REGULARIZATION METHODS FOR DIFFERENTIAL EQUATIONS AND THEIR NUMERICAL SOLUTION

By

Ping Lin

B. Sc. (Mathematics), Nanjing University, Nanjing, P.R.China, 1984

M. Sc. (Applied Mathematics), Nanjing University, Nanjing, P.R.China, 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

INSTITUTE OF APPLIED MATHEMATICS

DEPARTMENT OF MATHEMATICS

We accept this thesis as conforming

to the required standard

....

....

....

THE UNIVERSITY OF BRITISH COLUMBIA

December 1995

Abstract

Many mathematical models arising in science and engineering, including circuit and device simulation in VLSI, constrained mechanical systems in robotics and vehicle simulation, certain models in early vision and incompressible fluid flow, lead to computationally challenging problems of differential equations with constraints, and more particularly to high-index, semi-explicit differential-algebraic equations (DAEs). The direct discretization of such models in order to solve them numerically is typically fraught with difficulties. We thus need to reformulate the original problem into a better behaved problem before discretization. Index reduction with stabilization is one class of reformulations in the numerical solution of high index DAEs. Another class of reformulations is called regularization. The idea is to replace a DAE by a better behaved nearby system. This method reduces the size of the problem and avoids the derivatives of the algebraic constraints associated with the DAE. It is more suitable for problems with some sort of singularities in which the constraint Jacobian does not have full rank. Unfortunately, this method often results in very stiff systems, which accounts for its lack of popularity in practice.

In this thesis we develop a method which overcomes this difficulty through a combination of stabilization and regularization in an iterative procedure. We call it the sequential regularization method (SRM). Several variants of the SRM which work effectively for various circumstances are also developed. The SRM keeps the benefits of regularization methods and avoids the need for using a stiff solver for the regularized problem. Thus the method is an important improvement over usual regularization methods and can lead to improved numerical methods requiring only solutions to linear systems. The SRM also provides cheaper and more efficient methods than the usual stabilization methods for some choices of parameters and stabilization matrix. We propose the method first for linear index-2 DAEs. Then we extend the idea to nonlinear index-2 and index-3 problems. This is especially useful in applications such as constrained multibody systems which are of index-3. Theoretical analysis and numerical experiments show that the method is useful and efficient for problems with or without singularities.

While a significant body of knowledge about the theory and numerical methods for DAEs has been accumulated, almost none of it has been extended to partial differential-algebraic equations (PDAEs). As a first attempt we provide a comparative study between stabilization and regularization (or pseudo-compressibility) methods for DAEs and PDAEs, using the incompressible Navier-Stokes equations as an instance of PDAEs. Compared with stabilization methods, we find that regularization methods can avoid imposing an artificial boundary condition for the pressure. This is a feature for PDAEs not shared with DAEs. Then we generalize the SRM to the nonstationary incompressible Navier-Stokes equations. Convergence is proved. Again nonstiff time discretization can be applied to the SRM iterations. Other interesting properties associated with discretization are discussed and demonstrated.

The SRM idea is also applied to the problem of miscible displacement in porous media in reservoir simulation, specifically to the pressure-velocity equation. Advantages over mixed finite element methods are discussed. Error estimates are obtained and numerical experiments are presented.

Finally we discuss the numerical solution of several singular perturbation problems which come from many applied areas and regularized problems. The problems we consider are nonlinear turning point problems, a linear elliptic turning point problem and a second-order hyperbolic problem. Some uniformly convergent schemes with respect to the perturbation parameter are constructed and proved. A spurious solution phenomenon for the upwinding scheme is analyzed.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I am specially grateful to Dr. Uri Ascher, my thesis supervisor, who contributed significantly to the development of this thesis. That contribution ranged from fruitful suggestions and enlightening discussions to general guidelines and ideas about how to present the results.

I am also grateful to Drs. John Heywood, Tim Salcudean, Michael Ward and Brian Wetton, members of my supervisory committee, for their helpful comments and suggestions during this research and writing of the thesis.

Thanks are also extended to Drs. Robert O'Malley, Jr., Robert Miura and Jim Varah, external and university examiners of my thesis examining committee, for their many suggestions on modification of the thesis.

Chapter 1

Introduction

Many mathematical models arising in science and engineering, including circuit and device simulation in VLSI, constrained mechanical systems in robotics and vehicle simulation, certain models in early vision and incompressible fluid flow, lead to computationally challenging problems of differential equations with constraints, and more particularly to high-index, semi-explicit differential-algebraic equations (DAEs). The direct discretization of such models in order to solve them numerically is typically fraught with difficulties, and most methods proposed in the literature seek to circumvent this by employing combinations of problem reformulation, regularization [1] and special discretization techniques.

We will consider the regularization of such mathematical models and the numerical solution of the resulting regularized formulations. These formulations are often singular perturbation problems because they typically depend on a small parameter which provides a measure of the closeness between the regularized and the original problems. We will also apply our regularization method and idea to other relevant practical problems.

1.1 Regularization for Differential-Algebraic Equations (DAEs)

DAEs are special implicit ordinary differential equations (ODEs)

$$f(x'(t), x(t), t) = 0, \tag{1.1}$$

[1] The concept of regularization was introduced by Tikhonov (see [109]). Its idea is that one solves a better behaved nearby problem instead of solving the original problem to circumvent some sort of difficulties. See the next section for more details.

where the partial Jacobian matrix $f_y(y, x, t)$ is singular for all relevant values of its arguments. Here $x' = \frac{dx}{dt}$. An extension to partial differential equations is considered in the next subsection.

DAEs were motivated by applications like network analysis, circuit simulation and mechanical system simulation starting in the 1970's. They often arise as ordinary differential equations with additional variables and (equality) algebraic constraints. An extensive list of applications is given in [92].

In the 1980's, DAEs have developed into a highly topical subject of computational and applied mathematics. Contributions devoted to DAEs have appeared in various fields, such as applied mathematics, scientific computation, mechanical engineering, chemical engineering, system theory, etc. Frequently, other names have been assigned to DAEs, e.g. semistate equations, descriptor systems, singular systems. Gear ([51], 1971) proposed to handle DAEs numerically by backward differentiation formulas (BDF). For a long time DAEs had been considered not to differ essentially from regular implicit ODEs in general. Only since about 1980, because of computational results that could not be brought into line with the above supposition (e.g. Sincovec et al [103], 1981), the mathematical and particularly the numerical community have started investigating DAEs more thoroughly. With their famous paper, Gear, Hsu and Petzold ([52], 1981) started a discussion on DAEs that will surely be carried on for a while.

The structure of DAEs is very much related to the concept of index, which is a measure of the amount of singularity of the system. There are several ways to define index. The most popular one is called differential index, which is defined as the minimal number of analytical differentiations in t such that (1.1) can be transformed by algebraic manipulations into an explicit ordinary differential system (in the original unknowns)

$$x' = \phi(x, t)$$

which in turn is called the "underlying ODE ". There are several structural forms of DAEs which appear frequently in applications (see [92]). The differential index of these structural forms can be found by differentiating their algebraic constraints with respect to t (and substituting into the differential equations which complement the algebraic constraints). For instance,

- Semi-explicit index-1 system

$$\begin{aligned} x' &= f(x,y), &\quad (1.2a)\\ 0 &= g(x,y), &\quad (1.2b) \end{aligned}$$

  if g_y is invertible.

- Hessenberg index-2 system

$$\begin{aligned} x' &= f(x,y), &\quad (1.3a)\\ 0 &= g(x), &\quad (1.3b) \end{aligned}$$

  if $g_x f_y$ is invertible.

- Hessenberg index-3 system

$$\begin{aligned} x' &= f(x,y), &\quad (1.4a)\\ y' &= k(x,y,z), &\quad (1.4b)\\ 0 &= g(x), &\quad (1.4c) \end{aligned}$$

  if $g_x f_y k_z$ is invertible. Mechanical multibody systems with holonomic constraints are examples of Hessenberg index-3 DAEs.

In [56] it is pointed out that higher index (≥ 2) DAEs, in the natural function space formulations, lead to ill-posed [2] problems because they do not have the usual

[2] A problem is called ill-posed if it is not well-posed. A problem is called well-posed if it satisfies three conditions, i.e. the existence, uniqueness and stability of its solution. "Stability" means that the solution of the problem continuously depends on the "data", which may be initial data, boundary data, coefficients in the equation, values of the operator, etc.. Here high-index DAEs fail to satisfy a stability condition.

stability property of differential equations in general.

Example 1.1 *Consider (See [86])*

$$x' = y, \tag{1.5a}$$

$$0 = x - p(t), \tag{1.5b}$$

where $x(t)$ and $y(t)$ are scalar functions and $p(t)$ is a given function. This is a very simple index-2 DAE. The exact solution is $x = p(t)$, $y = p'(t)$. If we add a small perturbation $\delta \sin \omega t$, $\delta << 1$, to the right hand side of the second equation we have the exact solution

$$\bar{x} = p(t) + \delta \sin \omega t, \ \bar{y} = p'(t) + \delta\omega \cos \omega t.$$

Hence $\bar{y} - y = \delta\omega \cos \omega t$ could be very large if $\omega \gg 1/\delta$, i.e. the solution changes a lot under a small change in the right-hand side of the equation. □

A numerical method which is directly applied to a complex, ill-posed problem may generally fail. Therefore, to solve DAEs, we have to stabilize such problems to bring about continuous dependence on the "data" (or stability). One such approach is to change the formulation of the problem but not its solution, e.g. in Example 1.1 we differentiate (1.5b) once and obtain

$$x' = y \tag{1.6a}$$

$$0 = y - p'(t), \ x(0) = p(0). \tag{1.6b}$$

Now (1.6) is of index-1 and becomes a well-posed problem with the same solution as (1.5). We can solve (1.6) instead of (1.5) and gain well-posedness. However, such a direct index reduction procedure may cause the well-known drift difficulty (see [29]), i.e. the approximate solution of (1.6) may be far from satisfying the constraint (1.5b).

Hence, methods have been designed to prevent moving away from the constraints. Baumgarte's stabilization [17] and projection invariant methods (cf. [16]) are popular among such methods. Most of these approaches treat initial value problems, and only a few apply to boundary value problems. See [29, 58] for various numerical methods for initial value problems; [93, 7, 15] for boundary value problems; and [16, 9, 35] for a survey of various stabilization reformulations.

Another approach consists in adding some small perturbation terms (measured by a small positive parameter ϵ) to the given DAE. The perturbed problem is close to the original problem (if ϵ is small) and is well-posed. Such an approach is usually called regularization. This is a natural approach since the high-index DAE is ill-posed; indeed in [43] a well-known Tikhonov regularization algorithm (see [109]) was applied to solve DAEs. However, such a method seems so general that it is not sufficiently related to the special structure of the DAE. There are two types of regularization methods which are probably more interesting for DAE researchers. One is called parameterization. One such possibility, the pencil regularization, was given independently by Boyarintsev [24] (or see his newer book [25] published in English) and Campbell [32]. But the regularized problem is ensured to be well-posed only for constant coefficient cases. A further parameterization was proposed by März [85]. Her regularization is aimed at obtaining well-posed index-1 DAEs instead of obtaining well-posed ODEs. Heuristically, it seems evident that the DAE is less changed if it is transformed into an index-1 DAE rather than an ODE. März's regularization was proved to be well-posed for usual structural forms of DAEs. We refer to [59] for further results in this direction.

Another class of regularization uses the penalty idea (see [84, 91]). It originates from penalty methods for constrained optimization problems. Note that an algebraic equation in a DAE can be viewed as a constraint. This method seems more natural for DAEs. References [68, 70, 69] used the penalty regularization and singular perturbation theory to determine the solutions of DAEs when the initial or boundary values are given improperly (i.e. inconsistently). In Chapter 2 we will mention these methods again with a bit more detail and indicate that März's regularization is actually a kind of penalty method.

Because the regularization method requires fewer differentiations of the constraints it is perhaps more suitable for DAEs which have singularities, i.e. whose constraints do not have full rank, e.g. when the matrix $g_x f_y$ is singular at some isolated points in the index-2 system (1.3). These problems can be challenging for the methods that are usually employed and appear frequently in simulation of constrained mechanical systems. To our knowledge, there has not been a paper in the numerical analysis literature about this until the recent two preprints [11] and [94] (although a number of relevant papers appear in the mechanical engineering literature).

From a practical point of view, a number of codes which work well and efficiently (at least if the regularization parameter ϵ is not too small) are available for numerically solving the regularized problems. We also note that the regularization method requires less smoothness of the coefficients of the differential-algebraic problem than other stabilization methods. These are the advantages of the regularization method. The dominant disadvantage in the above regularization methods is that the parameter ϵ must be small enough to maintain the accuracy of the numerical method we use for the regularized problem at an acceptable level. Hence, a stiff solver is necessary. Typically in regularization methods, the parameter ϵ must be chosen both "large enough" and "small enough": large so that the regularized problem would behave significantly better than the original, and small so that its solution will not differ too much from that of the original problem.

We will present a class of new regularization methods, inspired by [19], which we call sequential regularization method (SRM). The SRM can be viewed as a combination of the penalty method with Baumgarte's stabilization in an iterative procedure;

see §2.4 or §3.1 for specific instances. It is applicable for DAEs with constraint singularities. Moreover, the regularization parameter ϵ in the method is not necessarily small. Thus, a nonstiff solver can be used for solving the regularized problems. Some variants of the SRM are discussed for index-2 and index-3 DAEs with the goal of simplifying the computations. We will apply the SRM to mechanical multibody systems as well.

1.2 Regularization for the Incompressible Navier– Stokes equations

As noted before, DAEs have become a highly topical subject of applied and numerical mathematics. However, there seems to be still a void in the literature about partial differential equations with constraints (PDAEs). A typical instance of such problems is the well-known incompressible Navier-Stokes equations:

$$\mathbf{u}_t + (\mathbf{u} \cdot \mathbf{grad})\mathbf{u} = \mu \Delta \mathbf{u} - \mathbf{grad} p + \mathbf{f}, \tag{1.7a}$$

$$div \mathbf{u} = 0, \tag{1.7b}$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{b} \quad , \quad \mathbf{u}|_{t=0} = \mathbf{a}, \tag{1.7c}$$

in a bounded two- or three-dimensional domain Ω and $0 \leq t \leq T$. Here $\mathbf{u}(\mathbf{x}, t)$ represents the velocity of a viscous incompressible fluid, $p(\mathbf{x}, t)$ the pressure, $\mathbf{f}$ the prescribed external force, $\mathbf{a}(\mathbf{x})$ the prescribed initial velocity, and $\mathbf{b}(t)$ the prescribed boundary values.

The system (1.7) can be seen as a partial differential equation with constraint (1.7b) with respect to the time variable t. Hence, we call it a PDAE. It is easily verified that it is of index-2 without singularities since the operator $div\mathbf{grad} = \Delta$ is invertible (under appropriate boundary conditions).

A huge number of methods have been designed to solve the nonstationary incompressible Navier-Stokes equations (1.7). Direct discretizations include finite difference and finite volume techniques on staggered grids (e.g. [65, 26, 66]), finite element methods using conformal and nonconformal elements (e.g. [54, 110, 63, 64]) and spectral methods (e.g. [33]). Another approach yielding many methods has involved some initial reformulation and/or regularization of the equations, to be followed by a discretization of the (hopefully) simplified system of equations. Examples of such methods include pseudo-compressibility methods, projection and pressure-Poisson reformulations (e.g. [36, 55, 72, 97, 102, 117]). The two types of regularizations we mentioned in §1.1 for DAEs were already proposed in the Navier-Stokes context quite a while ago (cf. [108, 72]). We are interested in the generalization of the SRM to this problem because the regularized problems can be made essentially nonstiff and then a more convenient difference scheme (e.g. an explicit scheme) in time is possible. Moreover, the method retains the benefits of the penalty method. For example, computations for the velocity $\mathbf{u}$ and the pressure p are uncoupled and an artificial boundary condition for calculating the pressure p is not necessary.

1.3 A Problem in Reservoir Simulation

The idea of the SRM can be applied to a reservoir-simulation problem — miscible displacement in porous media.

Miscible displacement is an enhanced oil-recovery process that has attracted considerable attention in the petroleum industry. It involves injection of a solvent at certain wells in a petroleum reservoir, with the intention of displacing resident oil to other wells for production. This oil may have been left behind after primary production by reservoir pressure and secondary production by waterflooding. The economics of the process can be precarious, because the chemicals it requires are expensive and the performance of the displacement is by no means guaranteed. Complex physical behavior in the reservoir will determine whether enough additional oil is recovered to make the expense worthwhile. A numerical simulation of the complex process undoubtedly plays an important role.

Mathematically, the process is described by a convection– dominated parabolic partial differential equation for each chemical component in the system. By summing up the component equations, one can obtain an equation that determines the pressure in the system; this nonlinear equation is elliptic or parabolic, according to whether the system is incompressible or compressible. Thus, in this problem one encounters elliptic, parabolic, and near–hyperbolic equations with complicated nonlinear behavior.

For simplicity, we consider the miscible displacement of one incompressible fluid by another in a porous reservoir $\Omega \subset \mathbf{R}^2$ over a time period $[0, T]$. Let $p(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$ denote the pressure and Darcy velocity of the fluid mixture, and $c(\mathbf{x}, t)$ the concentration of the invading fluid. Then the mathematical model is a strongly coupled nonlinear system of partial differential equations (see [44, 82]):

$$\mathbf{u} = -a(\mathbf{grad}p - \gamma\mathbf{grad}d), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{1.8a}$$

$$div\mathbf{u} = q(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{1.8b}$$

$$\phi\frac{\partial c}{\partial t} - div(D(\mathbf{u})\mathbf{grad}c) + \mathbf{u} \cdot \mathbf{grad}c = g(c), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{1.8c}$$

with the boundary conditions

$$\mathbf{u} \cdot \mathbf{n} = 0, \quad (\mathbf{x}, t) \in \Gamma \times [0, T], \tag{1.9a}$$

$$D(\mathbf{u})\mathbf{grad}c \cdot \mathbf{n} = 0, \quad (\mathbf{x}, t) \in \Gamma \times [0, T], \tag{1.9b}$$

and initial condition

$$c(\mathbf{x}, 0) = c_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{1.10}$$

where $a = a(\mathbf{x}, c)$ is the mobility of the fluid mixture, $\gamma = \gamma(\mathbf{x}, c)$ and $d(\mathbf{x})$ are the gravity and vertical coordinate, q is the imposed external rates of flow, $\phi(\mathbf{x})$ is the porosity of the rock, D is the coefficient of molecular diffusion and mechanical dispersion of one fluid into the other, $g = g(\mathbf{x}, t, c)$ is a known linear function of c representing sources, and $\mathbf{n}$ is the exterior normal to the boundary $\Gamma = \partial\Omega$.

The pressure-velocity equation (1.8a)–(1.8b) is elliptic (after eliminating $\mathbf{u}$). The concentration equation (1.8c) is parabolic, but normally convection- dominated. It is derived from the conservation of mass which involves the Darcy velocity of the fluid mixture, but the pressure variable does not appear in it. Thus a good approximation of the concentration equation requires accurate solution for the velocity variables. Mixed finite element methods have been applied to the pressure-velocity equation, which can yield velocity solutions one order more accurate than those obtained using corresponding finite difference and usual finite element methods [40, 41, 42, 46, 47, 120]. However, the finite dimensional spaces for the velocity and pressure need to satisfy the Babuska-Brezzi condition, and the resulting linear system does not have a positive definite coefficient matrix. Moreover, the number of degrees of freedom in the linear system doubles that of finite difference or finite element methods.

We are interested in designing an SRM for the pressure-velocity equation since the SRM formulation can produce as accurate a velocity approximation as the mixed finite element methods, and can avoid the above-mentioned problems in mixed finite element methods.

1.4 Regularization for Differential Equations without Constraints

Regularization methods are also used to treat differential equations without constraints when these equations have some sort of singularities or their solutions may have some sort of discontinuities. Examples are viscous solutions of hyperbolic conservation laws [76], shape from shading problems with singularities [34] and transition phenomena in semi-conductor device simulation [13]. A frequently considered problem is the following first-order partial differential equation

$$a(u,x,t)\frac{\partial u}{\partial t}+\sum_{i=1}^{n}b_i(u,x,t)\frac{\partial u}{\partial x_i}+c(u,x,t)=0 \tag{1.11}$$

with appropriate initial and boundary conditions. In general, (1.11) may have some kind of discontinuous solutions, e.g. a shock wave [76]. Or, if we consider the steady-state case, (1.11) may have singularities, i.e. b_i may vanish at some points, as in the shape from shading problem [34]. Some regularization techniques have been designed for solving (1.11). A popular one is

$$a(u,x,t)\frac{\partial u}{\partial t}+\epsilon d(u,x,t)\Delta u+\sum_{i=1}^{n}b_i(u,x,t)\frac{\partial u}{\partial x_i}+c(u,x,t)=0. \tag{1.12}$$

The regularized problem often has a physical meaning by itself, e.g. a time-dependent advection-diffusion equation with a small diffusion term. Another choice could be

$$\epsilon d(u,x,t)(\frac{\partial^2 u}{\partial t^2}-\Delta u)+a(u,x,t)\frac{\partial u}{\partial t}+\sum_{i=1}^{n}b_i(u,x,t)\frac{\partial u}{\partial x_i}+c(u,x,t)=0. \tag{1.13}$$

This has physical meanings as well, e.g. a traffic flow problem [118] or so-called overdamped vibration problems [104].

Thus the approximate resolution of (1.11) becomes that of the singular perturbation problem (1.12) or (1.13). Unlike the SRM, the regularization parameter ϵ has to be small in comparison with the mesh size to ensure that the solution of the regularized problem be a good approximation of that of the original problem. It is well-known that there are difficulties in solving these regularized problems numerically with small ϵ, e.g. the stability problem for the central difference scheme and the accuracy problem for the upwinding scheme in a boundary layer region in which the derivatives of the solution may be large (see [37, 61]). We will consider some special cases of (1.12) or (1.13) and focus mainly on uniformly convergent methods. We will discuss spurious solutions of a simple upwinding scheme as well.

1.5 Contribution of This Thesis

Our objectives are to propose and to investigate regularization methods for various differential equations with or without constraints. Most attention is paid to ordinary and partial differential equations with constraints (DAEs and PDAEs). We propose and analyze a regularization method called the sequential regularization method (SRM). A very important advantage of our regularization method (SRM) is that the problem after regularization need not be stiff. Hence explicit difference schemes can be used to avoid solving nonlinear systems and they make the computation much simpler. Improvements over stabilization methods and extra benefits for PDAEs are also achieved.

In Chapter 2, the SRM is presented for linear index-2 DAEs with or without constraint singularities. A complete theoretical analysis is performed for both cases. It is proved that the difference between the exact solution of a DAE and the corresponding iterate becomes $O(\epsilon^s)$ in magnitude at the sth iteration, at least away from the starting value of the independent variable t. Hence, the regularization parameter ϵ need not be very small so the regularized problems are less stiff. By some choice of parameters the regularized problems can be essentially nonstiff for any ϵ. As an example, a simple difference scheme for solving the regularized problems is investigated. Implementation techniques are discussed to get an approximation in the whole region for boundary value problems and to economize storage for initial value problems. Numerical experiments support our theoretical results. Numerical examples also show that usual stabilization methods do not work for problems with constraint singularities. Most parts of this chapter are taken from the paper [11].

In Chapter 3, we extend the SRM to nonlinear problems and to DAEs with index higher than 2. Again, nonstiffness of the regularized problems is achieved. Rather than having one "winning" method, this is a class of methods from which a number of variants are singled out as being particularly effective methods in certain circumstances of practical interest. In the case of no constraint singularity we prove convergence results. The method is also applied to constrained multibody systems. Numerical experiments confirm our theoretical predictions and demonstrate the viability of the proposed methods. Most parts of this chapter are taken from the paper [12].

In Chapter 4, we generalize the SRM to PDAEs, in particular, to the nonstationary Navier-Stokes equations. The convergence rate $O(\epsilon^s)$ at the sth iteration is again proved for this PDAE case in appropriate norms. The SRM not only avoids the stiffness of the regularized problems which always occurs in pseudo-compressibility methods but also avoids providing an unphysical pressure boundary condition which has to be imposed in stabilization methods. Discretization and implementation issues of the SRM are considered as well. In particular, a simple explicit difference scheme is analyzed and its stability is proved under the usual step size condition (independent of the regularization parameter ϵ) of explicit schemes. The stability result also indicates that the step size restriction can be relaxed as the viscosity becomes small. A numerical example is calculated to demonstrate these results. The SRM formulation is new in the Navier–Stokes context and it performs well. Most parts of the chapter are taken from the paper [79].

In Chapter 5, we apply the idea of the SRM to the simulation of miscible displacement in porous media. The problem is modeled by a nonlinear coupled system of two partial differential equations: the pressure-velocity equation and the concentration equation. Only the approximation of velocity is important for the approximation of concentration. The SRM idea is used for the pressure-velocity equation. An $O(\epsilon^s)$ error estimate at the sth SRM iteration is also proved. A Galerkin finite element method is used for the discretization of the SRM formulation. It is capable of producing as accurate a velocity approximation as the mixed finite element method. But unlike the mixed finite element method its stiffness matrix is symmetric positive definite and its finite element spaces need not satisfy the so-called Babuska-Brezzi condition. Most parts of this chapter are taken from the report [82].

Chapter 6 is devoted to numerical methods of some special cases of singular perturbation equations in the form of (1.12) or (1.13). Sections §6.1 and §6.3 describe a collection of papers [79, 115, 107] which reflects the author's earlier research interests. We believe that, by considering these special cases, we make steps towards the general problems (1.12) or (1.13) which are undoubtedly very difficult. Also, these special cases have practical meaning in themselves, hence, are worthwhile to be considered independently. In §6.1, we consider the one dimensional steady-state case of (1.12). §6.2 covers a special two-dimensional steady-state instance of (1.12) given in [121] to show that upwind schemes may lead to spurious solutions even for problems with very smooth solutions. We indicate that this is actually an ill-posed problem when ϵ is small. Hence, it is not strange that a direct discretization to the problem fails. In §6.3, we consider the linear one dimensional time-dependent case of (1.13). In this case, derivatives of the reduced problem may be discontinuous along the characteristic curves.

Finally, conclusions and possible future work are contained in Chapter 7.

Chapter 2

Sequential Regularization Methods for Differential Algebraic Equations

2.1 Motivation of the SRM for General High Index DAEs

The sequential regularization method (SRM) is motivated from the augmented Lagrangian method applied to constrained multibody systems (index-3 DAEs in general) by Bayo and Avello [19] and an earlier paper [20]. So we start this chapter by considering a mechanical system whose configuration is characterized by the generalized coordinates q. Let L be the system Lagrangian, defined by

$$L = T - V,$$

where T and V are the kinetic energy and the potential energy, respectively. Let Q represent non-conservative forces.

Usually the Lagrangian coordinates are not independent, but rather are interrelated through certain constraint conditions. When the connections between bodies are of holonomic type, these constraint conditions can be expressed mathematically in the following form:

$$\Phi(q,t) = 0. \tag{2.1}$$

Then Hamilton's principle leads to the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial q'}\right) - \frac{\partial L}{\partial q} + \Phi_q^T \lambda = Q, \tag{2.2}$$

where λ is a vector function whose components are Lagrange multipliers. For general multibody systems, (2.2) becomes

$$M(q)q'' + \Phi_q^T \lambda = f(q,q') \tag{2.3}$$

with common initial conditions, where M is the mass matrix and Φ_q is the Jacobian of the constraint equations. (2.3) and (2.1) form the Euler-Lagrange equations for a constrained multibody system. This is an index-3 DAE if Φ_q has full rank.

We have indicated that direct discretization would not be good in general for such a higher index DAE. One usual way to treat this problem is index reduction (to an index-1 DAE or an ODE). The most straightforward transformation of the DAE (2.3), (2.1) to an index-1 DAE involves replacing the constraint (2.1) with its second derivative plus initial conditions:

$$\Phi'' = \frac{d^2\Phi(q(t),t)}{dt^2} = 0, \tag{2.4a}$$

$$\Phi(q(0),0) = \frac{d\Phi(q(0),0)}{dt} = 0, \tag{2.4b}$$

(cf. (1.6)). However, this causes well-known drift difficulties, i.e. the numerical solution of (2.3), (2.4) may drift away from the original constraints (2.1) as time proceeds. Hence we have to look for stabilized index reduction methods. A very popular method called Baumgarte's method proposed in 1972 [17] is a generalization of (2.4). It replaces (2.4a) with the equation

$$\Phi'' + a\Phi' + b\Phi = 0, \tag{2.5}$$

where a and b are parameters chosen so that the roots of the polynomial

$$\sigma(\tau) = \tau^2 + a\tau + b = 0, \tag{2.6}$$

are both negative, i.e. the initial value problem for the differential equation (2.5) for Φ is asymptotically stable (see [16]). The system (2.5), (2.3) can be written in the form

$$\begin{bmatrix} M & \Phi_q^T \\ \Phi_q & 0 \end{bmatrix} \begin{bmatrix} q'' \\ \lambda \end{bmatrix} = \begin{bmatrix} f(q,q') \\ -(\Phi_q)'q' - (\Phi_t)' - a\Phi' - b\Phi \end{bmatrix}. \tag{2.7}$$

The matrix

$$\begin{bmatrix} M & \Phi_q^T \\ \Phi_q & 0 \end{bmatrix} \tag{2.8}$$

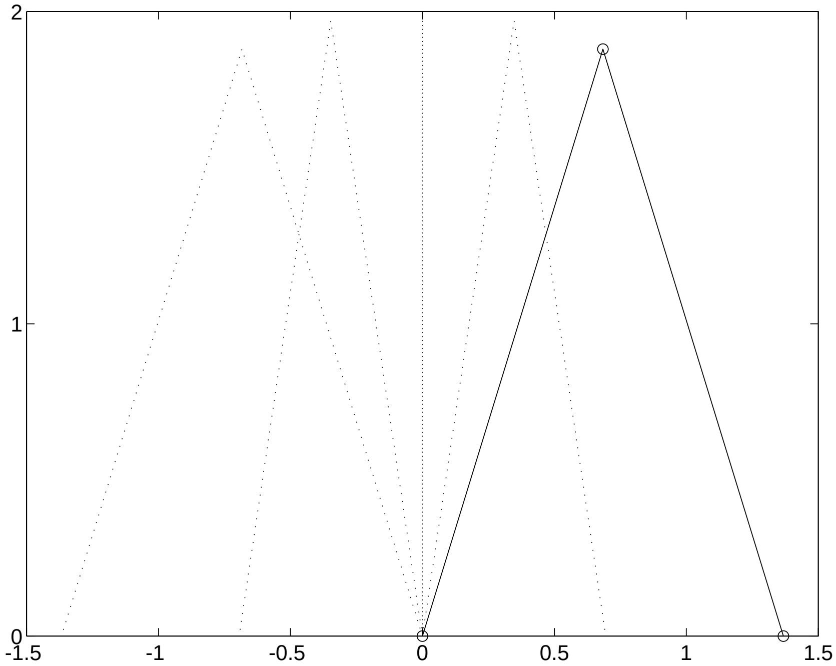

Figure 2.1: planar slider-crank: initial state in solid line, subsequent states in dotted lines

is nonsingular if Φ_q has full row rank. Hence (2.7) can be integrated using standard numerical integrators. If Φ_q is rank-deficient, we have a potential difficulty in solving (2.7). Baumgarte's method may not work then. The problem is called singular if Φ_q does not have full rank.

Example 2.1 *Consider two linked bars (see Fig. 2.1). One end of one bar is fixed at the origin, allowing only rotational motion in the plane. The other end of the other bar slides on the x-axis. The equations of motion form a nonlinear index-3 DAE*

$$\begin{aligned} p' &= v \\ Mv' &= f - G^T\lambda \\ g(p) &= 0 \end{aligned}$$

where x_i, y_i, ϕ_i are the coordinates of the center of mass of the ith bar, and

$$p = (x_1, y_1, \phi_1, x_2, y_2, \phi_2)^T.$$

If the left bar is strictly shorter than the right bar, then the Jacobian matrix G of the constraint functions of this problem has full rank. The problem is nonsingular. If the length of these two bars are the same, for example, each with length 2 and mass 1, then we have

$$M = diag\{1, 1, 1/3, 1, 1, 1/3\}$$

$$f = (0, -9.81, 0, 0, -9.81, 0)^T$$

$$g = \begin{pmatrix} x_1 - \cos\phi_1 \\ y_1 - \sin\phi_1 \\ x_2 - 2x_1 - \cos\phi_2 \\ y_2 - 2y_1 - \sin\phi_2 \\ 2\sin\phi_1 + 2\sin\phi_2 \end{pmatrix} \qquad G = g_p = \begin{pmatrix} 1 & 0 & \sin\phi_1 & 0 & 0 & 0 \\ 0 & 1 & -\cos\phi_1 & 0 & 0 & 0 \\ -2 & 0 & 0 & 1 & 0 & \sin\phi_2 \\ 0 & -2 & 0 & 0 & 1 & -\cos\phi_2 \\ 0 & 0 & 2\cos\phi_1 & 0 & 0 & 2\cos\phi_2 \end{pmatrix}$$

Clearly, as the mechanism moves left through the point where both bars are upright ($\phi_1 = \frac{\pi}{2}$, $\phi_2 = \frac{3\pi}{2}$) the last row of G vanishes at this one point and a singularity is obtained. When arriving at this point with no momentum, this is actually a bifurcation point where two subsequent motion configurations are possible. We will consider only the case where the sliding bar continues to slide along the x-axis past the singularity, and note that the solution is smooth in the passage through the singularity. □

In [19], Bayo and Avello proposed to solve the multibody system (2.3) using an augmented Lagrangian algorithm which is transplanted from the same method in the optimization context [5]. Their idea is to derive a modified formulation by adding to the expression of Hamilton's principle three terms:

- a fictitious potential

$$V^* = \sum_k \frac{1}{2}\alpha_k \omega_k^2 \Phi_k^2 \tag{2.9}$$

- a set of Rayleigh dissipative forces

$$G_k = -2\alpha_k \omega_k \mu_k \Phi_k' \tag{2.10}$$

- a fictitious kinematic energy term

$$T^* = \sum_k \frac{1}{2}\alpha_k \Phi_k'^2, \tag{2.11}$$

where each α_k is a very large real number (the penalty), and ω_k and μ_k represent the natural frequency and the damping ratio of the penalized system (mass, dashpot and spring) corresponding to the constraint $\Phi_k = 0$. Then we get a modified Euler-Lagrange equation

$$Mq'' + \Phi_q^T \alpha(\Phi'' + 2\Omega\mu\Phi' + \Omega^2\Phi) + \Phi_q^T \lambda^* = f(q, q') \tag{2.12}$$

or

$$(M + \Phi_q^T \alpha \Phi_q)q'' + \Phi_q^T \lambda^* = f(q, q') - \Phi_q^T \alpha((\Phi_q^T)'q' + (\Phi_t)' + 2\Omega\mu\Phi' + \Omega^2\Phi), \tag{2.13}$$

where α, Ω and μ are diagonal matrices that contain the values of the penalties, the natural frequencies and the damping ratios of the fictitious penalty systems assigned to each constraint condition. Because λ^* is not given in advance Bayo, Jalon and Serna [20] propose an iteration to solve (2.12) or (2.13) by comparing (2.12) with (2.3):

$$\lambda_s^* = \lambda_{s-1}^* + \alpha(\Phi'' + 2\Omega\mu\Phi' + \Omega^2\Phi)|_{q=q_{s-1}}, s = 1, 2, \cdots \tag{2.14}$$

to get an approximation of λ.

In [19] the authors called the whole procedure an augmented Lagrangian algorithm and claimed that the algorithm works well, however without any theoretical analysis.

In fact, the algorithm would certainly not work in general. In multibody motions q usually remains smooth even in passage through singular positions. However, as indicated in later sections, λ may be unbounded at the singularity. So the iteration (2.14), as an approximate procedure to obtain λ, is not appropriate in some cases. Also, Φ'' should not be included in (2.12) in the singular case because unbounded coefficients may appear in it. From (2.14) the formulation corresponds to Baumgarte's formulation since when λ_s^* gets close to λ_{s-1}^* (2.14) gets close to Baumgarte's stabilization. But Baumgarte's stabilization is not as good as some other stabilization techniques (see [8]). Moreover, in [19] the authors indicated that the iteration (2.14) is applied until $\|q_s'' - q_{s-1}''\| < \delta$, where δ is a user-specified tolerance. This criterion is perhaps applied because they did not know the convergence rate of their iterative procedure. We do not recommend this criterion because it causes not only unnecessary extra iterations but also makes a storage-saving implementation difficult (cf. §2.6).

Our aim is to construct a method for general DAEs which not only avoids the above shortcomings, but for which we also do not have a Lagrangian and Hamilton's principle. So another derivation of the algorithm is needed.

In this and the next chapter, we propose a class of algorithms motivated by the augmented Lagrangian method for more general DAEs of order ν,

$$x^{(\nu)} = f(x, x', \ldots, x^{(\nu-1)}, t) - B(x,t)y, \tag{2.15a}$$

$$0 = g(x,t). \tag{2.15b}$$

The DAE (2.15) has index $\nu+1$ if GB is nonsingular for all t, $0 \le t \le t_f$, where $G = g_x$. We are interested in the cases $\nu = 1$ or 2. The Euler-Lagrange equations (2.3) for mechanical systems with holonomic constraints are in this form with $\nu = 2$. The algorithm is derived by combining a modified penalty idea (a kind of regularization) given in [91] with stabilization techniques such as Baumgarte's stabilization or the stabilization analyzed in [8, 35] in an iterative procedure. We call the method the *sequential regularization method* (SRM) (cf. [11, 12]). The method is applicable for more general higher index DAEs. More importantly, it works for both boundary and initial value problems and is justified by a theoretical analysis. The number of iterations can be determined beforehand depending on the penalty parameter $\alpha = 1/\epsilon$, the mesh size h and the order of the method used. Since we specify the iteration number in advance we can design a procedure for the initial value case to perform the iteration without the need to store all previous approximate values.

The sequential regularization method is actually a functional iteration procedure in which the difference between the exact solution of a DAE and the corresponding iterate becomes $O(\epsilon^s)$ in magnitude at the sth iteration, at least away from the starting value of the independent variable (which we shall call 'time'). Hence, unlike the usual regularization, the perturbation parameter ϵ does not have to be chosen very small, so the regularized problems can be less stiff and/or more stable.

Next we will propose and analyze the SRM for the linear index-2 case with singularities. Numerical experiments are given to verify our theoretical results. Some simple difference schemes for the regularized problems and implementation issues for the SRM are also discussed.

2.2 Linear Index-2 Problems

We first write down the linear index-2 problem:

$$x' = A(t)x - B(t)y + q(t), \tag{2.16a}$$

$$0 = G(t)x + r(t) \equiv g(x,t), \tag{2.16b}$$

where $A(t)$, $B(t)$ and $G(t)$ are sufficiently smooth functions of $t, 0 \le t \le t_f$, $A(t) \in \mathbf{R}^{n_x \times n_x}$, $B(t) \in \mathbf{R}^{n_x \times n_y}$, $G(t) \in \mathbf{R}^{n_y \times n_x}$, and $n_y \le n_x$. We consider the DAE (2.16) subject to $n_x - n_y$ boundary conditions

$$\bar{B}_0 x(0) + \bar{B}_1 x(t_f) = \beta. \tag{2.17}$$

These boundary conditions are assumed to be such that they yield a unique solution for the ODE (2.16a) on the manifold given by (2.16b). In particular, if we were to replace (2.16b) by its differentiated form

$$0 = Gx' + G'x + r' = \frac{d}{dt} g(x,t), \tag{2.18a}$$

$$g(x(0),0) = G(0)x(0) + r(0) = 0, \tag{2.18b}$$

and use (2.18a) in (2.16a) to eliminate y and obtain n_x ODEs for x, then the boundary value problem for x with (2.17) and (2.18b) specified has a unique solution. In the initial value case $\bar{B}_1 = 0$, this means that (2.17) and (2.18b) can be solved uniquely for $x(0)$. We will give a more precise assumption in Lemma 2.1 below. The problem (2.16), (2.17) is of index-2 if GB is nonsingular for all t. However, here we allow GB to be singular at some isolated points of t. For simplicity of exposition, let us say that there is one singular point $t_*, 0 < t_* < t_f$. The inhomogeneities are $q(t) \in \mathbf{R}^{n_x}$ and $r(t) \in \mathbf{R}^{n_y}$. We are only interested in the kind of singularities as in Example 2.1, where the solution x of (2.16), (2.17) passes through the singularity smoothly.

Returning to Example 2.1 (where $B = M^{-1}G^T$), we can verify that, although the matrix $GM^{-1}G^T$ is singular at the singularity , the matrix $M^{-1}G^T(GM^{-1}G^T)^{-1}G$ is smooth for all t. Also, two types of singular constraints (i.e., with vanishing rows or with some rows linearly dependent at some points) mentioned in [2] both have a similar property. Thus, for the linear model (2.16), we assume accordingly:

Assumption 2.1 *The matrix function $P = B(GB)^{-1}G$ is smooth, or more precisely, P is continuous and P' is bounded near the singular point t_* where we define*

$$P(t_*) = \lim_{t \to t_*} (B(GB)^{-1}G)(t).$$

Because we are only interested in the case where (2.16) has a smooth solution for x (as is the case in Example 2.1), it is necessary to assume, in view of (2.16b):

Assumption 2.2 *The inhomogeneity $r(t)$ satisfies $r \in \mathcal{S}$, where*

$$\mathcal{S} = \{w(t) \in \mathbf{R}^{n_y} : Gz = w \textit{ for a smooth function } z(t).$$

We note that Assumptions 2.1 and 2.2 are satisfied automatically if GB is nonsingular for each t. On the other hand, neither $B(GB)^{-1}$ nor $(GB)^{-1}G$ alone are smooth near a singularity in general. We also indicate here that to formulate the SRM (see §2.4) we only need the continuity of P. The further requirement in Assumption 2.1 on the derivative of P is needed for the regularity of the solution (cf. Lemma 2.1 and (2.24)) and the stability proof for the regularized problems (cf. §2.5). This requirement can be avoided if we make a more general assumption about the regularity and stability of the original problem (cf. §3.1).

We consider both initial and boundary value problems. In §2.3 we briefly discuss the conditioning of the problem (2.16) with singularities. In §2.4 we derive the sequential regularization method. In §2.5 we estimate the error of the SRM. In §2.6, we consider some discretization and implementation issues for both initial and boundary value problems. Finally, in §2.7 several numerical examples demonstrate our theoretical results.

2.3 Problem Conditioning

Similarly to [15] and to the method of pseudo upper triangular decomposition (PUTD) described in [2] (cf. §10.6; with the difference that we do pivoting to interchange the row with the singularity of the lowest order and the current row when all the other rows vanish at some singular point), there exists a smooth matrix function $R(t) \in \mathbf{R}^{(n_x-n_y)\times n_x}$, which has full row rank and satisfies

$$RB = 0, \quad \text{for each } t, \ 0 \le t \le t_f,$$

where R can be taken to have orthonormal rows.

As in [15, 16], define the new variable

$$u = Rx, \ 0 \le t \le t_f. \tag{2.19}$$

Then, using (2.16b), the inverse transformation is given by

$$x = Su - B(GB)^{-1}r, \tag{2.20}$$

where

$$S = (I - B(GB)^{-1}G)R^T = (I-P)R^T.$$

By the assumptions at the beginning of this chapter, this transformation is well-defined. Differentiating (2.19) and using (2.16a) and (2.20) we obtain the essential underlying ODE (EUODE):

$$u' = (RA + R')Su - (RA + R')B(GB)^{-1}r + Rq. \tag{2.21}$$

Hence the underlying problem of (2.21) is

$$u' \;=\; (RA + R')Su + f, \tag{2.22a}$$

$$\bar{B}_0 S(0)u(0) + \bar{B}_1 S(t_f)u(t_f) = \beta_1. \tag{2.22b}$$

We make

Assumption 2.3 *The boundary value problem (2.22) is stable, i.e. there exists a moderate-size constant K such that*

$$\|u\| \le K(\|f\| + |\beta_1|),$$

where $\|u\| = max_t\{|u(t)|, \ 0 \le t \le t_f\}$.

Similarly to Theorem 2.2 of [15], we have

Lemma 2.1 *Let the DAE (2.16) have smooth coefficients, and assume that Assumptions 2.1 and 2.2 hold. If the EUODE (2.21) with boundary conditions (2.22b) has a unique solution, then there exists a unique solution for the x of problem (2.16)-(2.17) which is smooth. This implies the unique existence of a smooth By as well. Furthermore, if Assumption 2.3 holds then there is a constant K such that*

$$\|x\| \leq K(\|q\| + \|B(GB)^{-1}r\| + |\beta|),$$

$$\|x'\| \leq K(\|q\| + \|B(GB)^{-1}r\| + \|(B(GB)^{-1}r)'\| + |\beta|).$$

Remark 2.1 *For problem (2.16) without singularities, we can get*

$$\|x\| \leq K(\|q\| + \|r\| + |\beta|),$$

$$\|x'\| \leq K(\|q\| + \|r\| + \|r'\| + |\beta|)$$

as in [15, 16]. □

The difference between the situation here and in the nonsingular case is that the perturbation inhomogeneities r yield reasonably bounded perturbations in the solution x only if they are (in general) from the subspace *Range* (G).

From (2.16a) and (2.20), we can write

$$y = -(GB)^{-1}G(x' - Ax - q), t \in [0, t_*) \cup (t_*, t_f], \tag{2.23}$$

which could be unbounded at the singular point t_* (whereas By is bounded). Note that G could be replaced in (2.23) by any appropriate matrix Q with the same size as G, e.g. Q can be B^T.

Remark 2.2 *If B has full rank for each t, then*

$$(GB)^{-1}G = (B^T B)^{-1} B^T P.$$

Hence, $(GB)^{-1}G$ is smooth. Hence, there exists a unique solution for the y of problem (2.16)-(2.17) which is smooth and can be expressed as (2.23) for each t. Furthermore, using Lemma 2.1, we have in this case

$$\|y\| \leq K(\|q\| + \|B(GB)^{-1}r\| + \|(B(GB)^{-1}r)'\| + |\beta|). \tag{2.24}$$

In the general case, however, we will have to consider By, rather than y alone, in the theorems of the next section. $\square$

A Baumgarte stabilization applied to (2.16) consists of eliminating y according to (2.18),(2.23), and stabilizing (see (2.30) below for the usual form). This gives the ODE

$$x' = (I - B(GB)^{-1}G)(Ax + q) - B(GB)^{-1}(G'x + r') - \epsilon^{-1}B(GB)^{-1}(Gx + r) \tag{2.25}$$

where $\epsilon > 0$ is a parameter (cf. [17, 8]). If there are no singularities then it follows from the analysis in [16] that if Assumption 2.3 holds then the boundary value problem (2.25),(2.17),(2.18b) is also stable. In other words, the "initial value stabilization" works also for the boundary value case, because the new modes introduced by replacing (2.16b) with (2.18a) are separable and decaying, in agreement with the additional *initial* conditions (2.18b).

However, in the singular case (2.25) may not work because the terms $B(GB)^{-1}G'$ and $B(GB)^{-1}r'$ are in general unbounded. Therefore, we develop an iterative method in the next section which builds up an approximation to By and x that avoids going through unbounded quantities.

2.4 Derivation of the SRM

There are many discussions on regularization methods for DAEs. A direct regularization (cf. the pseudo-compressibility method in the Navier-Stokes context [108, 72, 97]) is:

$$x' = A(t)x - B(t)y + q(t), \tag{2.26a}$$

$$-\epsilon y' = G(t)x + r(t). \tag{2.26b}$$

This formulation is not popular because it requires conditions on A, B and G, for the purpose of stability of the system, and the existence of the first derivative of y, which is not necessarily true for the original problem (2.16) [86]. It may also change the properties of the original index-2 problem too much by jumping from index-2 to index-0 (ODE). It seems evident that a regularization with fewer changes of the original problem (e.g. from index-2 to index-1) might be better. The penalty method [84, 91, 68, 70] is such a method. It reads

$$x' = A(t)x - B(t)y + q(t), \tag{2.27a}$$

$$\epsilon E^{-1} y = G(t)x + r(t), \tag{2.27b}$$

where $E \in \mathbf{R}^{n_y \times n_y}$ is chosen such that BEG has non-negative eigenvalues. Hence, the system obtained by substituting (2.27b) into (2.27a) is generally stable. For example, we can choose, relying on Assumption 2.1, $E = (GB)^{-1}$(hence, $BEG = P$). Also, $E = (GB)^T$ could be a good choice in some circumstances. If $B = M^{-1}G^T$ for some positive definite matrix M (as in the case of mechanical systems) then it is possible to choose $E = I$. Advantages of these choices of E will be discussed in Chapter 3. For problems with singularities, we suggest using $E = (GB)^{-1}$ to avoid a turning point problem. Another approach is parameterization [85, 59]:

$$x' = A(t)x - B(t)y + q(t), \tag{2.28a}$$

$$0 = G(t)(x + \epsilon x') + r(t). \tag{2.28b}$$

Substituting (2.28a) to (2.28b), we get

$$\epsilon GBy = Gx + r + \epsilon GAx. \tag{2.29}$$

This implies that parameterization can be seen as an instance of the penalty method with $E = (GB)^{-1}$. Recently, [94] has reported a regularization for DAEs with singularities based on a formulation obtained by the trust-region method in numerical optimization. All these regularizations require the regularization parameter ϵ to be very small. Therefore a stiff solver is needed to solve the regularized problem. In this section, we derive a new regularization method which is called the *sequential regularization method* [11]. The SRM is an iterative procedure which combines the popular Baumgarte stabilization or other stabilizations with a modified penalty method. One purpose for doing so is that the regularized problems of the SRM can be non-stiff or at least less stiff. Hence, simple discrete schemes (e.g. explicit schemes) can be used. Other advantages of the method will be discussed in Chapter 3.

The Baumgarte stabilization of (2.16) reads (cf. (2.5))

$$\alpha_1 \frac{d}{dt} g(x,t) + \alpha_2 g(x,t) = 0, \; g(x(0),0) = 0. \tag{2.30}$$

Applying the idea of the penalty method to equations (2.16a) with constraints (2.30), we obtain

$$x' = A(t)x - B(t)y + q(t), \tag{2.31a}$$

$$y = y_0 + \frac{1}{\epsilon} E(\alpha_1 \frac{d}{dt} g(x,t) + \alpha_2 g(x,t)). \tag{2.31b}$$

where y_0 can be seen as an initial guess of the exact solution y_e of problem (2.16), (2.17). If we take $y_0 = y_e$ then the solution of problem (2.31), (2.17) is exactly the same as that of problem (2.16), (2.17). If we take $y_0 \equiv 0$ then (2.31) coincides with the penalty method (2.27). Given any initial guess $y_0(t)$, the solution, say $\{x_1, y_1\}$, of (2.31), (2.17) is an approximation of the exact solution $\{x_e, y_e\}$ of (2.16), (2.17). Using this solution y_1 as a new initial guess, we re-solve problem (2.31), (2.17). We expect that the solution obtained is a better approximation of the exact solution. Repeating the procedure, we invent the following iterative algorithm for solving (2.16):

For $s = 1, 2, \dots$, solve the ODE problem

$$x_s{}' = Ax_s - By_s + q \tag{2.32}$$

where

$$y_s = y_{s-1} + \frac{1}{\epsilon}E(\alpha_1 \frac{d}{dt}g(x_s, t) + \alpha_2 g(x_s, t)), \tag{2.33}$$

subject to the same boundary conditions (2.17). Note that $y_0(t)$ is a given initial iterate and that $\epsilon > 0$ is the regularization parameter.

We call this algorithm a *sequential regularization method* (SRM). Note that $x_s(t)$ and $y_s(t)$ are defined on the entire interval $[0, t_f]$ for each s. For the problem with singularities, the choice $E = (GB)^T$ generates turning point regularized problems which are complicated to solve and analyze. We thus choose $E = (GB)^{-1}$ for the singular case. Noting that $B(GB)^{-1}G'$ may be unbounded at the singularity we then choose $\alpha_1 = 0$ to avoid this term. Also, in practice we multiply (2.33) by B and keep track only of the approximations By_s to the bounded function By, since y may be unbounded at the singularity. We thus have an SRM variant for the singular case:

For $s = 1, 2, \dots$, solve the ODE problem

$$x_s{}' = Ax_s - By_s + q \tag{2.34}$$

where

$$By_s = By_{s-1} + \frac{1}{\epsilon}BEg(x_s, t), \tag{2.35}$$

subject to the boundary conditions (2.17).

If y is desired (at times other than at the singular point t_*) then it can be easily retrieved from By in a post-processing step.

2.5 Convergence Analysis of the SRM

We first prove a lemma which will be used to discuss the convergence of the SRM.

Lemma 2.2 *Let u, v be the solution of*

$$\begin{aligned} u' &= (RA + R')Su + S_1 v + f_1, && \text{(2.36a)} \\ \delta v' + \gamma v &= \epsilon S_2 u + \epsilon S_3 v + f_2, && \text{(2.36b)} \\ \bar{B}_0 S(0) u(0) + \bar{B}_1 S(t_f) u(t_f) &= \beta - S_4 v(0) - S_5 v(t_f),\ v(0) = v_0, && \text{(2.36c)} \end{aligned}$$

where all coefficients are bounded, $\delta = 1$ or $\delta = \epsilon$, γ is a positive constant and Assumption 2.3 holds. Then, for ϵ appropriately small or γ appropriately large, we have the following stability inequality

$$\|u\| \le K(\|f_1\| + \|f_2\| + |\beta| + |v_0|),$$

$$\|v\| \le K(\epsilon\|f_1\| + \|f_2\| + |\beta| + |v_0|),$$

where K is a positive constant.

Proof: Let $v = (v_1, \cdots, v_{n_y})^T$. From (2.36b), we easily have

$$|v_i| \le \frac{\epsilon}{\gamma}\|S_2\|\|u\| + \frac{\epsilon}{\gamma}\|S_3\|\|v\| + \frac{1}{\gamma}\|f_2\| + |v_0|,\ i = 1, \cdots, n_y.$$

Hence, taking the maximum of the left hand side for $1 \le i \le n_y$ and choosing small ϵ or large γ appropriately such that $\epsilon\|S_3\| < \gamma$, we get

$$\|v\| \le \frac{\epsilon}{\gamma - \epsilon\|S_3\|}\|S_2\|\|u\| + \frac{\|f_2\| + \gamma|v_0|}{\gamma - \epsilon\|S_3\|}. \tag{2.37}$$

By using Assumption 2.3, from (2.36a), there exists a positive constant K_1 such that

$$\|u\| \le K_1(\|S_1\|\|v\| + \|f_1\| + |\beta| + |S_4||v(0)| + |S_5||v(t_f|)$$

$$\le K_1((\|S_1\| + |S_5|)\|v\| + \|f_1\| + |\beta| + |S_4||v_0|)$$

$$\le \frac{K_1\epsilon(\|S_1\| + |S_5|)\|S_2\|}{\gamma - \epsilon\|S_3\|}\|u\| + \frac{K_1(\|S_1\| + |S_5|)(\|f_2\| + \gamma|v_0|)}{\gamma - \epsilon\|S_3\|} + K_1(\|f_1\| + |\beta| + |S_4||v_0|).$$

Hence, by choosing smaller ϵ or larger γ such that $\frac{K_1\epsilon(\|S_1\|+|S_5|)\|S_2\|}{\gamma-\epsilon\|S_3\|} < 1$, the stability inequality for u follows. Now the stability inequality for v follows from that for u and (2.37). □

Now we estimate the error of the SRM (2.34), (2.35).

Definition 2.1 *J is an integer such that*

$$y_0(0) = y_e(0),\ y_0'(0) = y_e'(0), \ldots, y_0^{(J)}(0) = y_e^{(J)}(0),$$

where $y_0(t)$ is the initial guess of the SRM iteration (2.34), (2.35) and $y_e(t)$ is the exact solution of the original problem (2.16). Set $J = -1$ if $y_0(0) \neq y_e(0)$. □

For initial value problems we may calculate ${y_e}^{(i)}(0)$, $i = 0, 1, \ldots$ in advance by using the ODE and its derivatives. For boundary value problems we have $J = -1$ in general since we don't know $y_e(0)$ beforehand.

Theorem 2.1 *Let the DAE (2.16) have sufficiently smooth coefficients, and assume that Assumptions 2.1, 2.2 and 2.3 hold. Then, for the solution of iteration (2.34),(2.35), we have the following error estimates (for J defined in Definition 2.1):*

$$\begin{aligned} x_s(t) - x_e(t) &= O(\epsilon^s) + O(\epsilon^{J+2} p_s(t/\epsilon) e^{-t/\epsilon}), \\ By_s(t) - By_e(t) &= O(\epsilon^s) + O(\epsilon^{J+1} p_s(t/\epsilon) e^{-t/\epsilon}), \end{aligned}$$

for $0 \le t \le t_f$ and $s \ge 1$. Here $p_s(\tau) \equiv 0$ if $s \le J+1$; otherwise $p_s(\tau)$ is a polynomial of degree $s - J - 2$ with generic positive coefficients and $|p_s(0)| = |(By_0)^{(J+1)}(0) - (By_e)^{(J+1)}(0)|$.

Proof: Let $u_s = Rx_s$ and $w_s = Px_s$. Similarly to (2.20), we have

$$x_s = Su_s + w_s. \tag{2.38}$$

Furthermore, using (2.34) we obtain

$$\begin{aligned} u_s' &= (RA + R')Su_s + (RA + R')w_s + Rq, & (2.39a) \\ \epsilon w_s' + w_s &= \epsilon(PA + P')Su_s + \epsilon(PA + P')w_s - \epsilon By_{s-1} & (2.39b) \\ &+ \epsilon Pq - B(GB)^{-1}r, \end{aligned}$$

subject to

$$\begin{aligned} \bar{B}_0 S(0)u_s(0) + \bar{B}_1 S(t_f)u_s(t_f) &= \beta - \bar{B}_0 w_s(0) - \bar{B}_1 w_s(t_f), & (2.40a) \\ w_s(0) &= -B(0)(G(0)B(0))^{-1}r(0). & (2.40b) \end{aligned}$$

The iteration (2.35) for By becomes

$$By_s = By_{s-1} + \frac{1}{\epsilon}(w_s + B(GB)^{-1}r). \tag{2.41}$$

The proof proceeds along familiar lines of singular perturbation analysis. According to [111, 112] we can construct the asymptotic expansion of w_s and u_s sequentially for $s = 1, 2, \ldots$, where we use Lemma 2.2 to estimate the remainders. Then, using (2.41) and (2.38), we get the asymptotic expansion of By_s and x_s respectively. Note that in these expansions the first terms are exactly x_e and By_e. This process eventually yields the proof of the theorem. □

To provide a better understanding about the sequential regularization method we give in §2.8 a detailed proof for the initial value case with no layers, $s \leq J + 1$. In that proof, the construction of the asymptotic expansion is directly for x and By. Moreover, the construction method we apply is somewhat different from [111, 112] and more relevant to the concept of DAEs.

Next, we consider the SRM (2.32), (2.33). For the initial value problem with $E = I$ and $B = G^T$, this corresponds to Algorithm (2.13), (2.14) of [19] for constrained mechanical systems (although they do it for the corresponding index-3 case) derived by a penalty-augmented Lagrangian formulation. Bayo and Avello have noted that under repetitive singular conditions this algorithm may lead to unstable behavior. For our index-2 case (2.16) with singularity, it appears to be impossible to choose a matrix E such that problem (2.32),(2.33) is always stable, even if we assume $B = G^T$. A numerical example in §2.7 will verify such instability phenomena even for the case of one singular point. However, for the case where constraints are without singularities, (2.33) (multiplied by B) may have a benefit over (2.35). That is, (2.33) yields an ODE problem for x_s which is essentially not a stiff problem. Take $E = (GB)^{-1}$ as before and rewrite (2.33) as

$$By_s = By_{s-1} + \frac{1}{\epsilon} BE(\alpha_1 \frac{d}{dt} g(x_s, t) + \alpha_2 g(x_s, t)). \tag{2.42}$$

Then we give the following error estimation for (2.32), (2.42):

Theorem 2.2 *Let the DAE (2.16) have sufficiently smooth coefficients, and assume that G has full rank and that Assumptions 2.1, 2.2 and 2.3 hold. Then for the solution of the iteration procedure (2.32), (2.42) with $\alpha_1 \neq 0$, we have the following error estimates:*

$$x_s - x_e = O(\epsilon^s),$$

$$By_s - By_e = O(\epsilon^s)$$

for $0 \leq t \leq t_f$ and $s = 1, 2, \ldots$. Note that no boundary layer terms appear here even for $J = -1$ (See Definition 2.1)!

Proof: Denote $u_s = Rx_s$ and $v_s = Gx_s$. Hence

$$x_s = Su_s + Fv_s, \tag{2.43}$$

where $S = (I - P)R^T$ and $F = B(GB)^{-1} = PG^T(GG^T)^{-1}$ are both sufficiently smooth. From (2.32),(2.42), we get

$$\begin{aligned} u_s' &= (RA + R')Su_s + (RA + R')Fv_s + Rq, \\ (\epsilon + \alpha_1)v_s' + \alpha_2 v_s &= \epsilon(G' + GA)Su_s + \epsilon(G' + GA)Fv_s + \epsilon GBy_{s-1} + \epsilon Gq - \alpha_1 r' - \alpha_2 r, \end{aligned}$$

with the corresponding boundary conditions, and

$$By_s = By_{s-1} + \frac{1}{\epsilon}B(GB)^{-1}(\alpha_1(v_s + r)' + \alpha_2(v_s + r)).$$

Repeating the procedure of the proof of Theorem 2.1 and using Lemma 2.2 again to estimate the remainder of the asymptotic expansion, we obtain

$$\begin{aligned} u_s - u_e &= O(\epsilon^s), \\ v_s - v_e &= O(\epsilon^s), \\ By_s - By_e &= O(\epsilon^s), \end{aligned}$$

where $u_e = Rx_e,\ v_e = Gx_e = -r$. Hence, using (2.43) and $x_e = Su_e + Fv_e$, we obtain

$$x_s - x_e = S(u_s - u_e) + F(v_s - v_e) = O(\epsilon^s).$$

□

Remark 2.3 *For the problem (2.32), (2.33) where GB is nonsingular, we have*

$$x_s - x_e = O(\epsilon^s), \;\; y_s - y_e = O(\epsilon^s).$$

This estimate also holds for $E = (GB)^T$. □

2.6 Discretization and Implementation Issues

The SRM iteration yields a sequence of ODE problems which are to be solved numerically. We only consider the most difficult case, i.e. (2.34), (2.35) with singularities.

Inserting (2.35) into (2.34), the ODE problem to be solved at the sth iteration is written as the singular-singularly-perturbed problem (see [112, 89])

$$\epsilon x_s' + BE(Gx_s + r) \;=\; \epsilon A x_s - \epsilon(By_{s-1} - q), \tag{2.44a}$$

$$\bar{B}_0 x_s(0) + \bar{B}_1 x_s(t_f) = \beta \quad , \quad G(0)x_s(0) + r(0) = 0. \tag{2.44b}$$

We consider finite difference (or collocation) discretizations of (2.44) on a mesh

$$\pi \;:\; 0 = t_0 < t_1 < \ldots < t_N = t_f$$

$$h_i = t_i - t_{i-1}, \qquad h = \max_{1\le i\le N} h_i$$

and denote by x_i^s, y_i^s the corresponding approximations of $x_s(t_i)$, $y_s(t_i)$, respectively. We now have essentially two small, positive parameters to choose: ϵ and h. We assume that h is chosen small enough so that the EUODE problem (2.22) may be considered as nonstiff and that the problem's coefficients are sufficiently smooth.

In the BVP case the situation is the familiar one, much like the iterative solution of a nonlinear boundary value ODE using quasilinearization (see, e.g., [14]). Each of the linear boundary value ODEs (2.44) is discretized on a mesh π using, say, a symmetric finite difference scheme or some other method. We expect, as $h \to 0$, convergence to the solution of (2.44) and our theory then applies for the entire numerical algorithm.

As an example, we give here a detailed analysis of the convergence of the backward Euler difference scheme for (2.44). A similar discussion and results can easily apply to the forward Euler difference scheme. The results for general higher order collocation schemes have been described in [11]. Now we write the backward Euler scheme of (2.44) as follows:

$$\epsilon \frac{x_i^s - x_{i-1}^s}{h_i} + B_i E_i (G_i x_i^s + r_i) \;=\; \epsilon A_i x_i^s - \epsilon(B_i y_{s-1}(t_i) - q_i), \tag{2.45a}$$

$$\bar{B}_0 x_0^s + \bar{B}_1 x_N^s \;=\; \beta,\; G_0 x_0^s + r_0 = 0, \tag{2.45b}$$

$$B_i y_i^s = B_i y_{s-1}(t_i) + \frac{1}{\epsilon} B_i E_i (G_i x_i^s + r_i), \tag{2.46}$$

where we represent the value $f(t_i)$ of a given function f at mesh point t_i by f_i. Multiplying (2.45) by R_i and P_i, respectively, and denoting

$$u_i^s = R_i x_i^s,\ w_i^s = P_i x_i^s$$

(then $x_i^s = S_i u_i^s + w_i^s$ since (2.38) holds), we have

$$\frac{u_i^s - u_{i-1}^s}{h_i} = R_i A_i (S_i u_i^s + w_i^s) + \epsilon \frac{R_i - R_{i-1}}{h_i}(S_{i-1}u_{i-1}^s + w_{i-1}^s) + R_i q_i, \tag{2.47a}$$

$$\begin{aligned} \epsilon \frac{w_i^s - w_{i-1}^s}{h_i} + w_i^s &= \epsilon P_i A_i (S_i u_i^s + w_i^s) + \epsilon \frac{P_i - P_{i-1}}{h_i}(S_{i-1}u_{i-1}^s + w_{i-1}^s) \\ &- \epsilon (B_i y_{s-1}(t_i) - q_i) - B_i E_i r_i,\ i = 1, \cdots, N, \end{aligned} \tag{2.47b}$$

$$\bar{B}_0(S_0 u_0^s + w_0^s) + \bar{B}_1(S_N u_N^s + w_N^s) = \beta,\ w_0^s = -B_0 E_0 r_0 = P_0 x^e(t_0). \tag{2.47c}$$

At first, we consider the stability of the following difference scheme corresponding to (2.47):

$$\begin{aligned} L_1^h u_i &\equiv \frac{u_i - u_{i-1}}{h_i} - R_i A_i S_i u_i - \frac{R_i - R_{i-1}}{h_i} S_{i-1} u_{i-1} = R_i A_i w_i \\ &+ \frac{R_i - R_{i-1}}{h_i} w_{i-1} + f_i, \end{aligned} \tag{2.48a}$$

$$\begin{aligned} L_2^h w_i &\equiv \epsilon \frac{w_i - w_{i-1}}{h_i} + w_i = \epsilon P_i A_i S_i u_i + \epsilon \frac{P_i - P_{i-1}}{h_i} S_{i-1} u_{i-1} \\ & \epsilon P_i A_i w_i + \epsilon \frac{P_i - P_{i-1}}{h_i} w_{i-1} + g_i,\ i = 1, \cdots, N, \end{aligned} \tag{2.48b}$$

$$\bar{B}_0 S_0 u_0 + \bar{B}_1 S_N u_N = \beta_1 = \beta - \bar{B}_0 w_0 - \bar{B}_1 w_N,\ w_0 = \beta_2. \tag{2.48c}$$

Using the discrete maximum principle for the difference operator L_2^h, i.e.

$$z_0 \geq 0 \text{ and } L_2^h z_i \geq 0 \Longrightarrow z_i \geq 0, \forall i, \tag{2.49}$$

we easily get

$$\max_{1 \leq j \leq i} |w_j| \leq M(|\beta_2| + \max_{1 \leq j \leq i} |L_2^h w_j|). \tag{2.50}$$

Defining $\|z\|_i = \max_{1 \leq j \leq i} |z_j|$ and using (2.48b), we have

$$\|w\|_i \leq M\epsilon(\|u\|_i + \|w\|_i) + M|\beta_2| + \|g\|_i$$

or

$$\|w\|_i \le M(\epsilon\|u\|_i + |\beta_2| + \|g\|_i). \tag{2.51}$$

Here M stands for a generic constant independent of i, ϵ and h. On the other hand, $L_1^h u_i$ is a one-step difference operator of (2.22a) — the underlying problem of (2.21). Using Assumption 2.3 and *Theorem 5.38* of [14], we obtain

$$\|u\|_\infty \le K_1(|\beta_1| + \max_{1\le j\le N} |L_1^h u_j|), \tag{2.52}$$

where $\|z\|_\infty = \max_{0\le j\le N} |z_j|$ and $K_1 = K + O(h)$. Here K is the stability constant defined in Assumption 2.3. Using (2.48a) and (2.48c) we have

$$\|u\|_\infty \le M(\|w\|_N + |\beta_2| + |\beta| + \|f\|_N). \tag{2.53}$$

Then, using (2.51), yields

$$\begin{aligned} \|u\|_\infty &\le M(\|f\|_N + \|g\|_N + \|\beta| + |\beta_2|) && (2.54a)\\ \|w\|_\infty &\le M(\epsilon\|f\|_N + \|g\|_N + \|\beta| + |\beta_2|). && (2.54b) \end{aligned}$$

We thus obtain the stability inequalities (2.54) for the difference scheme (2.48) or (2.47).

Now we discuss the convergence of (2.47). Using (2.54) , we have the estimates:

$$\begin{aligned} \|u_s(t_i) - u_i^s\|_\infty &\le M(\|\tau^u\|_N + \|\tau^w\|_N) && (2.55a)\\ \|w_s(t_i) - w_i^s\|_\infty &\le M(\|\tau^u\|_N + \|\tau^w\|_N), && (2.55b) \end{aligned}$$

where τ_i^u and τ_i^w are local truncation errors for the difference scheme (2.47) and they can be written as

$$\begin{aligned} \tau_i^u &= h_i\{u_s''(\xi_i^1) + R''(\xi_i^2)[S(t_{i-1})u_s(t_{i-1}) + w_s(t_{i-1})]\\ &+ R'(t_i)(Su_s + w_s)'(\xi_i^3)\} && (2.56a)\\ \tau_i^w &= \epsilon h_i\{w_s''(\eta_i^1) + P''(\eta_i^2)[S(t_{i-1})u_s(t_{i-1}) + w_s(t_{i-1})]\\ &+ P'(t_i)(Su_s + w_s)'(\eta_i^3)\}, && (2.56b) \end{aligned}$$

where $\xi_i^\mu, \eta_i^\mu \in (t_{i-1}, t_i), \mu = 1, 2, 3$. To bound the truncation error, we need the derivative estimates of u_s and w_s. From the asymptotic expansions of u_s and w_s (cf. *Theorem 2.1*)

$$u_s = u_e + \epsilon\bar{u}_{s1} + \epsilon^2(\bar{u}_{s2} + \tilde{u}_{s2}) + \cdots \tag{2.57a}$$

$$w_s = w_e + \epsilon(\bar{w}_{s1} + \tilde{w}_{s1}) + \epsilon^2(\bar{w}_{s2} + \tilde{w}_{s2}) + \cdots \tag{2.57b}$$

where $u_e = Rx_e$, $w_e = Px_e$, $\bar{u}_{sj}$ and $\bar{w}_{sj}$ are functions of regular expansions, $\tilde{u}_{sj}$ and $\tilde{w}_{sj}$ are boundary layer functions whose basic forms are $p(t/\epsilon)\exp(-t/\epsilon)$ (where p is some polynomial) and $\bar{u}_{sj} = \bar{w}_{sj} = 0$ for $j \leq s$ (since SRM iteration cancels out the lower terms of regular expansions). We can expect that

$$|u_s'|, |u_s''|, |w_s'| \leq M. \tag{2.58}$$

But

$$|w_s''| = O(\epsilon^{-1}\exp(-t/\epsilon) \leq \begin{cases} M & \text{if } t \gg \epsilon \\ M\epsilon^{-1} & \text{otherwise .} \end{cases} \tag{2.59}$$

Therefore

$$\|\tau^u\|_N = O(h), \ \|\tau^w\| = \begin{cases} O(\epsilon h) & \text{if } \epsilon \ll h_1 = t_1 - t_0 \\ O(h) & \text{otherwise .} \end{cases} \tag{2.60}$$

From (2.55), we thus have

$$\|u_s(t_i) - u_i^s\|_\infty \leq Mh \tag{2.61a}$$

$$\|w_s(t_i) - w_i^s\|_\infty \leq \begin{cases} M\epsilon h & \text{if } \epsilon \ll h_1 \\ Mh & \text{otherwise} \end{cases} \tag{2.61b}$$

i.e.

$$x_s(t_i) - x_i^s = S(t_i)(u_s(t_i) - u_i^s) + (w_s(t_i) - w_i^s) = O(h). \tag{2.62}$$

However, we can not generally get a good approximation for By_e in the whole region if ϵ is not very small compared with h_1 since in this case we generally have

$$\begin{aligned} B_i y_i^s &= B_i y_{s-1}(t_i) + \frac{1}{\epsilon} B_i E_i (G_i x_i^s + r_i) \\ &= B_i y_{s-1}(t_i) + \frac{1}{\epsilon} B_i E_i (G_i x_s(t_i) + r_i) + \frac{1}{\epsilon}(w_i^s - w_s(t_i)) \\ &= B_i y_e(t_i) + O(\epsilon + \exp(-t_i/\epsilon)) + O(h/\epsilon). \end{aligned} \tag{2.63}$$

Fortunately, we can get $O(h)$ accuracy locally, i.e. in a smooth region or away from the layer region, say for $t_{i_0} \le t \le t_f$. Indeed, considering an equidistant mesh for simplicity, from (2.47b), we have

$$\begin{aligned} L_1^h \Delta w_i^s = \epsilon \frac{\Delta w_i^s - \Delta w_{i-1}^s}{h} + \Delta w_{i-1}^s = \\ \epsilon P_i A_i (S_i \Delta u_i^s + \Delta w_i^s) + \epsilon \frac{P_i - P_{i-1}}{h} (S_{i-1} \Delta u_i^s + \Delta w_{i-1}^s) + O(\tau_i^w), \end{aligned} \quad (2.64)$$

where $\Delta w_i^s = w_s(t_i) - w_i^s$, $\Delta u_i^s = u_s(t_i) - u_i^s$ and we note that $\tau_i^w = O(\epsilon h)$ for $i_0 \le i \le N$. Using the discrete maximum principle for L_1^h in $t_{i_0} \le t \le t_f$ on the barrier function

$$z_i = |\Delta w_{i_0}^s| \lambda^{i-i_0} + \max_{i_0 \le i \le N} |\delta_i| \pm \Delta w_i^s,$$

where $\delta_i (= O(h))$ is the right-hand side of (2.64) and $\lambda = h/(\epsilon + h)$, and using (2.61), we get $z_i \ge 0$ or

$$\begin{aligned} |\Delta w_i^s| &\le |\Delta w_{i_0}^s| \lambda^{i-i_0} + \max_{i_0 \le i \le N} |\delta_i| \\ &\le M(h \lambda^{i-i_0} + \epsilon h). \end{aligned} \quad (2.65)$$

For $\epsilon = h^{1+\delta}$, $-1 < \delta < 1$,

$$\lambda^{i-i_0} = \exp(-(i - i_0) h^{\delta}) = \exp(-(t_i - t_{i_0})/h^{1-\delta}).$$

Taking i_1 such that $\exp(-(t_{i_1} - t_{i_0})/h^{1-\delta}) \le M h^{1+\delta} = M\epsilon$ when h is sufficiently small, we get

$$\lambda^{i-i_0} \le \exp(-(t_{i_1} - t_{i_0})/h^{1-\delta}) \le M\epsilon \ \text{ for } i \ge i_1,$$

i.e. $|\Delta w_i^s| \le M\epsilon h$, $\forall i \ge i_1$, and any ϵ satisfying $\epsilon = h^{1+\delta}$, $-1 < \delta < 1$. Combining this with (2.61b), we obtain

$$|\Delta w_i^s| \le M\epsilon h, \ \forall i \ge i_1. \quad (2.66)$$

Then, using (2.63) and (2.66), we have the local error estimate

$$|B_i y_i^s - B_i y_e(t_i)| \le M(h + \epsilon^s), \ \text{for } i_1 \le i \le N. \quad (2.67)$$

This means we can get good approximation in a region which is away from the initial layer.

Once an accurate SRM solution, say $\{x_i^*, B(t_i)y_i^*\}$, has been determined outside the initial layer it may be possible to obtain an accurate solution everywhere by applying a few SRM iterations numerically solving (2.44a) (changing BEG to $-BEG$) subject to the terminal value

$$x(t_f) = x_N^*, \tag{2.68}$$

and choosing By_0 satisfying $B(t_f)y_0(t_f) = B(t_f)y_N^*$. This procedure is feasible provided that the terminal value problem (2.44a),(2.68) is well-conditioned (which holds if the terminal value problem for the EUODE (2.22a) is well-conditioned).

For the IVP case, where (2.44b) reduces to

$$x(0) = \bar{x} \text{ given}, \tag{2.69}$$

we may, of course, proceed in the same way as for the BVP case. But now a few things are easier. Firstly, for this case we can calculate $By_e(0)$ and then choose By_0 to be exact at $t = 0$. In fact, as indicated earlier we can also do this for higher derivatives of By at the initial value by repeated differentiation of (2.16). Such a preparation of the initial iterate By_0 allows removing the layer error terms (or the condition $t_i \gg \epsilon$) in the error estimates (2.61).

Secondly, one may use a more convenient difference scheme to integrate the IVP (2.44a),(2.69). If the EUODE is sufficiently nonstiff to warrant use of a nonstiff integration method then this can be an attractive possibility here. Note, though, that $-h_i/\epsilon$ must be in the absolute stability region of the method (see (2.39b)). Thus, an explicit Runge-Kutta method of order p, for instance, may necessitate (at least) p SRM iterations in order for the error in the estimates of Theorem 2.1 to be of the same order as the error in the numerical approximation.

The most important difference between the IVP and BVP cases is that the iterative method described here does not appear to be necessarily optimal or even natural in the IVP context, certainly not from the storage requirement point of view: Note that the entire approximation of By_{s-1} on $[0, t_f]$ needs to be stored. The situation here is similar to that encountered with other functional iteration methods like waveform methods.

However, this difficulty can be resolved by rearranging the computation, assuming that the number of the SRM iterations, m, is chosen in advance. Thus, at each time step i, $1 \leq i \leq N$, we calculate sequentially the quantities $x_i^1, By_i^1, x_i^2, By_i^2, \ldots, x_i^m, By_i^m$. To do this using a one-step scheme, say, we need only the corresponding quantities locally, over the mesh subinterval $[t_{i-1}, t_i)$, and By_i^0. For the latter we may use, for instance, $y_i^0 \equiv y_e(0)$, i.e. $By_i^0 = B(t_i)y_0^0$, $0 \leq i \leq N$. The storage requirements are now independent of N and other typical IVP techniques such as local error control may be applied as well.

2.7 Numerical Experiments

We now present a few very simple examples to demonstrate our claims in the previous sections. Throughout this section we use a constant step size h and set $t_f = 1$. To make life difficult we choose h so that there is an i such that $t_i = t_*$ (if there is a singularity). In the implementation we monitor the size of the pivot in a Gauss elimination procedure for GB and slightly perturb t_i away from t_* when needed. At a given time t, we use $'ex'$ to denote the maximum over all components of the error in x^s while $'ey'$ denotes the maximum over all components of the error in By^s. Similarly, $'drift'$ denotes the maximum residual in the algebraic equations.

We first look at a boundary value problem.

Example 2.2 *Consider the DAE (2.16) with*

$$\begin{aligned} A &= \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}, \; B = -\begin{pmatrix} 0 \\ 1-2t \end{pmatrix}, \; q = \begin{pmatrix} -\sin t \\ 0 \end{pmatrix} \\ G &= \begin{pmatrix} 1-2t & 1-2t \end{pmatrix}, \; r = -(1-2t)(e^{-t} + \sin t) \end{aligned}$$

subject to

$$x_1(1) + x_2(0) = 1/e.$$

The exact solution is $x_e = \begin{pmatrix} e^{-t} & \sin t \end{pmatrix}$, $y_e = \frac{\cos t}{1-2t}$. *A singularity is located at* $t = 1/2$, *where* y_e *becomes infinite while* By_e *stays bounded. We start computing with the iterate* $y_0(t) \equiv 0$.

In Table 2.1 we list errors when using the midpoint scheme

$$\begin{aligned} \frac{x_i^s - x_{i-1}^s}{h_i} &= A_{i-\frac{1}{2}} x_{i-\frac{1}{2}}^s - By_{i-\frac{1}{2}}^s + q_{i-\frac{1}{2}} \\ By_{i-\frac{1}{2}}^s &= By_{i-\frac{1}{2}}^{s-1} + \epsilon^{-1} B_{i-\frac{1}{2}} E_{i-\frac{1}{2}} (G_{i-\frac{1}{2}} x_{i-\frac{1}{2}}^s + r_{i-\frac{1}{2}}) \end{aligned}$$

where $x_{i-\frac{1}{2}}^s = \frac{x_i^s + x_{i-1}^s}{2}$ *(but no such relation is necessary for* y^s*). We apply this scheme with* $h_i = h = .01$ *for various values of* ϵ.

It is indicated in [11] that this scheme has 2nd order accuracy in ex and in ey, except for the case $\epsilon \ll h$ *when the error's order in By drops to 1. This is evident in the error column for* $t = 1.0$. *Note also the* $O(\epsilon)$ *improvement per SRM iteration when this term dominates the error (i.e. when* $\epsilon^s \gg h^2$*).*

We note that the approximation for By at points within the initial layer is not accurate. To get a better approximation within the initial layer (i.e. near the initial point $t = 0$*), we solve a terminal value problem (2.44a) (changing BEG to* $-BEG$*), (2.68), as described in §2.6. Then we apply the SRM for the given problem with the improved values for* By_0. *In Table 2.2 we list the computed results after 3 iterations. They are obviously much better than the comparable ones in Table 2.1.*

□

ϵ	iteration	error at $\rightarrow$	t=.01	t=.1	t = .3	t=.5	t=1.0
1e-1	1	ex	.38e-1	.35e-1	.56e-1	.52e-1	.39e-1
		ey	.96	.40	.14	.34e-1	.66e-1
		drift	.87e-2	.49e-1	.51e-1	.0	.61e-1
	2	ex	.92e-2	.37e-1	.89e-2	.65e-2	.72e-2
		ey	.91	.96e-2	.14	.34e-1	.65e-2
		drift	.90e-2	.32e-1	.61e-2	.0	.72e-2
	3	ex	.94e-2	.19e-1	.12e-1	.63e-2	.15e-2
		ey	.87	.20	.30e-1	.43e-1	.85e-3
		drift	.86e-2	.16e-1	.43e-2	.0	.74e-3
1e-2	1	ex	.38e-2	.60e-2	.53e-2	.44e-2	.38e-2
		ey	.67	.30e-2	.40e-2	.48e-2	.62e-2
		drift	.65e-2	.80e-2	.38e-2	.0	.55e-2
	2	ex	.45e-2	.10e-3	.88e-4	.77e-4	.64e-4
		ey	.45	.32e-3	.70e-4	.44e-4	.23e-4
		drift	.44e-2	.15e-4	.14e-4	.0	.68e-4
	3	ex	.30e-2	.55e-5	.52e-5	.59e-5	.11e-4
		ey	.30	.18e-2	.26e-4	.18e-4	.67e-5
		drift	.29e-2	.17e-5	.26e-5	.0	.56e-5
1e-3	1	ex	.13e-2	.58e-3	.52e-3	.45e-3	.39e-3
		ey	.17	.47e-2	.41e-3	.49e-3	.62e-3
		drift	.17e-2	.79e-3	.38e-3	.0	.54e-3
	2	ex	.30e-3	.71e-4	.75e-5	.72e-5	.12e-4
		ey	.30e-1	.17e-1	.34e-4	.13e-4	.54e-5
		drift	.30e-3	.51e-4	.20e-5	.0	.65e-5
	3	ex	.65e-4	.15e-3	.70e-5	.70e-5	.12e-4
		ey	.70e-2	.33e-1	.12e-3	.14e-4	.56e-5
		drift	.69e-4	.11e-3	.21e-5	.0	.59e-5

Table 2.1: SRM errors for Example 2.2 using the midpoint scheme

Next we consider initial value problems.

Example 2.3 *Consider the same DAE as for Example 2.2 with the same exact solution but with initial values $x_1(0) = 1$, $x_2(0) = 0$ specified. From these initial conditions we can calculate $y(0) = 1$ in advance, and we choose the initial guess $y_0(t) \equiv 1$. Tables 2.3 and 2.4 display error results for $\epsilon = .1$ and $h = .001$ using the backward Euler and the forward Euler schemes, respectively. As explained in §2.6 we calculate all iterates at each step before proceeding to the next.*

These tables show a significant improvement with each SRM iteration and no strong initial layer effect, as predicted by theory.

□

ϵ	error at $\rightarrow$	t=.01	t=.1	t = .3	t=.5
1e-1	ex	.62e-2	.54e-2	.39e-2	.28e-2
	ey	.48e-2	.45e-2	.44e-2	.54e-2
5e-2	ex	.76e-3	.58e-3	.32e-3	.17e-3
	ey	.22e-2	.18e-2	.10e-2	.50e-3
1e-2	ex	.57e-4	.49e-4	.35e-4	.26e-4
	ey	.10e-3	.85e-4	.52e-4	.32e-4
1e-3	ex	.49e-4	.42e-4	.31e-4	.23e-4
	ey	.56e-4	.46e-4	.30e-4	.19e-4

Table 2.2: SRM errors for Example 2.2 using the shooting-back technique

iteration	error at $\rightarrow$	t=.001	t=.1	t = .3	t=.5	t=1.0
1	ex	.20e-5	.72e-2	.37e-1	.63e-1	.11
	ey	.20e-2	.12	.15	.12	.59e-1
	drift	.15e-5	.60e-2	.16e-1	.76e-4	.15
2	ex	.20e-5	.51e-2	.13e-1	.10e-1	.25e-2
	ey	.20e-2	.68e-1	.45e-2	.20e-1	.80e-2
	drift	.15e-5	.42e-2	.58e-2	.14e-4	.67e-2
3	ex	.20e-5	.35e-2	.23e-2	.16e-2	.76e-3
	ey	.20e-2	.32e-1	.26e-1	.10e-1	.37e-2
	drift	.15e-5	.29e-2	.12e-2	.10e-5	.12e-2

Table 2.3: SRM errors for Example 2.3 using backward Euler

Example 2.4 *Here we investigate the use of the modified formula (2.42) instead of (2.35). First, we solve the previous example numerically using (2.42). In Table 2.5 we record error values at the singularity point* $t = .5$ *after 3 SRM iterations, starting with* $y_0(t) \equiv 1$ *and using as before* $\epsilon = .1$ *and* $h = .001$ *(cf. Tables 2.5).*

From these results it is clear that the SRM with (2.42) does not work well when $\alpha_1 \neq 0$*: large errors in* By *are obtained near the singularity and these adversely affect the accuracy in* x *as well. However, the comparison changes when there is no singularity in the constraints: We now replace the algebraic constraint in Example 2.3 by*

$$x_1 + x_2 - e^{-t} - \sin t = 0$$

leaving everything else the same (including the singularity in B*). In Table 2.6 we record maximum errors in* x *and* By *over all mesh points (denote those 'exg' and*

iteration	error at →	t=.001	t=.1	t = .3	t=.5	t=1.0
1	ex	.50e-6	.71e-2	.36e-1	.63e-1	.11
	ey	.20e-2	.12	.15	.12	.60e-1
	drift	.50e-6	.60e-2	.16e-1	.76e-4	.15
2	ex	.50e-6	.51e-2	.12e-1	.10e-1	.44e-2
	ey	.20e-2	.68e-1	.41e-2	.20e-1	.70e-2
	drift	.50e-6	.42e-2	.58e-2	.14e-4	.67e-2
3	ex	.50e-6	.35e-2	.43e-2	.18e-2	.98e-3
	ey	.20e-2	.32e-1	.26e-1	.97e-2	.46e-2
	drift	.50e-6	.29e-2	.12e-2	.11e-5	.12e-2

Table 2.4: SRM errors for Example 2.3 using forward Euler

$(\alpha_1, \alpha_2) \to$	$(0,1)$		$(h,1)$		$(1,1)$	
method	ex	ey	ex	ey	ex	ey
backward Euler	.16e-2	.10e-1	.80e-3	.20e+1	.15	.37e+3
forward Euler	.18e-2	.96e-2	.25e-2	.18e+1	.64	.15e+4

Table 2.5: Errors near singularity using modified formula (2.42)

'eyg', respectively) for the starting iterates $y_0(t) \equiv 1$ *and* $y_0(t) \equiv 0$ *(the latter does not agree with the exact* $y_e(0)$*).*

	$(\alpha_1, \alpha_2) \to$	$(0,1)$		$(h,1)$		$(1,1)$	
y_0	method	exg	eyg	exg	eyg	exg	eyg
$\equiv 1$	backward Euler	.46e-2	.44e-1	.45e-2	.43e-1	.22e-3	.28e-3
	forward Euler	.45e-2	.44e-1	.44e-2	.43e-1	.19e-3	.23e-3
$\equiv 0$	backward Euler	.22e-1	.97	.22e-1	.94	.12e-3	.75e-3
	forward Euler	.22e-1	.97	.22e-1	.94	.19e-3	.75e-3

Table 2.6: Errors for problem without singularity using modified formula (2.42)

The modified method (corresponding to $(\alpha_1, \alpha_2) = (1,1)$ *in Table 2.6 is seen to work better for problems without singularities.*

□

The above calculations all agree with our theoretical results described in §2.5 and §2.6.

2.8 More about the Proof of Theorem 2.1

To provide a better understanding about the sequential regularization method we now give a detailed proof of Theorem 2.1 for the initial value case with no layers, $s \leq J+1$. J is some positive integer defined in Definition 2.1. In this proof, the construction of the asymptotic expansion is directly for x and By. Moreover, the construction method we apply is somewhat different from [111, 112] and more relevant to the concept of DAEs. The same idea is applied to prove the convergence of the SRM for Navier-Stokes equations in Chapter 4. For $s > J+1$, additional initial layer expansions have to be developed. However, the construction of these layer expansions is precisely the same as in [111, 112] and so it is omitted here. In case that (2.17) are initial conditions (i.e. $\bar{B}_1 = 0$) our assumptions imply that (2.17) together with (2.18b) specify $x(0)$, say

$$x(0) = \bar{x} \tag{2.70}$$

At first, consider the case $s = 1$ of (2.34),(2.35):

$$\epsilon x_1' + B(GB)^{-1}(Gx_1 + r) = \epsilon A x_1 - \epsilon B y_0 + \epsilon q,$$

with the initial conditions (2.70). This is a singular-singularly-perturbed problem (see [112, 89]). Let

$$x_1 = x_{10} + \epsilon x_{11} + \cdots + \epsilon^s x_{1s} + \cdots$$

Comparing the coefficients of like powers of ϵ, we thus have

$$B(GB)^{-1}Gx_{10} = -B(GB)^{-1}r \tag{2.71a}$$

$$B(GB)^{-1}Gx_{11} = -x_{10}' + Ax_{10} - By_0 + q, \tag{2.71b}$$

$$B(GB)^{-1}Gx_{1i} = -x_{1i-1}' + Ax_{1i-1},\ 2 \leq i \leq s+1, \tag{2.71c}$$

where (2.71a) satisfies (2.70) and (2.71b) and (2.71c) satisfy homogeneous initial conditions corresponding to (2.70). Now, (2.71a) has infinitely many solutions in general. To realize the construction, we should choose x_{10} to satisfy (2.71a) and to ensure that the solution of (2.71b) exists. We choose x_{10} to be the solution x_e of problem (2.16)-(2.17), i.e.

$$x_{10}' = Ax_{10} - By_e + q, \tag{2.72a}$$

$$0 = Gx_{10} + r, \tag{2.72b}$$

$$\bar{B}_0 x_{10}(0) = \beta. \tag{2.72c}$$

So $x_{10} = x_e$ and (2.71b) has the following form

$$B(GB)^{-1}Gx_{11} = B(y_e - y_0). \tag{2.73}$$

Now we choose x_{11} and a corresponding y_{01} to satisfy

$$x_{11}' = Ax_{11} - By_{01} \tag{2.74a}$$

$$Gx_{11} = GB(y_e - y_0), \tag{2.74b}$$

$$\bar{B}_0 x_{11}(0) = 0. \tag{2.74c}$$

Noting that $By_e = -x_e' + Ax_e + q$ is smooth, we have $GB(y_e - y_0) \in \mathcal{S}$. Hence, using Lemma 2.1, there exists a smooth solution x_{11} of (2.74), and x_{11} satisfies (2.73). Indeed, using (2.74b) and Definition 2.1, we have $G(0)x_{11}(0) = 0$, so $x_{11}(0) = 0$. And, from (2.74b) again,

$$(GB)^{-1}Gx_{11} = y_e - y_0, \text{ for each } t \in [0, t_*) \cup (t_*, t_f].$$

That is,

$$B(GB)^{-1}Gx_{11} = B(y_e - y_0), t \in [0, t_*) \cup (t_*, t_f]. \tag{2.75}$$

Taking the limit of (2.75) at t_*, we thus get that x_{11} satisfies (2.73) for each $t \in [0, t_f]$.

Moreover, from Definition 2.1, we have

$$y_{01}(0) = y_{01}'(0) = \cdots = y_{01}^{(s-1)}(0) = 0, \; s \leq J + 1.$$

Also we note that By_{01} is smooth.

Generally, supposing we have got x_{1i-1}, By_{0i-1} and

$$y_{0i-1}(0) = y'_{0i-1}(0) = \cdots = y_{0i-1}^{(s-i+1)}(0) = 0$$

for $i \geq 2$, we choose x_{1i}, y_{0i} satisfying

$$x'_{1i} = Ax_{1i} - By_{0i},$$

$$Gx_{1i} = (GB)y_{0i-1},$$

$$\bar{B}_0 x_{1i}(0) = 0.$$

By the same argument as before, we obtain that x_{1i} satisfies (2.71c) for $2 \leq i \leq s+1$, and

$$y_{0i}(0) = y'_{0i}(0) = \cdots = y_{0i}^{(s-i)}(0) = 0, \;\; s \leq J+1.$$

Also, By_{0i} is smooth. Next we denote the asymptotic solution

$$\bar{x}_{1s+1} = x_{10} + \epsilon x_{11} + \cdots + \epsilon^s x_{1s} + \epsilon^{s+1} x_{1s+1}$$

and

$$z_{1s+1} = x_1 - \bar{x}_{1s+1}.$$

Then

$$\epsilon z'_{1s+1} + Pz_{1s+1} = \epsilon A z_{1s+1} + \epsilon^{s+2}(-x'_{1s+1} + Ax_{1s+1}),$$

$$z_{1s+1}(0) = 0$$

Let $u_{1s+1} = Rz_{1s+1}$ and $w_{1s+1} = Pz_{1s+1}$. Hence, we have (cf. (2.38))

$$z_{1s+1} = Su_{1s+1} + w_{1s+1}$$

and

$$u'_{1s+1} = (RA + R')Su_{1s+1} + (RA + R')w_{1s+1} + O(\epsilon^{s+1}),$$

$$\epsilon w'_{1s+1} + w_{1s+1} = \epsilon(PA + P')Su_{1s+1} + \epsilon(PA + P')w_{1s+1} + O(\epsilon^{s+2}),$$

$$u_{1s+1}(0) = 0, \;\; w_{1s+1}(0) = 0.$$

Using Lemma 2.2, we get $w_{1s+1} = O(\epsilon^{s+2})$ and $u_{1s+1} = O(\epsilon^{s+1})$, i.e.

$$z_{1s+1} = O(\epsilon^{s+1}).$$

Therefore,

$$x_1 = x_{10} + \epsilon x_{11} + \cdots + \epsilon^s x_{1s} + O(\epsilon^{s+1}). \tag{2.76}$$

Noting $x_{10} = x_e$, we thus obtain

$$x_1 - x_e = O(\epsilon). \tag{2.77}$$

Then, by using (2.35),(2.76),(2.71),(2.72a) and (2.74a), it follows that

$$By_1 = By_0 + \frac{1}{\epsilon}B(GB)^{-1}(Gx_1 + r)$$

$$\begin{aligned} By_1 &= By_0 + \tfrac{1}{\epsilon}(Px_{10} + B(GB)^{-1}r + \epsilon Px_{11} + \cdots + \epsilon^s Px_{1s} + O(\epsilon^{s+1})) \\ &= By_e + \epsilon By_{01} + \cdots + \epsilon^{s-1}By_{0s-1} + O(\epsilon^s) \end{aligned} \tag{2.78}$$

or

$$By_1 - By_e = O(\epsilon). \tag{2.79}$$

Now we look at the second iteration $s = 2$ of (2.34), (2.35):

$$\epsilon x'_2 + B(GB)^{-1}(Gx_2 + r) = \epsilon Ax_2 - \epsilon By_1 + \epsilon q,$$

with initial conditions (2.70). Let

$$x_2 = x_{20} + \epsilon x_{21} + \epsilon^2 x_{22} + \cdots.$$

Noting that (2.78) gives us a series expansion for By_1 we obtain,

$$B(GB)^{-1}Gx_{20} = -B(GB)^{-1}r, \tag{2.80a}$$

$$B(GB)^{-1}Gx_{21} = -x'_{20} + Ax_{20} - By_e + q, \tag{2.80b}$$

$$B(GB)^{-1}Gx_{2i} = -x'_{2i-1} + Ax_{2i-1} - By_{0i-1}, \;\; 2 \le i \le s+1 \tag{2.80c}$$

Again, (2.80a) satisfies initial conditions (2.70) and (2.80b) and (2.80c) satisfy the corresponding homogeneous ones. As the case of $s = 1$, we choose $x_{20} = x_e$. We thus have

$$B(GB)^{-1}Gx_{21} = 0.$$

Then x_{21} is constructed to satisfy

$$x'_{21} = Ax_{21} - By_{11}, \tag{2.81a}$$

$$Gx_{21} = 0, \tag{2.81b}$$

$$\bar{B}_0 x_{21}(0) = 0. \tag{2.81c}$$

Obviously $x_{21} = 0$ since (2.81) is uniquely solvable for x_{21} by Lemma 2.1. In general, similarly to the case of $s = 1$, we choose x_{2i} satisfying

$$x'_{2i} = Ax_{2i} - By_{1i}, \tag{2.82a}$$

$$Gx_{2i} = -(GB)(y_{0i-1} - y_{1i-1}), \tag{2.82b}$$

$$\bar{B}_0 x_{2i}(0) = 0. \tag{2.82c}$$

for $2 \leq i \leq s+1$. By applying Lemma 2.2 and the same argument as in the case of $s = 1$ we get

$$x_2 = x_e + \epsilon x_{21} + \epsilon^2 x_{22} + \cdots + \epsilon^s x_{2s} + O(\epsilon^{s+1}) \tag{2.83}$$

or

$$x_2 - x_e = O(\epsilon^2). \tag{2.84}$$

Then, using (2.35),(2.80),(2.81a),(2.82a), (2.83) and (2.78), we conclude

$$\begin{aligned} By_2 &= By_1 + \frac{1}{\epsilon}B(GB)^{-1}(Gx_2 + r) \\ &= By_e + \epsilon^2 By_{12} + \cdots + \epsilon^{s-1} By_{1s-1} + O(\epsilon^s) \end{aligned} \tag{2.85}$$

or

$$By_2 - By_e = O(\epsilon^2) \tag{2.86}$$

We can repeat this procedure, and, by induction, complete the proof for $s \leq J+1$.

□

Chapter 3

SRM for Nonlinear Problems

In the previous chapter we derived the sequential regularization method and gave a detailed continuous and discrete analysis for linear index-2 DAEs. The method relates to a combination of stabilization and penalty-like methods. In this chapter we extend the method to nonlinear index-2 and index-3 problems ($\nu = 1$ and $\nu = 2$ in (2.15)), including constrained multibody systems. A number of variants are proposed, and particularly effective methods are singled out in certain circumstances. All results obtained here are certainly applicable to the linear case.

The chapter is organized as follows: In §3.1 we consider problems without constraint singularities. Two SRM variants are discussed. One variant involving $\frac{dg}{dt}$ (corresponding to (2.32), (2.33)) leads to nonstiff problems. Taking $E = I$ is particularly attractive. The other variant, corresponding to (2.32), (2.33) with $\alpha_1 = 0$, does not involve $\frac{dg}{dt}$. The choice $E = I$, if possible (otherwise one can choose $E = (GB)^T$), makes the computation particularly simple. Problems with constraint singularities are considered in §3.2. The SRM corresponding to (2.34), (2.35) is proposed for such problems. This variant works well in practice, but our proofs to date extend only to the linear case.

In §3.3 we analyze and discuss various methods for index-3 problems. A number of SRM variants are possible, combining regularization with Baumgarte's stabilization or with invariant stabilization [8]. Of particular interest, in case of no constraint singularity, are the methods (3.47) and (3.33)–(3.35) which corresponds to invariant stabilization. The choice $E = I$ leads to particularly simple iterations. A corresponding convergence result is given in Theorem 3.3. In case of a possible constraint singularity, the SRM (3.41) is recommended.

These methods are reformulated in §3.4 for the special case of multibody systems with holonomic constraints. The "winning" methods are (3.44)-(3.45) with $E = I$ for the nonsingular case and (3.46)-(3.47) for the case where the constraint Jacobian may have isolated rank deficiencies. In §3.5 we report the results of numerical experiments confirming our theoretical predictions and demonstrating the effectiveness of the proposed methods.

3.1 Nonlinear, Nonsingular Index-2 Problems

The nonlinear index-2 DAE ($\nu = 1$ in (2.15)) reads:

$$x' = f(x,t) - B(x,t)y, \tag{3.1a}$$

$$0 = g(x,t), \tag{3.1b}$$

where f, B and g are sufficiently smooth functions of $(x,t) \in \mathbf{R}^{n_x} \times [0,t_f]$, and $y \in \mathbf{R}^{n_y}$. We consider this DAE subject to $n_x - n_y$ boundary conditions

$$b(x(0), x(t_f)) = \beta\,. \tag{3.2}$$

These boundary conditions are assumed to yield a unique[1] and bounded solution for the ODE (3.1a) on the manifold given by (3.1b). Concretely, if we were to replace (3.1b) by its differentiated form (denoting $G = \frac{\partial g}{\partial x}$)

$$0 = Gx' + g_t \;\; (= \frac{dg}{dt}) \tag{3.3a}$$

$$g(x(0), 0) = 0 \tag{3.3b}$$

[1] locally unique, or isolated solution in a sufficiently large neighborhood would suffice.

and use (3.3a) in (3.1a) to eliminate y and obtain n_x ODEs for x, then the boundary value problem for x with (3.2) and (3.3b) specified has a unique solution. In the initial value case (i.e., when b is independent of $x(t_f)$), this means that (3.2) and (3.3b) can be solved uniquely for $x(0)$.

In this section, we consider the case where GB is nonsingular. Generalizing the idea in §2.4, we have the following SRM formulation for the nonlinear index-2 DAEs (3.1), (3.2): for $s = 1, 2, \ldots,$

$$x_s' = f(x_s, t) - B(x_s, t)y_s, \tag{3.4}$$

where

$$y_s = y_{s-1} + \frac{1}{\epsilon}E(x_s, t)(\alpha_1 \frac{d}{dt}g(x_s, t) + \alpha_2 g(x_s, t)), \tag{3.5}$$

subject to the boundary conditions (3.2) and (3.3b). Note that $y_0(t)$ is a given initial iterate which we assume is sufficiently smooth and bounded and that $\epsilon > 0$ is the regularization parameter. The regularization matrix E is nonsingular and has a uniformly bounded condition number; possible choices are $E = I$, $E = (GB)^{-1}$ and others (e.g. $E = (GB)^T$, cf. [11, 94]). We note that if we take $y_0 \equiv y$ then $x_1 \equiv x$, where x and y are the solution of (3.1). If we take $y_0 \equiv 0$, then one SRM iteration is the usual penalty method (cf. [84, 91, 69]). As customary for the penalty method, we assume:

Assumption 3.1 *The problem (3.4), (3.5),(3.2),(3.3b) has a unique solution and the solution is bounded if y_{s-1} is bounded.*

Assumption 3.1 is generally true for initial value problems. For general boundary value problems, we expect that it would hold for most practical cases since (3.4) (with (3.5) plugged in) may be seen as a perturbed problem of (3.1) according to the proof of Theorem 3.1 (see below), where the perturbation and its first derivative are both small if ϵ is small.

To analyze the SRM, we assume the following perturbation inequality: For $0 \leq t \leq t_f$,

$$\|\hat{x}(t) - x(t)\| \leq M \max_{0\leq\tau\leq t_f}(|\delta(\tau)| + |\delta'(\tau)|), \tag{3.6a}$$

$$\|\hat{y}(t) - y(t)\| \leq M \max_{0\leq\tau\leq t_f}(|\delta(\tau)| + |\delta'(\tau)|), \tag{3.6b}$$

where $\|\cdot\|$ is some l_p norm (say, the maximum norm), and $\hat{x}$ and $\hat{y}$ satisfy the following perturbed version of (3.1):

$$\hat{x}' = f(\hat{x}, t) - B(\hat{x}, t)\hat{y}, \tag{3.7a}$$

$$0 = g(\hat{x}, t) + \delta(t) \tag{3.7b}$$

with the same boundary conditions as (3.2). For initial value problems, (3.6) has been proved in [58], pp. 478-481. It is actually the definition of the perturbation index introduced in [58]. Furthermore, (3.6) also holds for boundary value problems if we impose some boundedness conditions on the corresponding Green's function (cf. [14]).

The case $\alpha_1 \neq 0$ in (3.5) is sufficiently different from the case $\alpha_1 = 0$ to warrant a separate treatment.

3.1.1 The case $\alpha_1 = 1$

Now we estimate the error of the sequential regularization method (3.4)-(3.5). We prove a theorem which says that the error after s SRM iterations is $O(\epsilon^s)$ (i.e., each iteration improves the error by $O(\epsilon)$) everywhere in t. This result coincides with that of the linear case.

Theorem 3.1 *Let all functions in the DAE (3.1) be sufficiently smooth and the above assumptions hold. Then, for the solution of iteration (3.4),(3.5) with $\alpha_1 \neq 0$, we have* *the following error estimates:*

$$\begin{aligned} x_s(t) - x_e(t) &= O(\epsilon^s), \\ y_s(t) - y_e(t) &= O(\epsilon^s), \end{aligned}$$

for $0 \le t \le t_f$ *and* $s \ge 1$.

Proof: Let $v_s = g(x_s, t)$. Then, from (3.4),

$$v_s' = G(x_s,t)x_s' + g_t(x_s,t) = G(x_s,t)f(x_s,t) - G(x_s,t)B(x_s,t)y_s + g_t(x_s,t).$$

Using (3.5), we thus have

$$(\epsilon(GBE)^{-1} + I)v_s' + \alpha_2 v_s = \epsilon(GBE)^{-1}(Gf + g_t) - \epsilon E^{-1} y_{s-1}, \tag{3.8a}$$
$$v_s(0) = 0. \tag{3.8b}$$

Therefore it is not difficult to get

$$v_s = g(x_s,t) = O(\epsilon),\ v_s' = g(x_s,t)' = O(\epsilon), \tag{3.9}$$

if y_{s-1} is bounded (which implies that x_s is bounded from Assumption 3.1).

For $s = 1$, we have

$$\begin{aligned} x_1' &= f(x_1,t) - B(x_1,t)y_1 \\ & g(x_1,t) = O(\epsilon),\ g(x_1,t)' = O(\epsilon) \end{aligned}$$

since y_0 is chosen to be bounded. From assumption (3.6), we immediately get

$$x_1 - x_e = O(\epsilon),\ y_1 - y_e = O(\epsilon). \tag{3.10}$$

Then it is easy to see that y_1 is bounded. So for $s = 2$, we obtain

$$\begin{aligned} x_2' &= f(x_2,t) - B(x_2,t)y_2 \\ & g(x_2,t) = O(\epsilon),\ g'(x_2,t) = O(\epsilon) \end{aligned}$$

By using assumption (3.6) again, this yields

$$x_2 - x_e = O(\epsilon).$$

Hence it can be verified, by substituting (3.3a),(3.1a) for the exact solution, that the right hand side of (3.8a) becomes $O(\epsilon^2)$. So, from (3.8), we can get

$$g(x_2, t) = O(\epsilon^2),\ g'(x_2, t) = O(\epsilon^2).$$

Applying assumption (3.6), it follows that

$$x_2 - x_e = O(\epsilon^2),\ y_2 - y_e = O(\epsilon^2). \tag{3.11}$$

This also gives the boundedness of y_2.

We can repeat this procedure, and, by induction, conclude the results of the theorem. □

From (3.8) it is clear that *there is no stiffness here*, so we can choose $\epsilon > 0$ very small, so small in fact that one SRM iteration would suffice for any desired accuracy, and discretize the regularized ODE, possibly using a nonstiff method like explicit Runge-Kutta. This gives a modified penalty method

$$[I + \epsilon^{-1}BEG]x' = f - By_0 - \epsilon^{-1}BE(g_t + \alpha_2 g) \tag{3.12}$$

where B, E, g etc, all depend on x, with the subscript $s = 1$ suppressed.

For the choice $E = (GB)^{-1}$, let $P = BEG = B(GB)^{-1}G$ be the associated projection matrix. Multiplying (3.12) by $\frac{1}{1+\epsilon^{-1}}P$ and by $I - P$, respectively, and then adding together, we have

$$x' = f - \frac{1}{1+\epsilon^{-1}}By_0 - \frac{\epsilon^{-1}}{1+\epsilon^{-1}}B(GB)^{-1}[Gf + g_t + \alpha_2 g]$$

Thus the iteration obtained is similar to Baumgarte's stabilization

$$x' = f - B(GB)^{-1}[Gf + g_t + \alpha_2 g] \tag{3.13}$$

In fact, the single SRM iteration tends to (3.13) in this case when $\epsilon \to 0$. Indeed, the parameter α_2 is the usual Baumgarte parameter, and choosing $\alpha_2 > 0$ obviously makes equation (3.8a) asymptotically stable for the drift v_s. As indicated in [12], for both of these methods we can apply post-stabilization instead, i.e. take $\alpha_2 = 0$ but stabilize after each discretization step [8, 9].

For reasons of computational expense, it may be better to choose $E = I$ in (3.12). The iteration obtained is simple, although a possibly large matrix (with a special structure) must be "inverted".

Example 3.1 *The choice of $E = I$ is utilized in Chapter 4 (see also [80]) for the time-dependent, incompressible Navier-Stokes equations governing fluid flow. The advantage gained is that no treatment of pressure boundary conditions is needed, unlike methods based on Baumgarte-type stabilizations which lead to the pressure-Poisson equation.* □

3.1.2 The case $\alpha_1 = 0$

For this case the drift equation (3.8) is clearly stiff for $0 < \epsilon \ll 1$. As in §2.5, we denote J such that

$$y_0(0) = y_e(0),\ y_0'(0) = y_e'(0), \ldots, y_0^{(J)}(0) = y_e^{(J)}(0), \tag{3.14}$$

where $J = -1$ if $y_0(0) \neq y_e(0)$, then we can prove the same result as Theorem 3.1 for $s \leq J + 1$. Note that we may choose y_0 satisfying (3.14) for some $m \geq 0$ by expressing y in terms of x at $t = 0$ for initial value problems. But this starting procedure generally does not work for boundary value problems. Hence we state and prove the theorem for initial value problems and comment on the boundary value case following the proof.

Theorem 3.2 *Let the assumptions of Theorem 3.1 plus (3.14) hold. In addition, suppose that the matrix function $E(x,t)$ has been chosen so that GBE is positive definite. Then, for the solution of iteration (3.4),(3.5) with $\alpha_1 = 0$, we have the following error estimates:*

$$\begin{aligned} x_s(t) - x_e(t) &= O(\epsilon^s), \\ y_s(t) - y_e(t) &= O(\epsilon^s), \end{aligned}$$

for $1 \leq s \leq J+1$ and $0 \leq t \leq t_f$.

Proof: We derive the result for the case $s \leq J+1 = 2$. Following the proof, we will comment on additional generalizations. The key is again the basic drift equation (3.8), which we rewrite here as

$$\epsilon v_s' + (GBE)v_s = \epsilon(Gf + g_t - GBy_{s-1}), \tag{3.15a}$$

$$v_s(0) = 0. \tag{3.15b}$$

where quantities are evaluated as before, at (x_s, t), unless otherwise noted.

For $s = 1$, given the boundedness of y_0 we obtain as before

$$v_1 = O(\epsilon)$$

To obtain a similar result for v_1', however, a different procedure from that of Theorem 3.1 is needed. Note that at $t = 0$, the condition $y_0(0) = y(0)$ implies

$$(Gf + g_t - GBy_0)|_{t=0} = 0$$

Hence from (3.15a), $v_1'(0) = 0$. Differentiating (3.15a) with respect to t and using $v_1 = O(\epsilon)$, we get

$$\epsilon v_1'' + (GBE)v_1' = O(\epsilon), \quad v_1'(0) = 0$$

and this yields

$$v_1' = O(\epsilon)$$

From assumption (3.6) we then get (3.10).

Subtracting (3.1a) from (3.4) and using (3.10) gives also

$$x_1' = x' + O(\epsilon)$$

and boundedness of y_1' is obtained from a differentiation of (3.5).

For $s = 2$, given the boundedness of y_1 (by (3.10)) and $y_1(0) = y(0)$, we get as for $s = 1$

$$v_2 = O(\epsilon), \quad v_2' = O(\epsilon)$$

and hence also

$$x_2 = x + O(\epsilon)$$

This yields that the right hand side of (3.15a) is $O(\epsilon^2)$, so

$$v_2 = O(\epsilon^2)$$

Now comes the delicate part. To obtain an $O(\epsilon^2)$ estimate also for v_2', so that the estimate (3.6) can be used to complete the proof, we differentiate the drift equation again, obtaining

$$\epsilon v_2'' + (GBE)v_2' = O(\epsilon^2) + \epsilon(Gf + g_t - GBy_1)'$$

and $v_2'(0) = 0$ obtained as for the $s = 1$ case. We are then left to show that

$$F(x_2, t) := (Gf + g_t - GBy_1)' = O(\epsilon) \tag{3.16}$$

For this purpose we must estimate v_1'' first. Using the condition $y_0'(0) = y'(0)$, and also $x_1(0) = x(0)$, $x_1'(0) = x'(0)$ (obtained from (3.4)), we can obtain

$$(Gf(x_1, t) + g_t(x_1, t) - GB(x_1, t)y_0)'|_{t=0} = 0$$

Hence $v_1''(0) = 0$ from (3.15a) once differentiated. Differentiating (3.15a) twice we now obtain precisely as when estimating v_1' above,

$$v_1'' = O(\epsilon)$$

The boundedness of all needed quantities can also be obtained in the same way as before. Finally, we note

$$v_1' = Gx_1' + g_t(x_1,t) = Gf(x_1,t) + g_t(x_1,t) - GB(x_1,t)y_1$$

Ready to show (3.16), we now write

$$F(x_2,t) = [(Gf(x_2,t)-Gf(x_1,t))+(g_t(x_2,t)-g_t(x_1,t))+(GB(x_1,t)-GB(x_2,t))y_1-v_1']'$$

Our previous estimates allow the conclusion that $x_2 = x_1 + O(\epsilon)$, $x_2' = x_1' + O(\epsilon)$, hence we can finally conclude the estimate (3.16) and obtain the result of the theorem for $s = 2$.

The proof proceeds in a similar manner for larger J. Generally, one needs the estimate $v_1^{(j)} = O(\epsilon)$, $1 \le j \le J+1$, and this necessitates (3.14). □

Remark 3.1 *The convergence result holds for all s (i.e. also for $s > J+1$, assuming sufficient smoothness) away from an initial layer of size $O(\epsilon)$ in t. This is so because E is chosen so that we can express the solution for small ϵ as a smooth outer solution which is bounded in terms of the right hand side as before, plus an initial layer of width $O(\epsilon)$. Conditions (3.14) then ensure that the layer error is bounded by $O(\epsilon^{J+1})$ for the first $J+1$ iterations.* □

Remark 3.2 *For boundary value problems, there is no obvious technique to ensure $J > -1$. For a given J, the results of Theorem 3.2 and Remark 3.1 can be extended as in Chapter 2. This requires a different proof technique, though. Basically, an asymptotic expansion for x_s and y_s is constructed, where the first term is the exact solution x, y. This latter proof technique follows more along traditional singular perturbation lines (see [112, 89, 69]), and is not as close to Theorem 3.1 and to DAE concepts.* □

Taking $\alpha_2 = 1$ without loss of generality, we obtain the iteration

$$x'_s = f - By_{s-1} - \epsilon^{-1}BEg(x_s, t) \tag{3.17}$$

This is a singular, singularly perturbed problem (so ϵ should not be taken extremely small compared to machine precision even if a stiff solver is being used). If GB is positive definite then we may choose $E = I$, and this yields a very simple iteration in (3.17) which avoids the inversion necessary in stabilization methods like Baumgarte's. However, if an explicit discretization method of order p is contemplated then approximately p SRM iterations like (3.17) are needed, because one must choose $\epsilon = O(h)$, where h is the step size.

3.2 Nonlinear, Singular Index-2 Problems

In this section we consider the nonlinear index-2 problem (3.1) with an isolated singular point $t^\star$, i.e. GB is singular at $t^\star$. For simplicity, we assume that B and g are independent of t. Denote $P(x) = B(GB)^{-1}G$. Motivated by constrained multibody systems (see Example 2.1), we assume $P(x)$ to be differentiable in t, but $\frac{\partial P}{\partial x}(x)$ may be unbounded. For this reason, we consider only the case $\alpha_1 = 0$ in this section (cf. Chapter 2 for the linear case). In the drift equation (3.8) we then have essentially the singularly perturbed operator $\epsilon v' + GBEv$ to consider. The choices of $E = I$ or $E = (GB)^T$ yield a turning point problem (i.e., at least one of eigenvalues of the matrix GBE vanishes at the point $t^\star$), which complicates the analysis, even in the linear case , and degrades the numerical performance as well in our experience. Therefore, we choose $E = (GB)^{-1}$. In the sequel we will be careful to evaluate the effect of E only when its singularity limit is well-defined, as e.g. in $P(x)$.

A direct generalization of the linear case in Chapter 2 would give the SRM formulation (3.4), (3.5) where instead of updating y (because y may be unbounded at

t^*) we update By by

$$B(x_s)y_s = B(x_{s-1})y_{s-1} + \frac{1}{\epsilon}B(x_s)(G(x_s)B(x_s))^{-1}g(x_s). \tag{3.18}$$

However, (3.18) needs to be modified, since we may have $RangeB(x_s) \neq RangeB(x_{s-1})$. So we use the projection $P(x_s)$ to move from $RangeB(x_{s-1})$ to $RangeB(x_s)$. Then we consider the following SRM formulation for singular problems:

$$x_s' = f(x_s, t) - B(x_s)y_s, \tag{3.19a}$$

$$B(x_s)y_s = P(x_s)B(x_{s-1})y_{s-1} + \frac{1}{\epsilon}B(x_s)(G(x_s)B(x_s))^{-1}g(x_s), \tag{3.19b}$$

where x_s satisfies the boundary condition (3.2).

If the assumptions given at the beginning of §3.1 and in Theorem 3.2 remain valid, then the result of Theorem 3.2 still holds. Unfortunately, for the singular problem, assumption (3.6) may not be true in general. To see this, consider one iteration, i.e. $s = 1$. The accuracy for the approximation of x depends on the extent that the bound (3.6a) holds. Numerical experiments show that we can get a good approximation of x near the singularity. But the situation for By is worse, and the bound (3.6b) often does not hold. Indeed, assume for the moment that we have a good, smooth approximation of x, say $x_s = \hat{x}$, i.e. (3.7) holds with $\delta, \delta' = O(\epsilon)$, and $B(\hat{x})\hat{y}$ is defined by (3.19b) for some $B(x_{s-1})y_{s-1}$. From (3.7) we have

$$B(\hat{x})\hat{y} = P(\hat{x})f(\hat{x}, t) + \eta, \tag{3.20}$$

where $\eta = B(\hat{x})(G(\hat{x})B(\hat{x}))^{-1}\delta'$. It is not difficult to find that the exact $B(x)y$ from (3.1) satisfies

$$B(x)y = P(x)f(x, t). \tag{3.21}$$

Yet, even if η is small, $B(\hat{x})\hat{y}$ may not be a good approximation of By because $\frac{\partial P}{\partial x}$ may be unbounded at the singular point so that $P(\hat{x})$ is not a good approximation of $P(x)$.

Example 3.2 *In (3.1) let* $x = (x_1, x_2)$, $g(x) = -\cos x_1 - \cos x_2$, *and* $G = B^T = \begin{pmatrix} \sin x_1 & \sin x_2 \end{pmatrix}$. *Then* $P(x) = (\sin^2 x_1 + \sin^2 x_2)^{-1} \begin{pmatrix} \sin^2 x_1 & \sin x_1 \sin x_2 \\ \sin x_1 \sin x_2 & \sin^2 x_2 \end{pmatrix}$. *Clearly, at a singular point* $x = (0, \pi)$, *the value of* P *depends on the direction from which it is approached. Thus,* $\frac{\partial P}{\partial x}$ *is unbounded, even though* P *is a differentiable function of* t.

Further letting $f = (\sin x_2 - \sin x_1)^{-1} \begin{pmatrix} \sin x_2 & 2\sin x_2 - \sin x_1 \end{pmatrix}^T$, *and given the initial conditions* $x_1(0) = -\pi/2$, $x_2(0) = \pi/2$, *the exact solution is*

$$x(t) = \begin{pmatrix} t - \pi/2 & t + \pi/2 \end{pmatrix}^T, \; y = (\sin x_2 - \sin x_1)^{-1}$$

Thus, as t *crosses* $t^* = \pi/2$, $y(t)$ *becomes unbounded, but*

$$By = (\sin x_2 - \sin x_1)^{-1} \begin{pmatrix} \sin x_1 & \sin x_2 \end{pmatrix}^T$$

remains bounded. However, it is easy to perturb $x(t)$ *slightly and smoothly in such a way that the perturbed* By *becomes unbounded near* $t = t^*$, *still satisfying (3.7) with a small* δ. $\Box$

Note that for the linear model problem (see Chapter 2), $P \equiv P(t)$ is independent of x. Hence we do not have the above difficulty in the linear case. For the nonlinear problem, the accuracy near the singular point is reduced and it no longer behaves like $O(\epsilon^s)$ for more than one iteration. However, we do expect $O(\epsilon^s)$ accuracy away from the singular point, assuming that no bifurcation or impasse point is encountered by the approximate solution .

3.3 SRM for Nonlinear Higher-index Problems

We now generalize the SRM to the more general problem (2.15). In particular, we consider the index-3 problem ($\nu = 2$). The Euler-Lagrange equations for multibody systems with holonomic constraints yield a practical instance of the problem. The

SRM formulations presented in this section are easy to generalize for more general problems (2.15) (index $\nu + 1$). The index-3 problem reads:

$$x'' = f(x, x', t) - B(x, t)y, \tag{3.22a}$$

$$0 = g(x, t), \tag{3.22b}$$

with given $2(n_x - n_y)$ boundary conditions,

$$b(x(0), x(t_f), x'(0), x'(t_f)) = 0. \tag{3.23}$$

The meaning of G, B and the stabilization matrix E below remain the same as in the index-2 problems considered in previous sections.

3.3.1 The case of nonsingular GB

We first use the idea from previous sections, viz. a combination of Baumgarte's stabilization with a modified penalty method, to derive the SRM for the nonlinear index-3 problem (3.22). Then we apply a better stabilization [8] to generate a new SRM which is expected to have better constraint stability. Finally, we seek variants which avoid evaluation of complicated terms in the second derivative of the constraints.

First consider, instead of (3.22b), the Baumgarte's stabilization

$$\alpha_1 \frac{d^2}{dt^2} g(x, t) + \alpha_2 \frac{d}{dt} g(x, t) + \alpha_3 g(x, t) = 0, \tag{3.24a}$$

$$g(x(0), 0) = 0, \ \frac{d}{dt} g(x(0), 0) = 0, \tag{3.24b}$$

where $\alpha_j, j = 1, 2, 3$ are chosen so that the roots of the polynomial

$$\sigma(\tau) = \sum_{j=1}^{3} \alpha_j \tau^{3-j}$$

are all negative. Following the procedure of previous sections, we can write down an SRM for (3.22): for $s = 1, 2, \ldots$ and y_0 given,

$$x_s'' = f(x_s, x_s', t) - B(x_s, t)y_s, \tag{3.25}$$

where x_s satisfies boundary conditions (3.23) and (3.24b) and y_s is given by

$$y_s = y_{s-1} + \frac{1}{\epsilon}E(x_s,t)(\alpha_1\frac{d^2}{dt^2}g(x_s,t) + \alpha_2\frac{d}{dt}g(x_s,t) + \alpha_3 g(x_s,t)). \tag{3.26}$$

It is not difficult to repeat the approach of §3.1 for the present case. Under assumptions similar to the index-2 case, i.e. (3.6) with a change to include $\delta''(\tau)$ at the right hand side (cf. [58]) and Assumption 3.1 with the addition that the derivative of the solution is also bounded, we readily obtain extensions of Theorems 3.1 and 3.2 for the cases $\alpha_1 \neq 0$ and $\alpha_1 = 0$ (with $\alpha_2 \neq 0$), respectively. We do not allow $\alpha_1 = \alpha_2 = 0$ since in this case equations (3.25),(3.26) have different asymptotic properties. Note that the SRM (3.25),(3.26) with $\alpha_1 = 0$ avoids computing g_{xx}; however, the iteration obtained now calls for solving problems which become stiff when ϵ gets small, and to avoid g_{xx} one should use a non-stiff discretization method. This formulation with $E = I$ and $\alpha_1 \neq 0$ is the same as that proposed in [20, 19] using the augmented Lagrangian method.

Another way to generalize the SRM to higher index problems is based on invariant stabilization. Its advantages over Baumgarte's stabilization have been discussed in [8, 9]. We thus prefer the way based on this stabilization. Theoretical evidence is also mentioned in Remark 3.4. We first describe this new stabilization. By two direct differentiations of the constraints (3.22b), we can eliminate y and get an ODE

$$x'' = \tilde{f}(x, x', t), \tag{3.27}$$

for which the original constraint (3.22b) together with its first derivative give an invariant. The idea of the method is to reformulate the higher index DAE (3.22) as a first order ODE (cf. (3.27)):

$$z' = \hat{f}(z,t) \tag{3.28}$$

with an invariant

$$0 = h(z,t), \tag{3.29}$$

where

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} x \\ x' \end{pmatrix}, \; \hat{f}(z,t) = \begin{pmatrix} z_2 \\ \tilde{f}(z,t) \end{pmatrix}, \; h(z(t),t) = \begin{pmatrix} g(x(t),t) \\ \frac{d}{dt}g(x(t),t) \end{pmatrix} \tag{3.30}$$

and to consider the stabilization families

$$z' = \hat{f}(z,t) - \gamma F(z,t)h(z,t), \tag{3.31}$$

where $F = D\tilde{E}$ for some appropriate matrix functions D and $\tilde{E}$ such that $\tilde{E}$ and HD are nonsingular and $H = h_z$. The ODE (3.31) coincides with Baumgarte's stabilization for the index-2 problem (3.1) with $D = B$ and $\tilde{E} = E = (HD)^{-1}$. One choice for D is $D = H^T$, but others will be mentioned below. Note that (3.31) has the same solution as the original problem (3.22) for any parameter value γ. Although the method has better constraint stabilization, both the evaluation of $\tilde{f}$ and that of H involve g_{xx} which may be complicated to calculate in practice.

Next, we present an SRM method based on invariant stabilization which avoids the computation of $\tilde{f}$. In fact, we can avoid g_{xx} altogether using the new stabilization. If we do not eliminate y by differentiations, $\hat{f}(z,t)$ in the stabilization (3.31) becomes

$$\hat{f}(z,t) = \begin{pmatrix} z_2 \\ f(z,t) - B(z_1,t)y \end{pmatrix}. \tag{3.32}$$

Since y is not known in advance, we use an iterative SRM procedure to calculate y as in [20, 11]. The solutions of the iterative procedure no longer satisfy (3.22) precisely. Hence the iterative procedure has to be a regularization procedure and the parameter in (3.31) is changed to $\gamma = \frac{1}{\epsilon}$ to emphasize that it must be chosen sufficiently large. These lead to the following SRM formulation (for simplicity of notation, we only consider the special case where B and g are independent of t):

$$z_s' = \begin{pmatrix} z_{1s} \\ z_{2s} \end{pmatrix}' = \begin{pmatrix} z_{2s} \\ f(z_s,t) - B(z_{1s})y_{s-1} \end{pmatrix} - \frac{1}{\epsilon}F(z_s)h(z_s), \tag{3.33}$$

where z_s satisfies boundary conditions (3.23), (3.24b) and $h = (g(z_1), G(z_1)z_2)^T$. Thus the Jacobian of h is

$$H = \begin{pmatrix} G(z_1) & 0 \\ L(z) & G(z_1) \end{pmatrix}, \quad \text{where} \quad L = z_2^T g_{xx}(z_1).$$

We choose D and $\tilde{E}$ so that

$$F = BE \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} = \begin{pmatrix} BE & 0 \\ 0 & BE \end{pmatrix} \tag{3.34}$$

where, as in §3.1, E is chosen such that GBE has non-negative eigenvalues. Updating y by

$$y_s = y_{s-1} + \frac{1}{\epsilon} E(z_{1s}) G(z_{1s}) z_{2s} \tag{3.35}$$

implies that the second part of the original index-3 system holds exactly, i.e.

$$z_{2s}' = f(z_s, t) - B(z_{1s}) y_s.$$

Next we analyze the convergence of (3.33)–(3.35). Again we assume that the solutions of (3.33), (3.23), (3.24b) exist uniquely and are bounded if y_{s-1} is bounded (see Assumption 3.1). Assumption (3.6) changes slightly: We first rewrite the system (3.22) as

$$z_1' = z_2, \tag{3.36a}$$

$$z_2' = f(z,t) - B(z_1) y, \tag{3.36b}$$

$$0 = g(z_1). \tag{3.36c}$$

Then we assume the following perturbation bound,

$$\|\hat{z}(t) - z(t)\| \leq M \max_{0 \leq \tau \leq t_f} (|\delta(\tau)| + |\delta'(\tau)| + |\delta''(\tau)| + |\theta(\tau)| + |\theta'(\tau)|), \tag{3.37a}$$

$$\|\hat{y}(t) - y(t)\| \leq M \max_{0 \leq \tau \leq t_f} (|\delta(\tau)| + |\delta'(\tau)| + |\delta''(\tau)| + |\theta(\tau)| + |\theta'(\tau)|), \tag{3.37b}$$

where $\hat{z}$ and $\hat{y}$ satisfy a perturbed problem of (3.36),

$$\hat{z}_1' = \hat{z}_2 + \theta(t), \tag{3.38a}$$

$$\hat{z}_2' = f(\hat{z}, t) - B(\hat{z}_1)\hat{y}, \tag{3.38b}$$

$$0 = g(\hat{z}_1) + \delta(t), \tag{3.38c}$$

with the same boundary conditions (3.23). Again, for initial value problems, (3.37) can be easily proved by following the technique presented in [58], and this can be extended for boundary value problems as well.

Similarly to the proof of Theorem 3.1, let $h(z_s) = \begin{pmatrix} v_s \\ w_s \end{pmatrix}$, where $v_s = g(z_{1s})$, $w_s = G(z_{1s})z_{2s}$. From (3.33), we get the drift equations (cf. (3.15))

$$\epsilon v_s' = -GBE(z_s)v_s + \epsilon w_s \tag{3.39a}$$

$$\epsilon v_s'' = -GBE(z_s)v_s' - LBE(z_s)v_s + \epsilon[Lz_{2s} + Gf(z_s) - GB(z_s)y_{s-1}] \tag{3.39b}$$

with the initial conditions $v_s(0) = 0$, $w_s(0) = 0$. Applying (3.39a) and then (3.39b) for $s = 1$, we obtain $v_1 = O(\epsilon), w_1 = O(\epsilon)$. Then (3.39a) further yields

$$v_1 = O(\epsilon^2), \quad v_1' = O(\epsilon)$$

Comparing (3.38) with (3.33)–(3.35), we have to bound

$$\delta = v_1, \quad \theta = -\epsilon^{-1}BEv_1$$

and their derivatives appearing in (3.37). We already have that

$$\delta = O(\epsilon^2), \ \delta' = O(\epsilon), \ \theta = O(\epsilon)$$

The procedure that follows continues to be similar to the one employed in the proof of Theorem 3.2, so we only sketch it here for $s = 1$. From (3.39a) we obtain $v_1'(0) = 0$. Using the condition $y_0(0) = y(0)$ gives

$$[Lz_{21} + Gf(z_1) - GB(z_1)y_0]|_{t=0} = 0$$

so, from (3.39b), also $w_1'(0) = 0$. Differentiating (3.39b) we get

$$\epsilon w_1'' + GBE(z_1)w_1' = O(\epsilon), \quad w_1'(0) = 0$$

hence $w_1' = O(\epsilon)$. Differentiating (3.39a) we next have

$$\epsilon v_1'' + GBE(z_1)v_1' = O(\epsilon^2), \quad v_1'(0) = 0$$

so $v_1' = O(\epsilon^2)$. This implies

$$\theta' = O(\epsilon), \quad \delta'' = O(\epsilon)$$

We can now use (3.37) and obtain the desired conclusion for $s = 1$,

$$z_1 = z + O(\epsilon), \; y_1 = y + O(\epsilon),$$

where $\{z, y\}$ is the exact solution of the index-3 problem. Then, continuing to follow the proof procedure of Theorem 3.2, we obtain:

Theorem 3.3 *Let all functions in the DAE (3.22) be sufficiently smooth and assume the above assumptions (particularly (3.37)) hold. Assume in addition that y_0 satisfies (3.14). Then, for the solution of iteration (3.33)–(3.35), the following error estimates hold:*

$$\begin{aligned} z_s(t) - z_e(t) &= O(\epsilon^s), && (3.40a) \\ y_s(t) - y_e(t) &= O(\epsilon^s) && (3.40b) \end{aligned}$$

for $1 \le s \le J + 1$.

Remark 3.3 *Extensions of this theorem to the boundary value case and to $s > J+1$ away from an initial layer are possible, similarly to the extensions for Theorem 3.2 contained in Remarks 3.1 and 3.2.* □

Remark 3.4 *We note that, unlike* Proposition 2.2 *of [8], we do not assume*

$$\|H(z)\hat{f}(z)\|_2 \leq \gamma_0 \|h(z)\|_2$$

to discuss the stability and accuracy of the constraints. Also, from (3.39), we see the difference of the constraint stability or accuracy between SRM formulations based on Baumgarte's stabilization and the new stabilization. For the former, we only have

$$v_1' = G(z_{11})z_{11}' = G(z_{11})z_{21} = w_1$$

So if we obtain $w_1 = O(\epsilon)$ *then* $v_1 = O(t\epsilon)$. *This can be much worse than what we get from (3.39a).* □

3.3.2 The case for constraint singularities

For the singular case GB may be singular at some isolated point t^* as described in the previous sections. The situation here is similar to that for index-2 problems. An examination of the drift equations (3.39) suggests that here, too, the choice $E = (GB)^{-1}$ is preferable to $E = I$ or $E = (GB)^T$. The iteration for y_s is modified as well. Still assuming for simplicity that g and B do not depend explicitly on t, this gives, in place of (3.33)–(3.35) the iteration

$$\begin{aligned} z_{1s}' &= z_{2s} - \frac{1}{\epsilon}B(GB)^{-1}g(z_s) && (3.41a)\\ z_{2s}' &= f(z_s,t) - \hat{y}_s && (3.41b)\\ \hat{y}_s &= P(z_s)\hat{y}_{s-1} + \frac{1}{\epsilon}P(z_s)z_{2s} && (3.41c) \end{aligned}$$

Also, as indicated in §3.2 for index-2 problems, we cannot expect $O(\epsilon^s)$ approximation near the singular point any more. But we do expect that (3.40) holds away from the singular point, because the singularity is in the constraint and the drift manifold is asymptotically stable (following our stabilization). A numerical example in §3.5 will show that we do get improved results by using SRM iterations for the singular problem.

3.4 SRM for Constrained Multibody Systems

Constrained multibody systems provide an important family of applications of the form (3.22) and (3.1). We consider the system

$$q' = v \tag{3.42a}$$

$$M(q)v' = f(q,v) - G(q)^T\lambda \tag{3.42b}$$

$$0 = g(q) \tag{3.42c}$$

where q and v are the vectors of generalized coordinates and velocities, respectively; M is the mass matrix which is symmetric positive definite; $f(q,v)$ is the vector of external forces (other than constraint forces); $g(q)$ is the vector of (holonomic) constraints; λ is the vector of Lagrange multipliers; and $G(q) = \frac{d}{dq}g$. For notational simplicity, we have suppressed any explicit dependence of M, f or g on the time t. We first consider the problem without singularities.

Corresponding to (3.22) in §3.3, we have $B = M^{-1}G^T$, so $GB = GM^{-1}G^T$. Other quantities like h and H retain their meaning from the previous section. In some applications it is particularly important to avoid terms involving g_{xx}, since its computation is somewhat complicated and may also easily result in mistakes and rugged terms. So [9] suggests post-stabilization using the stabilization matrix

$$F = M^{-1}G^T(GM^{-1}G^T)^{-1}\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \tag{3.43}$$

twice, instead of involving H, at the end of each time step or as needed. They find that this F performs very well in many applications. However, while this stabilization avoids the g_{xx} term in F, g_{xx} is still involved in obtaining $\tilde{f}$, although only through matrix-vector multiplications (see (3.27)). The SRM formulation (3.33)–(3.35) enables us to avoid the computation of $\tilde{f}$ in the absence of constraint singularities. For the multibody system (3.42) we write the iteration as follows:

For $s = 1, 2, \ldots$, find $\{q_s, v_s\}$ by

$$q'_s = v_s - \frac{1}{\epsilon} BE(q_s)g(q_s) \tag{3.44a}$$

$$v'_s = M^{-1}f(q_s, v_s) - B(q_s)\lambda_{s-1} - \frac{1}{\epsilon} BEG(q_s)v_s \tag{3.44b}$$

Then update λ by

$$\lambda_s = \lambda_{s-1} + \frac{1}{\epsilon} EG(q_s)v_s. \tag{3.45}$$

It is easy to see that in this SRM formulation the g_{xx} term is avoided completely. Moreover, since $GM^{-1}G^T$ is positive definite, we can choose $E = I$ in (3.44),(3.45), obtaining a method for which Theorem 3.3 applies, which avoids computing $(GM^{-1}G^T)^{-1}$. Although it requires an iterative procedure, a small number of iterations (p if an explicit discretization method of order p is used) typically provide sufficient accuracy. Numerical experiments will show the $O(\epsilon^s)$ error estimate.

Next we consider the singular problem, i.e. with the matrix $GM^{-1}G^T$ being singular at some isolated point t^*, $0 < t^* < t_f$. A typical example of singular multibody systems is the two-link slider-crank problem (see Example 2.1 and Figure 2.1) consisting of two linked bars of equal length, with one end of one bar fixed at the origin, allowing only rotational motion in the plane, with the other end of the other bar sliding along the x-axis.

Various formulations of the equations of motion for this problem appear, e.g., in [60, 19, 11, 12, 94]. In our calculations we have used the formulation in Example 2.1 (see [11]), to make sure that the problem is not accidentally too easy. It consists of 6 ODEs and 5 constraints, with the last row of the Jacobian matrix G vanishing when the mechanism moves left through the point where both bars are upright ($\phi_1 = \frac{\pi}{2}$, $\phi_2 = \frac{3\pi}{2}$, where x_i, y_i, ϕ_i are the coordinates of the center of mass of the ith bar).

The last row of G vanishes at this point and a singularity is obtained. We note that the solution is smooth in the passage through the singularity with a nonzero velocity.

When we attempt to integrate this system using a stabilization method like [8] which ignores the singularity, the results are unpredictable, depending on how close to the singular time point the integration process gets when attempting to cross it. In fact, radically different results may be obtained upon changing the value of the error tolerance. (Similar observations are made in [94].) In some instances a general purpose ODE code would simply be unable to "penetrate the singularity" and yield a solution which, after hovering around the upright (singular) position for a while, turns back towards the initial position (solid line in Figure 2.1). Such a pattern of motion may well look deceptively plausible.

Methods which do not impose the constraints on the position level (e.g. methods consisting of differentiating the constraints once and solving the resulting index-2 problem numerically, or of projecting only on the velocity-level constraint manifold) perform particularly poorly (cf. numerical results in [94]). This is easy to explain: The position-level constraint corresponds to ensuring that the two bars have equal length. If this is not strictly imposed in the process of numerical solution, inevitable numerical errors due to discretization may yield a model where the lengths are not close enough to being equal, and this leads to the lock-up phenomena described e.g. in [60], which have a vastly different solution profile.

We now wish to generalize the SRM to problem (3.42) with singularities since we have seen its success for the linear index-2 case. From the two-link slider crank problem, we find that, although $GM^{-1}G^T$ is singular at t^*, $P(q) \equiv M^{-1}G^T(GM^{-1}G^T)^{-1}G$ and $M^{-1}G^T(GM^{-1}G^T)^{-1}g$ are smooth functions of t for the exact solution or functions q satisfying the constraints, while $M^{-1}G^T(GM^{-1}G^T)^{-1}$, $M^{-1}G^T(GM^{-1}G^T)^{-1}G_q$ and the derivative $\frac{dP(q)}{dq}$ are not. Also, as indicated in [11], λ is no longer smooth, while $B\lambda$ is since we assume the solution q to be sufficiently smooth. We only include terms which are (most possibly) smooth in the SRM formulation.

Applying (3.41), we obtain the method

$$q_s' = v_s - \frac{1}{\epsilon}M^{-1}G^T(GM^{-1}G^T)^{-1}g(q_s), \tag{3.46a}$$

$$v_s' = M^{-1}f(q_s, v_s, t) - \hat{\lambda}_s, \tag{3.46b}$$

$$\hat{\lambda}_s = P(q_s)\hat{\lambda}_{s-1} + \frac{1}{\epsilon}P(q_s)v_s \tag{3.47}$$

As we indicated in §3.2, we do not expect $O(\epsilon^s)$ accuracy near the singular point. However, we do expect that the SRM iteration would improve the accuracy and that we still expect to get $O(\epsilon^s)$ accuracy away from the singular point. Numerical experiments in §3.5 will show such improvements.

3.5 Numerical Experiments

We now present a few examples to demonstrate our claims in the previous sections. Throughout this section we use a constant step size h. To make life difficult we choose h when we can so that there is an i such that $t_i = t^*$, namely, there is a mesh point hitting the singularity point t^*, for singular test problems. At a given time t, we use $'ex'$ to denote the maximum over all components of the error in x_s. Similarly, $'drift'$ denotes the maximum residual in the algebraic equations.

Example 3.3 *Consider the DAE (2.16),(2.17) with*

$$f = \begin{pmatrix} 1 - e^{-t} \\ \cos t + e^t \sin t \end{pmatrix}, \; B = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$
$$g = \frac{1}{2}(x_1^2 + x_2^2 - e^{-2t} - \sin^2 t).$$

subject to $x_1(0) = 1, \; x_2(0) = 0.$

The exact solution is $x_e = (\, e^{-t}, \;\; \sin t\,)$, $y_e = e^t$. *This is a problem without singularities.*

Using an explicit second order Runge-Kutta method with $h = 0.001$ we test various choices of E and α_1 (always taking $\alpha_2 = 1$) of the SRM formulation in §3.1. We list the computational results in Table 3.1. Observe that, for $\alpha_1 \neq 0$, the SRM works well

methods	ϵ	iteration	error at $\rightarrow$	t=.1	t=.5	t=1.0
$\alpha_1 = 1$	1e-8	1	ex	.11e-7	.94e-7	.19e-6
$E = I$			drift	.79e-8	.56e-7	.14e-6
$\alpha_1 = 1$	1e-8	1	ex	.11e-7	.92e-7	.18e-6
$E = (GB)^T$			drift	.78e-8	.53e-7	.14e-6
$\alpha_1 = 1$	1e-8	1	ex	.11e-7	.95e-7	.19e-6
$E = (GB)^{-1}$			drift	.80e-8	.58e-7	.15e-6
Baumgarte			ex	.45e-6	.16e-6	.35e-6
			drift	.40e-6	.70e-7	.29e-6
$\alpha_1 = 0$	5e-3	1	ex	.60e-2	.11e-1	.11e-1
$E = I$			drift	.54e-2	.80e-2	.13e-1
		2	ex	.11e-3	.26e-3	.22e-3
			drift	.96e-4	.20e-3	.27e-3
		3	ex	.32e-5	.65e-5	.46e-5
			drift	.29e-5	.47e-5	.54e-5
		4	ex	.26e-6	.23e-6	.28e-6
			drift	.13e-6	.51e-7	.12e-6
$\alpha_1 = 0$	5e-3	1	ex	.70e-2	.12e-1	.13e-1
$E = (GB)^T$			drift	.64e-2	.13e-1	.15e-1
		2	ex	.22e-3	.65e-3	.31e-3
			drift	.20e-3	.49e-3	.29e-3
		3	ex	.11e-4	.16e-4	.69e-5
			drift	.10e-4	.10e-4	.52e-5
		4	ex	.85e-6	.91e-7	.29e-6
			drift	.75e-6	.77e-6	.14e-6
$\alpha_1 = 0$	5e-3	1	ex	.51e-2	.66e-2	.10e-1
$E = (GB)^{-1}$			drift	.46e-2	.49e-2	.12e-1
		2	ex	.35e-4	.11e-3	.21e-3
			drift	.30e-4	.79e-4	.24e-3
		3	ex	.86e-6	.23e-5	.47e-5
			drift	.77e-6	.17e-5	.53e-5
		4	ex	.26e-6	.18e-6	.26e-6
			drift	.26e-7	.31e-7	.13e-6

Table 3.1: Errors for Example 3.3 using the explicit second order Runge-Kutta scheme

for various choices of E. Its error is as good as Baumgarte's method whose parameter corresponds to the α_2 of the SRM. For $\alpha_1 = 0$, we see that the error improves at a rate of about $O(\epsilon)$ for various choices of E, including $E = I$. (Observe the errors at $t = 1$; the error situation near $t = .1$ is different because of an initial layer.) Such an error *improvement continues until the accuracy of the second order explicit Runge-Kutta method, i.e. $O(h^2)$, is reached.* □

The next two examples are problems with singularities. In the index-2 case of the Baumgarte stabilization the worst term is $B(GB)^{-1}g_t$ for the type of the singularities in this paper. To show what happens when the Baumgarte method does not work well, we choose nonautonomous problems (i.e. $g_t \neq 0$) as index-2 singular examples.

Example 3.4 *Consider the nonlinear DAE (2.16) with*

$$\begin{aligned} f &= \begin{pmatrix} 1+(t-\frac{1}{2})e^t \\ 2t+(t^2-\frac{1}{4})e^t \end{pmatrix}, \; B = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ g &= \frac{1}{2}(x_1^2+x_2^2-(t-\frac{1}{2})^2-(t^2-\frac{1}{4})^2) \end{aligned}$$

subject to the initial condition $x_1(0) = -\frac{1}{2}$, $x_2(0) = -\frac{1}{4}$.

The exact solution is $x_e = (\, t-\frac{1}{2}, \;\; t^2-\frac{1}{4}\,)$, $y_e = e^t$. A singularity is located at $t^* = \frac{1}{2}$. Using this example we test the SRM formulations of §3.2. We list the computational results in Table 3.2, where we take $h = \epsilon = 0.001$ for the case of $\alpha_1 = 0$, and $h = 0.001$, $\epsilon = 10^{-10}$ for the case of $\alpha_1 \neq 0$, and use the explicit second order Runge-Kutta scheme to easily see the iteration improvement (Ij stands for results of the jth iteration).

From Table 3.2, we see the error's deterioration for the Baumgarte method and the SRM with $\alpha_1 \neq 0$. The SRM with $\alpha_1 = 0$ performs better in the singular case. □

Next we try an example in which y is unbounded at the singularity.

Example 3.5 *Consider the nonlinear DAE (3.1) with*

$$\begin{aligned} f &= \begin{pmatrix} -x_1+x_2-\sin(t)-(1+2t) \\ 0 \end{pmatrix}, \; B = \begin{pmatrix} 0 \\ x_1 \end{pmatrix} \\ g &= x_1^2+x_1(x_2-\sin(t)-1+2t), \end{aligned}$$

methods	error at →	t=.1	t = .3	t = .5	t= .7	t= 1.0
$\alpha_1 = 1$	ex	.39e-6	.13e-5	.12e-3	.14e-3	.76e-4
	drift	.24e-6	.16e-6	.10e-7	.39e-6	.75e-6
$\alpha_1 = 0$ (I1)	ex	.46e-3	.32e-3	.43e-4	.49e-3	.20e-2
	drift	.24e-3	.89e-4	.18e-8	.20e-3	.22e-2
$\alpha_1 = 0$ (I2)	ex	.81e-6	.11e-5	.41e-5	.29e-5	.68e-5
	drift	.24e-6	.30e-6	.15e-10	.13e-5	.76e-5
$\alpha_1 = 0$ (I3)	ex	.23e-6	.26e-6	.34e-6	.29e-6	.29e-6
	drift	.90e-9	.11e-8	.78e-13	.35e-8	.18e-7
$\alpha_1 = 0$ (I4)	ex	.23e-6	.26e-6	.36e-6	.27e-6	.29e-6
	drift	.47e-11	.33e-11	.10e-12	.29e-11	.28e-10
Baumgarte	ex	.43e-6	.45e-6	.34e-3	.39e-3	.21e-3
	drift	.24e-6	.16e-6	.61e-7	.24e-6	.75e-6

Table 3.2: Example 3.4 – bounded y and singularity at $t^* = .5$

subject to the initial condition $x_1(0) = 1,\ x_2(0) = 0$.

The exact solution is $x_e = (\,1 - 2t,\quad \sin t\,)$, $y_e = -\cos t/(1 - 2t)$. *Taking the same parameters and using the same method as before, we get the results listed in Table 3.3. Clearly, the SRM with* $\alpha_1 = 0$ *performs well for this situation, while*

methods	error at →	t=.1	t = .3	t = .5	t= .7	t= 1.0
SRM ($\alpha_1 = 0$)	ex	.40e-6	.25e-6	.14e-6	.46e-7	.60e-7
(I3)	drift	.25e-8	.76e-9	.16e-15	.28e-9	.40e-9
Baumgarte	ex	.49e-7	.15e-6	.93e+1	NaN	NaN
	drift	.39e-7	.59e-7	.52e+13	NaN	NaN

Table 3.3: Example 3.5 – unbounded y and singularity at $t^* = .5$

Baumgarte method blows up upon hitting the singularity. □

Our next example tests the formulation (3.33)–(3.35) or (3.44)–(3.45) for index-3 problems.

Example 3.6 *This example is made up from Example 2 in [9] (see Figure 3.1), which describes a two-link planar robotic system. We use the notation of (1.11). Let*

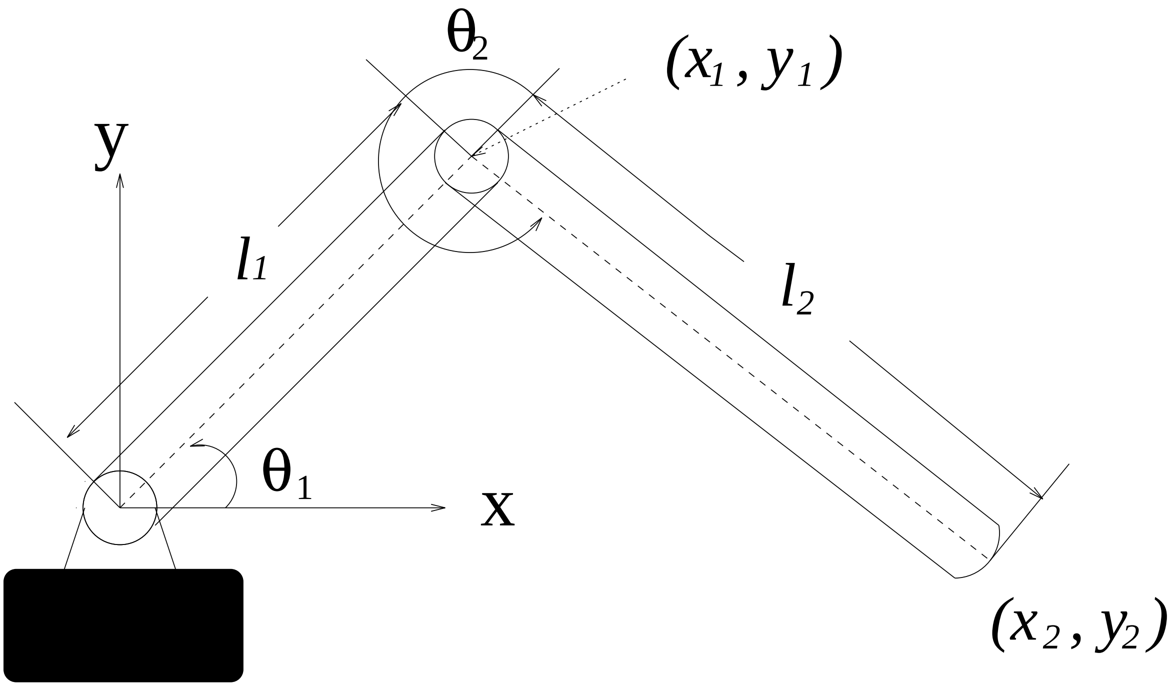

Figure 3.1: Two-link planar robotic system

$q = (\theta_1, \theta_2)^T$ and

$$M = \begin{pmatrix} m_1 l_1^2/3 + m_2(l_1^2 + l_2^2/3 + l_1 l_2 c_2) & m_2(l_2^2/3 + l_1 l_2 c_2/2) \\ m_2(l_2^2/3 + l_1 l_2 c_2/2) & m_2 l_2^2/3 \end{pmatrix},$$

where $l_1 = l_2 = 1$, $m_1 = m_2 = 3$ and $c_2 = \cos\theta_2$. The constraint equation is

$$g(q) = l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2) = 0.$$

We choose the force term

$$f = \begin{pmatrix} (l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2)) \cos t - 3 \sin t \\ l_2 \cos(\theta_1 + \theta_2) \cos t + (1 - \frac{3}{2} c_2) \sin t \end{pmatrix}$$

which yields the exact solution $\theta_1 = \sin t$, $\theta_2 = -2 \sin t$ and $\lambda = \cos t$. Because M is (symmetric) positive definite and $B = M^{-1} G^T$ we can take $E = I$ in the SRM formula (3.44)–(3.45). Again we use the second- order explicit Runge-Kutta scheme, and set $h = 0.001$, $\epsilon = 0.005$. The results are listed in Table 3.4, where eq and ev stand for maximum errors in q and $v = q'$, resp., and $pdrift$ and $vdrift$ stand for drifts at the *position and velocity level, resp. We see that the accuracy is improved significantly by the first two iterations. The third iteration is unnecessary here, because the error is already dominated by the Runge-Kutta discretization error. Qualitatively similar results are obtained for $E = (GB)^T$ and $E = (GB)^{-1}$. More interestingly, though, for $E = I$ we neither form nor invert $GM^{-1}G^T$, so a particularly inexpensive iteration is obtained.*

methods	ϵ	iteration	error at $\rightarrow$	t=.1	t=.5	t=1.0
$E = I$	5e-3	1	eq	.41e-4	.66e-3	.26e-2
			ev	.75e-2	.74e-2	.69e-2
			pdrift	.22e-4	.28e-4	.22e-4
			vdrift	.49e-2	.41e-2	.27e-2
		2	eq	.13e-6	.66e-6	.36e-6
			ev	.19e-5	.81e-6	.20e-4
			pdrift	.42e-9	.13e-7	.17e-6
			vdrift	.91e-7	.21e-5	.21e-4
		3	eq	.10e-6	.58e-6	.12e-5
			ev	.86e-6	.10e-5	.16e-5
			pdrift	.96e-11	.60e-9	.48e-8
			vdrift	.10e-8	.99e-7	.59e-6

Table 3.4: Errors for Example 3.6 using SRM (3.45)-(3.46)

□

Next we solve for the dynamics of the slider-crank mechanism described in Example 2.1. this is a nonlinear index-3 DAE with isolated, "smooth" singularities.

Example 3.7 *We take $h = \epsilon = 0.0001$ and use the explicit second order Runge-Kutta method again. Singularities are located at $(\phi_1, \phi_2) = (\frac{\pi}{2}, \frac{3\pi}{2})$ (i.e., each time the periodic solution passes this point). Corresponding to the case shown in [94], we choose $\phi_1(0) = \frac{7\pi}{4}$ and $\phi_1'(0) = 0$ and compute*

$$\theta_1 = \phi_1 - \frac{3\pi}{2},\ \theta_2 = \phi_2 + \frac{\pi}{2},$$

θ_1' and θ_2'. Using the formulation (3.47), (3.46), we calculate until $t = 70$ without any difficulty (see Figure 3.2).

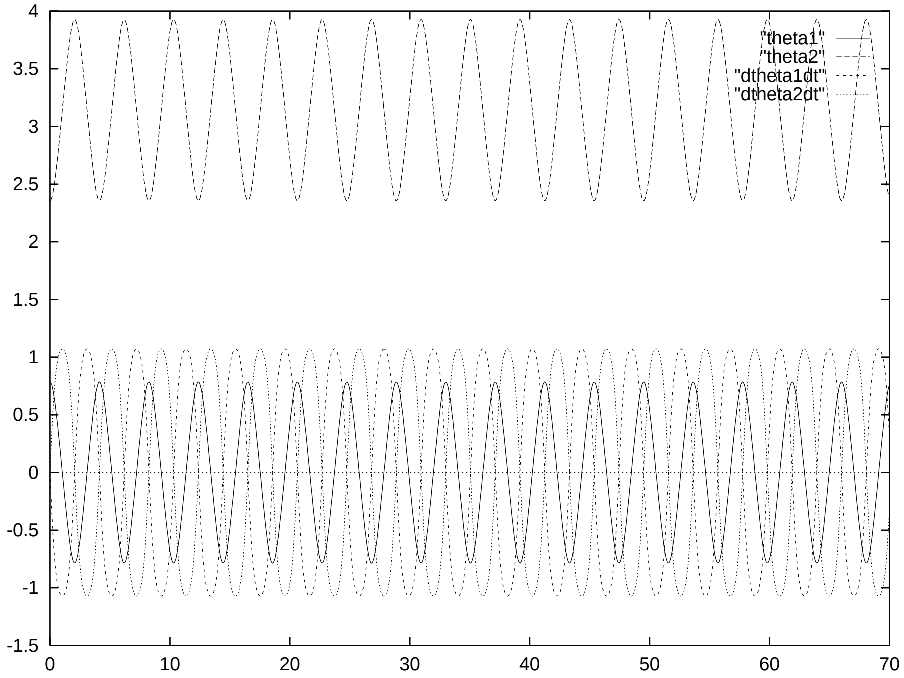

Figure 3.2: Solution for slider-crank problem with singularities

We also list the drift improvement as a function of the SRM iteration in Table 3.5.

iteration number	position drift at t=30	velocity drift at t = 30
1	.669e-8	.671e-4
2	.730e-11	.731e-7

Table 3.5: Drifts of the SRM for the slider-crank problem

If we use the SRM formulations considered in §§3.3 and 3.4 for problems withour singularities, or one of the usual stabilization methods with strict tolerances, the results become wildly different from the correct solution after several periods.

Next we calculate the acceleration of the slider end in the horizontal direction under the initial data $\phi_1(0) = \frac{\pi}{4}$ *and* $\phi_1'(0) = 2\sqrt{2}$. *The same problem was discussed in [19]. The result shown in [19] is not perfect since the maximum and minimum values in each period appear to differ. Our result looks better (see Figure 3.3).* □

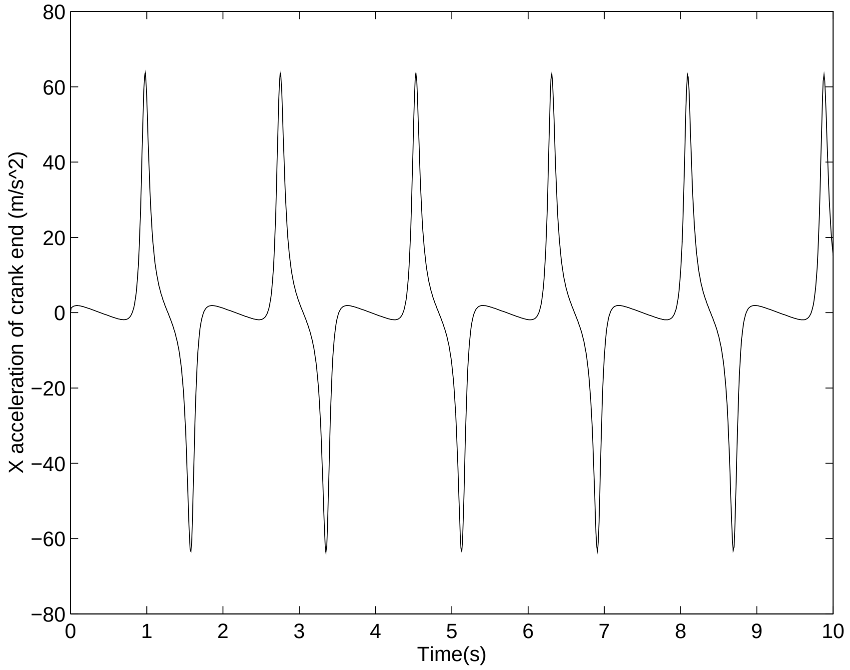

Figure 3.3: Acceleration of slider end

Chapter 4

SRM for the Nonstationary Incompressible Navier-Stokes Equations

4.1 DAE Methods for Navier-Stokes Equations

While a significant body of knowledge about the theory and numerical methods for DAEs has been accumulated, not much has been extended to partial differential-algebraic equations (PDAEs). The incompressible Navier-Stokes equations form, in fact, an example of a PDAE: to recall, these equations read

$$\mathbf{u}_t + (\mathbf{u}\cdot\mathbf{grad})\mathbf{u} = \mu\Delta\mathbf{u} - \mathbf{grad}p + \mathbf{f}, \tag{4.1a}$$

$$div\mathbf{u} = 0, \tag{4.1b}$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{b} \quad , \quad \mathbf{u}|_{t=0} = \mathbf{a}, \tag{4.1c}$$

in a bounded two- or three-dimensional domain Ω and the time interval $0 \le t \le T$. Here $\mathbf{u}(\mathbf{x},t)$ represents the velocity of a viscous incompressible fluid, $p(\mathbf{x},t)$ the pressure, $\mathbf{f}$ the prescribed external force, $\mathbf{a}(\mathbf{x})$ the prescribed initial velocity, and $\mathbf{b}(t)$ the prescribed velocity boundary values. The system (4.1) can be seen as a partial differential equation with constraint (4.1b) with respect to the time variable t. Comparing with the DAE form (3.1) p corresponds to y, the **grad** operator corresponds to the matrix B and the div operator corresponds to the Jacobian matrix G. It is easily verified that (4.1) has index-2 since the operator $div\mathbf{grad} = \Delta$ (corresponding to the matrix GB) is invertible (under appropriate boundary conditions). Indeed, the pressure-Poisson reformulation of (4.1) (see, e.g., [55]) corresponds to a direct index reduction of the PDAE, i.e. a differentiation of the constraint with respect to t followed by substitution of $\mathbf{u}_t$ from the momentum equations.

In this chapter we propose and analyze a *sequential regularization method* (SRM) for solving the incompressible Navier-Stokes equations. The method is defined as follows: with $p_0(\mathbf{x},t)$ an initial guess,

for $s = 1, 2, \cdots$, solve the problem

$$\begin{aligned} \epsilon(\mathbf{u}_s)_t &- \mathbf{grad}(\alpha_1(div\mathbf{u}_s)_t + \alpha_2 div\mathbf{u}_s) + \epsilon(\mathbf{u}_s \cdot \mathbf{grad})\mathbf{u}_s \\ &= \epsilon\mu\Delta\mathbf{u}_s - \epsilon\mathbf{grad}p_{s-1} + \epsilon\mathbf{f}, \qquad (4.2a) \\ & \quad\ \mathbf{u}_s|_{\partial\Omega} = \mathbf{b}, \mathbf{u}_s|_{t=0} = \mathbf{a}, \qquad (4.2b) \\ p_s &= p_{s-1} - \frac{1}{\epsilon}(\alpha_1(div\mathbf{u}_s)_t + \alpha_2 div\mathbf{u}_s). \qquad (4.2c) \end{aligned}$$

This method is an extension of the SRM which was proposed and analyzed in previous chapters for ordinary DAEs, especially for the index-2 DAEs (3.4), (3.5). Here we can take $E = I$ even for $\alpha_1 = 0$ because (4.1) corresponds to (3.1) with $B = G^T$. It is indicated in §3.1 that if we take $\alpha_1 \neq 0$ then certain restrictions on choosing the initial iterate (cf. (3.14)) do not apply and, more importantly, the equation for x_s is essentially not stiff if the original problem is not stiff. Hence, a non-stiff time integrator can be used for any regularization parameter ϵ. For the Navier-Stokes application (4.2) we therefore choose $\alpha_1 > 0$ so that we can still take ϵ to be very small even when we use an explicit time discretization. So one SRM iteration is often good enough. However, we should not ignore the choice $\alpha_1 = 0$ because §3.1 also indicates that with this choice the computation can be particularly simple. For (4.2), when $\alpha_1 > 0$, although we use explicit time discretization, a symmetric positive definite system relevant to the discretization of the operator $I + \frac{\alpha_1}{\epsilon}\mathbf{grad}\,div$ still needs to be inverted. If we take $\alpha_1 = 0$, then we do not need to solve any system to obtain the discrete solution. In this case, (4.2) is not stiff only for relatively large ϵ. So more than one SRM iterations are required generally. In the sequel, the convergence proof and discretization stability analysis in §4.3 and §4.4 are mainly for the case of $\alpha_1 > 0$. The discussion for the case of $\alpha_1 = 0$ can essentially be carried out in a similar way. We will remark on this case in §4.3 and §4.4 and provide a numerical verification in §4.4.

The importance of the treatment of the incompressibility constraint has long been recognized in the Navier-Stokes context. A classical approach is the projection method of [36], where one has to solve a Poisson equation for the pressure p with the zero Neumann boundary conditions which is, however, non-physical. Recently, a re-interpretation of the projection method in the context of the so-called *pressure stabilization methods*, or more generally, "pseudo - compressibility methods" has been given in [97]. Some convergence estimates for the pressure can be obtained (cf. [101, 96]). In his review paper [97], Rannacher lists some well known examples of "pseudo-compressibility methods" (which are actually regularization methods):

$$\begin{aligned} div\mathbf{u} + \epsilon p_t &= 0, \text{ in } \Omega \times [0,T), \quad p|_{t=0} = p_0, && \text{(artificial compressibility)} \\ div\mathbf{u} + \epsilon p &= 0, \text{ in } \Omega \times [0,T), && \text{(penalty method)} \\ div\mathbf{u} - \epsilon \Delta p &= 0, \text{ in } \Omega \times [0,T), \quad \tfrac{\partial p}{\partial n}|_{\partial\Omega} = p_0 && \text{(pressure stabilization).} \end{aligned}$$

If we generalize Baumgarte's stabilization to this PDAE example (4.1), we get

$$\mathbf{u}_t + (\mathbf{u} \cdot \mathbf{grad})\mathbf{u} = \mu \Delta \mathbf{u} - \mathbf{grad} p + \mathbf{f}, \tag{4.3a}$$

$$(div\mathbf{u})_t + \gamma\, div\mathbf{u} = 0. \tag{4.3b}$$

Eliminating $\mathbf{u}_t$ from (4.3), we obtain an equation for p:

$$-\Delta p + \gamma\, div\mathbf{u} - div\{(\mathbf{u} \cdot \mathbf{grad})\mathbf{u} - \mu \Delta \mathbf{u} - \mathbf{f}\} = 0.$$

We then find that this stabilization can be seen as a kind of pressure stabilization with $\gamma = \epsilon^{-1}$. Although it works, since we do not have a singularity here, it still sets up a non-physical boundary condition for the Poisson equation for p. Also, in this formulation, equations for $\mathbf{u}$ and p are not uncoupled.

In the SRM formulation (4.2) we do not need to set up boundary conditions for p. So it should be more natural than various pressure-Poisson formulations. This method relates to the idea of penalty methods but, unlike them, the regularized problems are not stiff for $\alpha_1 > 0$ or less stiff for $\alpha_1 = 0$ since we can choose ϵ to be relatively large. Hence, more convenient (nonstiff) methods can be used for time integration, and nonlinear terms can be treated easily. We will indicate in §4.4 that ϵ has little to do with the stability of the discretization there, i.e. the stability restriction is satisfied for a wide range of ϵ for $\alpha_1 > 0$. We also indicate there that, in the case of small viscosity, the usual time step restrictions for the explicit schemes can be loosened.

A similar procedure following [5] (Uzawa's iterative algorithm) in the framework of optimization theory and economics has actually appeared in the Navier-Stokes context for the stationary Stokes equations (i.e. without the nonlinear term and the time-dependent term in (4.1)) with $\alpha_1 = 0$ using the augmented Lagrangian idea, see Fortin and Glowinski [50]. (Also see [54] for a related discussion.) Note that, in their procedure, ϵ^{-1} in (4.2c) is replaced by a parameter ρ. They prove that $\rho = \epsilon^{-1}$ is approximately optimal. For the nonlinear case, they combine Uzawa's algorithm with a linearization iteration. They claim convergence but find it hard to analyze the convergence rate because their analysis depends on the spectrum of an operator which is non-symmetric in the nonlinear case. For the nonstationary case (4.1), the augmented Lagrangian method cannot be applied directly. Therefore, [50] first discretizes (4.1a) with respect to the time t (an implicit scheme is used). Then the problem becomes a stationary one in each time step. Hence, Uzawa's algorithm can be applied and converges in each time step. So, for the nonstationary case, their iterative procedure is, in essence, to provide a method to solve the time-discretized problem. Thus, their iterative procedure has little to do with time-discretization, or in other words, they still do time-discretization directly for the problem (4.1). Consequently an implicit scheme is always appropriate because of the constraints

(4.1b), and a linearization is always needed to treat the nonlinear case.

These properties are not shared by our method. We will prove that the convergence results of the SRM in the previous chapter still hold for the PDAE case (4.2). Hence, the solution sequence of (4.2) converges to the solution of (4.1) with the error estimate of $O(\epsilon^s)$ after the sth iteration. Therefore, roughly speaking, the rate is about $O(\epsilon)$. We prove the convergence results using the method of asymptotic expansions which is independent of the optimization theory and is also applicable to the steady-state case. In addition, when the finite element method is used, the difficulty of constructing test functions in a divergence-free space to decouple the $\mathbf{u}, p$ system can be avoided by using the formulation of the SRM.

We indicate here that, as many others do, we include the viscosity parameter μ in the error estimates. So the estimates could deteriorate when μ is very small. This is because we have an unresolved technical difficulty, associated with our inability to obtain an appropriate upper bound for the nonlinear term and with the weaker elliptic operator $\mu\Delta\mathbf{u}$ (which is dissipative) as $\mu \to 0$. In the SRM formulation, a supplementary dissipative term $-\alpha_2 \mathbf{grad}\, div \mathbf{u}_s$ is introduced without perturbing the solution. As indicated in [50] for the stationary case, the relative advantage of such methods may therefore become more apparent for small values of the viscosity.

The chapter is organized as follows: In §4.2 we define some preliminaries and discuss regularity properties of the solution of (4.2). The convergence of the SRM for Navier-Stokes equations is proved in §4.3. Finally, in §4.4 a simple difference scheme is discussed and some numerical experiments are presented. These numerical experiments are only exploratory in nature.

To summarize, our objective in this chapter is to present a method for the nonstationary Navier-Stokes equations from the viewpoint of DAE regularization, and to provide a way to apply a DAE method to PDAEs. It appears that such a formulation is new in the Navier-Stokes context and it is worthwhile because:

- Since ϵ need not be very small, the regularized problems in the sequence (4.2) are more stable/less stiff. So more convenient difference schemes, e.g. explicit schemes in time, can be used under theoretical assurance. If we take $\alpha_1 > 0$, this is also true for small ϵ.

- The problem of additional boundary conditions, which arises in the pressure-Poisson formulation and projection methods, does not arise here. Finite element methods can be used easily and the elements do not have to conform to the incompressibility condition to separate the variables $\mathbf{u}$ and p.

4.2 Preliminaries and the Properties of the Regularized Problems

Before we begin our analysis, we first describe some notation and assumptions. As usual, we use $\mathbf{L}^p(\Omega)$, or more simply $\mathbf{L}^p$, to denote the space of functions which are pth-power integrable in Ω, and

$$\|\mathbf{u}\|_p = (\int_\Omega \sum_{i=1}^n u_i^p \, d\mathbf{x})^{\frac{1}{p}}$$

as its norm, where $\mathbf{u} = (u_1, \cdots, u_n)$. We denote the inner product in $\mathbf{L}^2$ by $(\cdot, \cdot)$ and let $\|\cdot\| \equiv \|\cdot\|_2$. $\mathbf{C}^\infty$ is the space of functions continuously differentiable any number of times in Ω, and $\mathbf{C}_0^\infty$ consists of those members of $\mathbf{C}^\infty$ with compact support in Ω. $\mathbf{H}^m$ is the completion in the norm

$$\|\mathbf{u}\|_{\mathbf{H}^m} = (\sum_{0\le|\alpha|\le m} \|D^\alpha \mathbf{u}\|^2)^{\frac{1}{2}}.$$

We will consider the boundary conditions to be homogeneous, i.e. $\mathbf{b} \equiv \mathbf{0}$ in (4.1c), to simplify the analysis. Nevertheless, through the inclusion of a general forcing term, the results may be generalized to the case of nonhomogeneous boundary values. We are interested in the case that (4.1) has a unique solution and the solution belongs to $\mathbf{H}^2$, where the arbitrary constant which the pressure p is up to is determined by

$$\int_\Omega p(\mathbf{x}, \cdot) \, d\mathbf{x} = 0. \tag{4.4}$$

Hence, some basic compatibility condition is assumed (cf. [63]):

$$\mathbf{a}|_{\partial\Omega} = \mathbf{0}, \ \ div\mathbf{a} \equiv 0. \tag{4.5}$$

Furthermore, we assume

$$\sup_{t\in[0,T]} \|\mathbf{f}\| \leq M_1, \ \ \|\mathbf{a}\|_{\mathbf{H}^2} \leq M_1, \tag{4.6}$$

where M_1 is a positive constant.

We take p_0 in (4.2) satisfying (4.4). Hence, it is easy to see that p_s satisfies (4.4) for all s.

For simplicity, we only consider the two-dimensional case. We can treat the three-dimensional case in the same way, possibly with some more assumptions. Throughout the chapter M represents a generic constant which may depend on μ as we have explained in the introduction. We will also allow M to depend on the finite time-interval T since we are not going to discuss very long time behavior in this chapter.

At first, we write down some inequalities:

- Poincaré inequality:

$$\|\mathbf{u}\| \leq \gamma\|\mathbf{grad}\ \mathbf{u}\|, \ \ \text{if } \mathbf{u}|_\Omega = 0. \tag{4.7}$$

  More generally (see [87]), for $\mathbf{u} \in \mathbf{H}^1(\Omega)$

$$\|\mathbf{u}\| \leq C_\Omega(\|\mathbf{grad}\ \mathbf{u}\| + |\int_\Omega \mathbf{u}\ d\mathbf{x}|). \tag{4.8}$$

- Young's inequality:

$$abc \leq \frac{1}{p}a^p + \frac{1}{q}b^q + \frac{1}{r}c^r \tag{4.9}$$

  if $a, b, c > 0$, $p, q, r > 1$ and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$.

- Hölder's inequality:

$$\int_\Omega |f||g||h|d\mathbf{x} \leq \|f\|_p\|g\|_q\|h\|_r \tag{4.10}$$

  if $p, q, r > 1$ and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$.

- Sobolev's inequality in two-dimensional space:

$$\|\mathbf{u}\|_4 \le \gamma_1^{\frac{1}{4}} \|\mathbf{u}\|^{\frac{1}{2}} \|\mathbf{grad}\ \mathbf{u}\|^{\frac{1}{2}}, \tag{4.11}$$

  where $\gamma_1 = 2$ if $\Omega = \mathbf{R}^2$.

Suppose that $\mathbf{w}$ stands for the difference of two solutions of the SRM (4.2). Then $\mathbf{w}$ satisfies the homogeneous problem (4.27) (see next section). Hence, using the estimate in Lemma 4.2, uniqueness of the solution of the SRM (4.2) is easy to consider. The existence can be analyzed by following the standard existence argument for Navier-Stokes equations (e.g. [108, 62]) and that of penalized Navier-Stokes equations (e.g. [28]). Hence we assume the existence of the solution of the SRM and concentrate on the proof of the convergence of the method. Before we do so, we derive the following regularity results of the solution.

Lemma 4.1 *For the solution $\{\mathbf{u}_s, p_s\}$ of (4.2) there exists a constant ϵ_0 such that when $\epsilon \le \epsilon_0$ we have the following estimates*

$$\begin{aligned} &\|\mathbf{u}_s\|^2_{\mathbf{H}^1} + \int_0^T (\frac{\alpha_1}{\epsilon^2}\|(div\mathbf{u}_s)_t\|^2_{\mathbf{H}^1} + \frac{\alpha_2}{\epsilon^2}\|div\mathbf{u}_s\|^2_{\mathbf{H}^1} + \|(\mathbf{u}_s)_t\|^2 + \|\Delta\mathbf{u}_s\|^2 + \|p_s\|^2_{H^1})\ dt \\ &\le M[\|a\|^2_{\mathbf{H}^1} + \int_0^T (\|f\|^2 + \|\mathbf{grad}p_{s-1}\|^2)\ dt]. \end{aligned} \tag{4.12}$$

Proof: For simplicity of notation we denote $\mathbf{u}_s$ as $\mathbf{v}$ here. The proof for the case $\alpha_1 = 0$ is just the same as that in [28]. So we only consider the case $\alpha_1 > 0$. Hence, without loss of generality, we take $\alpha_1 = 1$ and $\alpha_2 = \alpha$. We then write (4.2) as

$$\begin{aligned} \mathbf{v}_t - \frac{1}{\epsilon}\mathbf{grad}((div\mathbf{v})_t + \alpha\, div\mathbf{v}) + (\mathbf{v}\cdot\mathbf{grad})\mathbf{v} & \\ = \mu\Delta\mathbf{v} - \mathbf{grad}p_{s-1} + \mathbf{f}, & \qquad (4.13a) \\ \mathbf{v}|_{\partial\Omega} = \mathbf{0}, \mathbf{v}|_{t=0} = \mathbf{a}, & \qquad (4.13b) \\ p_s = p_{s-1} - \frac{1}{\epsilon}((div\mathbf{v})_t + \alpha\, div\mathbf{v}). & \qquad (4.13c) \end{aligned}$$

The proof follows the ideas in [28]. Multiplying (4.13a) by $\mathbf{v}$ and integrating with respect to the space variables on the domain Ω, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\mathbf{v}\|^2 + \frac{1}{2\epsilon}\frac{d}{dt}\|div\mathbf{v}\|^2 + \frac{\alpha}{\epsilon}\|div\mathbf{v}\|^2 + \mu\|\mathbf{grad}\,\mathbf{v}\|^2 \\
&= -((\mathbf{v}\cdot\mathbf{grad})\mathbf{v},\mathbf{v}) - (\mathbf{grad}p_{s-1},\mathbf{v}) + (\mathbf{f},\mathbf{v}) \\
&\leq \frac{\alpha}{2\epsilon}\|div\mathbf{v}\|^2 + \frac{\gamma_1\epsilon}{2\alpha}\|\mathbf{v}\|^2\|\mathbf{grad}\,\mathbf{v}\|^2 + \frac{\mu}{2\gamma}\|\mathbf{v}\|^2 + \frac{\gamma}{\mu}\|\mathbf{grad}p_{s-1}\|^2 + \frac{\gamma}{\mu}\|\mathbf{f}\|^2,
\end{aligned}$$

where we use $-((\mathbf{v}\cdot\mathbf{grad})\mathbf{v},\mathbf{v}) = \frac{1}{2}((div\mathbf{v})\mathbf{v},\mathbf{v})$. Then let $c = \min(\frac{\alpha}{\epsilon}, \frac{\mu}{2\gamma})$ and $Y = \|\mathbf{v}\|^2 + \frac{1}{\epsilon}\|div\mathbf{v}\|^2$. Using Poincaré's inequality (4.7), we obtain

$$\frac{d}{dt}Y + cY + \frac{1}{2}(\mu - \frac{\gamma_1\epsilon}{\alpha}Y)\|\mathbf{grad}\,\mathbf{v}\|^2 \leq \frac{\gamma}{\mu}(\|\mathbf{f}\|^2 + \|\mathbf{grad}p_{s-1}\|^2). \tag{4.14}$$

Note that $Y(0) = \|\mathbf{a}\|^2$. Write (4.14) as

$$\frac{d}{dt}(\mu - \frac{\gamma_1\epsilon}{\alpha}Y) - \frac{\gamma_1\epsilon}{2\alpha}\|\mathbf{grad}\,\mathbf{v}\|^2(\mu - \frac{\gamma_1\epsilon}{\alpha}Y) \geq -\frac{\gamma\gamma_1\epsilon}{\mu\alpha}(\|\mathbf{grad}p_{s-1}\|^2 + \|\mathbf{f}\|^2).$$

Applying a standard technique for solving linear differential equations and taking ϵ appropriately small ($\leq \epsilon_0$) so that

$$\mu - \frac{\gamma_1\epsilon}{\alpha}Y(0) \geq \frac{\mu}{2} \quad\text{and}\quad \frac{\gamma\gamma_1\epsilon}{\mu\alpha}\int_0^T (\|\mathbf{f}\|^2 + \|\mathbf{grad}p_{s-1}\|^2)\,dt \leq \frac{\mu}{4},$$

we get

$$\mu - \frac{\gamma_1\epsilon}{\alpha}Y(t) \geq \frac{\mu}{4} \quad \forall t\in[0,T]. \tag{4.15}$$

Then, using the same technique and (4.14), we have

$$\begin{aligned}
Y &\leq \|\mathbf{a}\|^2\exp(-ct) + M\exp(-ct)\int_0^t (\|\mathbf{f}\|^2 + \|\mathbf{grad}p_{s-1}\|^2)\exp(cz)\,dz \\
&\leq M[\|\mathbf{a}\|^2 + \int_0^t (\|\mathbf{f}\|^2 + \|\mathbf{grad}p_{s-1}\|^2)\,dz]. \qquad (4.16)
\end{aligned}$$

Thus, (4.12) holds for $\|\mathbf{u}\|^2$. Integrating (4.14) directly and using (4.15) yields

$$\int_0^t \|\mathbf{grad}\,\mathbf{u}\|^2\,dz \leq M[\|\mathbf{a}\|^2 + \int_0^t (\|\mathbf{f}\|^2 + \|\mathbf{grad}p_{s-1}\|^2)\,dz]. \tag{4.17}$$

To prove other estimates for (4.12) we define the operator

$$A\mathbf{w} = -\frac{1}{\epsilon}\mathbf{grad}((div\mathbf{w})_t + \alpha\, div\mathbf{w}) - \mu\Delta\mathbf{w} = \mathbf{g}, \tag{4.18}$$

where $\mathbf{w}$ satisfies $\mathbf{w}|_{\partial\Omega} = \mathbf{0}$ and $\mathbf{w}|_{t=0} = \mathbf{a}$. Let

$$q = -\frac{1}{\epsilon}((div\mathbf{w})_t + \alpha\, div\mathbf{w}).$$

Then we have (noting $div\mathbf{w}|_{t=0} = 0$)

$$-\mu\Delta\mathbf{w} + \mathbf{grad}q = \mathbf{g}, \tag{4.19a}$$

$$div\mathbf{w} = -\epsilon\int_0^t q\exp(-\alpha(t-z))\, dz \tag{4.19b}$$

This is a general nonhomogeneous Stokes problem. Using the results described in [28] (or cf. [108]), we get

$$\|\Delta\mathbf{w}\| + \|\mathbf{grad}q\| \le M[\|\mathbf{g}\| + \epsilon\int_0^t \|\mathbf{grad}q\|\exp(-\alpha(t-z))\, dz]. \tag{4.20}$$

Applying Gronwall inequality, it is easy to obtain

$$\|\mathbf{grad}q\| \le M(\|\mathbf{g}\| + \epsilon\int_0^t \|\mathbf{grad}q\|\, dz) \tag{4.21}$$

and

$$\int_0^t \|\mathbf{grad}q\|^2\, dz \le M\int_0^t \|\mathbf{g}\|^2\, dz = M\int_0^t \|A\mathbf{w}\|^2\, dz. \tag{4.22}$$

It thus follows that

$$\|\Delta\mathbf{w}\| \le M(\|\mathbf{g}\| + \epsilon\int_0^t \|\mathbf{g}\|\, dz) = M(\|A\mathbf{w}\| + \epsilon\int_0^t \|A\mathbf{w}\|\, dz), \tag{4.23}$$

and then

$$\int_0^t \|\Delta\mathbf{w}\|^2\, dz \le M\int_0^t \|\mathbf{g}\|^2\, dz = M\int_0^t \|A\mathbf{w}\|^2\, dz. \tag{4.24}$$

From (4.19b) and (4.22), we thus have

$$\frac{1}{\epsilon^2}\int_0^t \|\mathbf{grad}\, div\mathbf{w}\|^2\, dz = \int_0^t \|\mathbf{grad}q\|^2\, dz \le M\int_0^t \|A\mathbf{w}\|^2\, dz. \tag{4.25}$$

Then

$$\frac{1}{\epsilon^2}\int_0^t \|\mathbf{grad}(div\mathbf{w})_t\|^2\,dz \le M\int_0^t \|A\mathbf{w}\|^2\,dz \tag{4.26}$$

follows from (4.18).

Now taking the scalar product of (4.13a) with $A\mathbf{v}$, we have

$$\begin{aligned}\frac{1}{2\epsilon}\|(div\mathbf{v})_t\|^2 + \frac{\alpha}{2\epsilon}\frac{d}{dt}\|div\mathbf{v}\|^2 + \frac{\mu}{2}\frac{d}{dt}\|\mathbf{grad}\,\mathbf{v}\|^2 + \|A\mathbf{v}\|^2 \\ = -((\mathbf{v}\cdot\mathbf{grad})\mathbf{v}, A\mathbf{v}) - (\mathbf{grad}p_{s-1}, A\mathbf{v}) + (\mathbf{f}, A\mathbf{v}).\end{aligned}$$

Note that

$$\begin{aligned}&-((\mathbf{v}\cdot\mathbf{grad})\mathbf{v}, A\mathbf{v}) \le \|\mathbf{v}\|_4\|\mathbf{grad}\,\mathbf{v}\|_4\|A\mathbf{v}\| \le \gamma_1^{\frac{1}{2}}\|\mathbf{v}\|^{\frac{1}{2}}\|\mathbf{grad}\,\mathbf{v}\|\|\Delta\mathbf{v}\|^{\frac{1}{2}}\|A\mathbf{v}\| \\ &\le \delta(\|A\mathbf{v}\|^2 + \|\Delta\mathbf{v}\|^2) + \frac{\gamma_1^2}{16\delta^3}(\|\mathbf{v}\|^2\|\mathbf{grad}\,\mathbf{v}\|^2)\|\mathbf{grad}\,\mathbf{v}\|^2 \\ &\le \delta[\|A\mathbf{v}\|^2 + M^2\|A\mathbf{v}\|^2 + M^2\epsilon^2(\int_0^t \|A\mathbf{v}\|\,dz)^2] + \frac{\gamma_1^2}{16\delta^3}(\|\mathbf{v}\|^2\|\mathbf{grad}\,\mathbf{v}\|^2)\|\mathbf{grad}\,\mathbf{v}\|^2,\end{aligned}$$

where we use (4.23) for the last inequality. Recall that we have estimates for $\|\mathbf{v}\|^2$ and $\int_0^t \|\mathbf{grad}\,\mathbf{v}\|^2\,dz$. Therefore, taking $\delta(1+M^2) < \frac{1}{4}$, it is not difficult to obtain

$$\begin{aligned}&\|\mathbf{grad}\,\mathbf{v}\| + \frac{1}{\epsilon}\int_0^t \|(div\mathbf{v})_t\|^2\,dz + \int_0^t \|A\mathbf{v}\|^2\,dz \\ \le\ & M[\|\mathbf{grad}\,\mathbf{a}\|^2 + \int_0^t (\|\mathbf{f}\|^2 + \|\mathbf{grad}p_{s-1}\|^2)\,dz + \epsilon^2\int_0^t \|A\mathbf{v}\|^2\,dz]\end{aligned}$$

Taking ϵ such that $M\epsilon^2 < 1$, we then get (4.12) for $\int_0^t \|A\mathbf{v}\|^2\,dz$ and $\|\mathbf{grad}\,\mathbf{v}\|$. Noting (4.25), (4.26), from (4.2c), (4.12) also holds for $\int_0^t \|\mathbf{grad}p_s\|^2\,dz$. Applying the inequality (4.8) and noting that p_s satisfies (4.4), yields the bound for $\int_0^t \|p_s\|^2\,dz$. Hence, from (4.2c) we can obtain (4.12) for $\int_0^t \|div\mathbf{v}\|^2\,dz$ and then $\int_0^t \|(div\mathbf{v})_t\|^2\,dz$. We thus complete the proof. □

From this lemma, we see that if we choose p_0 such that $\int_0^t \|\mathbf{grad}p_0\|^2\,dz$ is bounded then, by induction, all terms in the left of (4.12) are bounded for any given s.

4.3 Convergence of the SRM

In this section, we estimate the error of the SRM (4.2) in the solution of (4.1) by using the asymptotic expansion technique as in the proof of Theorem 2.1 of Chapter 2 (see §2.8). Note that, in the Navier Stokes context, the asymptotic expansion method was used in [54] to obtain a more precise estimate for a penalty method for the stationary Stokes equations and in [108] to calculate a slightly compressible steady-state flow. We will mainly consider the case $\alpha_1 > 0$. Hence, we take $\alpha_1 = 1$ and $\alpha_2 = \alpha$ for convenience. The result for $\alpha_1 = 0$ will be described in Remark 4.3. At first we discuss a couple of linear auxiliary problems. Then we go to the proof.

4.3.1 Two linear auxiliary problems

We discuss two linear problems in this section. One is

$$\begin{aligned} \epsilon \mathbf{w}_t \quad - \quad & \mathbf{grad}(div\mathbf{w})_t - \alpha \mathbf{grad}\, div\mathbf{w} + \epsilon(\mathbf{w} \cdot \mathbf{grad})\mathbf{U} \\ & + \epsilon(\mathbf{V} \cdot \mathbf{grad})\mathbf{w} = \epsilon\mu\Delta\mathbf{w} - \epsilon\mathbf{grad} q + \epsilon\mathbf{f}, \qquad (4.27a) \\ & \mathbf{w}|_{\partial\Omega} = \mathbf{0}, \; \mathbf{w}|_{t=0} = \mathbf{0}, \qquad (4.27b) \end{aligned}$$

where $\mathbf{U}$, $\mathbf{V}$ and q are given functions. The other is

$$\mathbf{w}_t \; + \; (\mathbf{V} \cdot \mathbf{grad})\mathbf{w} + (\mathbf{w} \cdot \mathbf{grad})\mathbf{V} = \mu\Delta\mathbf{w} - \mathbf{grad} p + f, \qquad (4.28a)$$

$$(div\mathbf{w})_t + \alpha\, div\mathbf{w} = g, \qquad (4.28b)$$

$$\mathbf{w}|_{\partial\Omega} = \mathbf{0}, \mathbf{w}|_{t=0} = \mathbf{a}, \qquad (4.28c)$$

where $\mathbf{V}$, g and $\mathbf{a}$ are given functions, $\mathbf{a}$ satisfies the compatibility conditions (4.5) and g satisfies (4.4). Now we show some properties of these two problems which will be used later in the proof of the convergence of SRM.

Lemma 4.2 *For the solution of problem (4.27), if* $\mathbf{U}$ *and* $\mathbf{V}$ *satisfy*

$$\| \cdot \|^2_{\mathbf{H}^1} + \int_0^T \| \cdot \|^2_{\mathbf{H}^2}\, dt \leq M, \qquad (4.29)$$

then we have the following estimate

$$\epsilon\|\mathbf{w}\|^2 + \|\mathrm{div}\mathbf{w}\|^2 \le M\epsilon \int_0^t (\|\mathbf{f}\|^2 + \|q\|^2)\, ds, \quad (4.30a)$$

$$\epsilon\|\mathbf{grad}\,\mathbf{w}\|^2 + \int_0^t (\epsilon\|\mathbf{w}_t\|^2 + \|\mathrm{div}\mathbf{w}_t\|^2)\, ds \ \le M\epsilon \int_0^t (\|\mathbf{f}\|^2 + \|q\|^2)\, ds. \quad (4.30b)$$

Proof: Multiplying (4.27a) by $\mathbf{w}$ and then integrating on the domain Ω yields

$$\begin{aligned}
&\frac{1}{2}\epsilon\frac{d}{dt}\|\mathbf{w}\|^2 + \frac{1}{2}\frac{d}{dt}\|div\mathbf{w}\|^2 + \alpha\|div\mathbf{w}\|^2 + \epsilon\mu\|\mathbf{grad}\,\mathbf{w}\|^2 \\
= \ & -\epsilon((\mathbf{w}\cdot\mathbf{grad})\mathbf{U},\mathbf{w}) - \epsilon((\mathbf{V}\cdot\mathbf{grad})\mathbf{w},\mathbf{w}) + \epsilon(q, div\mathbf{w}) + \epsilon(\mathbf{f},\mathbf{w}) \\
\le \ & \epsilon\|\mathbf{grad}\,\mathbf{U}\|\|\mathbf{w}\|_4^2 + \frac{\epsilon}{2}\|div\mathbf{V}\|\|\mathbf{w}\|_4^2 + \epsilon(q, div\mathbf{w}) + \epsilon(\mathbf{f},\mathbf{w}) \text{ (using (4.10))} \\
\le \ & \epsilon\gamma_1^{\frac{1}{2}}(\|\mathbf{grad}\,\mathbf{U}\| + \frac{1}{2}\|div\mathbf{V}\|)\|\mathbf{w}\|\|\mathbf{grad}\,\mathbf{w}\| + \epsilon(q, div\mathbf{w}) + \epsilon(\mathbf{f},\mathbf{w}) \text{ (using (4.11))} \\
\le \ & \frac{1}{2}\epsilon\mu\|\mathbf{grad}\,\mathbf{w}\|^2 + \frac{\epsilon\gamma_1}{2\mu}(\|\mathbf{grad}\,\mathbf{U}\| + \frac{1}{2}\|div\mathbf{w}\|)^2\|\mathbf{w}\|^2 + \epsilon(q, div\mathbf{w}) + \epsilon(\mathbf{f},\mathbf{w}),
\end{aligned}$$

where we have used $-\epsilon((\mathbf{V}\cdot\mathbf{grad})\mathbf{w},\mathbf{w}) = \frac{\epsilon}{2}((div\mathbf{V})\mathbf{w},\mathbf{w})$. Therefore, we have

$$\begin{aligned}
\frac{d}{dt}(\epsilon\|\mathbf{w}\|^2 \ &+ \ \|div\mathbf{w}\|^2) - C(t)(\epsilon\|\mathbf{w}\|^2 + \|div\mathbf{w}\|^2) \\
&\le -\epsilon\mu\|\mathbf{grad}\,\mathbf{w}\|^2 - (\alpha + C(t))\|div\mathbf{w}\|^2 + 2\epsilon(q, div\mathbf{w}) + 2\epsilon(\mathbf{f},\mathbf{w}) \\
&\le -\epsilon\mu\|\mathbf{grad}\,\mathbf{w}\|^2 + \epsilon\frac{\mu}{\gamma^2}\|\mathbf{w}\|^2 + \epsilon\frac{\gamma^2}{\mu}\|\mathbf{f}\|^2 + \epsilon\frac{1}{\alpha}\|q\|^2 \\
&\le \epsilon\frac{\gamma^2}{\mu}\|\mathbf{f}\|^2 + \epsilon\frac{1}{\alpha}\|q\|^2, \qquad (4.31)
\end{aligned}$$

where

$$C(t) = \frac{\gamma_1}{\mu}(\|\mathbf{grad}\,\mathbf{U}\| + \frac{1}{2}\|div\mathbf{V}\|)^2.$$

Noting that $\mathbf{w}|_{t=0} = 0$ and $div\mathbf{w}|_{t=0} = 0$, we thus get (4.30a).

Now, multiplying (4.27) by $\mathbf{w}_t$, then integrating with respect to $\mathbf{x}$ over Ω, we get

$$\begin{aligned}
&\epsilon\|\mathbf{w}_t\|^2 + \|div\mathbf{w}_t\|^2 + \frac{\alpha}{2}\frac{d}{dt}\|div\mathbf{w}\|^2 + \epsilon\mu\frac{d}{dt}\|\mathbf{grad}\,\mathbf{w}\|^2 = \\
&\quad \epsilon((\mathbf{w}\cdot\mathbf{grad})\mathbf{U},\mathbf{w}_t) + \epsilon((\mathbf{V}\cdot\mathbf{grad})\mathbf{w},\mathbf{w}_t) + \epsilon(q, div\mathbf{w}_t) + \epsilon(\mathbf{f},\mathbf{w}_t). \quad (4.32)
\end{aligned}$$

We use the inequalities listed in the previous section to estimate the right-hand side of (4.32) and have the following:

$$\begin{aligned}
\epsilon((\mathbf{w}\cdot\mathbf{grad})\mathbf{U},\mathbf{w}_t) &\le \frac{\epsilon}{4}\|\mathbf{w}_t\|^2 + M\epsilon(\|\mathbf{grad}\,\mathbf{w}\|^2+\|\mathbf{w}\|^2),\\
\epsilon((\mathbf{V}\cdot\mathbf{grad})\mathbf{w},\mathbf{w}_t) &\le \frac{\epsilon}{4}\|\mathbf{w}_t\|^2 + M\epsilon\sup_{\Omega}|\mathbf{V}|^2\|\mathbf{grad}\,\mathbf{w}\|^2,\\
\epsilon(\mathbf{f},\mathbf{w}_t) &\le \frac{\epsilon}{4}\|\mathbf{w}_t\|^2 + M\epsilon\|\mathbf{f}\|^2,\\
\epsilon(q,div\mathbf{w}_t) &\le \frac{\epsilon}{2}\|div\mathbf{w}_t\|^2 + M\epsilon\|q\|^2,
\end{aligned}$$

where the bounds of $\sup_\Omega|\mathbf{U}|$ and $\sup_\Omega|\mathbf{V}|$ can be obtained by using the inequality

$$\sup_{\Omega}|\cdot| \le M\|\Delta\cdot\|$$

(see, e.g. [119]). Then, similarly to the procedure for obtaining (4.30a), we obtain (4.30b). □

Next we consider problem (4.28).

Lemma 4.3 *There exists a solution for problem (4.28). Moreover, for the solution of (4.28), we have the following estimate:*

$$\|\mathbf{w}\|_{\mathbf{H}^1} + \int_0^T (\|\mathbf{w}\|^2_{\mathbf{H}^2} + \|\mathbf{w}_t\|^2 + \|p\|^2_{H^1})\,dt \le M \tag{4.33}$$

if $\int_0^T \|\mathbf{f}\|^2\,dt$ *and* $\int_0^T \|g\|^2_{H^1}\,dt$ *are bounded.*

Proof: First, we can solve $div\mathbf{w}$ from (4.28b) (noting that $div\mathbf{w}|_{t=0}=0$):

$$div\mathbf{w} = g_1, \tag{4.34}$$

where

$$g_1 = \exp(-\alpha t)\int_0^t g\exp(\alpha s)ds \tag{4.35}$$

satisfies (4.4) since g does. By applying *Corollary 2.4* in [54, p.23], the problem

$$\begin{aligned}
div\mathbf{w} &= g_1, &(4.36a)\\
\mathbf{w}|_{\partial\Omega} &= \mathbf{0} &(4.36b)
\end{aligned}$$

has many solutions. We pick one and denote it as $\mathbf{w}_p$. Then $\bar{\mathbf{w}} := \mathbf{w} - \mathbf{w}_p$ satisfies the linearized Navier-Stokes equations in the form of (4.28a) with a proper force term (denoted by $\bar{\mathbf{f}}$) and

$$div\bar{\mathbf{w}} = 0, \bar{\mathbf{w}}|_{\partial\Omega} = \mathbf{0} \text{ and } \bar{\mathbf{w}}|_{t=0} = \mathbf{a} - \mathbf{w}_p|_{t=0}.$$

By noting that $div\mathbf{w}_p|_{t=0} = g_1|_{t=0} = 0$, the basic compatibility conditions like (4.5) for $\bar{\mathbf{w}}$ are satisfied. From (4.35) and the assumption for g, $\int_0^T (\|g_1\|_{H^1}^2 + \|(g_1)_t\|^2)\, dt$ is bounded. Hence, based on the estimates for the solution of (4.36) (see [1] and [54]), it is not difficult to get

$$\|\mathbf{w}_p\|_{\mathbf{H}^1} + \int_0^T (\|(\mathbf{w}_p)_t\|^2 + \|\mathbf{w}_p\|_{\mathbf{H}^2}^2)\, dt \leq M. \tag{4.37}$$

Thus $\int_0^T \|\bar{\mathbf{f}}\|^2\, dt$ is bounded. Simulating the regularity argument of [62] or [63] (multiplying the linearized Navier-Stokes equations by $\bar{\mathbf{w}}$, $\bar{\mathbf{w}}_t$ and $\mathbf{P}\Delta\bar{\mathbf{w}}$ where $\mathbf{P}$ is a projection operator (cf. [62]), respectively), we can obtain

$$\|\bar{\mathbf{w}}\|_{\mathbf{H}^1} + \int_0^T (\|\bar{\mathbf{w}}\|_{\mathbf{H}^2}^2 + \|\bar{\mathbf{w}}_t\|^2 + \|p\|_{H^1})\, dt \leq M. \tag{4.38}$$

Therefore, (4.33) follows from (4.37) and (4.38). Using the estimate (4.38) and following a global existence argument (e.g. [62] or [108]), the existence of the solution for $\bar{\mathbf{w}}$ can be obtained. We thus have the results of the lemma . □

Remark 4.1 *The uniqueness of the solution of (4.28) follows from the standard argument for Navier-Stokes equations (cf. [108]).* □

4.3.2 The error estimate of SRM

In this section we prove the convergence of iteration (4.2) based on the same procedure described in the proof of Theorem 2.1 of Chapter 2 (see §2.8) . We describe our results in the following theorem.

Theorem 4.1 *Let* $\mathbf{u}$ *and* p *be the solution of problem (4.1), and* $\mathbf{u}_s$ *and* p_s *be the solution of problem (4.2) at the sth iteration. Then there exists a constant* ϵ_0 *such that when* $\epsilon \leq \epsilon_0$ *we have the following error estimates for all* $t \in [0, T]$*:*

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_s\|_{\mathbf{H}^1} &\leq M\epsilon^s, & (4.39a)\\ (\int_0^T \|p - p_s\|^2 \, dt)^{\frac{1}{2}} &\leq M\epsilon^s, & (4.39b) \end{aligned}$$

where T *is any given finite number and* $s = 1, 2, \cdots$.

Proof: At first, consider the case $s = 1$ of (4.2). Let

$$\mathbf{u}_1 = \mathbf{u}_{10} + \epsilon\mathbf{u}_{11} + \cdots + \epsilon^m \mathbf{u}_{1m} + \cdots$$

Comparing the coefficients of like powers of ϵ, we thus have

$$\begin{aligned} \mathbf{grad}((div\mathbf{u}_{10})_t + \alpha div\mathbf{u}_{10}) &= 0, & (4.40a)\\ \mathbf{grad}((div\mathbf{u}_{11})_t + \alpha div\mathbf{u}_{11}) &= (\mathbf{u}_{10})_t + (\mathbf{u}_{10} \cdot \mathbf{grad})\mathbf{u}_{10} &\\ -\mu\Delta\mathbf{u}_{10} &+ \mathbf{grad}p_0 - \mathbf{f}, & (4.40b)\\ \mathbf{grad}((div\mathbf{u}_{1i})_t + \alpha div\mathbf{u}_{1i}) &= (\mathbf{u}_{1i-1})_t + \sum_{j=1}^{i-1}(\mathbf{u}_{1j} \cdot \mathbf{grad})\mathbf{u}_{1i-1-j} &\\ -\mu\Delta\mathbf{u}_{1i-1} &, \quad 2 \leq i \leq m+1, & (4.40c) \end{aligned}$$

where (4.40a) satisfies (4.2b) and (4.40b) and (4.40c) satisfy the homogeneous initial and boundary conditions corresponding to (4.2b). Now (4.40a) has infinitely many solutions in general. We should choose $\mathbf{u}_{10}$ not only to satisfy (4.40a) but also to ensure that the solution of (4.40b) exists. A choice of $\mathbf{u}_{10}$ is the exact solution $\mathbf{u}$ of (4.1), i.e.

$$(\mathbf{u}_{10})_t + (\mathbf{u}_{10} \cdot \mathbf{grad})\mathbf{u}_{10} = \mu\Delta\mathbf{u}_{10} - \mathbf{grad}p + \mathbf{f}, \quad (4.41a)$$

$$(div\mathbf{u}_{10})_t + \alpha div\mathbf{u}_{10} = 0, \quad (4.41b)$$

$$\mathbf{u}_{10}|_{\partial\Omega} = 0, \ \mathbf{u}_{10}|_{t=0} = \mathbf{a}. \quad (4.41c)$$

Note that $div\mathbf{u}_{10}|_{t=0} = div\mathbf{a} = 0$ and p is taken to satisfy (4.4). So $\mathbf{u}_{10} \equiv \mathbf{u}$ and (4.40b) has the form

$$\mathbf{grad}((div\mathbf{u}_{11})_t + \alpha div\mathbf{u}_{11}) = \mathbf{grad}(p_0 - p). \tag{4.42}$$

Now we choose $\mathbf{u}_{11}$ and a corresponding p_{11} to satisfy

$$(\mathbf{u}_{11})_t + (\mathbf{u}_{10} \cdot \mathbf{grad})\mathbf{u}_{11} + (\mathbf{u}_{11} \cdot \mathbf{grad})\mathbf{u}_{10} = \mu\Delta\mathbf{u}_{11} - \mathbf{grad}p_{11}, \tag{4.43a}$$

$$(div\mathbf{u}_{11})_t + \alpha div\mathbf{u}_{11} = p_0 - p, \tag{4.43b}$$

$$\mathbf{u}_{11}|_{\partial\Omega} = 0,\ \mathbf{u}_{11}|_{t=0} = 0. \tag{4.43c}$$

Again we have $div\mathbf{u}_{11}|_{t=0} = 0$ and let p_{11} satisfy (4.4). According to Lemma 4.3, $\mathbf{u}_{11}$ and p_{11} exist .

Generally, supposing we have $\mathbf{u}_{1i-1}$, p_{1i-1} for $i \geq 2$, choose $\mathbf{u}_{1i}$, p_{1i} satisfying

$$\begin{aligned}(\mathbf{u}_{1i})_t &+ (\mathbf{u}_{10} \cdot \mathbf{grad})\mathbf{u}_{1i} + (\mathbf{u}_{1i} \cdot \mathbf{grad})\mathbf{u}_{10} \\ = \mu\Delta\mathbf{u}_{1i} &- \mathbf{grad}p_{1i} - \sum_{j=1}^{i-1}(\mathbf{u}_{1j} \cdot \mathbf{grad})\mathbf{u}_{1i-1-j},\end{aligned} \tag{4.44a}$$

$$(div\mathbf{u}_{1i})_t + \alpha div\mathbf{u}_{1i} = -p_{1i-1}, \tag{4.44b}$$

$$\mathbf{u}_{1i}|_{\partial\Omega} = 0,\ \mathbf{u}_{1i}|_{t=0} = 0, \tag{4.44c}$$

where we note that $div\mathbf{u}_{1i}|_{t=0} = 0$ and p_{1i} satisfies (4.4). Applying Lemma 4.3, all $\mathbf{u}_{1i}$ and p_{1i}, $i = 0, 1, \cdots$, exist and satisfy (4.33).

Next we estimate the remainder of the asymptotic expansion after the $(m+1)$th power of ϵ. Denote

$$\bar{\mathbf{u}}_{1m} = \mathbf{u}_{10} + \epsilon\mathbf{u}_{11} + \cdots + \epsilon^{m+1}\mathbf{u}_{1m+1} \tag{4.45}$$

($\bar{\mathbf{u}}_{1m}$ also satisfies (4.33)) and

$$\mathbf{w}_{1m} = \mathbf{u}_1 - \bar{\mathbf{u}}_{1m}. \tag{4.46}$$

Then $\mathbf{w}_{1m}$ satisfies

$$\epsilon(\mathbf{w}_{1m})_t - \mathbf{grad}(div\mathbf{w}_{1m})_t - \alpha\mathbf{grad}\,div\mathbf{w}_{1m}$$

$$+\epsilon(\mathbf{w}_{1m}\cdot\mathbf{grad})\mathbf{u}_1+\epsilon(\bar{\mathbf{u}}_{1m}\cdot\mathbf{grad})\mathbf{w}_{1m}=\epsilon\mu\Delta\mathbf{w}_{1m}-$$
$$\epsilon^{m+2}\{(\mathbf{u}_{1m+1})_t+\sum_{i=0}^{m+1}[(\mathbf{u}_{1i}\cdot\mathbf{grad})\mathbf{u}_{1m+1-i}]-\mu\Delta\mathbf{u}_{1m+1}\}, \qquad (4.47a)$$
$$\mathbf{w}_{1m}|_{\partial\Omega}=0,\ \mathbf{w}_{1m}|_{t=0}=0. \qquad (4.47b)$$

Using regularity we have for $\mathbf{u}_{1i}$, $\bar{\mathbf{u}}_{1m}$ and $\mathbf{u}_1$ (see (4.12)) and Lemma 4.2 , we obtain $\|\mathbf{w}_{1m}\|=O(\epsilon^{m+1})$ and $\|\mathbf{grad}\,\mathbf{w}_{1m}\|=O(\epsilon^{m+1})$. Therefore

$$\mathbf{u}_1=\mathbf{u}_{10}+\epsilon\mathbf{u}_{11}+\cdots+\epsilon^m\mathbf{u}_{1m}+O(\epsilon^{m+1}). \qquad (4.48)$$

in the $\mathbf{H}^1$-norm for the spatial variables. Noting $\mathbf{u}_{10}\equiv\mathbf{u}$, we thus obtain

$$\mathbf{u}_1-\mathbf{u}=O(\epsilon). \qquad (4.49)$$

Furthermore, according to Lemma 4.2, we have

$$\|div\mathbf{w}_{1m}\|=O(\epsilon^{m+\frac{3}{2}}),\ (\int_0^T\ \|(div\mathbf{w}_{1m})_t\|^2\ dt)^{\frac{1}{2}}=O(\epsilon^{m+\frac{3}{2}}).$$

Then, by using (4.2c),(4.48),(4.41b), (4.43b), (4.44b) and the estimates for $div\mathbf{w}_{1m}$ and $(div\mathbf{w}_{1m})_t$, it follows that

$$p_1=p+\epsilon p_{11}+\cdots+\epsilon^m p_{1m}+O(\epsilon^{m+\frac{1}{2}}) \qquad (4.50)$$

or

$$p_1-p=O(\epsilon). \qquad (4.51)$$

in the sense of the L^2-norm for both spatial and time variables, i.e. $(\int_0^T\ \|\cdot\|^2\ dt)^{\frac{1}{2}}$.

Now we look at the second iteration $s=2$ of (4.2). Let

$$\mathbf{u}_2=\mathbf{u}_{20}+\epsilon\mathbf{u}_{21}+\cdots+\epsilon^m\mathbf{u}_{2m}+\cdots.$$

Noting that (4.50) gives us a series expansion for p_1 we obtain

$$\mathbf{grad}((div\mathbf{u}_{20})_t+\alpha div\mathbf{u}_{20})\ =\ 0, \qquad (4.52a)$$

$$\begin{aligned}
\mathbf{grad}((div\mathbf{u}_{21})_t + \alpha div\mathbf{u}_{21}) &= (\mathbf{u}_{20})_t + (\mathbf{u}_{20}\cdot\mathbf{grad})\mathbf{u}_{20} \\
-\mu\Delta\mathbf{u}_{20} &+ \mathbf{grad}p - \mathbf{f}, \qquad (4.52b) \\
\mathbf{grad}((div\mathbf{u}_{2i})_t + \alpha div\mathbf{u}_{2i}) &= (\mathbf{u}_{2i-1})_t + \sum_{j=1}^{i-1}(\mathbf{u}_{2j}\cdot\mathbf{grad})\mathbf{u}_{2i-1} \\
-\mu\Delta\mathbf{u}_{2i-1} &- \mathbf{grad}p_{1i-1}, 2\le i\le m+1. \qquad (4.52c)
\end{aligned}$$

Again, (4.52a) is combined with the initial and boundary conditions (4.2b), and (4.52b) and (4.52c) are combined with the corresponding homogeneous ones. As in the case of $s=1$, we again choose $\mathbf{u}_{20}=\mathbf{u}$. We thus have

$$\mathbf{grad}((div\mathbf{u}_{21})_t + \alpha div\mathbf{u}_{21}) = 0. \qquad (4.53)$$

Then $\mathbf{u}_{21}$ is constructed to satisfy

$$(\mathbf{u}_{21})_t + (\mathbf{u}_{20}\cdot\mathbf{grad})\mathbf{u}_{21} + (\mathbf{u}_{21}\cdot\mathbf{grad})\mathbf{u}_{20} = \mu\Delta\mathbf{u}_{21} - \mathbf{grad}p_{21}, \qquad (4.54a)$$

$$(div\mathbf{u}_{21})_t + \alpha div\mathbf{u}_{21} = 0, \qquad (4.54b)$$

$$\mathbf{u}_{21}|_{\partial\Omega} = 0,\ \mathbf{u}_{21}|_{t=0} = 0. \qquad (4.54c)$$

Obviously $\mathbf{u}_{21}=0$ and $p_{21}=0$ is the solution of (4.54) and (4.4).

In general, similarly to the case of $s=1$, we choose $\mathbf{u}_{2i}, p_{2i}$ to satisfy

$$(\mathbf{u}_{2i})_t + (\mathbf{u}_{20}\cdot\mathbf{grad})\mathbf{u}_{2i} + (\mathbf{u}_{2i}\cdot\mathbf{grad})\mathbf{u}_{20} = \qquad (4.55a)$$

$$\mu\Delta\mathbf{u}_{2i} - \mathbf{grad}p_{2i} - \sum_{j=1}^{i-1}(\mathbf{u}_{2j}\cdot\mathbf{grad})\mathbf{u}_{2i-1-j}, \qquad (4.55b)$$

$$(div\mathbf{u}_{2i})_t + \alpha div\mathbf{u}_{2i} = p_{1i-1} - p_{2i-1}, \qquad (4.55c)$$

$$\mathbf{u}_{2i}|_{\partial\Omega} = 0,\ \mathbf{u}_{2i}|_{t=0} = 0 \qquad (4.55d)$$

for $2\le i\le m+1$, where p_{2i} satisfies (4.4). By the same procedure as for $s=1$ we obtain error equations similar to (4.47) with the addition of a remainder term $\mathbf{grad}(p_1-\bar{p}_{1m})$ in the right-hand side, where $\bar{p}_{1m}$ stands for the asymptotic expansion (4.50) of p_1. Applying Lemma 4.2 again, we get

$$\mathbf{u}_2 = \mathbf{u}_{20} + \epsilon\mathbf{u}_{21} + \cdots + \epsilon^m\mathbf{u}_{2m} + O(\epsilon^{m+1}). \qquad (4.56)$$

Noting $\mathbf{u}_{21} \equiv 0$,

$$\mathbf{u}_2 - \mathbf{u} = O(\epsilon^2). \tag{4.57}$$

Then, using (4.2b),(4.56),(4.54b) and (4.55c), we conclude

$$p_2 = p + \epsilon p_{21} + \cdots + \epsilon^m p_{2m} + O(\epsilon^{m+\frac{1}{2}}) \tag{4.58}$$

or

$$p_2 - p = O(\epsilon^2). \tag{4.59}$$

since $p_{21} \equiv 0$.

We can repeat this procedure, and, by induction for s (choosing m larger than s), conclude the results of the theorem. $\square$

Remark 4.2 *Corresponding to Theorem 2.1, we expect that the error estimates (4.39) also hold for the SRM (4.2) with $\alpha_1 = 0$, at least, away from $t = 0$.* $\square$

Remark 4.3 *In Theorem 4.1, we find that the result for p is in a weaker norm $\int_0^T \|\cdot\|^2\, dt$. This is because we have difficulty in estimating the first order time-derivative of the right-hand side of (4.47), or concretely, the term $\int_0^T \|(\mathbf{u}_{1m+1})_{tt}\|^2\, ds$. In [63]* (Corollary 2.1) *it is shown that $\int_0^T \|(\mathbf{u}_{1m+1})_{tt}\|^2\, ds$ may be unbounded as $t \to 0$ if we only assume the local compatibility conditions (4.5). In the case that this integral is bounded for $0 < t < T$, we can get*

$$\|p - p_s\| + \left(\int_0^t \|(p - p_s)_t\|^2\, ds\right)^{\frac{1}{2}} \le M\epsilon^s. \tag{4.60}$$

Otherwise, we only can expect that (4.60) holds away from $t = 0$ by following the argument in [63]. $\square$

Remark 4.4 *Multiplying (4.27) by $A\mathbf{w}$, where A is the operator defined by (4.18), and following the later steps of the proof of Lemma 4.1, we can get*

$$\frac{1}{\epsilon^2}\int_0^T (\|\mathbf{grad}(div\mathbf{w})_t\|^2 + \alpha\|\mathbf{grad}\, div\mathbf{w}\|^2)\, dt$$
$$\leq M\int_0^T (\|\mathbf{f}\|^2 + \|\mathbf{grad}\; q\|^2)\, dt.$$

Using this result to estimate the remainders of the asymptotic solutions in the proof of Theorem 4.1, we obtain

$$(\int_0^T \|p - p_s\|_{H^1}^2\, dt)^{\frac{1}{2}} \leq M\epsilon^s. \tag{4.61}$$

□

4.4 Discretization Issues and Numerical Experiments

In previous sections, we have proposed the SRM and performed some basic analysis on it. The SRM yields a sequence of PDEs which are to be solved numerically. The problem at the sth iteration can be written as:

$$\begin{aligned} \epsilon(\mathbf{u}_s)_t &- \mathbf{grad}(\alpha_1(div\mathbf{u}_s)_t + \alpha_2\, div\mathbf{u}_s) + \epsilon(\mathbf{u}_s\cdot\mathbf{grad})\mathbf{u}_s \\ &= \epsilon\mu\Delta\mathbf{u}_s + \epsilon\mathbf{r}_s, \qquad (4.62a) \\ &\quad \mathbf{u}_s|_{\partial\Omega} = \mathbf{0}, \mathbf{u}_s|_{t=0} = \mathbf{a}, \qquad (4.62b) \end{aligned}$$

where $\mathbf{r}_s(t)$ is the known inhomogeneity

$$\mathbf{r}_s = -\mathbf{grad}p_{s-1} + \mathbf{f}. \tag{4.63}$$

A variational formulation of (4.62) gives:

Find $\mathbf{u}_s \in \mathbf{H}_0^1$ such that

$$\begin{aligned} \epsilon\frac{d}{dt}(\mathbf{u}_s,\phi) + \alpha_1\frac{d}{dt}(div\mathbf{u}_s, div\phi) &+ \alpha_2(div\mathbf{u}_s, div\phi) \\ +\epsilon\mu(\mathbf{grad}\,\mathbf{u}_s, \mathbf{grad}\,\phi) &+ b(\mathbf{u}_s,\mathbf{u}_s,\phi) = \epsilon(\mathbf{r}_s,\phi),\ \forall\phi\in\mathbf{H}_0^1, \quad (4.64a) \\ \mathbf{u}_s|_{t=0} = \mathbf{a} \quad &, \quad div\mathbf{u}_s|_{t=0} = 0, \qquad (4.64b) \end{aligned}$$

where the trilinear form

$$b(\mathbf{u}, \mathbf{v}, \mathbf{w}) = ((\mathbf{u} \cdot \mathbf{grad})\mathbf{v}, \mathbf{w}).$$

From (4.64) we see that the finite element method in the spatial variables, combined with time discretizations, can be easily adopted. Note that we do not need to construct divergence-free test functions to separate the variables $\mathbf{u}$ and p. Nevertheless, in this section, we are not going to discuss finite element methods further. Some discussions on using the SRM with the finite element method will be given in the next chapter for a problem in reservoir simulations. Some numerical experiments are also given there. Here we only consider a very simple first-order difference scheme (forward Euler scheme in the time direction) in two dimensional space, as an initial attempt towards the discretization of the sequential regularization method for the PDAE. Concretely, we consider a rectangular domain such that an equidistant mesh can be used. Let $(u, v)^T$ stand for the approximation of $\mathbf{u}_s$, and let k, h_x, h_y denote step sizes in time and spatial direction, respectively. Without loss of generality, we assume that $h_x = h_y = h$ and that the domain is a unit square. Thus, mesh points can be expressed as

$$x_i = ih, i = 0, 1, \cdots, I;\ y_j = jh, j = 0, 1, , \cdots, J;\ t_n = nk, n = 0, 1, \cdots, N,\ \ N = [T/k].$$

The difference scheme reads:

$$\begin{aligned}
\epsilon u_{\dot t} - \alpha_1(u_{\bar x \dot x} + v_{\bar y \dot x})_{\dot t} \quad &= \quad \alpha_2(u_{\bar x \dot x} + u_{\bar y \dot x}) \\
-\epsilon(u u_{\bar x} + v u_{\bar y}) \quad &+ \quad \epsilon\mu(u_{\bar x \dot x} + u_{\bar y \dot y}) + \epsilon r_u, && (4.65\text{a}) \\
\epsilon v_{\dot t} - \alpha_1(u_{\bar x \dot y} + u_{\bar y \dot y})_{\dot t} \quad &= \quad \alpha_2(u_{\bar x \dot y} + u_{\bar y \dot y}) \\
-\epsilon(u v_{\bar x} + v v_{\bar y}) \quad &+ \quad \epsilon\mu(v_{\bar x \dot x} + v_{\bar y \dot y}) + \epsilon r_v, && (4.65\text{b}) \\
u|_{\partial\Omega} = 0,\ v|_{\partial\Omega} = 0, \qquad & \qquad u|_{t=0} = a_u,\ v|_{t=0} = a_v, && (4.65\text{c})
\end{aligned}$$

where

$$
\begin{aligned}
u &= u_{i,j}^n, \\
u_{\hat{t}} &= \frac{u_{i,j}^{n+1} - u_{i,j}^n}{k}, \\
u_{\dot{x}} &= \frac{u_{i+1,j}^n - u_{i,j}^n}{h}, \\
u_{\bar{x}} &= \frac{u_{i,j}^n - u_{i-1,j}^n}{h}.
\end{aligned}
$$

$u_{\dot{y}}$ and $u_{\bar{y}}$ can be defined accordingly and the definitions for v are similar.

Obviously, this is a first-order scheme explicit in time, where the nonlinear term is discretized somewhat arbitrarily. The scheme is easy to implement. Next we discuss its stability. For simplicity, we analyze the linear case (corresponding to the Stokes equations) first, and consider the full nonlinear equations (4.65) in Remark 4.7 below.

We write the linear case of (4.65) as follows :

$$
\begin{aligned}
\epsilon u_{\hat{t}} - \alpha_1(u_{\bar{x}\dot{x}} + v_{\bar{y}\dot{x}})_{\hat{t}} &= \alpha_2(u_{\bar{x}\dot{x}} + u_{\bar{y}\dot{x}}) + \epsilon\mu(u_{\bar{x}\dot{x}} + u_{\bar{y}\dot{y}}) + \epsilon r_u, && (4.66\text{a}) \\
\epsilon v_{\hat{t}} - \alpha_1(u_{\bar{x}\dot{y}} + u_{\bar{y}\dot{y}})_{\hat{t}} &= \alpha_2(u_{\bar{x}\dot{y}} + u_{\bar{y}\dot{y}}) + \epsilon\mu(v_{\bar{x}\dot{x}} + v_{\bar{y}\dot{y}}) + \epsilon r_v, && (4.66\text{b}) \\
u|_{\partial\Omega} = 0,\ v|_{\partial\Omega} = 0, &\qquad u|_{t=0} = a_u,\ v|_{t=0} = a_v. && (4.66\text{c})
\end{aligned}
$$

Here we take $\alpha_1 = 1$ and $\alpha_2 = \alpha$. The result for the case of $\alpha_1 = 0$ will be given in Remark 4.6.

The following theorem gives the stability estimate for (4.66) in the sense of the discrete $\mathbf{L}^2$-norm:

$$
\|w^h\|_h^2 = h^2 \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} (w_{i,j})^2, \tag{4.67}
$$

where $w^h = (w_{i,j}),\ i = 0, 1, \cdots, I-1,\ j = 0, 1, \cdots, J-1.$

Theorem 4.2 *Let u and v be the solution of (4.66) and*

$$
A = \epsilon(\|u\|_h^2 + \|v\|_h^2) + \|u_{\bar{x}} + v_{\bar{y}}\|_h^2 + \epsilon\mu(\|u_{\bar{x}}\|_h^2 + \|u_{\bar{y}}\|_h^2 + \|v_{\bar{x}}\|_h^2 + \|v_{\bar{y}}\|_h^2). \tag{4.68}
$$

If $\frac{\mu k}{h^2} \leq 1 - c$, *where* c *is any constant in* $(0,1)$, *then*

$$A + k \sum_{n=0}^{N-1} \|(u_{\bar{x}} + v_{\bar{y}})_{\bar{t}}\|_h^2 \leq M\epsilon \max_{0 \leq t_n \leq T} (\|r_u\|_h^2 + \|r_v\|_h^2) \tag{4.69}$$

where M *is a generic constant dependent on* μ *and* c.

Proof: We first write down the following identities and inequalities:

- some difference identities [75]:

$$(\phi\psi)_{\bar{x}} = \phi\psi_{\bar{x}} + \phi_{\bar{x}} E_x^{-1}\psi, \tag{4.70a}$$

$$(\phi\psi)_{\dot{x}} = \phi\psi_{\dot{x}} + \phi_{\dot{x}} E_x^{1}\psi, \tag{4.70b}$$

$$2\phi\phi_{\bar{t}} = (\phi^2)_{\bar{t}} - k(\phi_{\bar{t}})^2, \tag{4.70c}$$

$$\phi\phi_{\dot{x}\bar{x}} = (\phi\phi_{\dot{x}})_{\bar{x}} - (\phi_{\bar{x}})^2, \tag{4.70d}$$

  where the translation operator $E_x^i\phi(x,y,t) = \phi(x+ih,y,t)$.

- an difference inequality [75]:

$$h\|\phi_{\bar{x}}\|_h \leq 2\|\phi\|_h \tag{4.71}$$

- a discrete version of the Poincaré inequality (cf. [66]):

$$\|\phi\|_h^2 \leq \|\phi_{\bar{x}}\|_h^2 + \|\phi_{\bar{y}}\|_h^2 \tag{4.72}$$

  if ϕ satisfies homogeneous boundary conditions.

Multiplying (4.66a) by $au + bu_{\bar{t}}$ and (4.66b) by $av + bv_{\bar{t}}$ and adding, then summing for all (i,j), $i = 1, \cdots, I-1$, $j = 1, \cdots, J-1$ where we use the difference identities (4.70) (omitting lots of tedious algebraic manipulations), we obtain

$$\epsilon a(\|u\|_h^2+\|v\|_h^2)_{\bar{t}}+\frac{1}{2}(\alpha b+a)(\|u_{\bar{x}}+v_{\bar{y}}\|_h^2)_{\bar{t}}+\frac{1}{2}\epsilon\mu b(\|u_{\bar{x}}\|_h^2+\|u_{\bar{y}}\|_h^2+\|v_{\bar{x}}\|_h^2+\|v_{\bar{y}}\|_h^2)_{\bar{t}}$$
$$+\epsilon(b-ak)(\|u_{\bar{t}}\|_h^2+\|v_{\bar{t}}\|_h^2)+a\alpha\|u_{\bar{x}}+v_{\bar{y}}\|_h^2+(b-\frac{1}{2}(b\alpha+a)k)\|(u_{\bar{x}}+v_{\bar{y}})_{\bar{t}}\|_h^2$$
$$+\epsilon\mu a(\|u_{\bar{x}}\|_h^2+\|u_{\bar{y}}\|_h^2+\|v_{\bar{x}}\|_h^2+\|v_{\bar{y}}\|_h^2)-\frac{1}{2}\epsilon\mu bk(\|u_{\bar{x}\bar{t}}\|_h^2+\|u_{\bar{y}\bar{t}}\|_h^2+\|v_{\bar{x}\bar{t}}\|_h^2+\|v_{\bar{y}\bar{t}}\|_h^2)$$
$$\leq M\epsilon(\|r_u\|_h^2+\|r_v\|_h^2)+\frac{1}{2}\epsilon\mu a(\|u\|_h^2+\|v\|_h^2)+\frac{1}{2}\epsilon b\delta(\|u_{\bar{t}}\|_h^2+\|v_{\bar{t}}\|_h^2),$$

where $\delta>0$ can be chosen less than c/b. Applying (4.71) and (4.72), we get

$$h^2((\|u_{\bar{x}\bar{t}}\|_h^2+\|u_{\bar{y}\bar{t}}\|_h^2+\|v_{\bar{x}\bar{t}}\|_h^2+\|v_{\bar{y}\bar{t}}\|_h^2)\leq 4(\|u_{\bar{t}}\|_h^2+\|v_{\bar{t}}\|_h^2)$$

and

$$\|u\|_h^2+\|v\|_h^2\leq\|u_{\bar{x}}\|_h^2+\|u_{\bar{y}}\|_h^2+\|v_{\bar{x}}\|_h^2+\|v_{\bar{y}}\|_h^2,$$

respectively. Then we can choose a and b such that

$$b-ak-\mu\frac{k}{h^2}-\frac{1}{2}b\delta>0,\ b-\frac{1}{2}(b\alpha+a)k>0$$

and obtain

$$(A)_{\bar{t}}+d\epsilon\mu A+\|(u_{\bar{x}}+v_{\bar{y}})_{\bar{t}}\|_h^2\leq M\epsilon(\|r_u\|_h^2+\|r_v\|_h^2),$$

where d is a constant independent of k, h, ϵ and μ. From this inequality, it is not difficult to see that (4.69) holds. □

Remark 4.5 *From (4.69) of Theorem 4.2, we find that the value of ϵ will not affect the stability of the difference scheme. This means that the forward Euler scheme in the time direction works for any value of ϵ. Also, the time step restriction $k\leq(1-c)h^2/\mu$ is actually loosened in the case of small viscosity (or large Reynolds number) which people are often interested in. This implies that the explicit scheme (4.66) to which an appropriate discretization of the nonlinear term (see the next remark) is added works very well. It enables us not only to avoid the complicated iteration procedure for nonlinear equations but also to choose the time step fairly widely in the case of small viscosity.* □

Remark 4.6 *We have mentioned before that sometimes we may like to take* $\alpha_1 = 0$ *to avoid solving any algebraic system. Following the same procedure as the proof of Theorem 4.2, we can get the stability condition for the case of* $\alpha_1 = 0$. *That is,* $k \leq m\epsilon h^2$, *where* m *is a positive constant independent of* ϵ,h *and* μ. *We thus see that the stability of (4.66) with* $\alpha_1 = 0$ *depends on the parameter* ϵ. *This coincides with our experience with stiff problems discretized by explicit schemes. Fortunately, using the SRM, we do not need to take* ϵ *very small. So the time step restriction is not much worse than the usual one corresponding to an explicit scheme applied to a non-stiff problem.* $\square$

Remark 4.7 *For the nonlinear case (4.65), when the viscosity* μ *is not small, we expect similar results since the nonlinear term can be dominated by the viscous term. When the viscosity is small, however, the scheme (4.65) is unstable. Although numerical computations indicate that we do get better stability if we increase* α_2, *i.e. some kind of dissipation effect is obtained (we must note that such dissipation becomes small when the incompressibility condition is close to being satisfied), we suggest using spatial discretizations with better stability properties, e.g. upwinding schemes (cf. [102]), in the case of small viscosity.* $\square$

Remark 4.8 *Applying corresponding difference identities for a nonuniform mesh (see e.g. [106]), the results of Theorem 4.2 may be generalized to difference schemes (4.66) on a nonuniform mesh. Hence, the difference scheme may be used for problems defined on more general domains.* $\square$

Next we explain our theoretical results by calculating the solution of an artificial example.

Example 4.1 *Consider the Navier-Stokes equations (4.1) with the exact solution*

$\mathbf{u} = (u, v)^T$

$$
\begin{aligned}
u &= 50x^2(1-x)^2y(1-y)(1-2y)[1+\exp(-t)], \\
v &= -50y^2(1-y)^2x(1-x)(1-2x)[1+\exp(-t)], \\
p &= [-x(\frac{x}{2}+2) - y(\frac{y}{2}-2) + \frac{1}{3}][1+\exp(-t)]^2
\end{aligned}
$$

As indicated in §2.6 of Chapter 2, to carry out the SRM iterations, we do not need to store the entire approximation of p_{s-1} *on* $[0,T]$ *for calculating* $\mathbf{u}_s$. *Assuming that the number of the SRM iterations is chosen in advance, we can rearrange the computational order to make the storage requirements independent of* N, *where* N *represents the number of the mesh lines in the* t *direction. We first use constant steps* $k = 0.01$ *and* $h = 0.1$. *At a given time* t, *we use 'eu' to denote the absolute discrete* L^2*-error in* $\mathbf{u}_s$ *while 'ep' denotes the absolute discrete* L^2*-error in* p_s. *Table 4.1 summarizes the computational results of the difference scheme (4.65) with* $\alpha_1 = \alpha_2 = 1$ *and viscosity* $\mu = 0.1$.

ϵ	iteration	error at →	t = k	t = 1.0	t = 2.0	t = 3.0	t = 4.0	t = 5.0
5e-1	1	eu	4.65e-3	2.69e-1	1.55e-1	1.31e-1	1.15e-1	1.08e-1
		ep	2.49e-1	1.96e-1	1.57e-1	1.36e-1	1.25e-1	1.21e-1
	2	eu	2.16e-3	2.53e-2	3.12e-2	3.18e-2	3.12e-2	3.06e-2
		ep	1.80e-1	9.28e-2	7.37e-2	6.74e-2	6.48e-2	6.35e-2
	3	eu	2.15e-3	1.77e-2	2.28e-2	2.48e-2	2.55e-2	2.57e-2
		ep	1.80e-1	8.81e-2	6.69e-2	6.10e-2	5.91e-2	5.83e-2
1e-3	1	eu	2.14e-3	1.73e-2	2.21e-2	2.41e-2	2.48e-2	2.50e-2
		ep	1.80e-1	8.78e-2	6.61e-2	6.01e-2	5.82e-2	5.75e-2

Table 4.1: SRM errors for $\mu = 0.1$ without upwinding

We notice that the errors improve as the iteration proceeds until ϵ^s *reaches the discretization accuracy* $O(h)$, *where* s *is the number of iterations.*

For small viscosity, say $\mu = 0.001$, *the difference scheme (4.65) does not work. The errors blow up around* $t = 1$. *When we increase* α_2, *say to* 50, *we do get pretty good results around* $t = 1$; *however, the errors still blow up at a later time. This* *suggests that the scheme is not stable for small viscosity* μ*. So we next discretize the nonlinear term using the upwinding scheme given in [102]. For the case of small viscosity, e.g.* $\mu = 0.001$*, we get good results (see Table 4.2).*

ϵ	iteration	error at →	t = k	t = 1.0	t = 2.0	t = 3.0	t = 4.0	t = 5.0
5e-1	1	eu	4.66e-3	2.26e-1	2.50e-1	2.29e-1	2.13e-1	2.03e-1
		ep	2.58e-1	1.06e-1	6.67e-2	5.45e-2	5.12e-2	5.04e-2
	2	eu	2.16e-3	7.74e-2	8.78e-2	9.13e-2	9.34e-2	9.53e-2
		ep	1.84e-1	8.81e-2	6.22e-2	5.39e-2	5.11e-2	5.02e-2
	3	eu	2.14e-3	7.69e-2	8.71e-2	9.06e-2	9.29e-2	9.48e-2
		ep	1.83e-1	8.78e-2	6.21e-2	5.39e-2	5.11e-2	5.01e-2
1e-3	1	eu	2.14e-3	7.69e-2	8.72e-2	9.07e-2	9.29e-2	9.49e-2
		ep	1.83e-1	8.78e-2	6.21e-2	5.39e-2	5.11e-2	5.01e-2

Table 4.2: SRM errors for $\mu = 0.001$ with upwinding

Recall that according to Remark 4.5, in the case of small viscosity, the time step size can be increased to some extent without adverse stability effects. To demonstrate this, we take $k = h = 0.1$*, and* $\mu = 0.001$*. The numerical results in Table 4.3 support our claim.*

ϵ	iteration	error at →	t = k	t = 1.0	t = 2.0	t = 3.0	t = 4.0	t = 5.0
1e-3	1	eu	2.18e-2	8.61e-2	9.43e-2	9.70e-2	9.86e-2	9.99e-2
		ep	1.83e-1	8.83e-2	6.26e-2	5.42e-2	5.13e-2	5.03e-2

Table 4.3: SRM errors for $\mu = 0.001$ with a pretty large time step $k = h = 0.1$

Although we use explicit schemes for SRM (4.2) with $\alpha_1 > 0$*, we still have to solve a banded symmetric positive definite system. An alternative is to take* $\alpha_1 = 0$ *to avoid solving any algebraic systems. Table 4.4 shows the computational results of the difference scheme (4.65) with* $\alpha_1 = 0$ *and* $\alpha_2 = 1$*. We take viscosity* $\mu = 0.1$*,* $h = 0.1$ *and* $k = 0.0005$*. Good results are obtained except for the pressure near* $t = 0$ *(cf. Remark 4.4) .*

ϵ	iteration	error at $\rightarrow$	t = k	t = 1.0	t = 2.0	t = 3.0	t = 4.0	t = 5.0
5e-1	2	eu	5.64e-3	3.57e-2	2.94e-2	2.71e-2	2.62e-2	2.60e-2
		ep	2.92e-0	9.70e-2	7.03e-2	6.17e-2	5.87e-2	5.77e-2

Table 4.4: SRM errors for $\mu = 0.1$ with $\alpha_1 = 0$

Chapter 5

SRM for the Simulation of Miscible Displacement in Porous Media

5.1 Introduction

As explained in Chapter 1, miscible displacement occurs in the tertiary oil-recovery process which can enhance hydrocarbon recovery in the petroleum reservoir. Numerical simulation plays an important role for this process because solvents (or chemicals) are expensive and experiments are hardly possible. Mathematically, miscible displacement in porous media is modeled by a nonlinear coupled system of the pressure-velocity equation and the concentration equation with appropriate boundary and initial conditions. The pressure-velocity equation is elliptic, while the concentration equation is parabolic, but normally convection-dominated. Accuracy for velocity approximation is important to obtain a good approximation for concentration since the concentration equation only includes the velocity variable. Mixed finite element methods for the pressure-velocity equation have been applied for this purpose [40, 41, 42, 46, 47, 120]. We are interested in applying the idea of the SRM to this equation since it has benefits over mixed finite element methods.

The SRM formulation for the pressure-velocity equation is a direct application of the sequential regularization method for time-dependent problems (see previous chapters). The velocity variable is involved in the linear systems only and the pressure variable is obtained by substitutions (without solving any linear systems) at each iteration level. We notice that the same formulation can also be obtained from the augmented Lagrangian method (originated from Uzawa's algorithm) [50, 31]. Unlike the augmented Lagrangian method using spectral analysis to discuss the convergence rate for discretized problems, we use asymptotic methods directly for the differential problems. The asymptotic method is easier to use for more general and more complicated problems than spectral analysis because the latter is scarcely possible for non-symmetric operators. We will later prove that our iterative schemes can improve the error to $O(\epsilon^s)$ at the sth iteration level, where ϵ is a small positive number. In other words, the convergence rate of our iterative procedure is about $O(\epsilon)$. Theoretical convergence analysis and numerical experiments show that the number of iterations is extremely small, usually 2.

The organization of this chapter is as follows. In §5.2, we describe our SRM iteration for the time-discretized problem (the differential problem in spatial variables) and its advantages. In §5.3, we show its convergence . Then in §5.4 we give a fully-discretized scheme using the Galerkin finite element method for the problem formulation where the SRM is used for the pressure-velocity equation. Finally, in §5.5, we present numerical examples to demonstrate the effectiveness and accuracy of our method.

5.2 SRM Formulation

To recall, we again write down the model problem. Consider the miscible displacement of one incompressible fluid by another in a porous reservoir $\Omega \subset \mathbf{R}^2$ over a time period $[0, T]$. Let $p(x,t)$ and $\mathbf{u}(x,t)$ denote the pressure and Darcy velocity of the fluid mixture, and let c be the concentration of the invading fluid. Then the mathematical model is a coupled nonlinear system of partial differential equations

$$\mathbf{u} = -a(\mathbf{grad}p - \gamma\mathbf{grad}d), \quad (\mathbf{x},t) \in \Omega \times [0,T], \tag{5.1a}$$

$$div\mathbf{u} = q(\mathbf{x},t), \quad (\mathbf{x},t) \in \Omega \times [0,T], \tag{5.1b}$$

$$\phi\frac{\partial c}{\partial t} - div(D(\mathbf{u})\mathbf{grad}c) + \mathbf{u}\cdot\mathbf{grad}c = g(c), \quad (\mathbf{x},t) \in \Omega \times [0,T], \tag{5.1c}$$

with the boundary conditions

$$\mathbf{u} \cdot \mathbf{n} = 0, \quad (\mathbf{x}, t) \in \Gamma \times [0, T], \tag{5.2a}$$

$$D(\mathbf{u})\mathbf{grad} c \cdot \mathbf{n} = 0, \quad (\mathbf{x}, t) \in \Gamma \times [0, T], \tag{5.2b}$$

and the initial condition

$$c(\mathbf{x}, 0) = c_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{5.3}$$

where $a = a(\mathbf{x}, c)$ is the mobility of the fluid mixture (we will later denote it as $a(c)$), $\gamma = \gamma(\mathbf{x}, c)$ and $d(\mathbf{x})$ are the gravity and vertical coordinate (we will later denote γ as $\gamma(c)$), q is the imposed external rates of flow, $\phi(\mathbf{x})$ is the porosity of the rock, D is the coefficient of molecular diffusion and mechanical dispersion of one fluid into the other, $g = g(\mathbf{x}, t, c)$ is a known linear function of c representing sources, and $\mathbf{n}$ is the exterior normal to the boundary $\Gamma = \partial\Omega$.

We assume that the mobility is bounded below and above by positive constants

$$0 < m_0 \leq a(c) \leq M_0,$$

and its gradient is bounded above by a positive constant. For existence of p, we assume that the mean value of q is zero and for uniqueness we suppose p has mean value zero.

In recent years much attention has been devoted to the numerical simulation of this problem. In this chapter we are interested in solving the velocity-pressure equation (5.1a)-(5.1b) using the idea of the SRM for the time-discretized problem. The method considered herein is a direct application of the sequential regularization method (see previous chapters but without the sense of regularization) and is closely related to the augmented Lagrangian method [50, 31] (without the augmented Lagrangian framework). We will analyze the method using the technique presented in previous chapters, in particular Chapter 3. These analyses give the convergence of the iterative procedure and its convergence rate at the same time.

After a time discretization, we obtain the following system for $\mathbf{u}$ and p at the current time step:

$$\mathbf{u} = -a(\tilde{c})(\mathbf{grad}p - \gamma(\tilde{c})\mathbf{grad}d), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{5.4a}$$

$$div\mathbf{u} = q(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{5.4b}$$

where $\tilde{c}$ is an approximation of c assumed to be known. Taking ϵ to be a small positive number, we replace the system (5.1a), (5.1b), (5.2a) by the following iterative method: for $s = 1, 2, \cdots$, find $\{\mathbf{u}_s, p_s\}$ such that

$$a(\tilde{c})^{-1}\mathbf{u}_s - \frac{1}{\epsilon}\mathbf{grad}(div\mathbf{u}_s - q)$$
$$= -(\mathbf{grad}p_{s-1} - \gamma(\tilde{c})\mathbf{grad}d), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{5.5a}$$

$$\mathbf{u}_s \cdot \mathbf{n} = 0, \quad (\mathbf{x}, t) \in \Gamma \times [0, T], \tag{5.5b}$$

and

$$p_s = p_{s-1} - \frac{1}{\epsilon}(div\mathbf{u}_s - q), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{5.6}$$

where the initial guess p_0 is required to satisfy the zero mean value property: $\int p_0 dx = 0$. Thus p_s has mean value zero from (5.6) and (5.5b). We note that, by taking $p_0 \equiv 0$, each iteration is a kind of penalty method.

This iterative procedure has the following salient features:

1. We solve a small system (5.5) for the velocity $\mathbf{u}$, and obtain the pressure p from (5.6) directly. We will show that the accuracy of such a method is $O(\epsilon^s)$ at the sth iteration level. Note that the system (5.5) is well-posed since, unlike the usual penalty method, we need not take ϵ very small.

2. The velocity-pressure equation was recently solved by the mixed finite element method [40, 41, 42, 46, 47, 120], in which the discrete spaces for $\mathbf{u}$ and p need to satisfy the Babuska-Brezzi condition, and the resulting linear system has a nonpositive definite coefficient matrix. In our method, $\mathbf{u}$ and p are obtained from equations (5.5) and (5.6) separately, and compatibility conditions between the discrete spaces of $\mathbf{u}$ and p are not needed. Moreover, system (5.5) leads to a symmetric positive definite coefficient matrix.

3. When the standard finite element method [38, 39, 44, 45, 48, 98, 99] is applied for the pressure equation, the velocity needs to be obtained by finite differencing the pressure variable, which gives less accuracy. Note that the accuracy of the approximate velocity is important, since the concentration equation involves the velocity only. The velocity in our method is obtained directly, without finite differencing.

4. The discrete version of our SRM formulation (5.5)-(5.6) gives the same accuracy for the velocity as the mixed method and requires the solution of well-conditioned linear systems like Galerkin methods. Note that our numerical experiments will show that a few (much less than 10) iterations are usually enough for our iterative procedure.

5.3 Convergence Analysis

Before we begin our analysis, we first describe some notation to be used throughout the rest of the paper. As in Chapter 4, we use $L^p(\Omega)$, or simply L^p, to denote the space of functions whose pth power is integrable in Ω, with the norm

$$\|\mathbf{u}\|_p = \left(\int_\Omega \sum_{i=1}^{n} |u_i|^p \, dx\right)^{\frac{1}{p}},$$

where $\mathbf{u} = (u_1, \cdots, u_n)$. We will omit the subscript p in the norm notation when $p = 2$. $\mathbf{H}^m$ is the closure of $\mathbf{C}_0^\infty(\Omega)$ in the norm

$$\|\mathbf{u}\|_{\mathbf{H}^m} = \left(\sum_{0 \le |\alpha| \le m} \|D^\alpha \mathbf{u}\|^2 \right)^{\frac{1}{2}}.$$

In addition, we define the divergence space $\mathbf{H}(div) = \{\mathbf{w} \in L^2(\Omega)^2 : \ div\mathbf{w} \in L^2(\Omega)\}$ with the following norm:

$$\|\mathbf{u}\|_{\mathbf{H}(div)} = \left(\|\mathbf{u}\|^2 + \|div\mathbf{u}\|^2 \right)^{\frac{1}{2}}.$$

We shall denote by $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ the inner products in Ω and on Γ, respectively. For a normed linear space B with norm $\|\cdot\|_B$ and a sufficiently regular function $g : [\alpha, \beta] \to B$, we define

$$\|g\|_{L^2([\alpha,\beta];B)} = (\int_\alpha^\beta \|g(\cdot, t)\|_B^2 dt)^{\frac{1}{2}} \ \text{ and } \ \|g\|_{L^\infty([\alpha,\beta];B)} = \sup_{\alpha \le t \le \beta} \|g(\cdot, t)\|_B.$$

If $[\alpha, \beta] = [0, T]$, we simplify the notation as $\|g\|_{L^2(B)}$ and $\|g\|_{L^\infty(B)}$, respectively. We shall also denote generic constants by M and K , which may be different at different occurrences.

Before stating our convergence theorem, we first give a lemma.

Lemma 5.1 *There exists a unique solution $\{\mathbf{u}, p\}$ to the problem*

$$\mathbf{u} = -a(\tilde{c})(\mathbf{grad}p - \mathbf{f}), \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{5.7a}$$

$$div\mathbf{u} = q, \quad (\mathbf{x}, t) \in \Omega \times [0, T], \tag{5.7b}$$

$$\mathbf{u} \cdot \mathbf{n} = 0, \quad (\mathbf{x}, t) \in \Gamma \times [0, T], \tag{5.7c}$$

where p and q have mean value zero. Furthermore, there exits a constant M_1 such that the following estimates hold:

$$\|\mathbf{u}\|_{\mathbf{H}(div)} + \|p\|_{H^1} \le M_1[\|q\| + \|\mathbf{f}\|_{\mathbf{H}(div)}], \quad \forall t \in [0, T]. \tag{5.8}$$

Proof: Substituting (5.7a) into (5.7b) and (5.7c) we see that p satisfies a Poisson equation with Neumann boundary condition

$$\begin{aligned} &-div(a\mathbf{grad}p) = q - div(a\mathbf{f}), \quad \text{in } \Omega, \\ &a\frac{\partial p}{\partial \mathbf{n}} = a\mathbf{f}\cdot\mathbf{n} \quad \text{on } \Gamma. \end{aligned}$$

Noting the zero mean value of p and using the standard results for the Poisson equation, we obtain the uniqueness and existence of p and the estimate

$$\|p\|_{H^1} \leq M\left[\|q\| + \|\mathbf{f}\|_{\mathbf{H}(div)} + \|\mathbf{f}\cdot\mathbf{n}\|_{H^{-1/2}(\Gamma)}\right].$$

An application of the trace inequality [54](p.28)

$$\|\mathbf{f}\cdot\mathbf{n}\|_{H^{-1/2}(\Gamma)} \leq \|\mathbf{f}\|_{\mathbf{H}(div)}, \qquad \forall \mathbf{f} \in \mathbf{H}(div)$$

leads to the inequality (5.8) for p. Then the existence, uniqueness and the estimate (5.8) for $\mathbf{u}$ follow directly. □

We are now ready to describe our convergence theorem and prove it using a similar technique as Chapter 3.

Theorem 5.1 *Let $\{\mathbf{u}, p\}$ be the solution of system (5.4a), (5.4b), and (5.2a) and $\{\mathbf{u}_s, p_s\}$ the solution of (5.5)-(5.6). Then we have*

$$\|\mathbf{u} - \mathbf{u}_s\|_{\mathbf{H}(div)} + \|p - p_s\|_{H^1} \leq \left(\frac{M_1\epsilon}{1 - M_1\epsilon}\right)^s \|p_0 - p\|, \quad s = 1, 2, \cdots. \tag{5.9}$$

Here M_1 is from Lemma 5.1 and we assume $M_1\epsilon < \frac{1}{2}$.

Proof: We first consider the case $s = 1$ of (5.5)-(5.6). Write (5.5a) and (5.6) as

$$\begin{aligned} &a(\tilde{c})^{-1}\mathbf{u}_1 + \mathbf{grad}p_1 = \gamma(\tilde{c})\mathbf{grad}d \\ &div\mathbf{u}_1 - q = \epsilon(p_0 - p_1). \end{aligned}$$

Then subtracting equations (5.4a) and (5.4b), we have

$$a(\tilde{c})^{-1}(\mathbf{u}_1 - \mathbf{u}) + \mathbf{grad}(p_1 - p) = \mathbf{0} \tag{5.10a}$$

$$div(\mathbf{u}_1 - \mathbf{u}) = \epsilon(p_0 - p_1). \tag{5.10b}$$

Using Lemma 5.1, we obtain

$$\|\mathbf{u}_1 - \mathbf{u}\|_{\mathbf{H}(div)} + \|p_1 - p\|_{H^1} \leq M_1\epsilon\|p_0 - p_1\|. \tag{5.11}$$

Writing $p_0 - p_1 = p_0 - p + p - p_1$, we immediately have

$$\|\mathbf{u}_1 - \mathbf{u}\|_{\mathbf{H}(div)} + \|p_1 - p\|_{H^1} \leq \frac{M_1\epsilon}{1 - M_1\epsilon}\|p_0 - p\|. \tag{5.12}$$

Now we look at the second iteration $s = 2$. At first, we can get equations similar to (5.10a) and (5.10b):

$$a(\tilde{c})^{-1}(\mathbf{u}_2 - \mathbf{u}) + \mathbf{grad}(p_2 - p) = \mathbf{0} \tag{5.13a}$$

$$div(\mathbf{u}_2 - \mathbf{u}) = \epsilon(p_1 - p_2). \tag{5.13b}$$

Applying Lemma 5.1 again, we have

$$\begin{aligned} &\|\mathbf{u}_2 - \mathbf{u}\|_{\mathbf{H}(div)} + \|p_2 - p\|_{H^1} \leq M_1\epsilon\|p_1 - p_2\| \\ &\leq M_1\epsilon(\|p_1 - p\| + \|p_2 - p\|). \end{aligned} \tag{5.14}$$

Noting $M_1\epsilon < 1$ and using the estimate of $\|p_1 - p\|$ (see (5.12)), yield

$$\|\mathbf{u}_2 - \mathbf{u}\|_{\mathbf{H}(div)} + \|p_2 - p\|_{H^1} \leq (\frac{M_1\epsilon}{1 - M_1\epsilon})^2\|p_0 - p\|. \tag{5.15}$$

We can repeat this procedure, and by induction, conclude the results of the theorem. □

From Theorem 5.1 we see that the convergence rate of our iterative scheme (5.5)-(5.6) is about $O(\epsilon)$. This implies that the number of iterations needed to achieve a prescribed accuracy is very small. The fast convergence of our method makes it dramatically different from penalty-like methods.

5.4 The Galerkin Approximation and Its Error Estimates

In this section, we approximate the velocity-pressure iterative scheme (5.5) and the concentration equation (5.1c) by using the standard Galerkin method. We only provide a brief description about this approximation. More details are in [82].

Let $\mathbf{W} = \{\mathbf{w} \in \mathbf{H}(div) : \mathbf{w}\cdot\mathbf{n} = 0 \text{ on } \Gamma\}$. The variational form of (5.5) can be written into the following: find $\mathbf{u}_s : [0,T] \to \mathbf{W}$ such that

$$\begin{aligned}(a^{-1}(c)\mathbf{u}_s, \mathbf{w}) &+ \frac{1}{\epsilon}(div\mathbf{u}_s, div\mathbf{w}) \\ &= (p_{s-1} + \frac{1}{\epsilon}q, div\mathbf{w}) + (\gamma(c)\mathbf{grad}d, \mathbf{w}), \quad \forall \mathbf{w} \in \mathbf{W}. \end{aligned} \tag{5.16}$$

The weak form of the concentration equation (5.1c) reads: find $c : [0,T] \to \mathbf{H}^1(\Omega)$ such that

$$(\phi\frac{\partial c}{\partial t}, z) + (D(\mathbf{u})\mathbf{grad}c, \mathbf{grad}z) + (\mathbf{u}\cdot\mathbf{grad}c, z) = (g(c), z), \quad \forall z \in H^1(\Omega), \tag{5.17}$$

$$(c(x,0), z) = (c_0(x), z), \quad \forall z \in H^1(\Omega). \tag{5.18}$$

For $h_u > 0$ and an integer $k \ge 0$, with respect to the velocity-pressure equation, we introduce finite element spaces $\mathbf{W}_h \subset \mathbf{W}$ and $Y_h = \{y : y = div\mathbf{w} \text{ for } \mathbf{w} \in \mathbf{W}_h\}$ associated with a quasi-regular subdivision of Ω into triangles or rectangles of diameter less than h_u. Similarly, we denote by $Z_h \subset H^1(\Omega)$ the finite-dimensional space for the concentration equation with grid size h_c and approximation index l. Assume that the following approximation properties hold:

$$\inf_{\mathbf{w}_h \in \mathbf{W}_h} \|\mathbf{w} - \mathbf{w}_h\| \le K h_u^{k+1} \|\mathbf{w}\|_{\mathbf{H}^{k+1}}, \quad \mathbf{w} \in \mathbf{W}, \tag{5.19}$$

$$\inf_{\mathbf{w}_h \in \mathbf{W}_h} \|div(\mathbf{w} - \mathbf{w}_h)\| \le K h_u^{k+1} (\|\mathbf{w}\|_{\mathbf{H}^{k+1}} + \|div\mathbf{w}\|_{H^{k+1}}), \quad \mathbf{w} \in \mathbf{W}, \tag{5.20}$$

$$\inf_{z_h \in Z_h} \|z - z_h\| \le K h_c^{l+1} \|z\|_{H^{l+1}}, \quad z \in H^1(\Omega), \tag{5.21}$$

where K is a constant. The space $\mathbf{W}_h$ can be taken to be the vector part of the Raviart-Thomas [100] space of index k, or Brezzi-Douglas-Marini [30] space of index $k+1$.

Given a partition of $[0, T]$, $0 = t_0 < t_1 < \cdots < t_N = T$, we denote $J_n = [t_n, t_{n+1}]$, $\Delta t_n = t_{n+1} - t_n$, and $\Delta t = \max\{\Delta t_n\}$. Let $\{p_n, \mathbf{u}_n, c_n\}$ and $\{P_n, \mathbf{U}_n, C_n\}$ be $\{p, \mathbf{u}, c\}$ and its approximation at time level t_n. We define our approximation scheme at time $t_n (n = 0, 1, 2, \cdots)$ by the following.

Step 1: Given C_n, find $\{\mathbf{U}_n, P_n\} \in \mathbf{W}_h \times Y_h$ as follows. Take the initial guess $\bar{P}_0 = 0$. For $s = 1, 2, \cdots$, iteratively obtain $\bar{\mathbf{U}}_s \in \mathbf{W}_h$ and $\bar{P}_s \in Y_h$ such that

$$\begin{aligned}
&(a^{-1}(C_n)\bar{\mathbf{U}}_s, \mathbf{w}) + \frac{1}{\epsilon}(div\bar{\mathbf{U}}_s, div\mathbf{w}) \\
&\quad = (\bar{P}_{s-1} + \frac{1}{\epsilon}q, div\mathbf{w}) + (\gamma(C_n)\mathbf{grad}d, \mathbf{w}), \ \forall \mathbf{w} \in \mathbf{W}_h, \qquad (5.22a) \\
&\quad \bar{P}_s = \bar{P}_{s-1} - \frac{1}{\epsilon}(div\bar{\mathbf{U}}_s - q). \qquad (5.22b)
\end{aligned}$$

Let $\mathbf{U}_n = \bar{\mathbf{U}}_s$ and $P_n = \bar{P}_s$ for some integer s.

Step 2: When $\mathbf{U}_n$ is known, find $C_{n+1} \in Z_h$ such that

$$\begin{aligned}
&(\phi\frac{C_{n+1} - C_n}{\Delta t_n}, z) + (D(\mathbf{U}_n)\mathbf{grad}C_{n+1}, \mathbf{grad}z) + (\mathbf{U}_n \cdot \mathbf{grad}C_{n+1}, z) \\
&= (g(C_{n+1}), z), \qquad \forall z \in Z_h. \qquad (5.23)
\end{aligned}$$

Note that in step 1 of the scheme, the initial guess could be more efficiently taken as $\bar{P}_0 = P_{n-1}$ for $n \geq 1$ and $\bar{P}_0 = 0$ for $n = 0$. Our numerical experiments will show that the number of iterations can be generally taken to be $s = 2$ for the range of perturbation parameter $\epsilon = 10^{-3}$ to 10^{-5}.

The following error estimates of the above approximation are proved in [82].

Theorem 5.2 *Let $\{p, \mathbf{u}, c\}$ be the solution to Problem (5.1c)-(5.1b), and $\{P, \mathbf{U}, C\}$ the solution to the scheme (5.22)-(5.23) with s iterations at each time step. Then there exists a constant K, for Δt sufficiently small, such that the following error estimates hold at time step t_m ($m = 0, 1, 2, \cdots, N$):*

$$\|P_m - p_m\| + \|U_m - u_m\|_{H(div)} + \|C_m - c_m\|$$

$$
\begin{aligned}
&\leq K_s[\epsilon^s + h_c^{l+1}\|c\|_{L^\infty([0,t_m];H^{l+1})} + h_c^{l+1}\|c_t\|_{L^2([0,t_m];H^{l+1})} + \qquad (5.24)\\
&+h_u^{k+1}(\|\mathbf{u}\|_{L^\infty([0,t_m];\mathbf{H}^{k+1})} + \|div\mathbf{u}\|_{L^\infty([0,t_m];H^{k+1})}) + \Delta t(\|\mathbf{u}_t\|_{L^2([0,t_m];L^2)} + \|c_{tt}\|_{L^2([0,t_m];L^2)})],\\
&(\sum_{n=0}^{m-1} \Delta t_n \|C_{n+1} - c_{n+1}\|_1^2)^{1/2}\\
&\leq K_s[\epsilon^s + h_c^{l}\|c\|_{L^\infty([0,t_m];H^{l+1})} + h_c^{l+1}\|c_t\|_{L^2([0,t_m];H^{l+1})} + \qquad (5.25)\\
&+h_u^{k+1}(\|\mathbf{u}\|_{L^\infty([0,t_m];\mathbf{H}^{k+1})} + \|div\mathbf{u}\|_{L^\infty([0,t_m];H^{k+2})}) + \Delta t(\|\mathbf{u}_t\|_{L^2([0,t_m];L^2)} + \|c_{tt}\|_{L^2([0,t_m];L^2)})].
\end{aligned}
$$

This theorem tells us that for sufficiently small perturbation parameter ϵ, the error estimates for the velocity, pressure and concentration are optimal.

5.5 Numerical Experiments

In this section, we present some numerical examples to show how well our iterative scheme performs, and how the parameter ϵ affects the number of iterations needed and accuracy required. For simplicity, we will just consider the pressure-velocity equation, since the concentration equation has been analyzed previously [40, 41, 42, 45, 46, 47, 48, 98, 99, 120].

Consider the elliptic problem with Neumann boundary condition

$$
\begin{aligned}
&\mathbf{u} = -a(\mathbf{grad}p - \mathbf{f}), \quad \mathbf{x} \in \Omega,\\
&div\mathbf{u} = q(\mathbf{x}), \quad \mathbf{x} \in \Omega,\\
&\mathbf{u}\cdot\mathbf{n} = 0, \quad \mathbf{x} \in \Gamma,
\end{aligned}
$$

where Ω is a square and Γ its boundary. More general domains Ω will not present technical problems.

The approximation scheme takes the form: Find $\mathbf{U}^s \in \mathbf{W}_h$, for $s = 1, 2, \cdots$,

$$(a^{-1}\mathbf{U}^s, \mathbf{w}) + \tfrac{1}{\epsilon}(div\mathbf{U}^s, div\mathbf{w}) = (P^{s-1} + \tfrac{1}{\epsilon}q, div\mathbf{w}) + (\mathbf{f}, \mathbf{w}),\ \forall \mathbf{w} \in \mathbf{W}_h, \quad (5.26)$$

$$P^s = P^{s-1} - \tfrac{1}{\epsilon}(div\mathbf{U}^s - q). \quad (5.27)$$

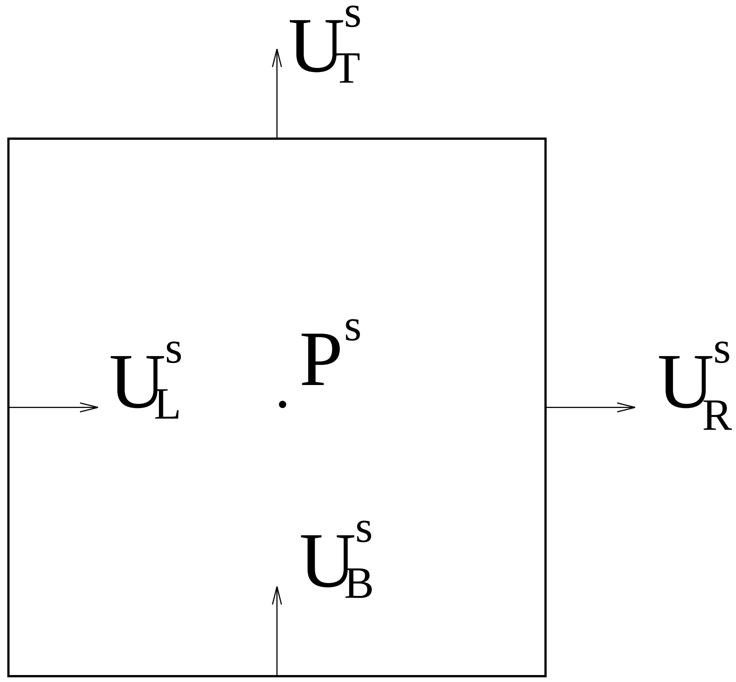

Figure 5.1: One element with velocity on each edge and pressure at the center

Partition the domain Ω into a set of squares of side length h. We take the space W_h to be the vector part of the Raviart-Thomas [100] space of index 0. Thus

$$\mathbf{W}_h = (\mathcal{P}_1 \otimes \mathcal{P}_0) \times (\mathcal{P}_0 \otimes \mathcal{P}_1),$$

where $\mathcal{P}_k$ is the set of one variable polynomials of order less than or equal to k. Consequently, the approximate pressure P^s lies in the space of piecewise constants. Partitioning the domain into triangles or rectangles or applying higher order approximation polynomials can be treated analogously.

Let U^s_α denote the constant value of the flux in the positive x or y-direction on the edge α, $\alpha =$ L, R, B, T (representing left, right, bottom, and top, respectively), of each element. See Figure 5.1. Consider $\mathbf{w}$ to be the basis function $(1 - x, 0)$ (on the standard reference square). Applying the trapezoidal rule to (5.26) we have

$$\begin{aligned}
&\frac{1}{2}a_L^{-1}U_L^s h^2 - \frac{U_R^s - U_L^s + U_T^s - U_B^s}{\epsilon h}\frac{1}{h}h^2 \\
&\quad = -\left[P^{s-1} + \frac{q_L + q_R + q_T + q_B}{4\epsilon}\right]\frac{1}{h}h^2 + \frac{1}{2}f_L h^2,
\end{aligned} \tag{5.28}$$

where q_α is the value of q at the middle point of edge α. Similarly, letting $\mathbf{w} = (x,0)$, $(0,1-y)$, $(0,y)$, and simplifying we have the following linear system for each element.

$$\begin{aligned}
&\epsilon h^2 a_L^{-1}U_L^s - 2(U_R^s - U_L^s + U_T^s - U_B^s) \\
&\quad = -2\epsilon h\left[P^{s-1} + \frac{q_L + q_R + q_T + q_B}{4\epsilon}\right] + \epsilon h^2 f_L,
\end{aligned} \tag{5.29a}$$

$$\begin{aligned}
&\epsilon h^2 a_R^{-1}U_R^s + 2(U_R^s - U_L^s + U_T^s - U_B^s) \\
&\quad = 2\epsilon h\left[P^{s-1} + \frac{q_L + q_R + q_T + q_B}{4\epsilon}\right] + \epsilon h^2 f_R,
\end{aligned} \tag{5.29b}$$

$$\begin{aligned}
&\epsilon h^2 a_B^{-1}U_B^s - 2(U_R^s - U_L^s + U_T^s - U_B^s) \\
&\quad = -2\epsilon h\left[P^{s-1} + \frac{q_L + q_R + q_T + q_B}{4\epsilon}\right] + \epsilon h^2 f_B,
\end{aligned} \tag{5.29c}$$

$$\begin{aligned}
&\epsilon h^2 a_T^{-1}U_T^s + 2(U_R^s - U_L^s + U_T^s - U_B^s) \\
&\quad = 2\epsilon h\left[P^{s-1} + \frac{q_L + q_R + q_T + q_B}{4\epsilon}\right] + \epsilon h^2 f_T.
\end{aligned} \tag{5.29d}$$

Note that equation (5.27) has the discrete version on each element:

$$P^s = P^{s-1} - \frac{1}{\epsilon}\left[\frac{U_R^s - U_L^s + U_T^s - U_B^s}{h} - \frac{q_L + q_R + q_T + q_B}{4}\right]. \tag{5.30}$$

From equations (5.29) we can easily form the element stiffness matrix. Then assembling all the element matrices and taking into account the boundary condition we obtain the stiffness matrix. The force vector can be obtained in an analogous way.

All velocity and pressure errors are measured for iterates $\{\mathbf{U}^s, P^s\}$ against the exact solution under the L^∞ norm, i.e.

$$\frac{\|\mathbf{U}^s - \mathbf{u}\|_\infty}{\|\mathbf{u}\|_\infty} \text{ and } \frac{\|P^s - p\|_\infty}{\|p\|_\infty}.$$

The initial guesses are always chosen to be zero, so all errors are 1.00 before the iterative procedure starts.

Example 5.1 *Let the velocity* $\mathbf{u}$ *and the pressure* p *satisfy* ($\mathbf{x} = (x, y)^T$)

$$\begin{aligned}
&div\mathbf{u} = \pi\left[\cos(\pi x) + \cos(\pi y)\right], \quad \mathbf{x} \in \Omega, \\
&\mathbf{u} = -\left[\mathbf{grad}p - \begin{pmatrix} \frac{1}{\pi} \\ -\frac{1}{\pi} \end{pmatrix}\right], \quad \mathbf{x} \in \Omega, \\
&\mathbf{u}\cdot\mathbf{n} = 0, \quad \mathbf{x} \in \Gamma,
\end{aligned}$$

where $\Omega = [0,1] \times [0,1]$, *and* $\Gamma = \partial\Omega$. *The true solutions for the velocity* $\mathbf{u}$ *and the pressure* p *are given by*

$$\begin{aligned}
\mathbf{u} &= \begin{pmatrix} \sin(\pi x) \\ \sin(\pi y) \end{pmatrix}, \\
p &= \frac{1}{\pi}(\cos(\pi x) + \cos(\pi y) + x - y).
\end{aligned}$$

The pressure p *and external flow rate* $q = div\mathbf{u}$ *are chosen in such a way that they both have mean value zero.*

Example 5.2 *Let the velocity* $\mathbf{u}$ *and the pressure* p *satisfy the nonhomogeneous problem*

$$\begin{aligned}
&div\mathbf{u} = ab^3(e^{by} - e^{bx}), \quad \mathbf{x} \in \Omega, \\
&\mathbf{u} = -a\left[\mathbf{grad}p - \begin{pmatrix} -x \\ y \end{pmatrix}\right], \quad \mathbf{x} \in \Omega, \\
&\mathbf{u}\cdot\mathbf{n} = g, \quad \mathbf{x} \in \Gamma,
\end{aligned}$$

where $\Omega = [0.5,1] \times [0.5,1]$, $\Gamma = \partial\Omega, a = 0.05$, *and* $b = 10$. *The function* g *is chosen such that the true solutions for the velocity* $\mathbf{u}$ *and the pressure* p *are given by*

$$\begin{aligned}
\mathbf{u} &= -a\begin{pmatrix} b^2e^{bx} + x \\ -b^2e^{by} - y \end{pmatrix}, \\
p &= b\left(e^{bx} - e^{by}\right).
\end{aligned}$$

	$\epsilon = 10^{-1}$		$\epsilon = 10^{-2}$		$\epsilon = 10^{-3}$	
iteration	velocity	pressure	velocity	pressure	velocity	pressure
0	1.00	1.00	1.00	1.00	1.00	1.00
1	1.14E-2	9.71E-3	1.53E-3	8.67E-4	2.69E-4	4.08E-5
2	2.69E-4	4.00E-5	1.29E-4	1.25E-4	1.28E-4	1.26E-4
3	1.29E-4	1.25E-4	1.28E-4	1.26E-4	-	-
4	1.28E-4	1.26E-4	-	-	-	-

Table 5.1: Numerical results for Example 5.1 with grid size = $\frac{1}{40}$

	$\epsilon = 10^{-4}$		$\epsilon = 10^{-5}$		$\epsilon = 10^{-6}$	
iteration	velocity	pressure	velocity	pressure	velocity	pressure
0	1.00	1.00	1.00	1.00	1.00	1.00
1	1.42E-4	1.16E-4	1.29E-4	1.25E-4	1.29E-4	1.26E-4
2	1.28E-4	1.26E-4	1.28E-4	1.26E-4	-	-

Table 5.2: Numerical results for Example 5.1 with grid size = $\frac{1}{40}$

For Example 5.1, the results with (uniform) grid size $\frac{1}{40}$ are shown in Tables 5.1 and 5.2. The results of Example 5.2 with (uniform) grid size $\frac{1}{40}$ and $\frac{1}{80}$ are shown in Tables 5.3. More examples and results can be found in [82].

From Tables 5.1 through 5.3 we conclude that our iterative method performs as well as the theory predicts. In particular, it can achieve the same accuracy as mixed finite element methods (at least for velocity), while the linear systems to be solved are symmetric and positive definite. Also, the computational work of our method is much smaller than that of mixed methods, since the number of iterations required is usually very small.

$\epsilon = 10^{-3}$	grid size $= \frac{1}{40}$		grid size $= \frac{1}{80}$	
iteration	velocity	pressure	velocity	pressure
0	1.00	1.00	1.00	1.00
1	1.18E-3	5.85E-3	3.23E-4	1.61E-3
2	1.02E-3	5.42E-3	2.00E-4	1.24E-3

Table 5.3: Numerical results for Example 5.2

Chapter 6

Numerical Methods of Some Singular Perturbation Problems

In this chapter, we discuss the numerical solutions of some singular perturbation problems which are all special cases of the regularizations (1.12) and (1.13), and have practical meanings in themselves as we indicated in §1.4. In §§6.1 and 6.3 we will construct and analyze uniformly convergent methods for these singular perturbation problems. So we first describe the definition of the method (cf. [90]).

Definition 6.1 *If u is the solution of a singular perturbation problem with a parameter ϵ and u^h is an approximation obtained using a* uniformly convergent method, *then there exist two constants ϵ_0 and h_0 independent of ϵ and h such that when $0 < \epsilon < \epsilon_0$ and $0 < h < h_0$ we have an inequality of the form*

$$\|u - u^h\| \leq Ch^p, \tag{6.1}$$

or in a weaker version (cf. [79])

$$\|u - u^h\| \leq C(h^p + \epsilon^r), \tag{6.2}$$

where $C > 0$, $p > 0$ and $r > 0$ are independent of ϵ and of the mesh width h, and $\|\cdot\|$ is some appropriate norm. □

We use M to represent generic (in the sense of $O(1)$) positive constants independent of ϵ and of the mesh width h. Some of these constants will also be denoted by m_0, m_1, m_2, M_0 and M_1, etc.

6.1 One Dimensional Quasilinear Turning Point Problems

In this section, we consider the following two-point boundary value problem with Dirichlet data at the endpoint:

$$-\epsilon u'' - b(x,u)u' + c(x,u) = 0, \; x \in I = [-1,1], \tag{6.3a}$$

$$u(-1) = U_-, \; u(1) = U_+, \tag{6.3b}$$

where $\epsilon \ll 1$ usually and $b(x,u)$ may be zero at some isolated points. Such problems are usually called turning point problems. Constructing uniformly convergent methods for problem (6.3) is generally very hard. We will consider a simpler case in which we know the point, say x^*, at which $b(x,u(x)) = 0$ and $b_x(x,u) \neq 0$ for all u in the vicinity of the solution. There are two types of turning point problems. They are called repulsive and attractive turning point problems corresponding to the negative and positive sign of $\frac{d}{dx}b(x,u(x))|_{x=x*}$, respectively.

The rest of the section is devoted to the two types of turning point problems and is actually a summary of two papers [79, 115] written by the author and his collaborator. We assume that the problem which we consider (in the form of (6.3)) has a unique solution. We also assume that the coefficients of problem (6.3) are sufficiently smooth (usually $C^2(I \times \mathbf{R})$ is enough).

6.1.1 A repulsive turning point problem

In this section, we assume

$$c_u(x,u) \geq c_0 \geq 0 \;\; \text{on} \;\; [-1,1] \times \mathbf{R} \tag{6.4}$$

Hence, a maximum principle holds for (6.3). By constructing a barrier function it is not difficult to get

$$\max_{-1 \leq x \leq 1} |u| \leq r, \tag{6.5}$$

where $r = \max_{-1\leq x\leq 1} |c(x,0)|/c_0 + max(|U_-|, |U_+|)$. Furthermore we make the following assumptions

$$b(0,u) = 0, b_x(0,u) < 0 \qquad \text{for} \quad |u| \leq r, \tag{6.6a}$$

$$b(x,u) \neq 0 \qquad \text{for} \quad |u| \leq r \quad \text{and} \quad x \neq 0. \tag{6.6b}$$

According the condition (6.6) there exists a small positive number δ such that

$$b_x(x,u) \leq -b_0 < 0 \qquad \text{for} \quad |x| \leq \delta \tag{6.7a}$$

$$b(x,u) \geq b_{-1} > 0 \qquad \text{for} \quad -1 \leq x \leq -\delta \tag{6.7b}$$

$$b(x,u) \leq -b_1 < 0 \qquad \text{for} \quad \delta \leq x \leq 1, \tag{6.7c}$$

where δ, b_0, b_{-1} and b_1 are positive constants independent of ϵ. We first give the bounds on the derivatives of the solution which is useful in the proof of uniform convergence:

$$|u^{(i)}(x)| \leq M(1+\epsilon^{-i}\exp(-m_0(x+1)/\epsilon) \qquad \text{for} \quad x \in [-1,\delta], \tag{6.8a}$$

$$|u^{(i)}| \leq M \qquad \text{for } |x| \leq \delta, \tag{6.8b}$$

$$|u^{(i)}(x)| \leq M(1+\epsilon^{-i}\exp(-m_0(1-x)/\epsilon) \qquad \text{for} \quad x \in [\delta,1], \tag{6.8c}$$

where δ is sufficiently small. (6.8a) and (6.8c) are a direct application of the result for problems without turning points which is derived by [113] and [77] independently. (6.8b) is proved in [79] by examining the Newton's sequence. This result means that the repulsive turning point problem does not have any interior layers.

Next we consider how the solution $v(x)$ of the reduced problem:

$$b(x,v)v' - c(x,v) \;=\; 0, -1 < x < 1, \tag{6.9a}$$

$$c(0,v(0)) \;=\; 0. \tag{6.9b}$$

approaches the solution $u(x)$. As in [21] (Remark 2.11), applying (6.8b) and results in [83], we have

$$|u(x) - v(x)| \leq M(\epsilon + \exp(-m_0(x+1)/\epsilon)) \qquad \text{for} \quad x \in [-1,-\delta], \tag{6.10a}$$

$$|u(x) - v(x)| \le M\epsilon \qquad \text{for } \; x \in [-\delta, \delta], \tag{6.10b}$$

$$|u(x) - v(x)| \le M(\epsilon + \exp(-m_0(1-x)/\epsilon)) \qquad \text{for } \; x \in [\delta, 1], \tag{6.10c}$$

Now we start to construct an almost uniformly convergent algorithm. In the construction, we combine the initial-value technique [67] (a modification is given in [81]) with the idea in [23]. We want to indicate here that [67, 81, 23] are all for problems without turning points.

We first rewrite (6.3) as

$$\begin{aligned} \epsilon u'' + (f(x,u))' - g(x,u) &= 0, \; x \in I = [-1,1], && (6.11a) \\ u(-1) = U_-, \; u(1) = U_+, & && (6.11b) \end{aligned}$$

where $f(x,u) = \int_0^u b(x,s)\,ds$ and $g(x,u) = c(x,u) - f_x(x,u)$. Integrating (6.11a), we get

$$\epsilon u' + f(x,u) = \int_0^x g(t,u(t))\,dt + K \text{ for } \; -1 < x < 1, \tag{6.12}$$

where the integration constant is $K = \epsilon u'(0)$. Let

$$E(x) = \int_0^x g(t,u(t))\,dt.$$

Then problem (6.3) is reduced to the following equivalent nonlinear initial value problems:

$$\begin{aligned} \epsilon u_1' + f(x,u_1) &= E(x) + K, \; -1 \le x < 0, && (6.13a) \\ u_1(-1) &= U_-; && (6.13b) \end{aligned}$$

and

$$\begin{aligned} \epsilon u_2' + f(x,u_2) &= E(x) + K, \; 0 < x \le 1, && (6.14a) \\ u_2(1) &= U_+. && (6.14b) \end{aligned}$$

Replacing $E(x)$ by

$$\bar{E}(x) = \int_0^x g(t, v(t))\, dt,$$

where $v(t)$ is the solution of the reduced problem (6.9), and neglecting K in (6.13a) and (6.14a), we obtain the approximate problems:

$$\begin{aligned} \epsilon y_1' + f(x, y_1) &= E(x),\ -1 \le x < 0, && (6.15a)\\ y_1(-1) &= U_-; && (6.15b) \end{aligned}$$

and

$$\begin{aligned} \epsilon y_2' + f(x, y_2) &= E(x),\ 0 < x \le 1, && (6.16a)\\ y_2(1) &= U_+. && (6.16b) \end{aligned}$$

It is proved in [79] that

$$|u_i(x) - y_i(x)| \le M\epsilon \ \text{ for } \ 0 < |x| \le 1 \ \text{ and } \ i = 1, 2. \qquad (6.17)$$

Finally we consider the numerical solution of (6.15) and (6.16). Note that by using the change of variable $\bar{x} = -x$ the numerical method for problem (6.15) will follow from that for problem (6.16), so we proceed considering only problem (6.16). For convenience we write (6.16) as

$$\begin{aligned} \epsilon y_2' - m_1 x y_2 &= -x e(x, y_2),\ 0 < x \le 1, && (6.18a)\\ y_2(1) &= U_+, && (6.18b) \end{aligned}$$

where $e(x, y) = d(x, y) + m_1 y$. In the expression $d(x, y) = (f(x, y) + \bar{E}(x))/x$ is bounded and m_1 is a positive constant to be determined below.

On the interval $[0, 1]$, introduce an arbitrary mesh $\{x_i, i = 1, 2, \cdots, N,$ with $x_N = 1\}$. Integrating problem (6.18) on the subinterval $[x_i, x_{i+1}]$, and replacing function

$e(x, y_2(x))$ by $e(x_{i+1}, y_2(x_{i+1}))$, suggests a difference scheme:

$$y^h_{2,i} = k_i y^h_{2,i+1} + (1-k_i)e(x_{i+1}, y^h_{2,i+1})/m_1, \tag{6.19a}$$

$$y^h_{2,N} = U_+, \ i = 1, 2, \cdots, N-1, \tag{6.19b}$$

where $k_i = \exp(-\frac{1}{2}m_1\epsilon^{-1}(x^2_{i+1} - x^2_i))$. According to the idea of [23], we can choose a suitable mesh

$$x_i = \begin{cases} 1 + (\epsilon/m)\ln(1 - N - i)h_1(1-\epsilon)), & i = N - N_1, \cdots, N, x_{N-N_1} = 1 - h_\epsilon \\ x_i, \max_i(x_i - x_{i-1}) = h_2, & i = 1, \cdots, N - N_1 - 1, \end{cases} \tag{6.20}$$

where $h_\epsilon = \epsilon|\ln\epsilon|/m$ and $h_1 = 1/N_1$. We have the following error estimate (see [79]):

Theorem 6.1 *Let $y_2(x_i)$ and $y^h_{2,i}$ be the solutions of problems (6.18) and (6.19), respectively, $i = 1, \cdots, N$. Under the mesh (6.20), taking*

$$m_1 \geq -f_u(x, u) = -b(x, u),$$

then we have

$$\max_{1 \leq i \leq N} |u_2(x_i) - u^h_{2,i}| \leq Mh, \tag{6.21}$$

where $h = max(h_1, h_2)$.

Applying the same method, we can also get a numerical solution $y^h_{1,i}, i = -N, \cdots, -1$, of problem (6.15) such that

$$\max_{-N \leq i \leq -1} |y_1(x_i) - y^h_{1,i}| \leq Mh. \tag{6.22}$$

Let

$$y^h_i = \begin{cases} y^h_{1,i} & \text{for } i = -N, \cdots, -1 \\ v(0) & \text{for } i = 0 \\ y^h_{2,i} & \text{for } i = 1, \cdots, N \end{cases} \tag{6.23}$$

be an approximation of the solution $u(x)$ of problem (6.3), where $v(0)$ can be solved from $c(0, v(0)) = 0$. Applying (6.17), (6.10b) , (6.21) and (6.22), we have

$$\max_{-N \leq i \leq N} |u(x_i) - y^h_i| \leq M(h + \epsilon). \tag{6.24}$$

Therefore, the numerical method we propose is an almost uniformly convergent method. In [79], a technique is given to modify the method to achieve uniform convergence. A numerical example is also given in [79].

6.1.2 An attractive turning point problem

In this section, we shall extend the results from [114] and [78]. We make the following assumptions:

$$\begin{aligned}
i)&\quad b(x,u) = xb_1(x,u), c(x,u) = xc_1(x,u) + \varepsilon c_2(x,u),\\
ii)&\quad b_1(x,u),\ c_k(x,u), \in C^2(I\times \mathbf{R}),\ k = 1,2,\\
iii)&\quad b_1(x,u) \geq b_* > 0,\quad x\in I,\quad u\in \mathbf{R},\\
iv)&\quad |c_{k,u}(x,u)| \leq c^*,\quad k=1,2,\quad x\in I,\quad u\in\mathbf{R}.
\end{aligned}$$

Moreover, we shall assume that ϵ is sufficiently small. Note that, unlike the previous section, the maximum principle may not hold since we do not assume (6.4) here. The corresponding reduced problem has a discontinuous solution consisting of two smooth curves, u_+ and u_-, which satisfy:

$$b_1(x,u_\pm)u_\pm' - c_1(x,u_\pm) = 0,\quad u_\pm(\pm 1) = U_\pm.$$

By μ denote $+\sqrt{\epsilon}$ and by $\|\cdot\|_\infty$ denote the maximum norm in $C(I)$. To consider uniformly convergent methods, we shall estimate the solution $u(x)$ and its derivatives and the quantities:

$$v_\pm \equiv (u-u_\pm)^{(k)}(x),\ \text{for } k = 0,1,2 \text{ and } x\in I_\pm,$$

where

$$I_- = [-1,0],\quad I_+ = [0,1].$$

We collect these estimates in the following theorem.

Theorem 6.2

$$|(xv_\pm(x))'| \quad\leq\quad M(\mu V(x)),\quad x\in I_\pm,$$

$$
\begin{aligned}
|(xv_\pm(x))''| &\leq M(\mu + \mu^{-1}V(x)), \quad x \in I_\pm, \\
\epsilon|u''(x)| &\leq M(\epsilon + V(x)), \quad x \in I, \\
\epsilon|u'''(x)| &\leq M(\mu + \mu^{-1}V(x)), \quad x \in I,
\end{aligned}
$$

where

$$V(x) = \exp(-\frac{m_0}{\mu}|x|),$$

with an positive constant m_0 independent of ε.

□

From the estimates we see that the attractive turning point problem has an interior layer but no boundary layers. This theorem is built on several lemmas whose proofs are very technical. To show how technical these proofs are we prove one lemma below. We refer to [115] for complete details.

Lemma 6.1 $|u(x)| \leq M$.

Proof: Let

$$p(x) = \begin{cases} |x|, & x \in I\backslash[-\mu, \mu]. \\ \frac{x^2}{2\mu} + \frac{\mu}{2}, & x \in [-\mu, \mu]. \end{cases}$$

We can easily verify that

$$p \in C^1(I), \ \max(|x|\ ,\ \mu) \geq p(x) \geq \frac{1}{2}\max(|x|\ ,\ \mu). \tag{6.25}$$

Next, consider the following Riccati initial value problem

$$P(\alpha) := \epsilon\alpha' + xb_*\alpha + M_0 p(x) + \epsilon\alpha^2 = 0, \tag{6.26}$$

$$\alpha(0) = 0. \tag{6.27}$$

It has a uniformly bounded solution. Indeed, by applying Newton's method to (6.26), (6.27) (cf.[88] or [79]), with the initial guess:

$$\alpha_0(x) = -\int_0^x \frac{M_0}{\epsilon} p(t) \exp(-\frac{b_*}{2\epsilon}(x^2 - t^2))dt,$$

the conditions of the Newton–Kantorovich theorem are satisfied since:

$$\begin{gathered}\|\alpha_0(x)\|_\infty \le M,\\ \|P'(\alpha_0)^{-1}\| \le M\mu,\\ \|\alpha_1-\alpha_0\|_\infty \le M\mu,\\ \|P''(\alpha)\| \le 2.\end{gathered}$$

Here, $\|\cdot\|$ is the operator norm corresponding to $\|\cdot\|_\infty$. Hence, there exists a solution $\alpha(x)$ to (6.26), (6.27), such that

$$\|\alpha-\alpha_0\|_\infty \le M\mu,$$

thus α is bounded uniformly in ϵ. Furthermore, we have

$$x\alpha(x) \le 0 \text{ for } x \in I. \tag{6.28}$$

Indeed, because of the maximum principle and

$$\epsilon\alpha' + xb_*\alpha = -M_0p(x) - \epsilon\alpha^2 \le 0, \quad \alpha(0)=0,$$

$\alpha(x) \le 0$ for $x \ge 0$ and $\alpha(x) \ge 0$ for $x \le 0$.

Let

$$\varphi(x) = \exp[\int_0^x \alpha(t)\,dt].$$

It holds that

$$0 < m_0 \le \varphi(x) \le M, \quad \varphi'(x) = \alpha(x)\varphi(x) = O(1). \tag{6.29}$$

Note that φ is a solution to the following equation:

$$\epsilon\varphi'' + xb_*\varphi' + M_0p(x)\varphi = 0.$$

Let us now consider an auxiliary problem:

$$\epsilon u'' + xb_1(x,u(x))u' - c(x,u) = 0, \tag{6.30}$$

$$u(-1) = U_-, \quad u(1) = U_+, \tag{6.31}$$

and let us make the transformation $u(x) = z(x)\varphi(x)$. Then $z(x)$ satisfies :

$$\epsilon z'' + [2\epsilon\alpha(x) + xb_1(x, u(x))]z' - \bar{c}(x, z) = 0, \tag{6.32}$$

$$z(-1) = \frac{U_-}{\varphi(-1)}, \quad z(1) = \frac{U_+}{\varphi(1)}, \tag{6.33}$$

where

$$\begin{aligned} \bar{c}(x, z) &= -[\epsilon\varphi''(x) + xb_1(x, u(x))\varphi'(x)]\varphi(x)^{-1}z + \varphi(x)^{-1}c(x, z\varphi(x)) \\ &= M_0 p(x) z - x[b_1(x, u(x)) - b_*]\alpha(x) z + \varphi(x)^{-1}c(x, z\varphi(x)). \end{aligned}$$

Choosing M_0 sufficiently large and using (6.25) and (6.28) we have

$$\begin{aligned} \bar{c}_z(x, z) &= M_0 p(x) - x\alpha(x)(b(x, u(x)) - b_*) + c_u(x, z\varphi(x)) \\ &\geq m_2 \max(|x|, \mu). \end{aligned}$$

Hence, the problem (6.32), (6.33) satisfies the maximum principle and therefore has a unique solution z, such that

$$|z(x)| \leq |z(-1)| + |z(1)| + \max_{x\in I} \frac{|\varphi^{-1}c(x, 0)|}{m_2 \max(|x|, \mu)} \leq M.$$

This shows that (6.30), (6.31) has a unique solution which is equal to u_ϵ. Then (6.29) completes the proof. $\square$

The numerical method is closely related to that from [114]. The same special non–equidistant mesh I^h is used. It has the following mesh points:

$$x_i = \lambda(t_i), \quad t_i = -1 + \frac{2i}{n}, \quad i = 0(1)n,$$

$$n = 2n_0, \quad n_0 \in \mathbf{N},$$

where

$$\lambda(t) = \begin{cases} \omega(t) := \frac{\beta\mu t}{\gamma - t}, & t \in [0, \alpha_0] \\ \pi(t) := \delta(t-\alpha_0)^3 + \frac{1}{2}\omega''(\alpha_0)(t-\alpha_0)^2 & \\ \quad +\omega'(\alpha_0)(t-\alpha_0) + \omega(\alpha_0), & t \in [\alpha_0, 1] \\ -\lambda(-t), & t \in [-1, 0] \end{cases}.$$

α_0 is an arbitrary parameter from (0,1),

$$\gamma = \alpha_0 + \mu^{\frac{1}{3}},$$

δ is determined from $\pi(1) = 1$, so that $\lambda \in C^2(I_\pm)$ and $\lambda \in C^1(I)$, and the parameter β is chosen from $(0, \gamma^{-1}(1-\alpha_0)^{-2}]$. It may look as if λ is artificial, but its part ω is a suitable rational approximation to the logarithmic function representing the inverse of the interior layer function $V(x)$ for $x \geq 0$. Then π is just a smooth extension of ω.

Let

$$h_i = x_i - x_{i-1}, \quad i = 1(1)n,$$

$$\hbar_i = \frac{1}{2}(h_i + h_{i+1}),$$

and let w^h denote a mesh function on $I^h \setminus \{-1, 1\}$, which will be identified with the $\mathbf{R}^{n-1}$–vector:

$$w^h = [w_1, w_2, \ldots, w_{n-1}]^T, \quad (w_i := w_i^h).$$

Moreover, let us introduce the following standard finite–difference operators:

$$D'_\pm w_i = \pm(w_{i\pm1} - w_i)/\hbar_i,$$

$$D'' w_i = [(w_{i-1} - w_i)/h_i + (w_{i+1} - w_i)/h_{i+1}]/\hbar_i.$$

We shall use the following discrete L^1–norm:

$$\|w^h\|_1^h = \sum_{i=1}^{n-1} \hbar_i |w_i|.$$

For all this cf. [114]. Finally, the constants M will now be independent of I^h as well.

Before discretizing the problem (6.3), as in the previous section, we rewrite (6.3a) in the following conservation form:

$$Tu := -\epsilon u'' - f(x, u)' + g(x, u) = 0, \quad x \in I, \tag{6.34}$$

where

$$f(x,u) = \begin{cases} f_-(x,u), & x \in I_- \\ f_+(x,u), & x \in I_+ \end{cases}, \quad g(x,u) = \begin{cases} g_-(x,u), & x \in I_- \\ g_+(x,u), & x \in I_+ \end{cases},$$

$$f_\pm(x,u) = \int_{u_\pm(x)}^{u} x b_1(x,s)ds,$$

$$\begin{aligned} g_\pm(x,u) &= c(x,u) - x b_1(x, u_\pm(x)) u'_\pm(x) + \int_{u_\pm(x)}^{u} (x b_1(x,s))_x ds \\ &= c(x,u) - x c_1(x, u_\pm(x)) + \int_{u_\pm(x)}^{u} (x b_1(x,s))_x ds. \end{aligned}$$

Then the discrete problem corresponding to (6.34), (6.3b) is given by:

$$T^h w_i = 0, \quad i = 1(1)n-1, \tag{6.35}$$

where

$$T^h w_i = \begin{cases} T_-^h w_i, & i = 1(1)n_0 \\ T_+^h w_i, & i = n_0 + 1(1)n-1 \end{cases},$$

$$T_\pm^h w_i = -\varepsilon D'' w_i - D'_\pm f_\pm(x_i, w_i) + g_\pm(x_i, w_i),$$

and where w_0 and w_n should be replaced by U_- and U_+ respectively. The discretization is a generalization of that for the mildly nonlinear case considered in [114] and [78].

Let us introduce the following assumption in addition to $i) - iv)$:

$$v) \;\; g_u(x,u) = (x b_1(x,u))_x + c_u(x,u) \geq g_* > 0, \quad x \in I, \quad u \in \mathbf{R}.$$

The following error estimate is proved in [115].

Theorem 6.3 *The discrete problem (6.35) has a unique solution w^h and the following estimate holds:*

$$\|w^h - u^h\|_1^h \leq M \frac{1}{n}[\mu + \exp(-n)],$$

where

$$u^h = [u(x_1), u(x_2), \ldots, u(x_{n-1})]^T.$$

□

A numerical example is given in [115].

6.2 Notes about Spurious Solutions of Upwind Schemes

6.2.1 Inadequacy of Yavneh's argument

First order upwind schemes have been found by several researchers to yield spurious solutions. Such behavior has been attributed to the excessive artificial viscosity introduced by the numerical scheme, to multiple solutions of the nonlinear set of algebraic equations that is obtained from the discretization, and to poor resolution by grids that are too coarse (cf. [121, 27]). A simple example in [61] also shows that an upwind scheme may lead to a spurious solution near a boundary layer. Yavneh in his Ph.D. thesis [121] (also see [27]) claimed that the spurious solution may occur even in cases where none of the above apply, and that such behavior is seen to occur even in the linear scalar advection-diffusion equation, despite the fact that the truncation-error tends to zero with the mesh-size throughout the entire domain, the solution being smooth everywhere. However, his proof is incomplete. Yavneh considers a linear advection-diffusion equation

$$-\epsilon\Delta u \quad + \quad \frac{1}{r^2}u_\theta = 0, \tag{6.36a}$$

$$u(a,\theta) = u_i \quad , \quad u(b,\theta) = u_o \tag{6.36b}$$

over the circular disk $0 < \theta \leq 2\pi, 0 < a < r < b$. Here

$$\Delta u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}$$

is the Laplacian in polar coordinates and $\epsilon > 0$. The unique solution (the uniqueness comes from the maximum principle) is

$$u = \frac{\ln[(\frac{r}{a})^{u_o}(\frac{b}{r})^{u_i}]}{\ln\frac{b}{a}}.$$

In order to see what might go wrong with the numerical solution of this problem, we rewrite the equation in Cartesian coordinates

$$Lu \quad = \quad -\epsilon\Delta u - \frac{y}{x^2+y^2}u_x + \frac{x}{x^2+y^2}u_y = 0, \tag{6.37a}$$

$$u|_{x^2+y^2=a} = u_i, u|_{x^2+y^2=b} = u_o, \tag{6.37b}$$

where $\Delta u = u_{xx} + u_{yy}$.

As is well known, discretization of this equation by central finite differences loses its stability when ϵ is small compared to the product of the mesh-size and the absolute value of either coefficient of the first derivatives in the equation. A common practice is to retain stability by increasing the absolute values of the coefficients of the discretized second derivatives. A particular method of this type is the first order upwind difference scheme. This sort of discretization gives a second-order central difference scheme to the following equation:

$$-\epsilon_1 \tilde{u}_{xx} - \epsilon_2 \tilde{u}_{yy} - \frac{y}{x^2+y^2}\tilde{u}_x + \frac{x}{x^2+y^2}\tilde{u}_y = 0 \tag{6.38}$$

with the same boundary conditions as (6.37b),where

$$\epsilon_1 = \epsilon + \frac{h|y|}{2(x^2+y^2)}, \ \epsilon_2 = \epsilon + \frac{h|x|}{2(x^2+y^2)}.$$

We plot the relationship of these continuous and discrete equations as follows:

$$\begin{array}{ccc} Problem\ (6.37) & \longleftarrow & Upwind\ scheme\ (first\ order) \\ \uparrow & & || \\ Problem\ (6.38) & \longleftarrow & Corresponding\ central\ scheme \end{array}$$

Yavneh considers the case $\epsilon < h/2r$ ($\epsilon \geq h/2r$ is not interesting since in this case the central difference scheme is stable) and shows that the difference between the solutions of problems (6.37) and (6.38) has a lower positive $O(1)$ bound throughout the domain except near the boundaries as the mesh-size tends to zero. Based on this result, he then claims that the first order upwind scheme yields a spurious solution of problem (6.37). From the above diagram, we see that the argument is incomplete since we do not know if the central difference scheme is a good approximation for problem (6.38). The proof of this fact is not that obvious since we no longer know if the solution of (6.38) is smooth. (6.38) is a partial turning point problem whose layer property could be complicated.

6.2.2 Our explanation

In this section, we show that the stability constant of problem (6.36) depends on ϵ. The smaller the parameter ϵ, the closer the problem is to being ill-posed. Hence, we can expect that any direct discretization (of course, including upwind schemes) on the problem would be unstable when ϵ is sufficiently small . Therefore, it is not strange that the first order upwind scheme fails for this problem.

Let's consider (6.36) with a forcing term $f(x, y)$. Suppose that $u(x, y)$ is a solution of the problem, i.e.

$$Lu = f(x, y) \tag{6.39}$$

and u satisfies boundary conditions (6.36b). We construct

$$u^*(x, y) = u(x, y) + (r^2 - a^2)(b^2 - r^2), \tag{6.40}$$

which satisfies the boundary conditions and

$$Lu^* = f(x, y) + \epsilon(12(x^2 + y^2) - 4(a^2 + b^2)). \tag{6.41}$$

In other words, if we make a small perturbation $\epsilon(12(x^2 + y^2) - 4(a^2 + b^2))$ to the right-hand side of the equation, the solution changes a lot (by $(r^2 - a^2)(b^2 - r^2)$). That means the problem (6.39),(6.36b) is not well-posed as ϵ is small. Then we may not expect its approximation (direct discretization) to be stable. Indeed we can prove that any difference scheme will not be stable (as ϵ is small).

Suppose we have a discretization (consistent with order h^r)

$$L_h u_h = f(x, y), B_h u_h = 0 \text{ (discrete BC)}$$

also

$$L_h u_h^* = f(x, y) + \epsilon(12(x^2 + y^2) - 4(a^2 + b^2))$$

$$B_h u_h^* = 0$$

Then

$$L_h(u_h - u_h^*) = \epsilon(12(x^2 + y^2) - 4(a^2 + b^2))$$

$$B_h(u_h - u_h^*) = 0$$

If (L_h, B_h) is stable, or more precisely, if the coefficient matrix A_h of the linear algebraic system corresponding to the discretization satisfies

$$\|A_h^{-1}\|_\infty \leq M, \tag{6.42}$$

where M is a generic positive constant independent of ϵ and h, we get

$$u_h - u = O(h^r)$$

$$u_h^* - u^* = O(h^r)$$

$$u_h - u_h^* = O(\epsilon)$$

Hence

$$u - u^* = O(\epsilon + h^r)$$

This is a contradiction since $u - u^* = (r^2 - a^2)(b^2 - r^2)$. So (L_h, B_h) can not be stable. In fact, we can show

Claim 1

$$\|A_h^{-1}\| \geq O(\frac{1}{max(\epsilon, h^r)}). \tag{6.43}$$

Proof: If $\epsilon \geq h^r$ and $\|A_h^{-1}\|_\infty = O(1/\epsilon^\delta)$, $\delta < 1$, then

$$u_h - u = O(\frac{h^r}{\epsilon^\delta}) \leq O(\epsilon^{1-\delta}),$$

$$u_h^* - u^* \leq O(\epsilon^{1-\delta}),\ u_h - u_h^* = O(\epsilon^{1-\delta}).$$

Therefore,

$$u - u^* = O(\epsilon^{1-\delta})$$

can be arbitrarily small. This is a contradiction in (6.40). So we must at least have $\|A_h^{-1}\|_\infty = O(1/\epsilon)$. On the other hand, if $\epsilon \le h^r$, we can prove that

$$\|A_h^{-1}\|_\infty \ge O(\frac{1}{h^r}).$$

Thus (6.43) is proved. $\square$

For the first order upwind scheme, we actually can prove

Claim 2

$$\|A_h^{-1}\| = O(\frac{1}{max(\epsilon, h)}). \tag{6.44}$$

Before we prove this claim, we first write down the scheme and then prove a discrete maximum principle. The scheme is

$$-\epsilon\frac{\frac{u_{i+1,j}-u_{i,j}}{h_{xi}} - \frac{u_{i,j}-u_{i-1,j}}{h_{xi-1}}}{\frac{1}{2}(h_{xi}+h_{xi-1})} - \epsilon\frac{\frac{u_{i,j+1}-u_{i,j}}{h_{yj}} - \frac{u_{i,j}-u_{i,j-1}}{h_{yj-1}}}{\frac{1}{2}(h_{yj}+h_{yj-1})}$$

$$\left\{\begin{array}{ll} -\frac{y_j}{x_i^2+y_j^2}\frac{u_{i+1,j}-u_{i,j}}{h_{xi}} + \frac{x_i}{x_i^2+y_j^2}\frac{u_{i,j}-ui,j-1}{h_{yj-1}} & (\text{for} x > 0, y > 0) \\ -\frac{y_j}{x_i^2+y_j^2}\frac{u_{i,j}-u_{i-1,j}}{h_{xi-1}} + \frac{x_i}{x_i^2+y_j^2}\frac{u_{i,j}-ui,j-1}{h_{yj-1}} & (\text{for} x > 0, y < 0) \\ -\frac{y_j}{x_i^2+y_j^2}\frac{u_{i+1,j}-u_{i,j}}{h_{xi}} + \frac{x_i}{x_i^2+y_j^2}\frac{u_{i,j+1}-ui,j}{h_{yj}} & (\text{for} x < 0, y > 0) \\ -\frac{y_j}{x_i^2+y_j^2}\frac{u_{i,j}-u_{i-1,j}}{h_{xi-1}} + \frac{x_i}{x_i^2+y_j^2}\frac{u_{i,j+1}-ui,j}{h_{yj}} & (\text{for} x < 0, y < 0) \end{array}\right\} = 0. \tag{6.45}$$

Lemma 6.2 *The difference scheme (6.45) has a discrete maximum principle, that is,*

$$L_h u_h \ge 0 \;\; and \;\; u_h|_{\Gamma_h} \ge 0 \Longrightarrow u_h \ge 0,$$

where Γ_h represents all the mesh points on the boundaries $x^2+y^2 = a^2$ and $x^2+y^2 = b^2$.

Proof: Suppose that $u_h \ge 0$ is not true. Then there exists an interior point (x_s, y_t) of the domain such that $u_{s,t} < 0$. Furthermore we can choose $u_{s,t}$ such that $u_{s,t} =$

$\min_{i,j} u_{i,j}$ and at least at one adjacent point $u_{i,j}$ of $u_{s,t}$ the rigorous inequality $u_{i,j} > u_{s,t}$ holds. Hence (for simplicity, we only consider $x > 0, y > 0$ case),

$$\begin{aligned} L_h u_h|_{s,t} &< -\epsilon \frac{\frac{u_{s,t}-u_{s,t}}{h_{xs}} - \frac{u_{s,t}-u_{s,t}}{h_{xs-1}}}{\frac{1}{2}(h_{xs}+h_{xs-1})} - \epsilon \frac{\frac{u_{s,t}-u_{s,t}}{h_{yt}} - \frac{u_{s,t}-u_{s,t}}{h_{yt-1}}}{\frac{1}{2}(h_{yt}+h_{yt-1})} \\ &\quad - \frac{y_t}{x_s^2+y_t^2}\frac{u_{s,t}-u_{s,t}}{h_{xs}} + \frac{x_s}{x_s^2+y_t^2}\frac{u_{s,t}-u_{s,t}}{h_{yt-1}} = 0. \end{aligned}$$

This is a contradiction to $L_h u_h \geq 0$, which completes the proof. □

Now we prove *Claim 2*.

Proof: We construct the barrier function

$$w_h = |u_o| + |u_i| + M_1 \frac{b^2 - x^2 - y^2}{max(\epsilon, h)} \max_{i,j} |L_h u_h|, \tag{6.46}$$

where $h = max(h_x, h_y)$, while $h_x = \max_i h_{xi},\ h_y = \max_j h_{yj}$. Then it is easy to verify that

$$L_h(w_h \pm u_h) \geq 0,\ w_h \pm u_h|_{\Gamma_h} \geq 0$$

when M_1 is sufficiently large. Using the discrete maximum principle we obtain

$$w_h \pm u_h \geq 0.$$

We thus have

$$|u_h| \leq w_h = |u_o| + |u_i| + M_1 \frac{b^2 - x^2 - y^2}{max(\epsilon, h)} \max_{i,j} |L_h u_h|.$$

This means

$$\|A_h^{-1}\|_\infty \leq M \frac{1}{max(\epsilon, h)}. \tag{6.47}$$

So (6.44) follows from (6.47) and *Claim 1*. □

Hence, the first order upwind scheme for problem (6.39),(6.36b) is not convergent when ϵ is small compared with mesh-size h. Of course, it produces a spurious solution in general. For Yavneh's example ($f(x,y) \equiv 0$ in (6.39)), numerical calculation verifies the appearance of the spurious solution. From the result of *Claim 2* we expect that a good second-order scheme would converge if ϵ is smaller than h but much larger than h^2. Numerical calculation in [121] verifies this too.

Note that, in polar coordinates, the first order upwind scheme converges since its truncation error has a factor ϵ (because the derivatives of u with respect to θ are all zero). In Cartesian coordinates, no such behavior exists.

6.3 A Linear Hyperbolic-Hyperbolic Singularly Perturbed Initial-Boundary Value Problem

Previous to this work, we discussed several singularly perturbed problems of hyperbolic type in joint papers [105, 106], constructed some difference schemes according to the properties of the problems, established discrete energy inequalities for the solutions of difference problems, and, based on the inequalities, proved that the difference schemes are uniformly convergent in the sense of the discrete energy norm. But in those papers the equations considered did not include a first derivative term with respect to the space variable x. Here we discuss a more complete initial-boundary value problem:

$$\begin{aligned} L_\epsilon u \quad &\equiv \quad \epsilon(u_{tt} - u_{xx}) + a(x,t)u_t + b(x,t)u_x + c(x,t)u \\ &= f(x,t),\ (x,t) \in G \equiv \{0 < x < l,\ 0 < t \leq T\} \qquad (6.48a) \\ &u(s,0) = \phi(x), u_t(x,0) = \psi(x),\ (0 \leq x \leq l) \qquad (6.48b) \\ &u(0,t) = 0,\ u(l,t) = 0,\ (0 \leq t \leq T) \qquad (6.48c) \end{aligned}$$

where $a(x,t), b(x,t), c(x,t), f(x,t), \phi(x)$ and $\psi(x)$ are sufficiently smooth functions and $a(x,t) \geq a_0 > 0$ for all $(x,t) \in \bar{G}$. Moreover, $b(x,t), f(x,t), \phi(x)$ and $\psi(x)$ satisfy the following compatibility conditions:

C1: $\phi(0) = 0,\ \psi(0) = 0,\ \phi(l) = 0,\ \psi(l) = 0;$

C2: $\phi''(0) = 0,\ b(0,0)\phi'(0) = f(0,0),\ \phi''(l) = 0,\ b(l,0)\phi'(l) = f(l,0).$

The subcharacteristics of the reduced operator

$$L_0 w = a(x,t)w_t + b(x,t)w_x + c(x,t)w \tag{6.49}$$

are timelike (cf.[53]) with respect to the characteristic directions of equation (6.48a), that is,

$$|b(x,t)|/a(x,t) < 1 \quad \text{for} \quad (x,t) \in \bar{G}. \tag{6.50}$$

For simplicity we assume that $b(x,t) \geq b_0 > 0$. Hence, (6.50) becomes

$$b(x,t) < a(x,t) \quad \text{for} \quad (x,t) \in \bar{G}. \tag{6.51}$$

The reduced problem of (6.48) is

$$L_0 u_0 \quad = \quad f(x,t), \tag{6.52a}$$

$$u_0(x,0) = \phi(x), \qquad u_0(0,t) = 0, \tag{6.52b}$$

where L_0 is defined as (6.49).

Therefore, the solution of problem (6.48) has boundary layer at $t = 0$ and $x = 1$. The reduced problem (6.52) is a first order hyperbolic initial-boundary problem in a semi-bounded region $D = \{(x,t),\ x \geq 0,\ t \geq 0\}$. According to [22], under the conditions $C1$ and $C2$, the first partial derivatives of the solution $u_0(x,t)$ of (6.52) are continuous, but the second derivatives are discontinuous along the characteristic line. In order to construct the asymptotic solution, it is necessary that $u_0(x,t) \in C^2(\bar{D})$. We will give a method to overcome the difficulty without additional compatibility conditions.

This section is a summary of the joint paper [107]. In what follows we first give an energy estimate of the solution of problem (6.48). Then, the asymptotic solution is constructed under the compatibility conditions $C1$ and $C2$ and uniform validity is proved in the sense of the energy norm. Thirdly, an exponentially fitted difference scheme for problem (6.48) is proposed and a discrete energy inequality is established.

Finally, we show that the discrete problem is uniformly convergent in the sense of discrete energy norm.

6.3.1 Construction of asymptotic solution and its remainder estimate

We first give the following energy inequality which will be used in estimating the remainder of the asymptotic expansion.

Lemma 6.3 *Let $u(x,t)$ be the solution of (6.48), and let the conditions we described previously be satisfied. Then as ϵ is sufficiently small, we have*

$$\|u\| + \epsilon\|u_t\| + \epsilon\|u_x\| \leq MK(G,\epsilon), \tag{6.53}$$

where

$$K(G,\epsilon) = \|f\|_{\bar{G}} + \|\phi\| + \epsilon\|\phi'\| + \epsilon\|\psi\|,$$
$$\|u\| = [\int_0^l u^2\, dx]^{\frac{1}{2}},\ \|f\|_{\bar{G}} = [\int_{\bar{G}} f^2\, dxdt]^{\frac{1}{2}}.$$

Proof: Multiplying equation (6.48a) by $2\epsilon a^{-1}u_t + u$, then integrating the obtained equation in the region $G_t = \{(x,s)|0 \leq x \leq l,\ 0 < s \leq t\}$, and performing the standard argument for the energy method, we thus have (6.53). $\square$

Next we construct the asymptotic solution. As we indicated before, under the compatibility conditions $C1$ and $C2$, the solution $u_0(x,t)$ of the reduced problem belongs to $C^1(D)$, but not $C^2(D)$. In order to realize the iterative procedure of the asymptotic solution, the continuity of the second derivatives of $u_0(x,t)$ is needed. However, this usually gives rise to the increase of the compatibility conditions. In this section, we construct the asymptotic solution under the conditions $C1$ and $C2$ without adding other ones. By making some transformations on (6.52) so the initial and boundary conditions become homogeneous and the right-hand function becomes equal to zero in the neighborhood of $t = 0$, we obtain a problem in which all compatibility conditions we need are satisfied. Then based on the transformed problem, an approximate problem for (6.48) is constructed and the solution of its reduced problem belongs to $C^2(D)$. Then an asymptotic expansion can be constructed.

Now we deal with problem (6.52). Let $u_0(x,t) = w_0(x,t) + \phi(x)$, noting that $\phi(0) = 0$, then

$$L_0 w_0 = F(x,t),\ w_0(x,0) = 0,\ w_0(0,t) = 0, \tag{6.54}$$

where operator L_0 is defined as (6.49) and $F(x,t) = f(x,t) - L_0\phi$. Introduce a function $\omega(y) \in C^\infty$ satisfying

$$\omega(y) = \begin{cases} 0 & (0 \le y \le \frac{1}{2}) \\ 1 & (y \ge 1) \end{cases}$$

and $0 \le \omega(y) \le 1$. Defining $\bar{F} = \omega(t/\delta)F(x,t)$, we have

$$F(x,t) - \bar{F}(x,t) = (1 - \omega(\frac{t}{\delta}))F(x,t) = \begin{cases} F(x,t) & (0 \le t \le \delta/2) \\ 0 & (t \ge \delta) \end{cases}$$

Therefore, $\bar{F}(x,t)$ and $F(x,t)$ have difference only in the region $\{0 \le t \le \delta,\ 0 \le x \le 1\}$. Replacing $F(x,t)$ by $\bar{F}(x,t)$, problem (6.54) is changed into

$$L_0\bar{w}_0 \ = \ \bar{F}(x,t) = \omega(\frac{t}{\delta})(f(x,t) - L_0\phi), \tag{6.55a}$$

$$\bar{w}_0(x,0) = 0,\ \bar{w}_0(0,t) = 0. \tag{6.55b}$$

Problem (6.55) for any $\delta > 0$ satisfies all compatibility conditions we need. Thus $\bar{w}_0$ is sufficiently smooth. Transforming back, i.e. letting $\bar{u}_0 = \bar{w}_0 + \phi$, we obtain an approximate problem for (6.52)

$$L_0\bar{u}_0 \ = \ \bar{f}(x,t) \tag{6.56a}$$

$$\bar{u}_0(x,0) = \phi(x) \quad , \quad \bar{u}_0(0,t) = 0, \tag{6.56b}$$

where

$$\bar{f}(x,t) = \omega(\frac{t}{\delta})f(x,t) + (1 - \omega(\frac{t}{\delta}))L_0\phi.$$

Here $\bar{u}_0(x,t)$ is sufficiently smooth since $\bar{w}_0$ and ϕ are. From the appendix of [22], we have

$$u_0 - \bar{u}_0 = O(\delta). \tag{6.57}$$

Now we can get an approximate problem of (6.48)

$$\begin{aligned} L_\epsilon \bar{u} &= \bar{f}(x,t), && (6.58a) \\ \bar{u}(0,x) = \phi(x), \bar{u}_t(0,x) = \psi(x) \quad &, \quad \bar{u}(0,t) = \bar{u}(l,t) = 0. && (6.58b) \end{aligned}$$

Using the energy inequality (6.53), we can get

$$\|u - \bar{u}\| = O(\delta^{\frac{1}{2}}). \; (0 \le t \le T) \tag{6.59}$$

Note that problem (6.49) is the reduced problem of (6.58). Hence, the reduced problem of (6.58) has enough smoothness such that the usual procedure to construct the asymptotic expansion can be performed. For example, we can construct the asymptotic expansion of (6.58) in the form of

$$\bar{u}(x,t) = \tilde{u}(x,t) + z,$$

where

$$\tilde{u}(x,t) = \bar{u}_0(x,t) + \epsilon v_0^{(0)}(x,\tau) + v_0^{(l)}(\xi,t) + \epsilon v_1^{(l)}(\xi,t), \;\; \tau = t/\epsilon, \;\; \xi = (l-x)/\epsilon,$$

$\bar{u}_0$ is the solution of (6.49) and $v_0^{(0)}, v_0^{(l)}$ and $v_1^{(l)}$ satisfy the following equations

$$\begin{aligned} (v_0^{(0)})_{\tau\tau} + a(x,0)(v_0^{(0)})_\tau &= 0, \\ (v_0^{(0)})_\tau(x,0) + (\bar{u}_0)_t(x,0) &= \psi(x), \lim_{\tau\to\infty} v_0^{(0)} = 0, \end{aligned}$$

$$\begin{aligned} (v_0^{(l)})_{\xi\xi} + b(l,t)(v_0^{(l)})_\xi &= 0, \\ v_0^{(l)}(0,t) + \bar{u}_0(l,t) = 0, &\qquad \lim_{\xi\to\infty} v_0^{(l)}(\xi,t) = 0 \end{aligned}$$

and

$$
\begin{aligned}
&(v_1^{(l)})_{\xi\xi} + b(l,t)(v_1^{(l)})_\xi = \\
&\qquad \xi b_x(l,t)(v_0^{(l)})_\xi + a(l,t)(v_0^{(l)})_t + c(l,t)v_0^{(l)} \\
&\qquad v_1^{(l)}(0,t) = 0, \lim_{\xi\to\infty} v_1^{(l)} = 0,
\end{aligned}
$$

respectively. Using the energy inequality (6.53) and estimating carefully, we have the following bound

$$\|z\| = \|\bar{u}(x,t) - \tilde{u}(x,t)\| \le M(\delta^{\frac{1}{2}} + \epsilon^{\frac{1}{2}} + \epsilon\delta^{-1} + \epsilon^2\delta^{-2}). \tag{6.60}$$

Denote the asymptotic expansion of problem (6.48) as

$$\tilde{u}_1(x,t) = u_0(x,t) + \epsilon v_0^{(0)}(x,\tau) + v_0^{(l)}(\xi,t) + \epsilon v_1^{(l)}(\xi,t).$$

Applying (6.57),(6.59) and (6.60) and taking $\delta = \epsilon^{\frac{2}{3}}$, we finally obtain

$$\|u(x,t) - \tilde{u}_1(x,t)\| \le M\epsilon^{\frac{1}{3}}. \tag{6.61}$$

Remark 6.1 *If the coefficients, right-hand side and initial values of problem (6.48) satisfy enough compatibility conditions such that the solution $u(x,t) \in C^3(G)$, then we can construct the asymptotic solution without needing the function $\omega(t/\delta)$. The asymptotic solution is still $\tilde{u}_1$ (replacing $\bar{u}_0$ by u_0 in the equations for $v_0^{(0)}$ and $v_0^{(l)}$). Moreover, we have the following remainder estimate*

$$|u(x,t) - \tilde{u}_1(x,t)| \le M\epsilon^{\frac{1}{2}}. \tag{6.62}$$

□

6.3.2 Difference scheme and its uniform convergence

Taking uniform meshes in the directions of x and t, we obtain a discrete region $\bar{G}_d = \{(x_i,t_j),\ i = 0,\cdots,N,\ j = 0,\cdots,[T/k],\ x_i = ih,\ t_j = jk\}$, where h and k

are mesh sizes and $Nh = l$. Denoting $u^d(x,t)$ as the approximate value of $u(x,t)$, we establish the difference scheme of problem (6.48) by using the exponential fitting idea:

$$L_\epsilon^{(h,k)} u^d(x,t) \equiv \gamma_1(x,t,k) u^d_{\bar{t}t}(x,t) - \gamma_2(x,t,h) u^d_{x\bar{x}}(x,t) + a(x,t) u^d_{\bar{t}}$$

$$+ b(x,t) u^d_{\bar{x}}(x,t) + c(x,t) u^d(x,t) = f(x,t), \ (x,t) \in G_d \tag{6.63a}$$

$$u^d(x,0) = \phi(x), u^d(x,k) - u^d(x,0) = k\psi(x), \tag{6.63b}$$

$$u^d(0,t) = u^d(l,t) = 0, \tag{6.63c}$$

where

$$\begin{aligned}
(x,t) &= (x_i, t_j), \ u^d_{x\bar{x}} = (u^d_x)_{\bar{x}}, \ u^d_{\bar{t}t} = (u^d_{\bar{t}})_{\bar{t}}, \\
\gamma_1(x,t,k) &= \frac{a(x,t)k \exp(-a(x,t)k/\epsilon)}{1 - \exp(-a(x,t)k/\epsilon)}, \\
\gamma_2(x,t,h) &= \frac{b(x,t)h \exp(-b(x,t)h/\epsilon)}{1 - \exp(-b(x,t)h/\epsilon)}, \\
u^d_{\bar{x}}(x,t) &= \frac{u^d(x,t) - u^d(x-h,t)}{h}, \\
u^d_{\bar{t}}(x,t) &= \frac{u^d(x,t) - u^d(x,t-k)}{k}, \\
u^d_x(x,t) &= \frac{u^d(x+h,t) - u^d(x,t)}{h}.
\end{aligned}$$

For this difference scheme we can establish the following discrete energy inequality.

Lemma 6.4 *Let $u^d(x,t)$ be the solution of (6.63) and let mesh sizes h and k satisfy inequality $b(x,t)h \le a(x,t)k$ for all $(x,t) \in \bar{G}_d$. Then when ϵ,h and k are sufficiently small we have*

$$\|u^d\|_s^2 + \|\gamma_1 u^d_{\bar{t}}\|_s^2 + \|\sqrt{\gamma_1\gamma_2} u^d_{\bar{x}}\|_s^2 \le MK(h,k,\epsilon), \tag{6.64}$$

where

$$\begin{aligned}
K(h,k,\epsilon) &= hk \sum_{j=1}^{J} \sum_{i=1}^{N} f^2 + \|\gamma_1 u^d_{\bar{t}}\|_1^2 + \|\sqrt{\gamma_1 \rho} u^d_{\bar{t}}\|_1^2 + \|\sqrt{\gamma_1\gamma_2} u^d_{\bar{x}}\|_1^2 + \|u^d\|_1^2, \\
\|v\|_s^2 &= h \sum_{i=1}^{N} v(ih, sk)^2, \ s = 1, \cdots, J, \ J = [T/k], \ \rho = k/\epsilon.
\end{aligned}$$

Proof: Essentially, the proof is a simulation of the continuous case but much more complicated because of the fitting factors γ_1 and γ_2. A sketch of the proof is in [107]. $\square$

Remark 6.2 *If fitting factors are not used, then the condition $b(x,t)h \le a(x,t)k$ can be removed in the proof.* $\square$

Next we prove the uniform convergence. We assume that enough compatibility conditions are satisfied so that the solution $u(x,t)$ of problem (6.48) belongs to $C^3(\bar{G})$. Based on the asymptotic expansion we may assert that the following estimates

$$|\frac{\partial^k u(x,t)}{\partial x^i \partial t^{k-i}}| \le M(\epsilon^{-i} + \epsilon^{1-(k-i)}),\ 0 \le k \le 3,\ 0 \le i \le k \tag{6.65}$$

hold. For convenience we also assume

$$c_1 k \le h \le c_2 k,\ c_1 > 0,\ a(x,t)/b(x,t) \ge c_2 > 0. \tag{6.66}$$

Using derivative estimate (6.65), it is not difficult to verify

$$\begin{aligned} L_\epsilon^{(h,k)}(u(x,t) - u^d(x,t)) &= O(\frac{h}{\epsilon^2} + \frac{k}{\epsilon}), \\ (u-u^d)|_{t=0} = 0,\ (u-u^d)|_{t=k} &= \min(\frac{k^2}{\epsilon}, k), \\ (u-u^d)|_{x=0} = 0,\ (u-u^d)|_{x=l} &= 0. \end{aligned}$$

Applying discrete energy inequality (6.64), yields

$$\|u - u^d\|_s \le M(\frac{h}{\epsilon^2} + \frac{k}{\epsilon}). \tag{6.67}$$

On the other hand, we can obtain

$$\begin{aligned} L_x^{(h,k)}(\tilde{u}_1 - u^d) &= L_\epsilon^{(h,k)}(u_0 + \epsilon v_0^{(0)} + v_0^{(l)} + \epsilon v_1^{(l)}) - f \\ &= O(\epsilon + h + k + \exp(-m_0 \frac{l-x}{\epsilon}),\ (m_0 > 0) \end{aligned}$$

and

$$(\tilde{u}_1 - u^d)|_{t=0} = O(\epsilon),\ (\tilde{u}_{1\bar{t}} - u^d_{\bar{t}})|_{t=k} = O(1),$$

$$(\tilde{u}_1 - u^d)|_{x=0} = O(\epsilon^n),\ (\tilde{u}_1 - u^d)|_{x=l} = 0,$$

where n is an arbitrary positive number. First making a change of variable such that the boundary value conditions become homogeneous, then using the discrete energy inequality (6.64), we have

$$\|\tilde{u}_1 - u^d\|_s \leq M(\sqrt{\max(\epsilon, h)} + k + \epsilon^n/h). \tag{6.68}$$

Hence, from the remainder estimate (6.62) of the asymptotic solution $\tilde{u}_1$, we obtain another error estimate

$$\|u - u^d\|_s \leq M(\sqrt{\max(\epsilon, h)} + k + \epsilon^n/h). \tag{6.69}$$

Combining (6.67) with (6.69) we achieve our uniformly convergent estimate.

Theorem 6.4 *Supposing that (6.66) is satisfied and the solution of (6.48) $u(x,t) \in C^3(\bar{G})$. Then the solution $u^d(x,t)$ of the discrete problem (6.63) uniformly converges to $u(x,t)$ in the sense of the discrete energy norm, i.e.*

$$\|u - u^d\|_s \leq Mh^{\frac{1}{5}},\ s = 0, 1, \cdots, J. \tag{6.70}$$

Proof: Using (6.69), (6.66) as $\epsilon^{5/2} \leq h$ and using (6.67), (6.66) as $\epsilon^{5/2} \geq h$, we have (6.70) immediately. □

Remark 6.3 *If we apply the remainder estimate (6.61) and corresponding procedure there we may discuss the uniform convergence under the compatibility conditions $C1$ and $C2$ only.* □

Chapter 7

Conclusion and Future Work

7.1 Summary and conclusions

The main focus of the thesis is on constrained ordinary differential equations (DAEs), constrained partial differential equations (PDAEs), and their applications. A new class of methods for solving high index DAEs has been developed, which we call the *Sequential Regularization Method* or SRM for short. These methods offer significant advantages over some known solution techniques, such as regularization and stabilization methods, and are applied to the nonstationary Navier-Stokes equations governing incompressible fluid flow and to a mathematical model of reservoir simulation.

High index DAEs (index ≥ 2) are usually difficult to discretize directly [29, 86]. We thus need to reformulate the original problem as a better behaved problem before discretization. Index reduction with stabilization is a popular reformulation for the numerical solution of semi-explicit high index DAEs. Another class of reformulations is regularization, where the DAE is replaced by a better behaved nearby problem. Such a method reduces the size of the system to be solved and avoids the derivatives of the algebraic constraints associated with the DAE problem. Regularization is particularly suitable for problems with certain singularities where the constraint Jacobian does not have full rank. Unfortunately, this approach often yields very stiff problems, which accounts for its lack of popularity in practice. The SRM is proposed to overcome this difficulty. It keeps the benefits of regularization methods and avoids the need to use stiff solvers for the regularized problems, because the regularization parameter does not need to be very small. Thus, we obtain an important improvement over usual regularization methods which leads to easier numerical methods (explicit time discretization for regularized problems). The SRM also provides cheaper and more efficient alternatives to the usual stabilization methods for some choices of parameters and stabilization matrix. We first propose and analyze the method for linear index-2 DAEs. Then we extend it to nonlinear index-2 and index-3 DAEs. This is especially useful in applications such as constrained multibody systems which are of index-3. Numerical experiments show that the method is useful and efficient for problems with and without singularities.

While a significant body of knowledge about the theory and numerical methods for DAEs has been accumulated, almost none has been extended to partial differential-algebraic equations (PDAEs). As a first attempt we provide a comparative study between stabilization and regularization (or pseudo-compressibility) methods for DAEs and PDAEs, using the Navier-Stokes equations as an instance of PDAEs. Compared with stabilization methods, regularization methods can avoid imposing an artificial boundary condition for the pressure p. This is a feature for PDAEs not shared with DAEs. We generalize the SRM to the nonstationary incompressible Navier-Stokes equations. Similar to DAEs, explicit schemes in the time direction can be used for the PDAE because of the reduced stiffness (taking the regularization parameter relatively large) or even essential nonstiffness obtained for some choice of parameters. Unlike usual regularization methods, the time step restriction for the explicit scheme can be independent of the regularization parameter ϵ. The time step restriction is further loosened for the case of small viscosity. A simple discretization (such as the forward Euler difference in time and a first-order scheme in space) is analyzed and implemented. Numerical results support our theoretical results. The method works for both two- and three-dimensional problems.

In recent years considerable attention has been devoted to numerical reservoir simulation, e.g. miscible displacement in porous media. We have applied the SRM idea to the pressure-velocity equation in its 2-dimensional mathematical model equations. This procedure is first analyzed at the differential level and then discretized by finite element methods. Theoretical analysis and numerical experiments show that this procedure converges at the rate of $O(\epsilon)$ per iteration, where ϵ is a small positive number. The fast convergence rate makes our iterative method dramatically different from penalty methods . In addition, the perturbation parameter ϵ does not have to be carefully chosen, unlike the case for other iterative methods. Indeed, our numerical experiments show that two iterations are usually enough for a variety of problems. Compared with mixed finite element methods, the discrete version of our scheme can provide the same accurate approximations for velocity and pressure, which is crucial in reservoir problems since velocity is intimately involved in the concentration equation. However, in contrast to mixed finite element methods, our scheme requires only the solution of symmetric positive definite linear systems which have a smaller number of degrees of freedom corresponding to the velocity variable. Since our method completely decoupled the velocity and pressure variables, the so-called Babuska-Brezzi condition is not needed in constructing the finite dimensional spaces for velocity and pressure. The method is easily applied to three-dimensional problems.

Another topic of this thesis is singular perturbation problems, which come from many applied areas and regularized problems. We discuss numerical solutions of several singular perturbation problems. Uniformly convergent schemes with respect to the perturbation parameter ϵ are constructed and analyzed for nonlinear repulsive and attractive turning point problems and a second-order hyperbolic problem. We are the first to be able to construct uniformly convergent schemes for these problems. Also, an interesting spurious solution phenomenon from an upwinding scheme is analyzed for an elliptic turning point problem. We find that the spurious solution is caused by a mild instability of the problem (the constant for the stability inequality is of $O(\frac{1}{\epsilon})$).

This type of instability is not as serious as supersensitivity [73, 74]. It can be handled by using higher order upwinding schemes as long as their accuracy $h^r \ll \epsilon$ when ϵ is not too small.

7.2 Discussion of future work

The results of this thesis can be extended in a number of directions.

Efficient simulation of kinematic chains with closed loops

The kinematic chain problem is an example of multibody systems, such as robot manipulators. Consider a chain consisting of n links. The numerical simulation of the problem is usually treated as two separate problems: (i) the forward dynamics problem for computing system accelerations, and (ii) the numerical integration problem for advancing the state in time. In recent years, many different algorithms have been proposed for solving the forward dynamics problem with tree structure (without closed loops), ranging in computational complexity from $O(n^3)$ (e.g. the composite rigid body method (CRBM) [116]) to $O(n)$ (e.g. the articulated–body method (ABM) [49]). However, it has never been possible to construct an $O(n)$ algorithm for the chain problem with many closed loops. The SRM (plus explicit discretization) opens up a way to do this. According to the idea of the SRM, we can remove the extra constraints caused by closed loops and incorporate them into external force terms of the system. The reformulated problem has the same structure as that of the problem without closed loops. We thus expect to develop and test an $O(n)$ algorithm for closed-loop chains.

Fully nonlinear DAEs

Consider a fully nonlinear index-ν DAE

$$x^{(\nu)} \quad = \quad f(t, x, x', \cdots, x^{(\nu-1)}, y),$$

$$0 = g(t, x),$$

(i.e. where algebraic variables y appear nonlinearly together with the differential variables x.) Some mechanical multibody systems are in this form with $\nu = 2$, e.g. the motion of a point mass on a parabolic orbit with the forces of gravitation and Coulomb friction (see [4]). The SRM can be applied to this problem. For instance, corresponding to the index-2 problem (with $\nu = 1$) we have: for $s = 1, 2, \cdots$,

$$\begin{aligned} x_s' &= f(t, x_s, y_s), \\ y_s &= y_{s-1} + \frac{1}{\epsilon} e(g(t, x_s)), \end{aligned}$$

where the function e should be chosen such that the matrix $f_y e_g g_x$ has positive eigenvalues.

Efficient simulation of real fluid problems using SRM

Because the computations for $\mathbf{u}$ and p are uncoupled and explicit time discretization can be used, we expect that the SRM incorporated with a finite difference or finite element discretization would provide a cheap and efficient method for simulating more realistic fluid problems. For long time simulations, a reinitialization technique (e.g. projecting back to the divergence free space after a number of time steps) may be useful. A comparative study between the SRM and other methods would be interesting.

Solving the system resulting from the discretization of the operator $I + \frac{\alpha_1}{\epsilon}\mathbf{grad}\, div$

This operator comes from using SRM to solve the nonstationary Navier-Stokes equations with $\alpha_1 \neq 0$. The system is easily made to be banded symmetric positive definite. Hence a direct method can be used to solve it. An interesting observation is that the usual iterative methods do not work well. This is probably due to the lack of ellipticity of the system. Some research on solving this problem using multigrid and domain decomposition techniques (at least for ϵ not too small) is about to be completed by Arnold, Falk and Winther [3]. Based on a technique described in [57], iterative methods (including multigrid) would be feasible under some pre-processing of the system (to increase the ellipticity). This was also suggested by W. Hackbusch in a private communication.

Bibliography

[1] C. Amirouche and V. Girault, Decomposition of vector spaces and application to the Stokes problem in arbitrary dimension, Czechoslovak Mathematical J. 44(1994), No. 119, pp. 109-140.

[2] F. Amirouche, *Computational Methods in Multibody dynamics*, Prentice-Hall, 1992.

[3] D. N. Arnold, *Private communication*, 1995.

[4] M. Arnold, *A perturbation analysis for the dynamical simulation of mechanical multibody systems*, Preprint 94/24, Universität Rostock, Fachbereich Mathematik, Germany, 1994.

[5] K. Arrow, L. Hurwicz and H. Uzawa, *Studies in Nonlinear Programming*, Stanford University Press, 1968.

[6] U. Ascher, *On some difference schemes for singular singularly-perturbed boundary value problems*, Numer. Math. 46 (1985), 1-30.

[7] U. Ascher, *On numerical differential algebraic problems with application to semiconductor device simulation*, SIAM J. Numer. Anal. 26(1989), 517-538.

[8] U. Ascher, H. Chin and S. Reich, *Stabilization of DAEs and invariant manifolds*, Numer. Math. 67 (1994), 131-149.

[9] U. Ascher, H. Chin, L. Petzold and S. Reich, *Stabilization of constrained mechanical systems with DAEs and invariant manifolds*, Journal of Mechanics of Structures and Machines 23 (1995), 135-157.

[10] U. Ascher, J. Christiansen and R.D. Russell, *Collocation software for boundary value ODEs*, Trans. Math. Software 7(1981), 209-222.

[11] U. Ascher and Ping Lin, *Sequential regularization methods for higher index DAEs with constraint singularities: Linear index-2 case*, SIAM J. Numer. Anal., to appear.

[12] U. Ascher and Ping Lin, *Sequential Regularization Methods for Nonlinear Higher Index DAEs*, SIAM J. Sci. Comput., to appear.

[13] U. Ascher, P.A. Markowich, P. Pietra, C. Schmeiser, *A phase plane analysis of transonic solutions for the hydrodynamic semiconductor model*, Math. Models & Methods in Applied Sciences 1(1991), 347-376.

[14] U. Ascher, R. Mattheij and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equation*, Prentice-Hall, 1988.

[15] U. Ascher and L. Petzold, *Projected implicit Runge-Kutta methods for differential-algebraic equations*, SIAM J. Numer. Anal. 28 (1991), 1097-1120.

[16] U. Ascher and L. Petzold, *Stability of Computational Methods for Constrained Dynamics Systems*, SIAM J. Sci. Comput. 14 (1993), 95-120.

[17] J. Baumgarte, *Stabilization of constraints and integrals of motion in dynamical systems*, Comp. Math. Appl. Mech. Eng. 1 (1972), 1-16.

[18] G. Bader and U. Ascher, *A new basis implementation for a mixed order boundary value ODE solver*, SIAM J. Scient. Stat. Comput. 8 (1987), 483-500.

[19] E. Bayo and A. Avello, *Singularity-free augmented Lagrangian algorithms for constrained multibody dynamics*, J. Nonlinear Dynamics 5 (1994), 209-231.

[20] E. Bayo, J. G. de Jalon and M. A. Serna, *A modified Lagrangian formulation for the dynamic analysis of constrained mechanical systems*, Computer Methods in Applied Mechanics and Engineering 71(1988), 183-195.

[21] A.E. Berger, H. Han and R.B. Kellogg, *A priori estimates and analysis of a numerical method for a turning point problem*, Math. Comp. 42(1984, 465-491.

[22] L.E. Bobisud, *The second initial-boundary value problem for a linear parabolic equation with a small parameter*, Mich. Math. J. 15(1969), 495-504.

[23] I.P. Boglaev, *On numerical integration of a singular- perturbed initial-value problem for ordinary differential equation*, Zh. Vychisl. mat. i mat. fiziki (Comp. Math. & Math. Physics) 25(1985), 1009-1022.

[24] Y. Boyarintsev, *Reguljarnyje i Singularnyje Sistemy Linejnych Obyknovennych Differencial'nych Uravnenij*, Nauka (Sibirskoje otdelenije), Novosibirsk, 1980.

[25] Y. Boyarintsev, *Methods of Solving Singular Systems of Ordinary Differential Equations*, John Wiley & Sons, 1992 (originally published in 1988 in Russian).

[26] A. Brandt, *Multigrid techniques: 1984 guide with applications to fluid dynamics*, The Weizmann Institute of Science, Rehovot, Israel, 1984.

[27] A. Brandt and I. Yavneh, *Inadequacy of first order upwind difference schemes for some recirculation flows*, J. Comp. Phys. 93(1991), 128-143.

[28] B. Brefort, J. M. Ghidaglia and R. Temam, *Attractors for the penalized Navier-Stokes equations*, SIAM J. Math. Anal. 19 (1988), 1-21.

[29] K. Brenan, S. Campbell and L. Petzold, *The numerical solution of higher index differential/algebraic equations by implicit Runge-Kutta methods*, SIAM J. Numer. Anal. 26 (1989).

[30] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47(1985), pp. 217-235.

[31] F. Brezzi, and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[32] S.L. Campbell, *Regularizations of linear time varying singular systems*, Automatica 20(1984), 365-370.

[33] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, Spectral Methods in Fluid Dynamics, Springer-Verlag, 1987.

[34] P. Carter, *Computational methods for the shape from shading problem*, Ph.D thesis, University of British Columbia, 1993.

[35] H. S. Chin, *Stabilization methods for simulations of constrained multibody dynamics*, Ph.D thesis, University of British Columbia, 1995.

[36] A.J. Chorin, *Numerical solution of the Navier-Stokes equations*, Math. Comp. 22 (1968), 745-762.

[37] E.P. Doolan, J.J.H. Miller and W.H.A. Schilders, *Uniform Numerical Methods for Problems with Initial and Boundary Layers*, Dublin, Boole Press, 1980.

[38] J. Douglas, Jr., *The numerical solution of miscible displacement in porous media*, in: Computational Methods in Nonlinear Mechanics (J. T. Oden, Ed.), Amsterdam, North Holland, 1980.

[39] J. Douglas, Jr., *Simulation of miscible displacement in porous media by a modified method of characteristics procedure*, in: Lecture Notes in Math. 912, Springer-Verlag, (1982).

[40] J. Douglas, Jr., R. E. Ewing and M. F. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Numer. Anal. 17(1983), pp. 17-33.

[41] J. Douglas, Jr, R. E. Ewing and M. F. Wheeler, *A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media.* RAIRO Numer. Anal. 17(1983), pp. 249-265.

[42] R. C. Duran, *On the approximation of miscible displacement in porous media by a method of characteristics combined with a mixed method*, SIAM J. Numer. Anal. 25(1988), pp. 989-1001.

[43] H.W. Engl, M. Hanke and A. Neubauer, *Tikhonov regularization of nonlinear differential-algebraic equations*, Institutsbericht Nr. 385, Johannes-Kepler-Universität, Institut für Mathematik, Linz, 1989.

[44] R. E. Ewing (Ed.), *The Mathematics of Reservoir Simulation*, SIAM Philadelphia, 1983.

[45] R. E. Ewing and T. F. Russell, *Efficient time-stepping methods for miscible displacement problems in porous media*, SIAM J. Numer. Anal. 19(1982), pp. 1-67.

[46] R. E. Ewing, T. F. Russell and M. F. Wheeler, *Simulation of miscible displacement using mixed methods and a modified method of characteristics*, SPE 12241, 7th SPE Symposium on Reservoir Simulation, San Francisco, 1983.

[47] R. E. Ewing, T. F. Russell and M. F. Wheeler, *Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics*, Comp. Meth. Appl. Mech. Engrg. 47(1984), pp. 73-92.

[48] R. E. Ewing and M. F. Wheeler, *Galerkin Methods for miscible displacement problems in porous media*, SIAM J. Numer. Anal. 17(1980), pp. 351-365.

[49] R. Featherstone, *Robot Dynamics Algorithms*, Kluwer Academic Publishers, Norwell, MA, 1987.

[50] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, Studies in Mathematics and Its Applications 15, North-Holland, 1983.

[51] C.W. Gear, *The simultaneous numerical solution of differential-algebraic equations*, IEEE Trans. Circuit Theory, CT-18(1971), 89-95.

[52] C.W. Gear, H.H. Hsu and L. Petzold, *Differential- algebraic equations revisited*, Proc. ODE Meeting, Oberwolfach, Germany, 1981.

[53] R. Geel, *Singular Perturbations of Hyperbolic Type*, Mathematical Center Tracts 98, Amsterdam, 1978.

[54] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin-Heidelberg 1986.

[55] P.M. Gresho, *Some current issues relevant to the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg. 87 (1987), 201-252.

[56] E. Griepentrog and R. März, *Differential-algebraic equations and their numerical treatment*, Teubner Texte zur Mathematik 88, Leipzig, 1986.

[57] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, Heidelberg 1985.

[58] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II*, Springer-Verlag, 1991.

[59] M. Hanke, *Regularization of differential -algebraic equations revisited*, Preprint Nr. 92-19, Humboldt- Univ. Berlin, Fachbereich Mathematik, 1992.

[60] E. J. Haug, *Computer-Aided Kinematics and Dynamics of Mechanical Systems* Volume I: Basic Methods, Allyn and Bacon, 1989.

[61] P.W. Hemker, A Numerical Study of Stiff Two-point Boundary Problems, Mathematical Center Tracts 80, Amsterdam 1977.

[62] J.G. Heywood, *The Navier-Stokes equations: On the existence, regularity and decay of solutions*, Indiana Univ. Math. J., 29 (1980), 639-681.

[63] J.G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier-Stokes problem, I. Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal. 19(1982), 275-311.

[64] J.G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier-Stokes problem, IV: Error analysis for second-order time discretization*, SIAM J. Numer. Anal. 27 (1990), 353-384.

[65] C. Hirsch, *A general analysis of two-dimensional convection schemes*, VKI Lecture Series 1991-02 on Computational Fluid Dynamics, Von Karman Institute, Brussels, Belgium, 1991.

[66] T.Y. Hou and B.T.R. Wetton, *Convergence of a finite difference scheme for the Navier-Stokes equations using vorticity boundary conditions*, SIAM J. Numer. Anal., 29 (1992), 615-639.

[67] M.K. Kadalbajoo and Y.N. Reddy, *Initial-value technique for a class of nonlinear singular perturbation problems*, J. Optim. Theory and Appls. 53(1987), 395-406.

[68] L. Kalachev and R. O'Malley, Jr., *The regularization of linear differential-algebraic equations*, SIAM J. Math. Anal., 1996, to appear.

[69] L. Kalachev and R. O'Malley, Jr., *Regularization of nonlinear differential-algebraic equations*, SIAM J. Math. Anal. 25 (1994), 615-629.

[70] L. Kalachev and R. O'Malley, Jr., *Boundary value problems for differential-algebraic equations*, Num. Func. Anal. Appl. 16 (1995), 363-378.

[71] M. Knorrenschild, *Differential/algebraic equations as stiff ordinary differential equations*, SIAM J. Numer. Anal. 29 (1992), no. 6, 1694-1715.

[72] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Pure and Applied Mathematics, Vol. 136, Academic Press, 1989.

[73] J. Laforgue and R.E. O'Malley, *On the motion of viscous shocks and the supersensitivity of their steady-state limits*, Methods and Applications of Analysis 2(1994), 465-487.

[74] J.Y. Lee and M.J. Ward, *On the asymptotic and numerical analyses of exponentially ill-conditioned singularly perturbed boundary value problems*, Studies in Applied Math. 94 (1995), 271-326.

[75] M. Lees, *A priori estimates for the solutions of difference approximations to parabolic partial differential equations*, Duke Math J. 27 (1960), 297-311.

[76] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1990.

[77] P. Lin, *Numerical solutions of some singular perturbation problems*, M. Sc. thesis, Nanjing University, Nanjing, China, 1987. (in Chinese)

[78] P. Lin, *Remarks on a mildly nonlinear turning point problem*, Applied Math. Letters 6(1993), 29-32.

[79] P. Lin, *A numerical method for quasilinear singular perturbation problems with turning points*, Computing 46(1991),155-164.

[80] P. Lin, *A sequential regularization method for time-dependent incompressible Navier-Stokes equations*, SIAM J. Numer. Anal., to appear.

[81] P. Lin and Y. Su, *Numerical solution of quasilinear singularly perturbed ordinary differential equations without turning points*, Appl. Math. Mech. (English Ed.) 10(1989).

[82] P. Lin and D. Q. Yang, *An iterative perturbation method for the pressure equation in the simulation of miscible displacement in porous media*, SIAM J. Sci. Comput., submitted.

[83] J. Lorenz, *Combinations of initial and boundary value methods for a class of singular perturbation problems*, In: P.W. Hemker and J.J.H. Miller (eds.), Numerical Analysis of Singular Perturbation Problems, London: Academic Press, 1979, 295-315.

[84] P. Lotstedt, *On a penalty function method for the simulation of mechanical systems subject to constraints*, Royal Inst. of Tech. TRITA-NA-7919, Stockholm, 1979.

[85] R. März, *On tractability with index 2*, Preprint 109, Humboldt-Univ. Berlin, Sektion Mathematik, 1986.

[86] R. März, *Numerical methods for differential- algebraic equations, Part I: Characterizing DAEs*, Preprint Nr. 91-32/I, Humboldt-Universität zu Berlin, 1991.

[87] J. Necas, *Les methodes directes en theorie des equations elliptiques*, Academia, Prague, 1967.

[88] K. Niijima, *On the behavior of solutions of a singularly perturbed boundary value problem with a turning point*, SIAM J. Math. Anal. 9(1978), 298-311.

[89] R. O'Malley, Jr., *Singular Perturbation Methods for Ordinary Differential Equations*, Springer, New York, 1991.

[90] E. O'Riordan and M. Stynes, *A globally uniformly convergent finite element method for a singularly perturbed elliptic problem in two dimensions*, Math. Comp. 57(1991), 47-62.

[91] K. Park and J. Chiou, *Stabilization of computational procedures for constrained dynamical systems*, J. of Guidance 11 (1988), 365-370.

[92] L.R. Petzold, *A brief history of numerical methods for differential-algebraic equations*, Lecture notes in the meeting of the 50th anniversary of the journal " Mathematics of Computations", 1993.

[93] L.R. Petzold and P. Lötstedt, *Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications*, SIAM J. Sci. Stat. Comput. 7(1986), 720-733.

[94] L.R. Petzold, Yuhe Ren and T. Maly, *Numerical solution of differential-algebraic equations with ill-conditioned constraints*, Technical Report 93-59, University of Minnesota, 1993

[95] P. Rabier and W. Rheinboldt, *On the computation of impasse points of quasilinear differential algebraic equations*, Math. Comp. 62 (1994), no. 205, 133-154.

[96] R. Rannacher, *On Chorin's projection for the incompressible Navier-Stokes equations*, In: Rautmann, et al. (eds.): Navier-Stokes equations: Theory and numerical methods. Proc. Oberwolfach Conf., 19.-23.8.1991, Springer, Heidelberg 1992.

[97] R. Rannacher, *On the numerical solution of the incompressible Navier-Stokes equations*, Z. angew. Math. Mech. 73(1993), 203-216.

[98] T. F. Russell, *Finite elements with characteristic finite element method for a miscible displacement problem*, SPE 10500, Proc. 6th SPE Symposium on Reservoir Simulation, Dallas, 1982.

[99] T. F. Russell, *Time stepping along characteristics with incomplete iteration for a Galerkin approximation of miscible displacement in porous media*, SIAM J. Numer. Anal. 22 (1985), pp. 970-1003.

[100] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in: I. Galligani and E. Magenes, Eds., Mathematical Aspects of the Finite Element Method, Lecture Notes in Mathematics 606, Springer-Verlag, Berlin and New York, 1977, pp. 292-315.

[101] J. Shen, *On error estimates of projection methods for the Navier-Stokes equations: First order schemes*, SIAM J. Numer. Anal. 29 (1992), 57-77.

[102] D. Sidilkover and U.M. Ascher, *A multigrid solver for the steady state Navier-Stokes equations using the pressure-Poisson formulation*, Technical Report 94-3, Dept. of Computer Science, University of British Columbia, 1994.

[103] R.F. Sincovec, A.M. Erisman, E.L. Yip and M.A. Epton, *Analysis of descriptor systems using numerical algorithms*, IEEE Trans. Aut. Control, AC-26(1981), 139-147.

[104] D.R. Smith and J.T. Palmer, *On the behavior of the solution of the telegraphist's equation for large absorption*, Arch. Rat. Mech. Anal. 39(1970), 146-157.

[105] Y. Su and P. Lin, *Numerical solution of singular perturbation problem for hyperbolic differential equation with characteristic boundaries*, Proceedings of the BAIL V Conference, Shanghai, Boole Press, Dublin(1988), 20-24.

[106] Y. Su and P. Lin, *Uniform difference scheme for a singularly perturbed linear 2nd order hyperbolic problem with zero-th order reduced equation*, Applied Math. Mech. 11 (English Ed.)(1990), No. 4.

[107] Y. Su and P. Lin, *An exponentially fitted difference scheme for the hyperbolic-hyperbolic singularly perturbed initial-boundary problem*, Applied Math. Mech. 12 (English Ed.)(1991), 237-245.

[108] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1977.

[109] A.N. Tikhonov and V.Ya. Arsenin, *Methods for Solving Ill-posed Problems*, Nauka, Moscow, 1979.

[110] S. Turek, *Tools for simulating nonstationary incompressible flow via discretely divergence-free finite element models*, Preprint, Universität Heidelberg, May 1992.

[111] A. Vasil'eva and V. Butuzov, *Asymptotic Expansion of Solutions of Singularly Perturbed Equations*, Nauka, Moscow, 1973.

[112] A. Vasil'eva and V. Butuzov, *Singularly Perturbed Equations in the Critical Case*, MRC Technical Summary Report # 2039, 1980.

[113] R. Vulanović, *A uniform numerical method for quasilinear singular perturbation problems without turning points*, Computing 41(1989), 97-106.

[114] R. Vulanović, *On numerical solution of a mildly nonlinear turning point problem*, RAIRO Math. Model. Numer. Anal. 24(1990), 765-784 .

[115] R. Vulanović and P. Lin, *Numerical solution of quasilinear attractive turning point problems*, Computers Math. Applic. 23(1992), 75-82.

[116] M.W. Walker and D.E. Orin, *Efficient dynamic computer simulation of robotic mechanisms*, Journal of Dynamic Systems, Measurement, and Control 104 (1982), 205-211.

[117] B. Wetton, *Error analysis for Chorin's original fully discrete projection method and regularizations in space and time*, Preprint, Institute of Applied Math, University of British Columbia, 1995.

[118] G.B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

[119] W. Xie, *A sharp pointwise bound for functions with L^2-Laplacians and zero boundary values on arbitrary three-dimensional domains*, Indiana University Math. J. 40 (1991), 1185-1192.

[120] D. Q. Yang, *Mixed methods with dynamic finite element spaces for miscible displacement in porous media*, J. Comput. Appl. Math. 30(1990), pp. 313-328.

[121] I. Yavneh, *Multigrid techniques for incompressible flows*, Ph. D. thesis, The Weizmann Institute of Science, Israel, 1991.