F^i)$ of Program (31) also solves the following linear program defined by a suitable matrix $A^i \in \partial G(\hat{z}_F^i, 0)$,

$$\begin{array}{ll} \min & t \\ \text{s.t.} & \left\{ \begin{array}{l} G(\hat{z}_F^i, 0) + A^i(z_F - \hat{z}_F^i, 0) \leq te, \\ z_k \geq 0, k \in F, \qquad z_j = i, \\ z_k \leq 1, k \in F \cap \{1, \ldots, p\}, \end{array} \right. \end{array} \tag{32}$$

Then, to complete the lifting procedure we just have to solve

$$\begin{array}{ll} \min & \hat{\xi}_F z_F \\ s.t. & \left\{ \begin{array}{l} G(\hat{z}^i_F, 0) + A^i(z_F - \hat{z}^i_F, 0) \leq \hat{t}e, \\ z_k \geq 0, k \in F, \qquad z_j = i, \\ z_k \leq 1, k \in F \cap \{1, \ldots, p\}, \end{array} \right. \end{array} \tag{33}$$

and, as explained in the proof of Proposition 3 we are now able to define $\hat{\xi}^i = (\hat{\xi}_F, \hat{\xi}^i_{F'}) \in \partial f(\hat{x})$ such that $\hat{\xi}^i z \geq \beta$, for every $z \in \bar{P} \cap \{z : z_j = i\}$. Since $z_{F'} \geq 0$, for every $z \in \bar{P}_j$, then again

$$\hat{\xi} = (\hat{\xi}_F, \max_{i=0,1}(\hat{\xi}^i_{F'}))$$

is a subgradient of the function f at $\hat{x}$ such that $\hat{\xi}x \geq \beta$, for every $x \in \bar{P}_j$. Moreover, since $\hat{\xi}\bar{x} < \beta$ then we have found a separating hyperplane.

References

[1] Egon Balas, Sebastián Ceria, and Gérard Cornuéjols. A lift-and-project cutting plane algorithm for mixed 0-1 programs. *Math. Programming*, 58(3, Ser. A):295–324, 1993.

[2] E. Balas, S. Ceria, and G. Cornuejols. Mixed 0-1 programming by lift-and-project in a branch-and-cut framework. *Management Science*, 42(9):1229–1246, Sep 1996.

[3] M. Duran and I. Grossmann. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, 36:307–339, 1986.

[4] C. Floudas. *Nonlinear and Mixed-integer Optimization.* Oxford University Press, 1995.

[5] B. Gavish. Topological design of computer communication networks - the overall design problem. *European Journal of Operational Research*, 58:149–172, 1992.

[6] J. Soares. *Disjunctive Convex Optimization.* PhD thesis, Graduate School of Business, Columbia Univeristy, Jun 1998.

[7] Robert A. Stubbs and Sanjay Mehrotra. A branch-and-cut method for 0-1 mixed convex programming. *Math. Program.*, 86(3, Ser. A):515–532, 1999.

[8] M. Turkay and I. Grossmann. Disjunctive programming techniques for the optimization of process systems with discontinuous investment costs - multiple size regions. *Industrial & Engineering Chemistry Research*, 35:2611–2623, 1996.