"Babeş-Bolyai" University Cluj Napoca
Faculty of Mathematics and Informatics

REGULARIZATION METHODS IN THE NUMERICAL ANALYSIS OF SOME DEGENERATE PARABOLIC EQUATIONS

doctoral thesis

Advisors
Prof. Dr. Gheorghe Coman
Prof. Dr. Willi Jäger

Candidate
Iuliu Sorin Pop

1998

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Degenerate parabolic equations appeared in the literature as mathematical models for several phenomena in physics, chemistry, biology or economy. The simplest example in this sense is the porous medium equation describing the flow of an ideal gas in a homogeneous porous medium, where the diffusion is a power-like function. More complex situation arise in petroleum reservoir and groundwater aquifer simulations. In this case convective or source terms appear naturally. Phase change problems corresponding to processes of heat transfer involving melting or solidification lead to equations of the same type, which are more complicated when the liquid phase is allowed to move and an internal generation or absorption of heat is present.

Thinking of the classical heat equation, if some singularities appear in the initial data, these are smoothed as the solution evolves in time. This is due to the infinite speed of propagation, which tells that any local perturbation of the solution will influence it in any of the subsequent moments and in the whole domain. The situation is fairly the same if nonlinear but regular parabolic equations are considered. This alleviates a numerical approximation of the solution of such kind of problems.

The above frame changes when the degenerate case is considered. Roughly speaking, this means the equation may change its parabolic character, for example, into an elliptic or even hyperbolic one. This phenomenon appears in the points where diffusion vanishes and may be influenced by the properties of the solution and consequently the equation behaves like a regular parabolic one in some subdomains but manifests other characters outside of them. The interfaces separating the domains of regularity - sometimes called also free boundaries - have finite speed of propagation, unlike the (nonlinear) heat equation. Moreover, these boundaries are not generally known in advance and have to be determined together with the solution. Often, the interfaces cannot be described by smooth curves (or surfaces) since cusps or indefinite (mushy) regions may appear.

Resulting from the features mentioned above, the solutions lacks regularity across the interface. The singularities do not smooth out as time evolves and, in fact, they may even develop, giving the problem a strongly nonlinear character. This fact is mirrored also in the numerical approximation of the solution pointing out the necessity of adequate algorithms being able to deal both with the free boundary and the singularities of the solution.

1.1 Overview

Here we deal with a fairly small class of free and moving boundary problems having a special unifying mathematical structure. This results from the equivalence of their classical formulations, as differential equations with suitable explicit conditions at the prescribed and unknown boundaries, to a variational formulation. The interface is hidden in the nonlinear diffusion, so the problem is brought to a fixed domain formulation. In fact, the problems considered in the subsequent chapters have the following form

Problem P:

$$\begin{aligned} \partial_t u - \nabla \cdot (\nabla \beta(u) + F(u)) = r(u), &\qquad \text{in } Q_T \equiv (0,T) \times \Omega, \\ u(0,x) = u_0(x), &\qquad \text{in } \Omega, \\ u = u_D, &\qquad \text{on } \partial\Omega, \end{aligned}$$

where $\beta : \mathbb{R} \to \mathbb{R}$ is an increasing function, Ω a bounded domain in $\mathbb{R}^d (d \geq 1)$ with a Lipschitz continuous boundary and T a finite and fixed time. A simple calculation shows that $\beta'(u)$ stands for the diffusion coefficient, which is nonnegative because of the assumptions on β. However, $\beta'(u)$ may also vanish for some values of u - the slow diffusion phenomenon, giving the above equation a degenerate hyperbolic character. Similarly, if β' is unbounded in some points, the problem becomes elliptic there. This can be better noticed if the new unknown $\theta \equiv \beta(u)$ is considered, so the time derivative of u becomes $(\beta^{-1})'(\theta)\partial_t\theta$. Thus, the values of u for which β' is infinite turn to points where the time derivative vanishes - the fast diffusion case, so the transformed equation becomes elliptic.

The analysis of existence and - eventually - uniqueness of solutions can be done avoiding an interference of the unknown free boundary. Because of the strongly nonlinear character of the problem, a solution in the classical sense can be seldom found and therefore it is understood rather in a weak sense. Several works are devoted to the problem of existence, uniqueness or regularity of solutions. Among them we mention [75], [18], [36], [1], [5], [52], [25], [97], [76], [45] or [48]. In the beginning, the interest was focussed especially on model problems without convection or reaction and in this case not only existence, but also uniqueness of solutions has been proven. However, many of the problems in the degenerate parabolic class are modelling phenomena where reaction or convection is present and therefore these terms have been taken into consideration more often in the last decade. Then the parabolic problem could change to a hyperbolic one and uniqueness may even not hold true. However, this is not the point of interest of the (numerical) analysis in the forthcoming chapters and therefore we always assume the problem has a unique solution.

Explicit analytical solutions for degenerate parabolic equations have been found only in some particular cases. Therefore, the numerical approximation of solutions is extremely important. But the algorithms considered for this purpose should take into account the features mentioned before, particularly the singularities of the solution along the free boundaries. Because the difficulties appear especially near the free boundaries, one possibility is given by the algorithms trying to approximate first these interfaces and solving then numerically the more regular problems in the regions of non-degeneracy (as done, e.g., in [26]). The tracking procedure is acceptable if the interfaces are smooth hypersurfaces, but even then the problem complicates in two or three spatial dimensions. A similar idea was considered in [68], [69] and [70], for predicting a region including the free boundaries. The spatial discretization is strongly refined in that part of the domain, resulting in an adaptive algorithm.

The fixed domain formulation suggests some procedures for obtaining a numerical approximation of the solution in the whole domain, while the interface comes out implicitly. Depending on the way the lack of regularity is treated, the resulting algorithms can be classified into non-regularization and regularization ones. In the first case the solution of the original problem is approximated directly at semi-discrete times, without performing any modification in the equation. In this setting, the theory of nonlinear semigroups of contractions in Banach spaces ([16]) leads not only to existence, uniqueness and global regularity results, but is also the starting point for semi-discrete numerical methods. A direct application of the definition of the semigroup is the Crandall-Ligget method ([23]) corresponding to the simple backward difference method, having the $O(\tau^{1/2})$ convergence order. If $\tau = T/n$ stands for the time step, the resulting algorithm for the above problem reads

$$\begin{aligned} u^k - u^{k-1} = &\ \tau\nabla\cdot(\nabla\beta(u^k) + F(u^k)) + \tau r(u^k), \\ u^k|_{\partial\Omega} = &\ u_D(k\tau) \end{aligned}$$

for $k = \overline{1,n}$, together with the initial data given above. Here u^k approximates the solution at the time $t_k = k\tau$. The framework is maintained in [29], where the analysis based on Green operators is extended to a fully discrete scheme. An improved convergence order ($O(\tau)$) was proven recently in [85] for the special case of a semigroup generated by a subgradient in a Hilbert space. In connection with a Galerkin finite element spatial discretization, the analysis was extended afterwards to a fully discrete scheme in [86]. Algorithms related to the backward Euler method are considered also in [68], [69] (in connection with a two dimensional adaptive spatial discretization), [4] (for a Raviart-Thomas mixed finite element method), [37] (the resulting nonlinear problems being solved with preconditioned Newton methods), [28] (where a regularization step is hidden in the separate treatment of the degeneracy regions), [35] (for a fully discrete finite volume method), [71] or [72] (together with apriori error estimates in an adaptive approach). The resulting schemes are nonlinear. Explicit methods have been proposed in [22] and [97], while a linear semi-implicit scheme is analysed in [2].

The schemes in the second category include also a regularization step. In this case the properties of the solution of the problem written above are taken into account. Hence, since u itself is less regular than $\beta(u)$, it is more convenient to consider $\theta \equiv \beta(u)$ as the main unknown. But the degenerate problem cannot be written in terms of θ because this involves also derivatives for $u \equiv \beta^{-1}(\theta)$, which may explode in the degeneracy points. Therefore, a natural remedy to this problem is the regularization, which can be done either by controlling the derivatives of β for getting strictly positive lower bounds and finite upper ones, or by approximating the initial degenerate problem through a non-degenerate one. The algorithms in the first group are akin to the nonlinear Chernoff formula ([17]). This gives rise to some (linear) relaxation schemes ([14]), among which the simplest one is analysed in [60]

$$\begin{aligned} \theta^k - \beta(u^{k-1}) = &\ \frac{\tau}{\mu}\nabla\cdot(\nabla\theta^k + F(u^{k-1})) + \frac{\tau}{\mu}r(u^{k-1}), \\ \theta^k|_{\partial\Omega} = &\ \beta(u_D(k\tau)), \\ u^k = &\ u^{k-1} + \mu(\theta^k - \beta(u^{k-1})), \end{aligned}$$

where the relaxation parameter μ satisfies some stability conditions. Fully discrete counterparts are considered in [74] and [70]. Since the relaxation parameter does not depend on the variables t or x, the scheme above is linear but the accuracy is altered especially around the free boundaries. This drawback is avoided through nonlinear versions of the scheme proposed in [46], [47] or [54]. Even though these schemes can be applied to a larger class of problems, we give here only a simplified version adapted to Problem P,

Scheme JK:

$$\begin{aligned} \lambda_k(\theta^k - \beta(u^{k-1})) =& \tau\nabla\cdot(\nabla\theta^k + F(u^{k-1})) + \tau r(u^{k-1}), \\ \theta^k|_{\partial\Omega} =& \beta(u_D(k\tau)), \end{aligned} \tag{1.1}$$

where $\lambda_k \in L^\infty(\Omega)$ satisfies the convergence condition

$$\frac{\alpha}{L_\beta} \leq \lambda_k \leq \min\left\{K, \frac{\beta^{-1}(\beta(u^{k-1}) + \alpha(\theta^k - \beta(u^{k-1}))) - u^{k-1}}{\theta^k - \beta(u^{k-1})}\right\}. \tag{1.2}$$

Here $K, \alpha \in (0,1)$ are parameters of the method, while L_β stands for the Lipschitz constant of β. Now $u(k\tau)$ is approximated through the relaxation step

$$u_k = u^{k-1} + \lambda_k(\theta^k - \beta(u^{k-1})). \tag{1.3}$$

For solving the resulting nonlinear (regular) elliptic problems in (1.1), an iterative procedure is considered in [47] and [54]. Taking initially $\lambda_{k,0} = \min\left\{K, \dfrac{\alpha}{\beta'(u^{k-1})}\right\}$ in (1.1), the first approximation $\theta^{k,1}$ for θ^k can be obtained solving the resulting linear problem. Applying this solution in (1.2) yields a further approximation $\lambda_{k,1}$ and the procedure can be continued.

Another regularization approach perturbs the degenerate problem in order to obtain a regular one. Any discretization method appropriate for parabolic problems may be applied then, but the analysis should rely only on those estimates for the solution of the modified problem which depend regularly on the perturbation parameter. Alike the methods related to the nonlinear Chernoff formula, if the resulting elliptic problems are nonlinear, these can be solved efficiently due to the regularization step. Moreover, linear semi-implicit schemes are also available in this framework ([67], [88] or [79]). Most of the numerical algorithms in this group start with a regularization obtained through the perturbation of the nonlinearity β or its inverse. This idea was first used in the numerical analysis in [50], [84] and [32]. A linear scheme based on the approximation of β was proposed in [88]. Related to this, the scheme in [33] combines the regularization procedure with a split of the operator into its hyperbolic part and the nonlinear diffusion.

A similar approach consists in the modification of the inverse of the nonlinearity function (β^{-1}), together with the transformation of the equation in terms of the more regular unknown, $\theta \equiv \beta(u)$. Numerical schemes using this idea are analysed in [63], [66] and [73]. Still in this frame, the particular relation between the enthalpy and temperature in the Stefan problem was exploited in [67] for proving the convergence of a linear scheme. A similar nonlinear approach is considered in [53] for approximating more complex problems including both slow and fast diffusion.

A different possibility to approximate the degenerate equation by regular ones is offered by the maximum principle. Applied for the analysis of the porous medium equation for the first time in [75], this technique may be useful for obtaining numerical schemes for some classes of problems ([98], [79]).

Hereafter we use the standard notation (see, e.g. [57], [59] or [99]). In particular, $L^2(\Omega)$ contains the functions which are square integrable w.r.t the Lebesgue integral on the domain Ω, while $H^1(\Omega)$ requests the same also for the (generalized) derivatives of first order. $H_0^1(\Omega)$ is a subset of $H^1(\Omega)$ whose elements have zero boundary values (in the sense of traces). If g lies in $H^1(\Omega)$, $g + H_0^1(\Omega)$ are the elements u of $H^1(\Omega)$ for which $u - g$ belongs to $H_0^1(\Omega)$. The dual of $H_0^1(\Omega)$ is denoted by $H^{-1}(\Omega)$. In case of time dependent functions, the elements of $L^2(I; X)$ are functions with values in the normed space X which are square integrable (in the sense of Bochner) on the time interval I, while for $H^1(I; X)$ the same holds for the derivatives of first order with respect to t. We let $(\cdot, \cdot)$ stand for the inner product on $L^2(\Omega)$, or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, $\|\cdot\|$ for the norm in $L^2(\Omega)$. In case these notations are ambiguous, $\|\cdot\|_{k,p,\Omega}$ denotes the norm in $H^{k,p}(\Omega)$ (the functions having the derivatives up to the order k in the space $L^p(\Omega)$), or $\|\cdot\|_{H^k(I;X)}$ gives the norm in $H^k(I; X)$. In addition, we often write u or $u(t)$ instead of $u(t, x)$ and use C to denote a generic positive constant independent of the discretization parameters (h and τ) and ε, a small parameter involved in the regularization process.

1.2 Results

The aim of this work is an investigation of some numerical methods for degenerate parabolic equations. All the schemes considered here rely on regularization techniques and solve the equation in the better unknown $\theta \equiv \beta(u)$. The time discretization is done by one-step methods, higher order ones not being justified theoretically because of the properties of the solution. For the spatial discretization an upwind box scheme is considered. The usual setting in previous papers dealing with the numerical analysis of such kind of problems (a Lipschitz continuity with respect to $\beta(u)$ of the convection and reaction - as requested, for example, in [60], [66], [73], [74], [67], [88], [89], [4], [51]) is extended here to be alike the one in [97] without affecting the convergence results. Concretely, the following relation should be fulfilled for any reals u and v

$$|r(u) - r(v)|^2 + |F(u) - F(v)|^2 \leq C(u - v)(\beta(u) - \beta(v)).$$

In the following chapter maximum principle based numerical schemes are investigated. As mentioned before, even though this regularization technique was applied long time ago in the analysis of some degenerate parabolic equations, we are not aware of some numerical methods exploiting this possibility. Based on the the idea presented in [98], here we extend the numerical analysis for a class of problems which include convection and reaction terms, in the setting mentioned above (see also [77]). Three semi-discrete numerical schemes are considered (an implicit one, a semi-linear and a semi-implicit linear one), and they are justified by the maximum principle proven there. For solving the resulting nonlinear elliptic problems iterative methods are studied, and these are useful also in showing existence and uniqueness of the semi-discrete solutions. Applying the techniques proposed in [63] and [29], error estimates are obtained for showing the convergence of the methods. In the general framework, these are at least as good as the ones for other schemes. In particular cases, based on the recent result in [85], the estimates can be proven to be optimal.

The above analysis is extended afterwards to the fully discrete case. In order to maintain the maximum principle at this level, stable spatial discretization methods have to be considered. The upwind box scheme proposed in the last part of the chapter respects the philosophy of the upwind technique for finite differences. Since these methods are of lower order, the upwind approach is combined with the classical box scheme for gaining in accuracy in the region not dominated by convection, but without loosing the stability. However, the theoretical convergence of this procedure still has to be proven.

The setting described before is maintained in the third chapter. Here a regularization technique relying on the perturbation of the nonlinearity β is considered for generating numerical schemes alike the ones studied in the previous chapter. The resulting schemes are akin to the ones appearing in the references mentioned before, in a slightly restrictive framework. The maximum principle is not essential anymore, but makes the analysis of the schemes easier. The error estimates are similar to those obtained for the maximum principle based schemes, including the optimal results. The complete discretization is performed in the same manner as before and the error analysis is satisfactory, improving - for example - the results obtained for the linear scheme in [89].

Finally some numerical examples are presented, including model problems describing the gas diffusion in porous media or phase transfers (melting and solidification). Moreover, the algorithms are tested also on a diffusion-transport process in a heterogeneous unsaturated porous medium. The theoretical results are confirmed by the examples.

At this point, I would like to express my deep gratitude to Prof. Willi Jäger for guiding me in this interesting topic and giving me the possibility to work at the Interdisciplinary Center for Scientific Computing of the University of Heidelberg. This work was done during my stay in Heidelberg, supported by the Deutsche Forschungsgemeinschaft through the Sonderforschungsbereich 359 and partially by the Deutscher Akademischer Austauschdienst. The research group met there helped me a lot for solving many of the problems which appeared during my studies. I am grateful to the members of the Department of Numerical Analysis at the "Babeş-Bolyai" University and especially to Prof. Gheorghe Coman for the permanent support and encouragement during my stay in Heidelberg. Many of the ideas comprised here are owing to the collaboration with Dr. Wen-An Yong. I am thankful to Dr. Nicolas Neuss for many helpful discussions and observations and his kind assistance with the software *UG*, also to Mr. Stefan Schnadt for his remarks. The suggestions of Prof. J. Kačur and Dr. M. Slodička were useful for clarifying some aspects concerning this work. The images obtained in the three dimensional simulations have been produced by Mr. Cătălin Dârţu with his volume rendering visualization program. My parents have encouraged and stimulated me permanently to do this work and I want to thank them for this. Last, but not least, nothing would have been possible without abusing of the patience and understanding of my beloved wife Valeria.

Chapter 2

A maximum principle based approach

This chapter contains the presentation of a maximum principle based numerical approach to a certain class of degenerate parabolic equations. It is based on the maximum principle, since in the solution of the resulting problem stays away from the values where degeneracy takes place. This represents an alternative to the usual regularization methods, consisting in the modification of the nonlinearity for avoiding a vanishing diffusion. After performing a shift of the data, both linear and nonlinear discretization schemes are proposed and analysed. The resulting error estimates are optimal.

2.1 Basic setting

Let Ω be a bounded domain in $\mathbb{R}^d (d \geq 1)$ with a Lipschitz continuous boundary and $Q_T \equiv (0,T) \times \Omega$ with $0 < T < \infty$ is fixed. We deal here with the following nonlinear degenerate parabolic problem

Problem P:

$$\begin{aligned} \partial_t u - \nabla \cdot (\nabla \beta(u) + F(u)) = r(u), & \qquad \text{in } \ Q_T \equiv (0,T) \times \Omega, \\ u(0,x) = u_0(x) \geq 0, & \qquad \text{in } \ \Omega, \\ u = 0, & \qquad \text{on } \ \partial\Omega, \end{aligned} \tag{2.1}$$

$\beta : \mathbb{R} \to \mathbb{R}$ being a strictly increasing smooth function. By degeneracy we mean a vanishing diffusion, namely $\beta'(u) = 0$ for some u. Throughout this chapter we assume, without loss of generality, that $\beta(0) = 0$. Moreover, a unique degeneracy point is allowed.

The maximum principle based approach relies on the perturbation of the data, which allows us to avoid the difficulties due to the degeneracy. This technique, introduced in [75] for the classical porous medium equation, has been used for the analysis of more general scalar equations (see, for example, the recent works [48] or [45]), but, to our knowledge, numerical algorithms relying on it have not been analysed so far.

More precisely, the degeneracy is assumed to appear in 0. As shown in the papers cited above, if positive solutions are sought, the degeneracy in the equation can be overcomed by a (local) perturbation of the initial and boundary data - namely a shift with a small parameter. In this way, the resulting data are away from the degeneracy value. The maximum principle obeyed by such equations guarantees that the resulting problem has a solution taking values away from the degenerate point. Thus, the initial degenerate problem is reduced to a regular parabolic one, which can be solved numerically by different methods. Based on the idea presented in [98], this chapter enlarges the class of problems taken into consideration there and in [79].

Throughout this chapter some assumptions on β are necessary. These, together with the ones on F, r and u_0 are given below.

(A1) β is Lipschitz and differentiable, $\beta(0) = 0$, $\beta'(u) \geq 0$ and β' may vanish at a single point at most (which is assumed to be 0).

(A2) $u_0 \geq 0, \quad u_0 \in L^\infty(\Omega)$ and $\beta(u_0) \in H_0^1(\Omega)$.

(A3) $r : \mathbb{R} \to \mathbb{R}$ and $F : \mathbb{R} \to \mathbb{R}^d$ are continuous in u and satisfy the condition

$$|r(u) - r(v)|^2 + |F(u) - F(v)|^2 \leq C(u - v)(\beta(u) - \beta(v))$$

for any $u, v \in \mathbb{R}$, where $C > 0$ does not depend on x, t, u and v. Moreover, it is assumed here that r is positive for all positive arguments and the graphs of both functions contain the origin, hence $r(0) = 0$ and $F(0) = \bar{0}$.

In some situations it will be necessary to give the constants appearing in the assumptions (A1) - (A3) explicitly. The Lipschitz constant of both β and β' (where this will be needed) is denoted by L_β, while the growth of F and r is controlled by C_F, respectively C_r.

The last part of the assumption (A2) is not essential in the analysis of the implicit scheme, but is requested for proving the stability of the other two schemes. In some cases the Lipschitz continuity of β' is requested supplementary. Moreover, if the convective term is linearized, the derivability of any of the components $F_i, i = \overline{1, d}$ of F is needed. Then, the assumption (A3) ensures that the derivatives - denoted by f_i, with $f = (f_1, \ldots, f_d)$ -

are (uniformly) bounded.

Remark 2.1.1 *The solutions we are interested in here are positive. As mentioned before, in this case a maximum principle based regularization can be applied in order to obtain regular problems. In (A2) the initial data are positive and therefore the same property holds true for the solution of Problem P.*

Remark 2.1.2 *In order to simplify the analysis we have not considered the dependence of β, F and r on the variables x and t. This situation can be treated in a similar manner under some additional assumptions (see, e.g., [53]). Moreover, the gradient of $\beta(u)$ in (2.1) may be multiplied by a symmetric and positive definite matrix.*

Remark 2.1.3 *In (A3) we have weakened the usual assumption on F and r, namely the Lipschitz continuity with respect to (w.r.t.) $\beta(u)$*

$$|r(u) - r(v)| + |F(u) - F(v)| \leq C|\beta(u) - \beta(v)|.$$

The weaker hypotheses in (A3) is enough for obtaining the uniqueness of the solution (see [1] or [97]).

Remark 2.1.4 *Non-homogeneous Dirichlet or natural boundary conditions fit into the framework here only if they provide a positive solution. But this restriction is fulfilled in most of the cases of practical interest.*

The assumption in (A1) implies that β has an inverse β^{-1}, which is continuously differentiable everywhere excluding 0. Then there are two positive constants such that

$$0 < C_1(\varepsilon) \leq (\beta^{-1})'(\theta) \leq C_2(\varepsilon) < \infty \tag{2.2}$$

for any real θ in $[\beta(\varepsilon), \beta\left(Me^{CT}\right)]$, if this interval is included in the range of β. Here M is a large constant satisfying $\theta^0 \leq \beta(M) - \beta(\varepsilon)$ almost everywhere (a.e.) and C is greater than the constant in (A3). Moreover, $C_1(\varepsilon)$ often depends on ε in a regular way. To see this, we refer to the typical porous medium case, where $\beta(u) = u^m (m \geq 1)$. This kind of dependence holds true if β is Lipschitz continuous at least on bounded intervals. So we will use C_1 to denote $C_1(\varepsilon)$ and this is assumed in what follows.

Since the problem in (2.1) is degenerate, the concept of solution should be understood in a weaker sense, one possibility being given below.

Problem WP. u is called a solution of the problem in (2.1) iff

$$u \in H^1(0,T;H^{-1}(\Omega)), \quad \beta(u) \in L^2(0,T;H_0^1(\Omega)), \quad u(0) = u_0 \text{ (in } H^{-1})$$

and for all $\varphi \in L^2(0,T;H_0^1(\Omega))$ the equation holds true

$$\int_0^T (\partial_t u(t), \varphi(t))\, dt + \int_0^T (\nabla\beta(u(t)) + F(u(t)), \nabla\varphi(t))dt = \int_0^T (r(u(t)), \varphi(t))dt. \quad (2.3)$$

Existence, uniqueness and boundedness of the solution for the above problem is studied in several papers (see [1], [25], [49], [52], [76] and the references therein). Since the regularization approach considered here makes use of the maximum principle, we assume that the solution is uniformly bounded (a.e.) in the whole cylinder Q_T. This implies better regularity properties for u. Particularly, we get $u \in C(0,T;H^{-1}(\Omega))$, thus Problem WP can be reformulated in a stronger sense than it was considered in [1] and the initial condition holds in $H^{-1}(\Omega)$.

2.2 Time discretization

The main goal here is the investigation of some approximation schemes for Problem P. Because $\beta(u)$ is more regular then u itself, it is more convenient to consider the equation in the unknown $\theta = \beta(u)$ and discretize it correspondingly. This approach was adopted in several papers (see, for example [60], [46], [47], [53], [70]) and generates effective numerical schemes which can be analysed easier.

The time discretization is based essentially on the first order Euler method. A higher order method is effective only if the solution is sufficiently regular, which - in general - is not the case. Moreover, the time step is kept fixed for the sake of simplicity, but the analysis can be performed in the same manner also for non-constant steps. This last case becomes interesting especially in connection with an aposteriori error control (as done, e. g., in [8], [71], [72]). If n is an integer and $\tau = T/n$ the time step, the Euler implicit scheme can be written formally as

Scheme MTI:

$$\begin{aligned} \beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}) &= \tau\nabla \cdot (\nabla\theta^k + F(\beta^{-1}(\theta^k)) + \tau r(\beta^{-1}(\theta^k)), \\ \theta^k|_{\partial\Omega} &= \beta(\varepsilon) \end{aligned} \quad (2.1)$$

for $k = \overline{1,n}$ with $\theta^0 = \beta(u_0) + \beta(\varepsilon)v_0$, where ε is an artificial positive small number and $v_0 \in H^1(\Omega)$ an artificial function. Here θ^k approximates $\theta(t_k) = \beta(u(t_k))$ $(t_k = k\tau)$ and therefore $u(t_k)$ is approximated by $\beta^{-1}(\theta^k)$.

The choice of v_0 depends on the given initial data $u_0(x)$, so that

$$v_0|_{\{u_0 \geq \varepsilon\}} = 0, \;\; v_0 \leq 1 \text{ and } \beta(u_0) + \beta(\varepsilon) v_0 \geq \beta(\varepsilon). \tag{2.2}$$

Here the inequalities hold almost everywhere. A typical example is

$$v_0 := \left[1 - \frac{\beta(u_0)}{\beta(\varepsilon)} \right]_+ , \tag{2.3}$$

where $[x]_+ = x$ if $x \geq 0$, otherwise $[x]_+ = 0$. Here the initial data are perturbed only locally, improving by this the efficiency of the method. However, all the results remain valid also for a global perturbation (for example, when $v_0 \equiv 1$).

Scheme MTI is nonlinear. It can be simplified successively up to a linear version. First, $(\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1})) \cdot \nabla \theta^k$ can replace the convective term in (2.1), where f is the derivative of F. This makes sense only in the connection with the maximum principle which, under the assumptions made below, guarantees that θ^k stays above $\beta(\varepsilon)$ for any k, so $(\beta^{-1})'(\theta^k)$ does not explode. An explicit treatment of this part cannot be accepted because we want to maintain the bound from below. Thus we are led to the scheme

Scheme MTC:

$$\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}) = \tau \Delta \theta^k + \tau (\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1})) \cdot \nabla \theta^k + \tau r(\beta^{-1}(\theta^k)), \tag{2.4}$$

together with the same initial and boundary data as before. The reaction term in the semi-discrete equations can be treated in an explicit manner, replacing $r(\beta^{-1}(\theta^k))$ by $r(\beta^{-1}(\theta^{k-1}))$.

Now we can give a linear approximation scheme for Problem P,

Scheme MTL:

$$\begin{aligned} \sigma_{k-1}(\theta^k - \theta^{k-1}) =& \tau \Delta \theta^k + \tau (\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1})) \cdot \nabla \theta^k + \tau r(\beta^{-1}(\theta^{k-1})), \\ \theta^k|_{\partial\Omega} =& \beta(\varepsilon), \\ \sigma_k =& (\beta^{-1})'(\theta^k) \end{aligned} \tag{2.5}$$

for $k = \overline{1, n}$, where $\theta^0 = \beta(u_0) + \beta(\varepsilon) v_0$ and $\sigma_0 = (\beta^{-1})'(\theta^0)$.

Remark 2.2.1 *The linearization of the convection is done by multiplying the gradient of θ with the 'speed' $(\beta^{-1})'(\theta) f(\beta^{-1}(\theta))$, which can be interpreted as the derivative of F w.r.t $\theta \equiv \beta(u)$. This term may generally go to infinity because of the derivative of the inverse of β. In connection with the maximum principle, this situation is avoided due to the lower* *bound for θ. Taking into account the assumption we have made in (A3) on F, a simple calculation shows that for any θ above $\beta(\varepsilon)$ the following holds*

$$|(\beta^{-1})'(\theta)f(\beta^{-1}(\theta))| \leq \sqrt{C_F C_2(\varepsilon)}.$$

If F is Lipschitz continuous w.r.t. $\beta(u)$, $C_2(\varepsilon)$ doesn't appear in the bound above.

It is worth here to mention some aspects concerning the schemes. First, the necessity of the regularization step becomes clear in the simplified versions of the nonlinear scheme. Moreover, in this way only one unknown ($\theta = \beta(u)$) can be considered and for the nonlinear schemes the convergence of some iterative procedures can be studied. For the non-regularized approach u and $\beta(u)$ have to be computed simultaneously, or the nonlinearity is treated on the fully discrete level (see, e.g., [28], [29], [51], [71], [72], [86]).

The regularization technique relies on the shift of the data, therefore the boundary values become nonzero. Thinking at Scheme MTI, in order to turn back to homogeneous Dirichlet boundary conditions, $\psi = \theta - \beta(\varepsilon)$ can be accepted as the unknown function. Then the nonlinearity function β becomes $\beta_\varepsilon(v) = \beta(v + \varepsilon) = \beta(u)$, its derivative being small, but nonzero in the origin. Therefore this approach is not far from the one in the Jäger - Kačur algorithm (Scheme JK, [46], [47]), where a cut-off procedure is applied - if the β' vanishes, it is replaced by a small nonzero value - and convection or reaction are discretized explicitly. There, after computing θ, the value of u is given by the relaxation step in (1.3). For a particular choice of the function λ in (1.2) and taking α equal to 1 there - this value not being allowed theoretically, but practical computations show that the algorithm works good also in this case - the relaxation step becomes

$$u_k = \beta^{-1}(\theta^k).$$

There is still an essential difference between the two approaches, even when the particular form of the Scheme JK is considered. By the shift of the data, a cut-off procedure becomes unnecessary because of the maximum principle. Analogous, Scheme MTL is related to the one proposed in [88], under the same remark.

As in the continuous case, it is necessary to give weak forms of the semi-discrete approximation schemes.

Problem WT. For any $1 \leq k \leq n$, find $\theta^k \in H_0^1(\Omega) + \beta(\varepsilon)$ such that for all $\varphi \in H_0^1(\Omega)$ one of the equations below (each corresponding to one of the schemes mentioned before) hold true.

Problem WMTI.

$$(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \varphi) + \tau(\nabla\theta^k + F(\beta^{-1}(\theta^k)), \nabla\varphi) = \tau(r(\beta^{-1}(\theta^{\underline{k}})), \varphi) \tag{2.6}$$

for Scheme MTI, or

Problem WMTC.

$$\begin{aligned} & (\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \varphi) + \tau(\nabla\theta^k, \nabla\varphi) \\ = \; & \tau((\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1})) \cdot \nabla\theta^k + r(\beta^{-1}(\theta^{\underline{k}})), \varphi) \end{aligned} \tag{2.7}$$

if the convection is linearized, respectively

Problem WMTL.

$$\begin{aligned} & (\sigma_k(\theta^k - \theta^{k-1}), \varphi) + \tau(\nabla\theta^k, \nabla\varphi) \\ = \; & \tau((\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1})) \cdot \nabla\theta^k + r(\beta^{-1}(\theta^{k-1})), \varphi) \end{aligned} \tag{2.8}$$

for the linear scheme.

The initial data - θ^0 - has been already chosen in Scheme MTI, while Scheme MTL contains the definition of σ_k. $\underline{k}$ stands either for k or for $k-1$, depending on the way the reaction term is treated.

2.2.1 The elliptic problems

As stated before, the maximum principle plays a crucial role in the discretization method. For Problem P, this property has been studied in several papers ([75], see also [45], [48] and the references therein). Our task here is to consider the elliptic problems resulting from the time discretization.

For each k less then n we define

$$V_k = \left\{ \varphi \in \beta(\varepsilon) + H_0^1(\Omega) : \beta(\varepsilon) \le \varphi \le \beta\left(Me^{\bar{C}k\tau}\right), \quad a.e. \right\}, \tag{2.9}$$

where $\bar{C} > 0$ will be chosen below. It is clear that for any $k > 0$, V_k includes any of the previous sets. We also note that V_k is convex and closed due to the Riesz theorem.

The maximum principle

The first step in the analysis of the schemes consists in proving that if a sequence of (weak) solutions for any of the problems arising in the semi-discrete approximations exists, it belongs (elementwise) to $\{V_k\}_{k=0}^n$. This will be shown by mathematical induction. The proof is fairly the same for any of the schemes considered below, therefore we consider here only the implicit case. The assumption (A2) guarantees that for M suitably chosen (e.g. $M = \|u_0\|_\infty$), the initial data is in V_0. Now, assuming that θ^{k-1} is given in V_{k-1}, our aim is to show that if a solution θ^k of Problem WMTI exists, then it belongs to V_k.

In order to make the presentation easier, the indices are skipped for the moment. We consider the auxiliary problem

Problem AUX. Find $\theta \in H_0^1(\Omega) + \beta(\varepsilon)$ such that

$$\begin{aligned} & (\beta^{-1}(\theta) - \beta^{-1}(\chi), \varphi) + \tau(\nabla\theta, \nabla\varphi) \\ = \; & \tau((\beta^{-1})'(\psi) f(\beta^{-1}(\psi)) \cdot \nabla\theta + r(\beta^{-1}(\theta)), \varphi) \end{aligned} \tag{2.10}$$

holds for all $\varphi \in H_0^1(\Omega)$.

Here ψ and χ are chosen in the spaces defined above, namely $\psi \in V_k$ and $\chi \in V_{k-1}$ for some $k > 0$ and the solution should belong to V_k. This is shown in the following lemma

Lemma 2.2.1 *Assume (A1), (A3), $\psi \in V_k$, $\chi \in V_{k-1}$ and $r(u) \geq 0$ for all u. Then, if τ is smaller than τ_0 - a constant given in the proof - and a solution of Problem AUX exists, it belongs to V_k.*

Proof. We start with the proof of the lower bound because this is essential in the whole chapter. This is done by reductio ad absurdum, following the ideas used in [39] for the proof of the weak maximum principle. Assume that

$$\inf_{\Omega} \theta < \beta(\varepsilon),$$

(here the infimum should be understood as the essential one).

Let φ be a positive (a.e.) function in $H_0^1(\Omega)$. Since $r \geq 0$ and $\chi \geq \beta(\varepsilon)$, for any $\delta < \varepsilon$, the monotonicity of β leads to the inequality

$$0 \leq (\beta^{-1}(\chi) - \delta, \varphi) + \tau(r(\beta^{-1}(\theta)), \varphi)$$

Thus, if θ solves Problem AUX, the following holds true

$$(\beta^{-1}(\theta) - \delta, \varphi) + \tau(\nabla\theta, \nabla\varphi) \geq \tau((\beta^{-1})'(\psi) f(\beta^{-1}(\psi)) \cdot \nabla\theta, \varphi)$$

for all $\varphi \geq 0$ in $H_0^1(\Omega)$. Taking $\varphi = [\beta(\delta) - \theta]_+ \in H_0^1(\Omega)$ in the above inequality yields

$$\begin{aligned} & \int_{\theta<\beta(\delta)} (\beta^{-1}(\theta) - \delta)(\beta(\delta) - \theta) + \tau \int_{\theta<\beta(\delta)} \nabla\theta \cdot \nabla(\beta(\delta) - \theta) \\ \geq \; & \tau \int_{\theta<\beta(\delta)} (\beta(\delta) - \theta)(\beta^{-1})'(\psi) f(\beta^{-1}(\psi)) \cdot \nabla\theta. \end{aligned}$$

Denoting by Ω_δ the support of $[\beta(\delta)-\theta]_+$ and applying the Cauchy inequality, this relation becomes

$$\begin{aligned}&\int_{\Omega_\delta}(\beta^{-1}(\theta)-\delta)(\theta-\beta(\delta))+\tau\|\nabla\theta\|^2_{0,2,\Omega_\delta}\\ \leq\ &\tau\|\theta-\beta(\delta)\|_{0,2,\Omega_\delta\cap supp\{\nabla\theta\}}\|(\beta^{-1})'(\psi)f(\beta^{-1}(\psi))\cdot\nabla\theta\|_{0,2,\Omega_\delta},\end{aligned}$$

where $supp\{\nabla\theta\}$ is the support of $\nabla\theta$. Since $\psi\geq\beta(\varepsilon)$, the assumptions made on β in (A1), (2.2) and the boundedness of f give

$$\|\nabla\varphi_\delta\|^2_{0,2,\Omega_\delta}\leq C\sqrt{C_2(\varepsilon)}\|\varphi_\delta\|_{0,2,\Omega_\delta\cap supp\{\nabla\theta\}}\|\nabla\varphi_\delta\|_{0,2,\Omega_\delta},$$

where $[\beta(\delta)-\theta]_+$ was replaced by φ_δ. Hence, the following holds true (for any $\delta<\varepsilon$)

$$\|\nabla\varphi_\delta\|_{0,2,\Omega_\delta}\leq C\sqrt{C_2(\varepsilon)}\|\varphi_\delta\|_{0,2,\Omega_\delta\cap supp\{\nabla\theta\}}.$$

Next, the Sobolev embedding theorem is applied. If $d>2$ (where d stands for the dimension of the domain Ω), because φ_δ lies in $H^1_0(\Omega)$, we get

$$\|\varphi_\delta\|_{0,\frac{2d}{d-2},\Omega}\leq C\|\nabla\varphi_\delta\|_{0,2,\Omega}\leq C\sqrt{C_2(\varepsilon)}\|\varphi_\delta\|_{0,2,\Omega_\delta\cap supp\{\nabla\theta\}}.$$

Now, the Hölder inequality for the last term yields

$$\|\varphi_\delta\|_{0,\frac{2d}{d-2},\Omega}\leq C\sqrt{C_2(\varepsilon)}\left(meas\{\Omega_\delta\cap supp\{\nabla\theta\}\}\right)^{\frac{1}{d}}\|\varphi_\delta\|_{0,\frac{2d}{d-2},\Omega},$$

where the inclusion $\Omega_\delta\subset\Omega$ was used. Therefore

$$meas\{\Omega_\delta\cap supp\{\nabla\theta\}\}\geq C(\varepsilon), \tag{2.11}$$

where $C>0$ does not depend on δ.

Analogous, if $d=2$, the same property of the above set can be obtained. This shows that the essential infimum of θ is finite. Moreover, since the constant in (2.11) does not depend on δ, the inequality must hold as δ tends to $\inf\limits_\Omega\theta$. That is, the function θ must attain its infimum in Ω on a set of positive measure, where at the same time its gradient vanishes (since the function is constant almost everywhere there). This contradicts the inequality in (2.11) and therefore the assumption on the infimum of θ is false.

A similar argument shows that θ has a finite essential supremum. The equality in (2.10) can be rewritten as

$$\begin{aligned}&(\beta^{-1}(\theta)-L,\varphi)+\tau(\nabla\theta,\nabla\varphi)\\ =\quad&(\beta^{-1}(\chi)-L,\varphi)+\tau((\beta^{-1})'(\psi)f(\beta^{-1}(\psi))\cdot\nabla\theta+r(\beta^{-1}(\theta)),\varphi),\end{aligned}$$

for any real number L. Assume now that the essential supremum of θ lies above $\beta(Me^{\bar{C}k\tau})$, then choose an arbitrary $L > Me^{\bar{C}k\tau}$ and everything follows as before if the forthcoming holds true

$$(\beta^{-1}(\chi) - L, [\theta - \beta(L)]_+) + \tau(r(\beta^{-1}(\theta)), [\theta - \beta(L)]_+) \leq (\beta^{-1}(\theta) - L, [\theta - \beta(L)]_+).$$

Since $\chi \in V_{k-1}$ and θ is positive (this being proven before), the assumption (A2) on r yields

$$\begin{aligned} & (\beta^{-1}(\chi) - L, [\theta - \beta(L)]_+) + \tau(r(\beta^{-1}(\theta)), [\theta - \beta(L)]_+) \\ \leq \quad & (Me^{\bar{C}(k-1)\tau} - L, [\theta - \beta(L)]_+) + \tau C(\beta^{-1}(\theta), [\theta - \beta(L)]_+), \end{aligned}$$

where C is any constant above $\sqrt{C_r L_\beta}$ which will be fixed below. Now, if $\tau C \leq 1$, the desired inequality is implied by

$$(Me^{\bar{C}(k-1)\tau} - L, [\theta - \beta(L)]_+) + \tau C(L, [\theta - \beta(L)]_+) \leq 0,$$

so that it is enough to take C such that $Me^{\bar{C}(k-1)\tau} + \tau CL \leq L$ for any $L \geq Me^{\bar{C}k\tau}$. Let C be of the form $C = \alpha\bar{C}$, with $\alpha > 0$. Our inequality becomes

$$Me^{\bar{C}(k-1)\tau} \leq L(1 - \alpha\bar{C}\tau)$$

and it suffices if this is fulfilled for $L = Me^{\bar{C}k\tau}$. This situation can be achieved for any $\alpha \in \left(\frac{\sqrt{C_r L_\beta}}{\bar{C}}, 1\right)$, but implies a restriction on the time step τ (namely $\tau \leq \frac{1-\alpha}{\alpha\bar{C}}$).

Remark 2.2.2 *For r we have requested supplementary the global positiveness. This was necessary in order to obtain the lower bounds if the reaction is discretized implicitly, leading to a problem of the same type as the auxiliary one. Replacing $r(\beta^{-1}(\theta))$ by $r(\beta^{-1}(\phi))$ (with $\phi \in V_k$), the above assumption can be relaxed to a more natural one, namely $r(u) \geq 0$ only for positive u. Generally, this can be obtained by a standard trick for parabolic problems, where u is replaced by ve^{Ct} (C being a constant specified below).*

Remark 2.2.3 *The upper bound induces some restriction on the time step, which does not depend on k. If the reaction term is considered explicitly, there is no need to impose anything on τ. Moreover, in this case the constant $\bar{C}$ can be taken exactly $C_r\sqrt{L_\beta}$. In fact, since stability is not affected by the explicit treatment of the reaction term, this way is recommendable for practical computations.*

Lemma 2.2.1 can be used in establishing the maximum principle for the schemes MTC or MTL. The proof is simpler for the implicit discretization since we can deal with the convection term as follows. Under the assumption (A3), for any real number δ, a vector valued function $G^\delta : \mathbb{R} \to \mathbb{R}^d$ can be defined by

$$G_i^\delta(\theta) = \int_{\beta(\delta)}^{\theta} F_i(\beta^{-1}(s))ds, \tag{2.12}$$

for all $i = \overline{1,d}$. Therefore we have $G_i^\delta(\beta(\delta)) = 0$ and $\nabla_x \cdot G^\delta(\theta) = F(\beta^{-1}(\theta))\nabla_x\theta$. Now, for $\delta = \varepsilon$, by taking $\varphi = [\beta(\varepsilon) - \theta]_+$ in (2.6) the convective part vanishes due to the Gauß integration formula and the proof of the lower bound follows directly from the inequality

$$\begin{aligned} 0 \leq\ & \textstyle\int_{\theta<\beta(\varepsilon)}(\beta^{-1}(\theta) - \varepsilon)(\theta - \beta(\varepsilon)) + \tau\|\nabla\theta\|^2_{\Omega_\varepsilon} \\ =\ & \textstyle\int_{\theta<\beta(\varepsilon)}(\beta^{-1}(\chi) - \varepsilon + \tau r(\beta^{-1}(\theta)))(\theta - \beta(\varepsilon)). \end{aligned}$$

Because of the assumptions on χ and r, the last term is negative, so that either $\theta \equiv \beta(\varepsilon)$, or $meas\{\Omega_\varepsilon\} = 0$, both leading to the desired lower bound. The upper one results in a similar fashion.

Remark 2.2.4 *Problem P may be degenerate only at $u = 0$. The above maximum principle guarantees that if the (initial and boundary) data are away from 0, then the solution of any of the semi-discrete approximation defined before stays always away from 0 at all time steps. This is just the underlying idea of our approach to treat the degeneracy.*

Remark 2.2.5 *The question of existence and uniqueness of a solution for Problem AUX (which is related to the nonlinear elliptic problems arising in the time discretization) is not a direct consequence of the nonlinear Lax-Milgram lemma ([99]) since the form involved here is unbounded on the whole space H^1. Relying on the maximum principle shown above, a solution will be obtained further in the subset V_k, while uniqueness is given by comparison arguments.*

Linearization of the nonlinear schemes

Having now the maximum principle, we can continue with a linearization procedure for the nonlinear schemes WMTI and WMTC. $\theta^k \in V_k$ is obtained by iterative methods. To do so, let K be a constant which will be given below and define, for $\psi, \varphi \in H_0^1(\Omega)$ and $\underline{\psi} \in V_k$,

$$\begin{aligned} a_K(\psi, \varphi; \underline{\psi}) =\ & K(\psi, \varphi) + \tau(\nabla\psi, \nabla\varphi) - \tau((\beta^{-1})'(\underline{\psi})f(\beta^{-1}(\underline{\psi})) \cdot \nabla\psi, \varphi), \\ l_K(\chi; \varphi) =\ & K(\chi, \varphi) + (\beta^{-1}(\theta^{k-1}) - \beta^{-1}(\chi + \beta(\varepsilon)), \varphi) + \tau(r(\beta^{-1}(\chi + \beta(\varepsilon))), \varphi), \end{aligned}$$

which are linear and bounded. Denoting by

$$W_k = V_k - \beta(\varepsilon) \subset H_0^1(\Omega),$$

the translation of V_k with $\beta(\varepsilon)$, the iterative scheme is induced through the operator $T : W_k \to W_k$ giving the solution of the following problem

Problem PISm: Let $\psi \in W_k$. Find $T\psi \in W_k$ such that

$$a_K(T\psi, \varphi; \theta^{k-1}) = l_K(\psi; \varphi) \tag{2.13}$$

for all $\varphi \in H_0^1(\Omega)$.

Now the first iteration can be defined as

Iteration ISm:

$$\psi^{i+1} = T\psi^i \tag{2.14}$$

for $i \geq 0$ and $\psi^0 = \theta^{k-1} - \beta(\varepsilon) \in W_{k-1}$.

An alternative which is more appropriate for practical purposes (see [46], [47] or [53]) reads

Iteration IJK:

$$\begin{aligned}
& \bar{\theta}^i \in \beta(\varepsilon) + H_0^1(\Omega), \\
& (\sigma(\bar{\theta}^{i-1}, \theta^{k-1})(\bar{\theta}^i - \theta^{k-1}), \varphi) + \tau(\nabla\bar{\theta}^i, \nabla\varphi) - \tau((\beta^{-1})'(\underline{\theta}) f(\beta^{-1}(\underline{\theta})) \cdot \nabla\bar{\theta}^i, \varphi) \\
= \;& \tau(r(\beta^{-1}(\underline{\theta})), \varphi), \\
& \sigma(\bar{\theta}^i, \theta^k) = \int_0^1 \left(\beta^{-1}\right)' (s\bar{\theta}^i + (1-s)\theta^k) ds
\end{aligned} \tag{2.15}$$

for all $\varphi \in H_0^1(\Omega)$ and $i \geq 1$, where $\bar{\theta}^0 = \theta^{k-1}, \sigma(\bar{\theta}^0, \theta^{k-1}) = \left(\beta^{-1}\right)' (\theta^{k-1})$. Some aspects concerning this iteration are discussed at the end of the subsection. Now we concentrate on Iteration ISm, which needs a more rigorous argumentation. This will help us in the proof of existence and uniqueness for the solutions of the semi-discrete problems. These properties cannot be obtained as a direct consequence of a nonlinear Lax-Milgram lemma since the form used in the definition of Problem AUX - which is similar to the nonlinear problems WMTI or WMTC - is bounded only on a subset of $H_0^1(\Omega)$. For the linearized scheme MTL, the maximum principle guarantees that the classical Lax-Milgram lemma can be applied for proving the existence and uniqueness of a solution.

If K satisfies

$$K \geq C_2(\varepsilon), \quad \text{and} \quad K \geq \tau\frac{C_F}{2}C_2(\varepsilon) \tag{2.16}$$

(the last inequality being implied by the first one for τ reasonably small), remembering the bounds given in Remark 2.2.1, the coercivity of a_K follows as a consequence of the Cauchy inequality

$$\begin{aligned} a_K(\psi, \psi; \underline{\psi}) = &\quad K\|\psi\|^2 + \tau\|\nabla\psi\|^2 - \tau\sqrt{C_F C_2(\varepsilon)}\|\nabla\psi\|\|\psi\| \\ \geq &\quad (K - \tau\tfrac{C_F}{2}C_2(\varepsilon))\|\psi\|^2 + \tfrac{\tau}{2}\|\nabla\psi\|^2. \end{aligned}$$

If $\psi \in W_k$, the Lax-Milgram lemma can be applied to get a unique solution $T\psi \in H_0^1(\Omega)$ of Problem PISm, which is a linear problem and therefore can be solved easily. In this way, we have defined an operator T from W_k to $H_0^1(\Omega)$. In fact, a similar proof as the one given for Lemma 2.2.1 leads to the following result

Lemma 2.2.2 *Assume (A1), (A3) and $\theta^{k-1} \in V_{k-1}$. Then $TW_k \subset W_k$.*

Remark 2.2.6 *Problem PISm is related to Scheme WMTC. The same holds also if θ^{k-1} in (2.13) is replaced by $\psi + \beta(\varepsilon)$. This can be used to obtain an iterative scheme for the implicit time discretization method.*

Having defined the iterative scheme ISm, a convergence result for the corresponding sequence is necessary. Below we will show that, under the restrictions in (2.16), T is a contraction mapping on the closed set W_k with an appropriate norm, so θ^k can be taken as

$$\theta^k = \beta(\varepsilon) + \lim_{i\to\infty} \psi^i.$$

In this case, assuming the above limit makes sense, it is easy to see that we have obtained a solution in V_k of Problem WMTC. The uniqueness of the solution in $\beta(\varepsilon) + H_0^1(\Omega)$ will be discussed below, together with the fully nonlinear scheme, WMTI.

The existence of $\lim_{i\to\infty} \psi^i$ in W_k can be immediately seen by applying the fixed point theorem to T. This statement is supported by the following lemma

Lemma 2.2.3 *Assume (A1), (A3) and $\theta^{k-1} \in V_{k-1}$. If K and τ satisfy the inequalities in (2.16) and $\tau C_2(\varepsilon) \leq C$ for a constant C which is determined below, then there is a norm on $H_0^1(\Omega)$ equivalent to the usual one, such that T is contractive on the closed set W_k.*

Proof. Lemma 2.2.2 shows that W_k is preserved by T. Note that

$$\|\varphi\|_K^2 \equiv (K - \tau\frac{C_F}{2}C_2(\varepsilon))\|\varphi\|^2 + \frac{\tau}{2}\|\nabla\varphi\|^2$$

is a norm on $H_0^1(\Omega)$ and $\|\varphi\|_K \leq \sqrt{a_K(\varphi,\varphi;\theta_{k-1})}$. With this new norm, T is a contraction mapping on the closed subset W_k, as follows from (2.13)

$$\begin{aligned}|a_K(T\psi_1 - T\psi_2,\varphi;\theta_{k-1})| = \; & |K\,(\psi_1-\psi_2,\varphi) - (\beta^{-1}(\psi_1+\beta(\varepsilon)) - \beta^{-1}(\psi_2+\beta(\varepsilon)),\varphi)\,| \\ & +\tau(r(\beta^{-1}(\psi_1+\beta(\varepsilon))) - (r(\beta^{-1}(\psi_2+\beta(\varepsilon))),\varphi) \\ \leq \; & |\,((K-(\beta^{-1})'(\chi))(\psi_1-\psi_2),\varphi)\,| + \tau\sqrt{C_r C_2(\varepsilon)}\|\psi_1-\psi_2\|\|\varphi\| \\ \leq \; & (K - C_1 + \tau\sqrt{C_r C_2(\varepsilon)})\|\psi_1-\psi_2\|\|\varphi\|,\end{aligned}$$

for any $\psi_1,\psi_2 \in W_k$ and χ between $\psi_1+\beta(\varepsilon)$ and $\psi_2+\beta(\varepsilon)$. In the above inequalities the mean value theorem, the growth condition on r and the positiveness of ψ_1,ψ_2 have been used. Hence, if $\tau(\frac{C_F}{2}C_2(\varepsilon)+\sqrt{C_r C_2(\varepsilon)}) \leq C_1$, since

$$\|\varphi\|_K \geq \sqrt{K - \tau\frac{C_F}{2}C_2(\varepsilon)}\|\varphi\|,$$

by taking $\varphi = T\psi_1 - T\psi_2$ the proof is completed.

Remark 2.2.7 *The restrictions on τ are due to the convective and reaction terms. If the Lipschitz continuity in $\beta(u)$ of F and r is assumed, the limitation on the time step becomes milder (namely $C\tau < 1$). Moreover, in the absence of the above terms, there is no need to impose anything on τ (see [98]).*

Remark 2.2.8 *As resulting from the above proof, the convergence rate of Iteration ISm is of order $1-O(1/C_2(\varepsilon))$, which is close to 1 when ε is small. But this approach is useful in obtaining the existence of the solutions of the nonlinear problems. The reduced convergence order can be observed also in practical computations, therefore we have considered also Iteration JK, which is more interesting from the practical point of view. In this case the theoretical restriction on τ is more severe than the one in Lemma 2.2.3 ([53]).*

Remark 2.2.9 *Iteration ISm relies on the operator T, which has been used in [90], pp. 96. But we show that T is a contraction mapping, at least for the present problems, in the setting defined in the proof of Lemma 2.2.3. This fact simplifies the context, while the monotonicity of T can lead to an alternative proof similar to the one in [90].*

The above results show that Problem WMTC has a unique solution in V_k. But in the frame set in Lemma 2.2.3, the uniqueness holds in the whole $\beta(\varepsilon) + H_0^1(\Omega)$.

Lemma 2.2.4 *Assume (A1), (A3) and $\theta^{k-1} \in V_{k-1}$. If K and τ satisfy the inequalities in (2.16) and $\tau C_2(\varepsilon) \leq C$ for a suitable constant C, Problem WMTC has at most one solution $\theta^k \in \beta(\varepsilon) + H_0^1(\Omega)$*

Proof. We consider here only the implicit discretization of the reaction term, the proof in the explicit case being identical. Let $\psi, \theta \in \beta(\varepsilon) + H_0^1(\Omega)$, then $\psi - \theta \in H_0^1(\Omega)$. Assuming that both solve Problem WMTC with the same θ^{k-1}, testing with $\varphi = \psi - \theta$ and subtracting the corresponding equalities for both solutions yields

$$\begin{aligned}
& (\beta^{-1}(\psi) - \beta^{-1}(\theta), \psi - \theta) + \tau\|\nabla(\psi - \theta)\|^2 \\
= \; & \tau((\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1})) \cdot \nabla(\psi - \theta) + r(\beta^{-1}(\psi)) - r(\beta^{-1}(\theta)), \psi - \theta) \\
\leq \; & \tau\sqrt{C_F C_2(\varepsilon)}\|\nabla(\psi - \theta)\|\|\psi - \theta\| + \tau\sqrt{C_r}(\beta^{-1}(\psi) - \beta^{-1}(\theta), \psi - \theta)^{\frac{1}{2}}\|\psi - \theta\| \\
\leq \; & \tfrac{\tau}{2}\|\nabla(\psi - \theta)\|^2 + \tfrac{1}{2}(\beta^{-1}(\psi) - \beta^{-1}(\theta), \psi - \theta) + (\tfrac{\tau}{2} C_F C_2(\varepsilon) + \tfrac{\tau^2}{2} C_r)\|\psi - \theta\|^2,
\end{aligned}$$

where the monotonicity of β has been used. This, remembering the relation in (2.2), leads to

$$\left(\frac{C_1}{2} - \frac{\tau}{2} C_F C_2(\varepsilon) + \frac{\tau^2}{2} C_r\right)\|\psi - \theta\|^2 + \frac{\tau}{2}\|\nabla(\psi - \theta)\|^2 \leq 0,$$

which, if τ is small enough (satisfying $\tau^2 C_r + \tau C_F C_2(\varepsilon) \leq C_1$), shows that ψ and θ coincide a.e. in Ω.

Remark 2.2.10 *The proof of uniqueness in W_k for the implicit discretization scheme WMTI is similar (the assumption in (A3) on F has to be taken into account). In this case, the restriction on τ changes to $C\tau \leq 1$, which is more convenient.*

As mentioned in the Remark 2.2.6, Iteration ISm can be used to get the solution of Problem WMTI. More precisely, the Lax-Milgram lemma can be applied again to get a unique $T\psi \in H_0^1(\Omega)$ satisfying

$$a_K(T\psi, \varphi; \psi + \beta(\varepsilon)) = l_K(\psi; \varphi) \qquad \forall \varphi \in H_0^1(\Omega) \tag{2.17}$$

for each $\psi \in W_k$. Also Lemma 2.2.2 holds true, but unfortunately we have not been able to obtain strong convergence in H^1 without additional assumptions on the convection term F. However, an uniform bound for the H^1 norm of the sequence $\{\psi^i, i \geq 0\}$ can be easily derived. This follows by taking $\varphi = T\psi$ in (2.17), giving

$$\begin{aligned}
& K\|T\psi\|^2 + \tau\|\nabla T\psi\|^2 - \tau((\beta^{-1})'(\psi + \beta(\varepsilon)) f(\beta^{-1}(\psi + \beta(\varepsilon))) \cdot \nabla T\psi, T\psi) \\
= \; & K(\psi, T\psi) + (\beta^{-1}(\theta^{k-1}) - \beta^{-1}(\psi + \beta(\varepsilon)), T\psi) + \tau(r(\beta^{-1}(\psi + \beta(\varepsilon))), T\psi).
\end{aligned}$$

Recalling the maximum principle, an upper bound ($\bar{C}$) depending on ε, the supremum in V_k and the measure of Ω - but not on $T\psi$ - can be obtained in the right hand side above. The last term in the left hand side can be bounded by

$$|\tau((\beta^{-1})'(\psi + \beta(\varepsilon)) f(\beta^{-1}(\psi + \beta(\varepsilon))) \cdot \nabla T\psi, T\psi)| \leq \frac{\tau}{2}\|\nabla T\psi\|^2 + C,$$

where the constant C has the same properties as $\bar{C}$. This shows the uniform boundedness of $\{\psi^i, i \geq 0\}$ in H^1. Compactness arguments in connection with the uniqueness of the solution guarantee the convergence of the above sequence weakly in $H^1(\Omega)$, hence strongly in $L^2(\Omega)$. The statements above can be included in the following lemma

Lemma 2.2.5 *Assume (A1), (A3) and $\theta^{k-1} \in V_{k-1}$. If K and τ satisfy the inequalities in (2.16) and $\tau \leq C$ for a constant C, Problem WMTI has at most one solution $\theta^k \in \beta(\varepsilon) + H_0^1(\Omega)$. Moreover,*

$$\theta^k = \beta(\varepsilon) + \lim_{i \to \infty} \psi^i,$$

and convergence takes place weakly in H^1, thus strongly in L^2.

As seen in Remark 2.2.8, the iteration scheme discussed up to now has a reduced convergence rate, but relying on it we have been able to show the existence and uniqueness of the solution for the semi-discrete equations. Iteration JK is more appropriate for practical purposes. Without giving the proof - which is similar to the one for Lemma 2.2.1 - if $\theta^{k-1} \in V_{k-1}$, then all the elements of the sequence of solutions $\{\bar{\theta}^i\}_{i=0}^{\infty}$ given by this iteration belong to V_k. Therefore, the regularization based on the maximum principle works in this case too.

In particular, taking $\underline{\theta} = \theta^{k-1}$ above leads to an iterative method for Problem WMTC, while the choice $\underline{\theta} = \bar{\theta}^{i-1}$ gives the alternative for Problem WMTI. This method of linearization is proposed in [46] and [47], where a cut-off approach for the regularization is necessary. The function σ here is similar to μ in Scheme JK, in the setting mentioned before. Again, since the problems WMTC and WMTI admit unique solutions, compactness arguments show that $\bar{\theta}^i$ converges weakly in H^1 (and therefore strongly in L^2) to the semi-discrete solution θ_k. Moreover, Theorem 5.9 of [53] can be applied here in order to obtain the order of convergence in L^2 provided the time step is small enough, namely $\tau^{\frac{\alpha}{2}} C_2(\varepsilon) \leq C$. Here $\alpha \leq 1$ comes from the Hölder continuity of the solution, namely $\theta^k \in C^{0,\alpha}$ ([58]). As mentioned before, this restriction is merely theoretically, practical computations showing that Iteration JK is efficient. Most of our numerical examples in the last chapter were computed using this approach.

2.2.2 Error estimates for the semi-discrete approximation

In this section we will show the convergence of the maximum principle based approach by proving error estimates for the schemes MTI, MTC and MTL written in the variational form. To do so, some stability properties are necessary.

We start here by stating three elementary identities to be used in what follows, which are valid for all $a_k, b_k \in \mathbb{R}^q \quad (q \geq 1)$

$$2\sum_{k=1}^{m} a_k(a_k - a_{k-1}) = |a_m|^2 - |a_0|^2 + \sum_{k=1}^{m} |a_k - a_{k-1}|^2, \tag{2.18}$$

$$2\sum_{k=1}^{m}\sum_{j=1}^{k} a_k a_j = \left|\sum_{k=1}^{m} a_k\right|^2 + \sum_{k=1}^{m} |a_k|^2, \tag{2.19}$$

$$\sum_{k=1}^{m} a_k(b_k - b_{k-1}) = a_m b_m - a_0 b_0 - \sum_{k=1}^{m}(a_k - a_{k-1})b_{k-1}. \tag{2.20}$$

The apriori estimates

The following theorem establishes the stability of the implicit scheme.

Theorem 2.2.6 *Assume (A1), (A2) and (A3). Then, for $p \leq n$, if θ^k solves Problem WMTI, we have*

$$\tau \sum_{k=1}^{p} \|\nabla\theta^k\|^2 \leq C, \tag{2.21}$$

$$\begin{aligned}&\sum_{k=1}^{p}(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}) + \sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 \\ &+\tau\|\nabla\theta^p\|^2 + \tau\sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2 \leq C\tau C_2(\varepsilon).\end{aligned} \tag{2.22}$$

Proof. By taking $\varphi = \theta^k - \beta(\varepsilon) \in H_0^1(\Omega)$ in Problem WMTI and summing up for $k = \overline{0,p}$ we get

$$\begin{aligned}&\sum_{k=1}^{p}(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \beta(\varepsilon)) + \tau\sum_{k=1}^{p} \|\nabla\theta^k\|^2 \\ =\ & -\tau\sum_{k=1}^{p}(F(\beta^{-1}(\theta^k)), \nabla\theta^k) + \tau\sum_{k=1}^{p}(r(\beta^{-1}(\theta^k)), \theta^k - \beta(\varepsilon)).\end{aligned} \tag{2.23}$$

Now, each of the resulting terms is estimated separately. To begin with, observe that

$$\int_a^b s(\beta^{-1})'(s)ds \leq b(\beta^{-1}(b) - \beta^{-1}(a)), \quad \int_0^b s(\beta^{-1})'(s)ds \geq \frac{C_1}{2}b^2$$

hold true for any reals a, b (due to the properties of β). Hence, for a.e. $x \in \Omega$, it follows that

$$\begin{array}{rl} & \sum_{k=1}^p (\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}))\theta^k = \sum_{k=1}^p \theta^k \int_{\theta^{k-1}}^{\theta^k} (\beta^{-1})'(s)ds \\ \geq & \sum_{k=1}^p \int_{\theta^{k-1}}^{\theta^k} s(\beta^{-1})'(s)ds = \int_0^{\theta^p} s(\beta^{-1})'(s)ds - \int_0^{\theta^0} s(\beta^{-1})'(s)ds \\ \geq & C(\theta^p)^2 - \beta^{-1}(\theta^0)\theta^0. \end{array}$$

Since $u_0, \theta^0 \in L^\infty$, integrating over Ω the above inequality gives

$$\sum_{k=1}^p (\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k) \geq C\|\theta^p\|^2 - (\beta^{-1}(\theta^0), \theta^0) \geq C\|\theta^p\|^2 - C \geq -C.$$

Using the maximum principle in Lemma 2.2.1, the remaining part of the first sum can be bounded as follows

$$\left| \sum_{k=1}^p (\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \beta(\varepsilon)) \right| = \left| (\beta^{-1}(\theta^p) - \beta^{-1}(\theta^0), \beta(\varepsilon)) \right| \leq C\beta(\varepsilon).$$

Next, the first term in the right hand side vanishes (as in (2.12)). For the last one, since θ^k is uniformly bounded (also independently on k)

$$\left| \sum_{k=1}^p (r(\beta^{-1}(\theta^k)), \theta^k - \beta(\varepsilon)) \right| \leq \tau \sum_{k=1}^p \|r(\beta^{-1}(\theta^k))\| \|\theta^k - \beta(\varepsilon)\| \leq C.$$

Since ε is small, we have $\beta(\varepsilon) \leq C$. The above inequalities give the first part of the conclusion.

For the second estimate, φ in (2.6) is replaced by $\theta^k - \theta^{k-1} \in H_0^1(\Omega)$. Recalling the identity in (2.18) and summing again over k from 1 to p yields

$$\begin{aligned} (I) + (II) :=& \sum_{k=1}^p (\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}) \\ &+ \frac{\tau}{2} \left(\|\nabla\theta^p\| - \|\nabla\theta^0\| + \sum_{k=1}^p \|\nabla(\theta^k - \theta^{k-1})\|^2 \right) \\ =& -\tau \sum_{k=1}^p (F(\beta^{-1}(\theta^k)), \nabla(\theta^k - \theta^{k-1})) \\ &+ \tau \sum_{k=1}^p (r(\beta^{-1}(\theta^k)), \theta^k - \theta^{k-1}) =: (III) + (IV). \end{aligned} \tag{2.24}$$

For (I), because of the assumptions on β, the following inequality holds true

$$(I) \geq \frac{C_1}{2} \sum_{k=1}^p \|\theta^k - \theta^{k-1}\|^2 + \frac{1}{2} \sum_{k=1}^p (\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}).$$

Since $\nabla\cdot F(\beta^{-1}(\theta)) = f(\beta^{-1}(\theta))(\beta^{-1})'(\theta)\nabla\theta$, recalling the boundedness of f from Remark 2.2.1, for (III) we have

$$\begin{aligned} |(III)| \leq & \ \tau\sum_{k=1}^{p} |(f(\beta^{-1}(\theta^k))(\beta^{-1})'(\theta^k)\nabla\theta^k, \theta^k - \theta^{k-1})| \\ \leq & \ \tau\sum_{k=1}^{p} \sqrt{C_F C_2(\varepsilon)}\|\nabla\theta^k\|\|\theta^k - \theta^{k-1}\|. \end{aligned}$$

Applying here the inequality $ab \leq \eta a^2 + \frac{1}{4\eta}b^2$ for $\eta = \frac{2\tau}{C_1}$ and using the first estimates gives

$$|(III)| \leq \frac{C_1}{8}\sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 + C\tau C_2(\varepsilon).$$

Now, recalling the maximum principle again, the last term leads in a similar manner to

$$|(IV)| \leq \frac{C_1}{8}\sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 + C\tau.$$

Finally, since $\theta_0 \in H^1(\Omega)$, the inequalities above show the remaining part of the theorem.

Remark 2.2.11 *It does not make any difference for the apriori estimates if the reaction term is considered explicitly.*

Remark 2.2.12 *The large constant $C_2(\varepsilon)$ appears in the last stability estimate due to the condition imposed in (A3) on F. As we will see below, this does not affect the error estimates. In the case F is Lipschitz continuous in $\beta(u)$, since $C_2(\varepsilon)$ disappears in (2.22), the apriori estimates become optimal (as obtained, e.g. in [73]).*

Remark 2.2.13 *For the analysis of Scheme MTC it will be useful to have apriori estimates without $C_2(\varepsilon)$. These can be obtained by a different handling of (III)*

$$\begin{aligned} |(III)| \leq & \ \tau\sum_{k=1}^{p} |(F(\beta^{-1}(\theta^k)), \nabla(\theta^k - \theta^{k-1}))| \\ \leq & \ C\tau\sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\| \leq C + \frac{\tau}{4}\sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2, \end{aligned}$$

where the uniform bounds for θ^k have been used again. The rest of the proof follows as before and leads to

$$\begin{aligned} & \sum_{k=1}^{p}(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}) + \sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 \\ & \qquad + \tau\|\nabla\theta^p\|^2 + \tau\sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2 \ \leq C. \end{aligned}$$

In this case the results are worser (since τ is lost), but still enough for getting the error estimates.

In a similar manner, for Scheme MTC we have

Theorem 2.2.7 *Assume (A1), (A2) and (A3). Then, for $p < n$, if θ^k solves Problem WMTC, there are constants C independent on p, τ and ε such that*

$$\tau \sum_{k=1}^{p} \|\nabla\theta^k\|^2 \leq C C_2(\varepsilon), \tag{2.25}$$

$$\begin{aligned} &\sum_{k=1}^{p}(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}) + \sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 \\ &+\tau\|\nabla\theta^p\|^2 + \tau \sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2 \leq C\tau C_2(\varepsilon)^2. \end{aligned} \tag{2.26}$$

Proof. The proof is almost identical to the one for Theorem 2.2.6. The only difference appears when dealing with the convection. For (2.25), the first term in the right hand side of the relation in (2.23) is bounded as follows

$$\begin{aligned} & \tau \textstyle\sum_{k=1}^{p} |((\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1}))\nabla\theta^k, \theta^k - \beta(\varepsilon))| \\ \leq\ & \tau\sqrt{C_F C_2(\varepsilon)} \textstyle\sum_{k=1}^{p} \|\nabla\theta^k\| \|\theta^k - \beta(\varepsilon)\| \\ \leq\ & \dfrac{\tau}{2} \textstyle\sum_{k=1}^{p} \|\nabla\theta^k\|^2 + C C_2(\varepsilon), \end{aligned}$$

and the first inequality above is obtained as before.

For the second estimate, the same steps can be done to get a similar relation to the one in (2.24). The estimate for (III) reads

$$\begin{aligned} |(III)| =\ & \tau \textstyle\sum_{k=1}^{p}((\beta^{-1})'(\theta^{k-1}) f(\beta^{-1}(\theta^{k-1}))\nabla\theta^k, \theta^k - \theta^{k-1}) \\ \leq\ & \dfrac{C_1}{8} \textstyle\sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 + C\tau^2 C_2(\varepsilon) \textstyle\sum_{k=1}^{p} \|\nabla\theta^k\|^2 \\ \leq\ & \dfrac{C_1}{8} \textstyle\sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 + C\tau C_2(\varepsilon)^2, \end{aligned}$$

the last inequality is due to the relation in (2.25). Proceeding as in the former case we arrive to the remaining estimate.

Remark 2.2.14 *As for the implicit case, optimal estimates are obtained if F is Lipschitz continuous in $\beta(u)$. Then, the large constant $C_2(\varepsilon)$ disappears in (2.25) and (2.26). Similarly, an explicit discretization of the reaction term yields the same results.*

Finally we turn our attention to the linear scheme. In this case we have to assume more on β, namely the Lipschitz continuity of its derivative. Still, reasonable estimates are obtained only if one of the following additional hypotheses hold true

(A1)$'$ $\quad\beta : \mathbb{R} \to \mathbb{R}$ is a convex function,

or

(A3)$'$ $\quad F : \mathbb{R} \to \mathbb{R}^d$ is Lipschitz continuous in $\beta(u)$.

Even though (A1)$'$ and (A3)$'$ do not have too much in common, at least one of them is essential for the stability. More precise, the apriori estimates which can be obtained without any of the above supplementary assumptions are useless - depending exponentially on $C_2(\varepsilon)$ - so the error estimates for Scheme MTL are irrelevant.

Theorem 2.2.8 *Assume (A1), (A1)$'$, (A2) and (A3). Then, for $p < n$, if θ^k solves Problem WMTL, we have*

$$\tau \sum_{k=1}^{p} \|\nabla \theta^k\|^2 \le C C_2(\varepsilon), \tag{2.27}$$

$$\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})\|^2 + \tau\|\nabla\theta^p\|^2 + \tau \sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2 \le C\tau C_2(\varepsilon)^{\frac{3}{2}}. \tag{2.28}$$

Proof. The proof for Theorem 2.2.6 can be followed in this case too. If we take $\varphi = \theta^k - \beta(\varepsilon) \in H_0^1(\Omega)$ in (2.8) and sum up the resulting equalities for $k = \overline{0,p}$, the only difference to the relation in (2.23) consists in the term

$$(\bar{I}) := \sum_{k=1}^{p} (\sigma_{k-1}(\theta^k - \theta^{k-1}), \theta^k - \beta(\varepsilon)). \tag{2.29}$$

This can be rewritten as

$$\begin{aligned} (\bar{I}) = \ & \textstyle\sum_{k=1}^{p}((\sigma_{k-1} - \sigma(\theta^k, \theta^{k-1}))(\theta^k - \theta^{k-1}), \theta^k - \beta(\varepsilon)) \\ & + \textstyle\sum_{k=1}^{p}(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \beta(\varepsilon)) =: (I_1) + (I_2), \end{aligned}$$

where $\sigma(\theta^k, \theta^{k-1})$ is defined with Iteration IJK. Since β if convex and β' positive,

$$sign\{(\beta^{-1})'(\theta^{k-1}) - (\beta^{-1})'(s\theta^{k-1} + (1-s)\theta^k)\} = -sign\{\theta^{k-1} - \theta^k\}.$$

Hence, remembering that $\theta^k > \beta(\varepsilon)$, it follows that (I_1) is positive. The bounds for (I_2) are now derived as above.

Next, repeating the steps in the proof of Theorem 2.2.7, we need an estimate for (III). Because $\sigma_{k-1} = (\beta^{-1})'(\theta^{k-1})$, it follows that

$$
\begin{aligned}
|(III)| = &\ \tau \textstyle\sum_{k=1}^{p} (\sqrt{\sigma_{k-1}} f(\beta^{-1}(\theta^{k-1}))\nabla\theta^k, \sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})) \\
\leq &\ \frac{1}{4} \textstyle\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})\|^2 + C\tau^2 \sqrt{C_2(\varepsilon)} \sum_{k=1}^{p} \|\nabla\theta^k\|^2 \\
\leq &\ \frac{1}{4} \textstyle\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})\|^2 + C\tau C_2(\varepsilon)^{\frac{3}{2}}.
\end{aligned}
$$

Here the boundedness of f and the inequality in Remark 2.2.1 have been used. The rest of the proof is identical to the previous ones.

Because of the convexity of β, the estimates obtained above are slightly better then in Theorem 2.2.7. Without this assumption, the Lipschitz continuity of F is necessary. In this case the estimates are given in the following theorem

Theorem 2.2.9 *Assume (A1), (A2), (A3) and (A3)'. Then, for $p < n$, if θ^k solves Problem WMTL, the following holds*

$$
\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})\|^2 + \tau\|\nabla\theta^p\|^2 + \tau \sum_{k=1}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2 \leq C\tau. \tag{2.30}
$$

Proof. Testing with $\theta^k - \theta^{k-1}$ in (2.8), the only thing we have to deal with is again (III). Here, since F is Lipschitz in $\beta(u)$, the constant C_F is an upper bound also for $(\beta^{-1})'(\theta) f(\beta^{-1}(\theta))$, so

$$
\begin{aligned}
|(III)| = &\ C\tau \textstyle\sum_{k=1}^{p} \|\nabla\theta^k\| \|\sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})\| \\
\leq &\ \frac{1}{4} \textstyle\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1}}(\theta^k - \theta^{k-1})\|^2 + C\tau^2 \sum_{k=1}^{p} \|\nabla\theta^k\|^2.
\end{aligned}
$$

The desired estimate is a consequence of the discrete Gronwall lemma.

Besides we need the following lemma.

Lemma 2.2.10 *For $u \geq 0$, define*

$$
u^\varepsilon = \beta^{-1}(\beta(u) + \beta(\varepsilon)).
$$

If β satisfies (A1), then

$$
0 < u^\varepsilon - u \leq \varepsilon + C_2(\varepsilon)\beta(\varepsilon), \tag{2.31}
$$

where $C_2(\varepsilon)$ is defined in (2.2).

Proof. Because β^{-1} as well as β is increasing and $\varepsilon > 0$, we have

$$u = \beta^{-1}(\beta(u)) < \beta^{-1}(\beta(u) + \beta(\varepsilon)) \equiv u^{\varepsilon}.$$

On the other hand, it follows from Taylor's theorem that

$$u^{\varepsilon} \equiv \beta^{-1}(\beta(u) + \beta(\varepsilon)) = \beta^{-1}(\beta(\max(u, \varepsilon))) + \beta(\min(u, \varepsilon))(\beta^{-1})'(\theta),$$

where θ is a convex combination of $\beta(u) + \beta(\varepsilon)$ and $\beta(\max(u, \varepsilon))$, which are both not less than $\beta(\varepsilon)$ due to $u \geq 0$. Thus, it is clear that

$$\beta^{-1}(\beta(\max(u, \varepsilon))) \leq u + \varepsilon, \quad \beta(\min(u, \varepsilon)) \leq \beta(\varepsilon) \quad \text{and} \quad (\beta^{-1})'(\theta) \leq C_2(\varepsilon)$$

and hence

$$u^{\varepsilon} \leq u + \varepsilon + C_2(\varepsilon)\beta(\varepsilon).$$

Remark 2.2.15 *If β satisfies*

$$\varepsilon \cdot \inf \{\beta'(x) : x \in [\varepsilon, M]\} \geq C\beta(\varepsilon),$$

which is true if $\beta(u) = u^m$ with $m \geq 1$, then the above inequality become

$$0 < u^{\varepsilon} - u \leq C\varepsilon.$$

Similarly, in case β is super-linear - satisfying $\beta(u) + \beta(v) \leq \beta(u + v)$ for any u, v - which holds true again for the typical case mentioned above, Lemma 2.2.10 becomes

$$0 < u^{\varepsilon} - u \leq \varepsilon.$$

Both cases will simplify the error estimates.

Error estimates for the implicit method

Now we turn to the estimates for the error in the semi-discrete case. To do so let us set for any function f integrable in time and defined in Q_T

$$\bar{f}^k := \frac{1}{\tau} \int_{(k-1)\tau}^{k\tau} f(s, \cdot) ds,$$

if $k \geq 1$ and $\bar{f}^0 := f(0, \cdot)$. We will make use of the following notations

$$e_u^k := \bar{u}^k - \beta^{-1}(\theta^k), \quad e_u^{\varepsilon,k} := \overline{u^\varepsilon}^k - \beta^{-1}(\theta^k), \quad e_\theta^k := \overline{\beta(u)}^k - \theta^k, \quad e_\theta^{\varepsilon,k} := \overline{\beta(u^\varepsilon)}^k - \theta^k,$$

where $k \geq 0$. Using the Lemma 2.2.10, because of the maximum principle, we have for any $k \geq 0$

$$\begin{aligned} e_u^k \leq e_u^{\varepsilon,k} &\leq e_u^k + \varepsilon + C_2(\varepsilon)\beta(\varepsilon), \\ e_\theta^k \leq e_\theta^{\varepsilon,k} &\leq e_\theta^k + \beta(\varepsilon). \end{aligned} \tag{2.32}$$

Moreover, remembering the definition of θ^0 in (2.1), Lemma 2.2.10 shows that initially these errors satisfy

$$\begin{aligned} -\varepsilon + C_2(\varepsilon)\beta(\varepsilon) \leq e_u^0 &\leq 0 \leq e_u^{\varepsilon,0} \leq C_2(\varepsilon)\beta(\varepsilon), \\ -\beta(\varepsilon) \leq e_\theta^0 &\leq 0 \leq e_\theta^{\varepsilon,0} \leq \beta(\varepsilon). \end{aligned} \tag{2.33}$$

Because of the definition of u^ε, for $k \geq 0$

$$\nabla e_\theta^k = \nabla e_\theta^{\varepsilon,k}$$

holds true.

The analysis below combines the approach in [60] with the ones in [29] and [73]. We have not considered the possibility given by the non-degeneracy property of the solution (which was taken into account first in [64], [66] and then in several papers like [2] or [28]), since this kind of results are shown up to now only in some particular cases (see, e.g. [6] for the porous medium equation, [65] for Stefan problems, or [27]).

In the sequel $G : H^{-1}(\Omega) \to H_0^1(\Omega)$ denotes the Green operator defined by

$$(\nabla G\psi, \nabla\varphi) = (\psi, \varphi), \tag{2.34}$$

for all $\varphi \in H_0^1(\Omega)$, where ψ is taken in $H^{-1}(\Omega)$. In the right hand side $(\cdot, \cdot)$ stands for the duality pairing between H^{-1} and H_0^1. Obviously, G is linear. Due to the Poincaré-Friedrichs inequality, the following

$$\|\psi\|_{-1} = \sup_{\varphi \in H_0^1,\, \varphi \neq 0} \frac{|(\psi, \varphi)|}{\|\nabla\varphi\|},$$

is a norm in $H^{-1}(\Omega)$ equivalent to the usual one. Recalling the definition of the operator G we can easily obtain

$$\|\nabla G\psi\|^2 = (\nabla G\psi, \nabla G\psi) = (\psi, G\psi) \leq \|\psi\|_{-1}\|\nabla G\psi\|,$$

hence $\|\nabla G\psi\| \leq \|\psi\|_{-1}$. The inequality in the inverse sense is a direct consequence of the same definitions. Moreover, if $\psi \in L^2(\Omega)$, the inequality

$$\|\psi\|_{-1} \leq C\|\psi\|$$

is a direct consequence of the inequality mentioned above. Thus we have shown that

$$\|\nabla G\psi\| = \|\psi\|_{-1}, \quad \|\psi\|_{-1} \leq C\|\psi\| \tag{2.35}$$

(where the last inequality applies only if $\psi \in L^2(\Omega)$).

Now we can proceed with the error estimates for the implicit scheme. These are obtained in the following theorem.

Theorem 2.2.11 *Assume (A1), (A2) and (A3) and let u be the weak solution of Problem WP and θ^k solves for each $k > 0$ Problem WMTI, then*

$$\sup_{k=\overline{1,n}} \|e_u^{\varepsilon,k}\|_{-1}^2 + \int_0^T (\beta(u^\varepsilon(t)) - \theta_\Delta(t), u^\varepsilon(t) - \beta^{-1}(\theta_\Delta(t)))dt + \|\beta(u) - \theta_\Delta\|_{L^2(Q)}^2$$
$$\leq C\left\{\tau + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2\right\},$$

where $\theta_\Delta(t) = \theta^k$ for $t \in (t_{k-1}, t_k]$ and $k = \overline{1,n}$.

Proof. Let $\chi|_I$ be the characteristic function of the time interval $I \subset [0,T]$. Choosing $\varphi\chi|_{[t_{j-1},t_j]}$ for an arbitrary $\varphi \in H_0^1(\Omega)$ as a test function in (2.3) leads to the semi-discrete equations satisfied by the continuous solution

$$\begin{aligned}(u(t_j) - u(t_{j-1}), \varphi) + &\left(\nabla \int_{t_{j-1}}^{t_j} \beta(u(t))dt + \int_{t_{j-1}}^{t_j} F(u(t))dt, \nabla\varphi\right) \\ &= \left(\int_{t_{j-1}}^{t_j} r(u(t))dt, \varphi\right),\end{aligned} \tag{2.36}$$

for all $\varphi \in H_0^1(\Omega)$ and $j > 0$. Subtracting (2.6) from (2.36), taking $\varphi = Ge_u^{\varepsilon,j} \in H_0^1$ in the resulting difference and summing up for $j = \overline{1,k}$ yields

$$\begin{aligned}(I_1) + (I_2) := &\sum_{j=1}^k (u(t_j) - u(t_{j-1}) - \beta^{-1}(\theta^j) + \beta^{-1}(\theta^{j-1}), Ge_u^{\varepsilon,j}) \\ &+ \tau\sum_{j=1}^k (\nabla e_\theta^{\varepsilon,j}, \nabla Ge_u^{\varepsilon,j}) \\ = &-\sum_{j=1}^k \left(\int_{t_{j-1}}^{t_j} F(u(t)) - F(\beta^{-1}(\theta^j))dt, \nabla Ge_u^{\varepsilon,j}\right) \\ &+ \sum_{j=1}^k \left(\int_{t_{j-1}}^{t_j} r(u(t)) - r(\beta^{-1}(\theta^j))dt, Ge_u^{\varepsilon,j}\right) =: (I_3) + (I_4).\end{aligned} \tag{2.37}$$

Now we have to estimate each of the terms in (2.37). First we consider (I_1) and decompose it as follows.

$$\begin{aligned}(I_1) = &\quad \textstyle\sum_{j=1}^k (u(t_j) - \bar{u}^j - u(t_{j-1}) + \bar{u}^{j-1}, Ge_u^{\varepsilon,j}) \\ &+ \textstyle\sum_{j=1}^k (\bar{u}^j - \overline{u^\varepsilon}^j - \bar{u}^{j-1} - \overline{u^\varepsilon}^{j-1}, Ge_u^{\varepsilon,j}) \\ &+ \textstyle\sum_{j=1}^k (e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}, Ge_u^{\varepsilon,j}) \\ =: &\ (I_{11}) + (I_{12}) + (I_{13}).\end{aligned}$$

Recalling the elementary identity in (2.20), since $u(t_0) = \bar{u}^0$, (I_{11}) can be rewritten as

$$(I_{11}) = (u(t_k) - \bar{u}^k, Ge_u^{\varepsilon,k}) - \sum_{j=2}^{k} (u(t_{j-1}) - \bar{u}^{j-1}, Ge_u^{\varepsilon,j} - Ge_u^{\varepsilon,j-1}) =: (I_{111}) + (I_{112}).$$

Because $\partial_t u \in L^2(0,T;H^{-1}(\Omega))$, using the inequality $ab \leq \dfrac{a^2}{\delta} + \dfrac{\delta b^2}{4}$ with $\delta > 0$ we have

$$\begin{aligned}|(I_{111})| \leq & \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \left|(u(t_k) - u(t), Ge_u^{\varepsilon,k})\right| dt \leq \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \left| \int_t^{t_k} (\partial_s u(s), Ge_u^{\varepsilon,k}) ds \right| dt \\ \leq & \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \int_t^{t_k} \|\partial_s u\|_{-1} \|\nabla Ge_u^{\varepsilon,k}\| ds dt \leq \tau^{\frac{1}{2}} \|\partial_t u\|_{L^2(t_{k-1},t_k;H^{-1}(\Omega))} \|e_u^{\varepsilon,k}\|_{-1} \\ \leq & \tau\delta_{111} \|\partial_t u\|^2_{L^2(t_{k-1},t_k;H^{-1}(\Omega))} + \frac{1}{4\delta_{111}} \|e_u^{\varepsilon,k}\|^2_{-1},\end{aligned} \tag{2.38}$$

with $\delta_{111} > 0$ given below. The estimation of (I_{112}) goes on similarly

$$\begin{aligned}|(I_{112})| \leq & \frac{1}{\tau}\sum_{j=2}^k \int_{t_{j-2}}^{t_{j-1}} \left|(u(t_{j-1}) - u(t), Ge_u^{\varepsilon,j} - Ge_u^{\varepsilon,j-1})\right| dt \\ \leq & \tau^{\frac{1}{2}} \sum_{j=2}^k \|\partial_t u\|_{L^2(t_{j-2},t_{j-1};H^{-1}(\Omega))} \|e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}\|_{-1} \\ \leq & \tau\delta_{112} \|\partial_t u\|^2_{L^2(0,t_{k-1};H^{-1}(\Omega))} + \frac{1}{4\delta_{112}} \sum_{j=2}^k \|e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}\|^2_{-1},\end{aligned} \tag{2.39}$$

The bounds for (I_{12}) are obtained in the same manner.

$$\begin{aligned}(I_{12}) = &\ (\bar{u}^k - \overline{u^\varepsilon}^k, Ge_u^{\varepsilon,k}) - (\bar{u}^0 - \overline{u^\varepsilon}^0, Ge_u^{\varepsilon,0}) - \textstyle\sum_{j=1}^k (\bar{u}^{j-1} - \overline{u^\varepsilon}^{j-1}, Ge_u^{\varepsilon,j} - Ge_u^{\varepsilon,j-1}) \\ =: &\ (I_{121}) + (I_{122}) + (I_{123}).\end{aligned}$$

Employing the relations in (2.32), (2.33), Lemma 2.2.10 yields

$$\begin{aligned}
|(I_{121})| \leq& \|\bar{u}^k - \overline{u^\varepsilon}^k\| \|Ge_u^{\varepsilon,k}\| \leq C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))\|e_u^{\varepsilon,k}\|_{-1} \\
\leq& C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2\delta_{121} + \frac{1}{4\delta_{121}}\|e_u^{\varepsilon,k}\|_{-1}^2, \\
|(I_{122})| \leq& \|\bar{u}^0 - \overline{u^\varepsilon}^0\| \|Ge_u^{\varepsilon,0}\| \\
\leq& C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))\|e_u^{\varepsilon,0}\|_{-1} \leq C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2, \\
|(I_{123})| \leq& \sum_{j=1}^{k} \|\bar{u}^{j-1} - \overline{u^\varepsilon}^{j-1}\| \|Ge_u^{\varepsilon,j} - Ge_u^{\varepsilon,j-1}\| \\
\leq& C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon)) \sum_{j=1}^{k} \|e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}\|_{-1} \\
\leq& C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2\delta_{123} + \frac{1}{4\delta_{123}} \sum_{j=1}^{k} \|e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}\|_{-1}^2.
\end{aligned} \tag{2.40}$$

Here we have used the boundedness of Ω and C is a generic constant.

Using the identity in (2.18) and the properties of G, the estimate of (I_{13}) is a simple matter.

$$\begin{aligned}
(I_{13}) =& \sum_{j=1}^{k} (\nabla Ge_u^{\varepsilon,j} - \nabla Ge_u^{\varepsilon,j-1}, \nabla Ge_u^{\varepsilon,j}) \\
=& \frac{1}{2}\left(\|e_u^{\varepsilon,k}\|_{-1}^2 - \|e_u^{\varepsilon,0}\|_{-1}^2 + \sum_{j=1}^{k} \|e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}\|_{-1}^2 \right).
\end{aligned} \tag{2.41}$$

Now we proceed with the second term in (2.37).

$$\begin{aligned}
(I_2) =& \ \tau \textstyle\sum_{j=1}^{k} (e_\theta^{\varepsilon,j}, e_u^{\varepsilon,j}) \\
=& \ \textstyle\sum_{j=1}^{k} (\int_{t_{j-1}}^{t_j} (\beta(u^\varepsilon(t)) - \theta^j)dt, \frac{1}{\tau}\int_{t_{j-1}}^{t_j} (u^\varepsilon(s) - \beta^{-1}(\theta^j))ds) \\
=& \ \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta(u^\varepsilon(t)) - \theta^j, u^\varepsilon(t) - \beta^{-1}(\theta^j))dt \\
& + \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta(u^\varepsilon(t)) - \theta^j, \frac{1}{\tau}\int_{t_{j-1}}^{t_j} (u^\varepsilon(s) - u(s))ds)dt \\
& - \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta(u^\varepsilon(t)) - \theta^j, u^\varepsilon(t) - u(t))dt \\
& + \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta(u^\varepsilon(t)) - \theta^j, \frac{1}{\tau}\int_{t_{j-1}}^{t_j} (u(s) - u(t))ds)dt \\
=:& \ (I_{21}) + (I_{22}) + (I_{23}) + (I_{24}).
\end{aligned}$$

For (I_{21}), recalling (2.2) we have

$$\begin{aligned}(I_{21}) \geq & \frac{1}{2}\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}(\beta(u^\varepsilon(t))-\theta^j, u^\varepsilon(t)-\beta^{-1}(\theta^j))dt \\ & + \frac{C_1}{2}\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}\|\beta(u^\varepsilon(t))-\theta^j\|^2 dt.\end{aligned} \tag{2.42}$$

The estimates of (I_{22}) and (I_{23}) are identical and can be done as for (I_{123}).

$$\begin{aligned}|(I_{22})|, |(I_{23})| \leq & C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}\|\beta(u^\varepsilon(t))-\theta^j\|dt \\ \leq & C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))\tau^{\frac{1}{2}}\sum_{j=1}^{k}\left(\int_{t_{j-1}}^{t_j}\|\beta(u^\varepsilon(t))-\theta^j\|^2 dt\right)^{\frac{1}{2}} \\ \leq & \frac{C_1}{2\delta_2}\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}\|\beta(u^\varepsilon(t))-\theta^j\|^2 dt + \frac{C\delta_2}{2}(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2.\end{aligned} \tag{2.43}$$

In order to obtain upper bounds for (I_{24}) we repeat the steps for (I_{112}). Remembering the apriori estimates in Theorem 2.2.6 it follows that $\theta^j - \beta(\varepsilon) \in L^2(t_{j-1}, t_j; H_0^1(\Omega))$. Because $\partial_t u \in L^2(0,T; H^{-1}(\Omega))$ and $\beta(u) \in L^2(0,T; H_0^1(\Omega))$, we have

$$\begin{aligned}|I_{24}| \leq & \frac{1}{\tau}\left|\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}\int_{t_{j-1}}^{t_j}\left(\beta(u^\varepsilon(t))-\theta^j, \int_t^s \partial_r u(r)dr\right)dsdt\right| \\ \leq & \sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}\int_{t_{j-1}}^{t_j}\|\nabla(\beta(u^\varepsilon(t))-\theta^j)\|\|\partial_r u(r)\|_{-1}drdt \\ \leq & \tau\sum_{j=1}^{k}\|\partial_t u\|_{L^2(t_{j-1},t_j;H^{-1}(\Omega))}\|\nabla(\beta(u^\varepsilon(t))-\theta^j)\|_{L^2(t_{j-1},t_j;H_0^1(\Omega))} \\ \leq & C\tau.\end{aligned} \tag{2.44}$$

Considering now the right hand side in (2.37), (I_3) can be splitted into two terms.

$$\begin{aligned}(I_3) = & -\textstyle\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}(F(u(t))-F(u^\varepsilon(t)), \nabla Ge_u^{\varepsilon,j})dt \\ & -\textstyle\sum_{j=1}^{k}\int_{t_{j-1}}^{t_j}(F(u^\varepsilon(t))-F(\beta^{-1}(\theta^j)), \nabla Ge_u^{\varepsilon,j})dt =: (I_{31}) + (I_{32}).\end{aligned}$$

For the first term, because of the assumption (A3) on F, Lemma 2.2.10 can be used again in order to get

$$
\begin{aligned}
|(I_{31})| \leq & \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|F(u(t)) - F(u^\varepsilon(t))\| \|\nabla G e_u^{\varepsilon,j}\| dt \\
\leq & C \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1} \int_{t_{j-1}}^{t_j} (u(t) - u^\varepsilon(t), \beta(u(t)) - \beta(u^\varepsilon(t)))^{\frac{1}{2}} dt \\
\leq & C\tau\beta(\varepsilon)^{\frac{1}{2}} (\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^{\frac{1}{2}} \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1} \\
\leq & \frac{C}{\delta_{31}} \beta(\varepsilon)(\varepsilon + C_2(\varepsilon)\beta(\varepsilon)) + C\tau\delta_{31} \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1}^2.
\end{aligned}
\tag{2.45}
$$

Analogous, for (I_{32}) we can obtain

$$
\begin{aligned}
|(I_{32})| \leq & C \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1} \int_{t_{j-1}}^{t_j} (u^\varepsilon(t) - \beta^{-1}(\theta^j), \beta(u^\varepsilon(t)) - \theta^j)^{\frac{1}{2}} dt \\
\leq & C\tau^{\frac{1}{2}} \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1} \left(\int_{t_{j-1}}^{t_j} (u^\varepsilon(t) - \beta^{-1}(\theta^j), \beta(u^\varepsilon(t)) - \theta^j) dt \right)^{\frac{1}{2}} \\
\leq & \frac{C}{4\delta_{32}} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (u^\varepsilon(t) - \beta^{-1}(\theta^j), \beta(u^\varepsilon(t)) - \theta^j) dt + C\tau\delta_{32} \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1}^2.
\end{aligned}
\tag{2.46}
$$

Proceeding exactly in the same manner as before, the inequalities in (2.35) lead to the same estimates for (I_4). Inserting all the inequalities in (2.38) - (2.46) in (2.37), recalling (2.33) and choosing the δ's properly, we get

$$
\begin{gathered}
\|e_u^{\varepsilon,k}\|_{-1}^2 + \textstyle\sum_{j=1}^k \|e_u^{\varepsilon,j} - e_u^{\varepsilon,j-1}\|_{-1}^2 \\
+ \textstyle\sum_{j=1}^k \int_{t_{j-1}}^{t_j} (\beta(u^\varepsilon(t)) - \theta^j, u^\varepsilon(t) - \beta^{-1}(\theta^j)) dt + C_1 \sum_{j=1}^k \int_{t_{j-1}}^{t_j} \|\beta(u^\varepsilon(t)) - \theta^j\|^2 dt \\
\leq C\tau \|\partial_t u\|_{L^2(0,t_k;H^{-1}(\Omega))}^2 + C\tau + C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2 + C\tau \textstyle\sum_{j=1}^k \|e_u^{\varepsilon,j}\|_{-1}^2.
\end{gathered}
$$

Since $u \in H^1(0, T; H^{-1}(\Omega))$, applying the discrete Gronwall inequality yields

$$
\begin{aligned}
\|e_u^{\varepsilon,k}\|_{-1}^2 + & \int_0^{t_k} (\beta(u^\varepsilon) - \theta_\Delta, u^\varepsilon - \beta^{-1}(\theta_\Delta)) + \|\beta(u^\varepsilon) - \theta_\Delta\|_{L^2(0,t_k;L^2(\Omega))}^2 \\
\leq & C \left(\tau + (\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2 \right)
\end{aligned}
\tag{2.47}
$$

for any $k \geq 0$.

Finally, we notice that

$$
\begin{aligned}
\|\beta(u^\varepsilon) - \theta^j\| = & \ \|\beta(u) - \theta^j + \beta(\varepsilon)\| \\
\geq & \ |\|\beta(u) - \theta^j\| - \|\beta(\varepsilon)\|| \geq |\|\beta(u) - \theta^j\| - C\beta(\varepsilon)|
\end{aligned}
$$

and therefore

$$2\|\beta(u^\varepsilon) - \theta^j\|^2 \geq \|\beta(u) - \theta^j\|^2 - C\beta^2(\varepsilon).$$

This, together with (2.47), gives the desired result.

Remark 2.2.16 *In the situation described in Remark 2.2.15, the errors in Theorem 2.2.11 are bounded by $C(\tau + \varepsilon^2)$, with some positive constant C.*

Remark 2.2.17 *The above error estimates show that Scheme MTI behaves at least as good as the algorithms in [60], [73], [46], [4] or [28]. In fact, the schemes in [60], [73] or [4] are proved to have a convergence order $\tau^{1/2}$, while the order $\tau^{1/4}$ is demonstrated in [42], for the method proposed in [46]. For the nonlinear scheme in [28] the same error behaves asymptotically like $\tau^{0.3}$. The possibility to obtain an optimal convergence rate will be discussed later.*

Remark 2.2.18 *An alternative proof is given in [98]. This relies exclusively on the method proposed in [60] and can be applied only if the Problem P does not contain a convective term. Having now the result in Theorem 2.2.11, this procedure can be repeated in order to get error estimates involving $H^1(\Omega)$ norms (see, e.g., [73]).*

Remark 2.2.19 *The same result is obtained if the reaction term r is discretized explicitly. To see this, we notice that the last term in (2.37) becomes*

$$\begin{aligned}(\bar{I}_4) = & \sum_{j=1}^k \left(\int_{t_{j-1}}^{t_j} r(u(t)) - r(\beta^{-1}(\theta^{j-1}))dt, Ge_u^{\varepsilon,j}\right) \\ = & (I_4) + \tau \sum_{j=1}^k (r(\beta^{-1}(\theta^{j-1})) - r(\beta^{-1}(\theta^j)), Ge_u^{\varepsilon,j}).\end{aligned}$$

Now, the assumption on r in (A3) can be used to estimate the last sum

$$|(\bar{I}_4)| \leq |(I_4)| + C\tau \sum_{j=1}^k \|e_u^{\varepsilon,j}\|_{-1} (\beta^{-1}(\theta^j) - \beta^{-1}(\theta^{j-1}), \theta^j - \theta^{j-1})^{\frac{1}{2}}.$$

Because of the apriori estimates in (2.22) (but in the form mentioned in Remark 2.2.13 for avoiding a τ - ε dependence) we get

$$|(\bar{I}_4)| \leq |(I_4)| + C\tau + C\tau \sum_{j=1}^k \|e_u^{\varepsilon,j}\|_{-1}^2$$

and the rest follows as before.

Remark 2.2.20 *The error estimates for u are obtained in the $H^1(0,T;H^{-1}(\Omega))$ norm. This can be improved if an inequality of the form*

$$(\beta^{-1}(\theta)-\beta^{-1}(\psi))(\theta-\psi)\geq C(\beta^{-1}(\theta)-\beta^{-1}(\psi))^p$$

holds true for a positive constant C and an exponent $p>1$. Then, using the estimate for the scalar product in (2.47), an error estimate in the better $L^p(Q)$ norm can be obtained. For example, if $\beta(u)=u^m$, recalling (2.32), the estimate reads

$$\|u-\beta^{-1}(\theta_\Delta)\|^{m+1}_{L^{m+1}(Q_T)}\leq C\left\{\tau+\varepsilon^2+(C_2(\varepsilon)\beta(\varepsilon))^2\right\}$$

(see, e.g. [73], [28]).

Error estimates with linearized convection

The next result applies to the simplified nonlinear scheme.

Theorem 2.2.12 *In the setting of Theorem 2.2.11, the following estimates can be obtained for Scheme MTC*

$$\begin{aligned}\sup_{k=\overline{1,n}}\|e_u^{\varepsilon,k}\|^2_{-1}+\int_0^T(\beta(u^\varepsilon(t))-\theta_\Delta(t),u^\varepsilon(t)-\beta^{-1}(\theta_\Delta(t)))dt+\|\beta(u)-\theta_\Delta\|^2_{L^2(Q)}\\ \leq C\left\{\tau C_2(\varepsilon)^3+\varepsilon^2+(C_2(\varepsilon)\beta(\varepsilon))^2\right\}.\end{aligned}$$

Proof. The steps in proving Theorem 2.2.11 can be repeated here. First, the estimate in (2.44) becomes $C\tau C_2(\varepsilon)$ when F is not Lipschitz in $\beta(u)$, but this does not affect the final result. The main difference appears when dealing with the convection part, where we get

$$\begin{aligned}(\bar I_3)=(I_{31})+(I_{32})\quad&+\sum_{j=1}^k\int_{t_{j-1}}^{t_j}(F(\beta^{-1}(\theta^{j-1}))-F(\beta^{-1}(\theta^j)),\nabla Ge_u^{\varepsilon,j})dt\\ &-\sum_{j=1}^k\int_{t_{j-1}}^{t_j}((\beta^{-1})'(\theta^{j-1})f(\beta^{-1}(\theta^{j-1}))\nabla(\theta^{j-1}-\theta^j),Ge_u^{\varepsilon,j})dt,\end{aligned}$$

the last terms being denoted by (I_{33}) and (I_{34}). The estimates for (I_{31}) and (I_{32}) in the previous theorem are valid in this case too. As in Remark 2.2.19, (I_{33}) gives

$$\begin{aligned}|(I_{33})|\leq\ & C\tau\sum_{j=1}^k\|e_u^{\varepsilon,j}\|_{-1}(\beta^{-1}(\theta^j)-\beta^{-1}(\theta^{j-1}),\theta^j-\theta^{j-1})^{\frac12}\\ \leq\ & C\tau\sum_{j=1}^k(\beta^{-1}(\theta^j)-\beta^{-1}(\theta^{j-1}),\theta^j-\theta^{j-1})+C\tau\sum_{j=1}^k\|e_u^{\varepsilon,j}\|^2_{-1}.\end{aligned}$$

Recalling Remark 2.2.1, (I_{34}) can be bounded as follows

$$\begin{aligned}|(I_{34})|\leq\ & C\tau\sqrt{C_2(\varepsilon)}\sum_{j=1}^k\|\nabla(\theta^j-\theta^{j-1})\|\|Ge_u^{\varepsilon,j}\|\\ \leq\ & C\tau\sum_{j=1}^k\|\nabla Ge_u^{\varepsilon,j}\|^2+C\tau C_2(\varepsilon)\sum_{j=1}^k\|\nabla(\theta^j-\theta^{j-1})\|^2.\end{aligned}$$

Using the apriori estimates in Theorem 2.2.7, the rest of the proof is identical to the one for Theorem 2.2.11.

Remark 2.2.21 *As before, there is no difference when r is discretized explicitly. In fact, the remarks for the implicit case can be done here.*

Remark 2.2.22 *In case F is Lipschitz continuous in $\beta(u)$, the error estimates are identical to the ones for the implicit scheme. The factor $C_2(\varepsilon)^3$ multiplying the time step τ appears only due to the behaviour of f.*

Error estimates for the linear scheme

Now we turn our attention to the linear scheme, for which the following estimates can be obtained

Theorem 2.2.13 *In the setting of Theorem 2.2.11, assuming additionally the Lipschitz continuity of β' and the convexity of β, if θ^k solves for each $k > 0$ Problem WMTL, then*

$$\sup_{k=\overline{1,n}} \|e_u^{\varepsilon,k}\|_{-1}^2 + \int_0^T (\beta(u^\varepsilon(t)) - \theta_\Delta(t), u^\varepsilon(t) - \beta^{-1}(\theta_\Delta(t)))dt + \|\beta(u) - \theta_\Delta\|_{L^2(Q)}^2$$
$$\leq C\left\{\tau C_2(\varepsilon)^{\frac{7}{2}} + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2\right\},$$

where θ_Δ has been defined in Theorem 2.2.11.

Proof. The proof is similar to the ones above. Following the same steps, the relation in (2.37) becomes

$$\begin{aligned}
&\sum_{j=1}^k (u(t_j) - u(t_{j-1}) - \sigma_{j-1}(\theta^j - \theta^{j-1}), Ge_u^{\varepsilon,j}) + \tau \sum_{j=1}^k (\nabla e_\theta^{\varepsilon,j}, \nabla Ge_u^{\varepsilon,j}) \\
= &- \sum_{j=1}^k \left(\int_{t_{j-1}}^{t_j} F(u(t)) - F(\beta^{-1}(\theta^j))dt, \nabla Ge_u^{\varepsilon,j}\right) \\
&- \tau \sum_{j=1}^k \left(F(\beta^{-1}(\theta^j)) - F(\beta^{-1}(\theta^{j-1})), \nabla Ge_u^{\varepsilon,j}\right) \\
&+ \tau \sum_{j=1}^k \left((\beta^{-1})'(\theta^{j-1}) f(\beta^{-1}(\theta^{j-1})) \nabla(\theta^j - \theta^{j-1}), Ge_u^{\varepsilon,j}\right) \\
&+ \sum_{j=1}^k \left(\int_{t_{j-1}}^{t_j} r(u(t)) - r(\beta^{-1}(\theta^{j-1}))dt, Ge_u^{\varepsilon,j}\right).
\end{aligned} \tag{2.48}$$

Subtracting and adding in the first sum $\beta^{-1}(\theta^j) - \beta^{-1}(\theta^{j-1}) = \sigma(\theta^j, \theta^{j-1})(\theta^j - \theta^{j-1})$ ($\sigma(\theta^j, \theta^{j-1})$ being defined in (2.15)) the only thing we have to deal with in the left hand side is the following term

$$T \equiv \sum_{j=1}^{k} (\beta^{-1}(\theta^j) - \beta^{-1}(\theta^{j-1}) - \sigma_{j-1}(\theta^j - \theta^{j-1}), Ge_u^{\varepsilon,j}).$$

The Lipschitz continuity of β' yields

$$\begin{aligned}|\sigma(\theta^j,\theta^{j-1}) - \sigma_{j-1}| &\equiv \int_0^1 (\beta^{-1})'(s\bar{\theta}^j + (1-s)\theta^{j-1}) - (\beta^{-1})'(\theta^{j-1})ds \\ &\leq L_\beta C_2(\varepsilon)^2 \sigma_{j-1}|\theta^j - \theta^{j-1}|.\end{aligned}$$

This, together with the stability result in Theorem 2.2.8 and the maximum principle (showing that $\|Ge_u^{\varepsilon,j}\| \leq C$) gives

$$|T| \leq C\tau C_2(\varepsilon)^{\frac{7}{2}}.$$

Now everything follows as before. Recalling the apriori estimates in Theorem 2.2.8, for (I_{33}) in the previous proof we get

$$\begin{aligned}|(I_{33})| \leq\ & C\tau \sum_{j=1}^{k} (\beta^{-1}(\theta^j) - \beta^{-1}(\theta^{j-1}), \theta^j - \theta^{j-1}) + C\tau \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1}^2 \\ \leq\ & C\tau C_2(\varepsilon) \sum_{j=1}^{k} \|\theta^j - \theta^{j-1}\|^2 + C\tau \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1}^2 \\ \leq\ & C\tau C_2(\varepsilon) \sum_{j=1}^{k} \|\sqrt{\sigma_{j-1}}(\theta^j - \theta^{j-1})\|^2 + C\tau \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1}^2 \\ \leq\ & C\tau^2 C_2(\varepsilon)^{\frac{5}{2}} + C\tau \sum_{j=1}^{k} \|e_u^{\varepsilon,j}\|_{-1}^2.\end{aligned}$$

Similarly, (I_{34}) yields

$$|(I_{34})| \leq C\tau C_2(\varepsilon)^{\frac{5}{2}},$$

thus the difference due to (I_{24}) will not affect the result and the error estimates are obtained as before.

Remark 2.2.23 *Like in Theorem 2.2.9, in case F is Lipschitz in $\beta(u)$, Scheme MTL is stable without assuming the convexity of β. The same steps give slightly improved error estimates*

$$\begin{aligned}\sup_{k=\overline{1,n}} \|e_u^{\varepsilon,k}\|_{-1}^2 + \int_0^T (\beta(u^\varepsilon(t)) - \theta_\Delta(t), u^\varepsilon(t) - \beta^{-1}(\theta_\Delta(t)))dt + \|\beta(u^\varepsilon) - \theta_\Delta\|_{L^2(Q)}^2 \\ \leq C\left\{\tau C_2(\varepsilon)^2 + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2\right\}.\end{aligned}$$

However, we cannot obtain a result similar to the implicit scheme.

Remark 2.2.24 *The linearized case has a reduced convergence order, depending on the properties of β' and the choice of ε. For example, if $\beta(u) = u^m$, the errors behave like*

$\tau^{1/2m}$, which is quite bad when m goes to infinity. But practical computations show that the above estimates are quite pessimistic since in most of the cases the results obtained with this scheme were close to the ones generated by the nonlinear ones. The main advantage of Scheme MTL is due to its linearity, thus no iterations are needed, which compensates its theoretically worser convergence properties.

Optimal estimates

In this part, Problem P is considered without convection and reaction

Problem P′:

$$\begin{aligned} \partial_t u - \Delta\beta(u) &= 0, && \text{in } \ Q_T \equiv (0,T)\times\Omega, \\ u(x,0) = u_0(x) &\geq 0, && \text{in } \ \Omega, \\ u &= 0, && \text{on } \ \partial\Omega, \end{aligned} \tag{2.49}$$

the assumptions on β and u_0 remaining the same.

One development of semigroup theory for the problem above proceeds by considering implicit Euler approximations and solving a sequence of degenerate elliptic problems to estimate the solution of the parabolic problem at discrete times. The Crandall-Ligget theorem in [23] establishes the convergence of this procedure together with an order $O(\tau^{1/2})$. In spite of numerical evidence, it was long believed that this is the optimal rate of convergence of the backward Euler method for degenerate parabolic problems. Recently Rulla (in [85]) has shown that in the special case of a semigroup generated by a sub-gradient in a Hilbert space, this order of convergence becomes $O(\tau)$. This result was used in several papers (see, e.g., [51], [72], [86]) in order to establish better convergence results for other algorithms based on first order time discretization schemes, or for fully discrete approximations.

Since β is maximal monotone, supposing its range is all of $\mathbb{R}$, the operator $-\Delta\beta(\cdot)$ is maximal monotone on $H^{-1}(\Omega)$. In fact, it is the sub-gradient of

$$\varphi(u) = \begin{cases} \int_\Omega j(u(x))dx, & \text{if } u, j(u) \in L^1(\Omega) \\ +\infty, & \text{otherwise,} \end{cases}$$

where j is convex and lower semicontinuous such that β is its subdifferential - in our situation, its derivative - [16], Theorem 17.

Recall now the Euler implicit discretization for Problem P′,

Scheme EI:

$$\begin{aligned} u^k - u^{k-1} =& \tau\Delta\beta(u^k), \\ \beta(u^k)|_{\partial\Omega} =& 0 \end{aligned} \tag{2.50}$$

for $k = \overline{1,n}$ with u_0 from (2.49). In this case, u^k approximates $u(k\tau)$. Replacing u^k with $\beta^{-1}(\theta^k)$, Scheme MTI is identical to the one above up to the initial and boundary data. In (2.1) these are subject to a small shift in order to get a regular parabolic problem, which is not the case for (2.49). The result in [85] establishing the optimal rate of convergence for the implicit scheme reads

Theorem 2.2.14 *Let $\{u^k, k = \overline{0,n}\}$ be the solutions of the implicit scheme EI and u_τ, respectively θ_τ the piecewise constant functions interpolating $\{u^k, k = \overline{1,n}\}$ and $\{\beta(u^k), k = \overline{1,n}\}$. Then, for any $k \geq 0$,*

$$\begin{aligned} \sup_{k=\overline{1,n}} \|u(t_k) - u^k\|_{-1}^2 + \textstyle\int_0^T (u(t) - u_\tau(t), \beta(u(t)) - \theta_\tau(t))dt \\ + \tau\|\beta(u) - \theta_\tau\|^2_{L^2(0,T;H_0^1(\Omega))} \leq C\tau^2, \end{aligned}$$

where C depends on the H^1 norm of $\beta(u_0)$.

It is worth here to notice that the second term in the above inequality is positive due to the monotonicity of β. Now, if u_0 is a positive $L^\infty(\Omega)$ function, a semi-discrete maximum principle can be obtained easily, thus u^k stays above 0 for any $k > 1$. Based on the previous estimates, we can get similar results for Scheme MTI.

Theorem 2.2.15 *In the setting of Theorem 2.2.11, if θ^k solves for each $k > 0$ Problem WMTI without convection or reaction, then*

$$\|\beta(u) - \theta_\Delta\|^2_{L^2(Q)} + \tau\|\beta(u) - \theta_\Delta\|^2_{L^2(0,T;H_0^1(\Omega))} \leq C\left\{\tau^2 + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2\right\}.$$

Proof. First, Scheme EI is written in a weak form

$$(u^j - u^{j-1}, \varphi) + \tau(\nabla\beta(u^j), \nabla\varphi) = 0$$

(for all $\varphi \in H_0^1(\Omega)$, with $j = \overline{1,n}$). We want to compare the approximations yielded by the two implicit schemes, EI and MTI. Subtracting (2.6) from the above equality, summing up for $j = \overline{1,k}$, taking $\varphi = \beta(u^{\varepsilon,k}) - \theta^k \equiv \beta(u^k) + \beta(\varepsilon) - \theta^k \in H_0^1$ in the resulting difference and summing up again for $k = \overline{1,n}$ gives

$$
\begin{aligned}
(I_1) + (I_2) :=& \sum_{k=1}^{n} (u^k - \beta^{-1}(\theta^k), \beta(u^{\varepsilon,k}) - \theta^k) \\
&+ \tau \sum_{k=1}^{n} \sum_{j=1}^{k} (\nabla(\beta(u^{\varepsilon,j}) - \theta^j), \nabla(\beta(u^{\varepsilon,k}) - \theta^k)) \\
&= \sum_{k=1}^{n} (u^0 - \beta^{-1}(\theta^0), \beta(u^{\varepsilon,k}) - \theta^k) =: (I_3).
\end{aligned} \tag{2.51}
$$

We have to estimate each of the terms above. First, (I_1) can be decomposed in

$$
(I_1) = \sum_{k=1}^{n} (u^{\varepsilon,k} - \beta^{-1}(\theta^k), \beta(u^{\varepsilon,k}) - \theta^k) + \sum_{k=1}^{n} (u^k - u^{\varepsilon,k}, \beta(u^{\varepsilon,k}) - \theta^k) =: (I_{11}) + (I_{12}).
$$

Recalling the relation in (2.2), (I_{11}) gives

$$
(I_{11}) \geq C_1 \sum_{k=1}^{n} \|\beta(u^{\varepsilon,k}) - \theta^k\|^2 \geq \frac{C_1}{2} \sum_{k=1}^{n} \|\beta(u^k) - \theta^k\|^2 - \frac{C}{\tau} \beta(\varepsilon)^2, \tag{2.52}
$$

where the argument ending the proof for Theorem 2.2.11 has been applied here too.

As a consequence of Lemma 2.2.10 and of the maximum principle for the solution of the implicit scheme, (I_{11}) can be bounded as follows

$$
\begin{aligned}
|(I_{12})| \leq& C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon)) \sum_{k=1}^{n} \|\beta(u^{\varepsilon,k}) - \theta^k\| \\
\leq& \frac{C}{\tau}(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2 + \frac{C_1}{8} \sum_{k=1}^{n} \|\beta(u^k) - \theta^k\|^2.
\end{aligned} \tag{2.53}
$$

For (I_2) we make use of the identity in (2.19) in order to obtain

$$
(I_2) = \frac{\tau}{2} \sum_{k=1}^{n} \|\nabla(\beta(u^{\varepsilon,k}) - \theta^k)\|^2 + \frac{\tau}{2} \|\nabla \sum_{k=1}^{n} (\beta(u^{\varepsilon,k}) - \theta^k)\|^2. \tag{2.54}
$$

Now, because of (2.33), (I_3) can be bounded as (I_{12}). Inserting all the inequalities in (2.52) - (2.54) in (2.51) and multiplying everything with τ we arrive at

$$
\frac{C_1}{4} \sum_{k=1}^{n} \tau \|\beta(u^k) - \theta^k\|^2 + \frac{\tau^2}{2} \sum_{k=1}^{n} \|\nabla(\beta(u^{\varepsilon,k}) - \theta^k)\|^2 \leq C(\varepsilon + C_2(\varepsilon)\beta(\varepsilon))^2. \tag{2.55}
$$

The remaining part of proof can be completed using the above relation, Theorem 2.2.14 and a property of norms ($\|a + b\| \leq \|a\| + \|b\|$).

Remark 2.2.25 *This result shows that if the implicit Euler discretization has a linear order of convergence, the same holds also for Scheme MTI.*

Remark 2.2.26 *The error estimates for the linear scheme MTL can be improved in a similar manner. In fact, in the results stated in Theorem 2.2.13, the time step τ appears with the exponent 2.*

Remark 2.2.27 *Even though the above estimates can be extended to other cases (for example, monotone perturbations of a subgradient, [85]), they cannot cover the whole range of problems considered in this chapter. Therefore we have maintained the analysis in the previous sections and complete it here with the optimal results, which are a simple consequence of the striking result of Rulla.*

2.3 Full discretization

Up to now we have been concerned exclusively on the time discretization, which generates at each time step an elliptic problem. In order to get approximate solutions for Problem P, we have to proceed with the spatial discretization, leading to an algebraic system.

There are several possibilities to approximate the resulting elliptic problems WMTI, WMTC or WMTL. The simplest way consists in applying a finite difference discretization also to the spatial derivatives, as done (for example) in [97] - where an explicit method was used in connection with the theory of semigroups, or in older papers like [61]. Generally, it is not so easy to apply this approach in case Ω is a complicated domain, or to non-Dirichlet boundary conditions. A natural extension of the finite differences are the finite volumes ([3]). In its strict sense, this method was considered for degenerate parabolic equations in [35], leading to convergence proofs based on compactness arguments which exclude any estimates for the error.

Another spatial discretization consists in the finite elements. For theoretical aspects regarding this method we refer to [19] or [20]. This technique - in its different variants - is widely used today because of the advantages resulting from the variational formulation of the discrete problems and its applicability to arbitrary domains (in relation with unstructured and adaptively refined meshes). The classical piecewise linear (and continuous) Galerkin finite element method was analysed in [84] and [50]. Combined with the optimality results in [85], this analysis was extended in [86]. A Raviart-Thomas mixed finite element method ([81]) is considered in [4].

The standard finite element approach creates several problems. First, it requires a severe restriction on the discretization parameters in order to get stable solutions w.r.t. maximum norm. This stands out especially in our situation, where the zero order terms resulting from the time discretization are multiplied with factors which become extremely large in the degeneracy region. Next, the evaluation of the integrals corresponding to the lowest order terms creates additional difficulties. This is why the lumped mass version of the finite element method ([44], [92]) was adopted in many papers dealing with such types of problems (see, e.g, [8], [22], [29], [43], [71], [68], [69], [73], [74], [89]). This allows an important gain in stability, provided the convection terms are discretized in an appropriate way.

Our approach relies on the box method ([40]), which is a combination between the lumped mass finite elements and the finite volumes. Concretely, the solution is obtained in the space of piecewise linear continuous functions, while the dual space consists on the piecewise constant ones, but on a dual mesh (the boxes). In case the convection is not present, an equivalence between the box scheme and the lumped mass finite element approach can be proven in particular cases ([40], [15]). Unfortunately we were not able to obtain a convergence proof for Problem P with convection. Moreover, since a box method has been applied, the complete discretization applies to convective terms in a divergence form, which is not the case for the schemes MTC and MTL. Even the analysis can be carried out in a similar manner for the schemes in the divergence form, we have not considered this situation because we could not prove a maximum principle under reasonable assumptions. With the upwind procedure proposed below, this property holds at the fully discrete level.

2.3.1 Assumptions on the triangularization

We consider S_h a decomposition of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh-size (the diameter of the largest triangle in S_h). N_h represents the number of vertices. This decomposition is always assumed regular, meaning that the ratio of the diameter of $T \in S_h$ to the diameter of the largest sphere inscribed in T is bounded independently on h. Moreover, $\overline{\Omega} = \cup_{T \in S_h} T$, hence Ω is polygonal. Therefore the effect of the approximation of a non-polygonal domain by a finite element decomposition is neglected, allowing us to avoid an excess of technicalities (a complete analysis in this sense can be found, e.g. in [29] or [73]). In what follows V_h denotes the piecewise linear finite element space defined for

S_h, so $V_h = \{\varphi \in C(\overline{\Omega}) : \varphi|_T$ is linear for all $T \in S_h\}$. If a Dirichlet boundary condition is imposed, the elements satisfying it are contained in $V_h(g) = \{\varphi \in V_h : \varphi|_{\partial\Omega} = g|_{\partial\Omega}\}$. The corresponding discrete semi-inner product is defined by

$$(\chi, \varphi)_h \equiv \sum_{T \in S_h} \int_T I_h(\chi\varphi) dx = \sum_{T \in S_h} \frac{meas(T)}{d+1} \sum_{k=0}^{d} \chi(A_k^T)\varphi(A_k^T), \tag{2.1}$$

where the last equality holds if χ and φ are piecewise linear continuous functions. I_h stands for the local linear interpolation operator and $A_k^T, k = \overline{0, d}$ are the vertices of the element T. Restricted to V_h, $(\cdot, \cdot)_h$ is an inner product equivalent to $(\cdot, \cdot)$ satisfying ([80])

$$\|\chi\|^2 \leq (\chi, \chi)_h \equiv \|\chi\|_h^2 \leq C\|\chi\|^2 \tag{2.2}$$

for any $\chi \in V_h$, where $\|\cdot\|$ is the usual L^2 norm in Ω and C does not depend on h. Following [22], the effect of numerical integration is measured by

$$|(\chi, \psi) - (\chi, \psi)_h| \leq Ch^{1+s}\|\chi\|_s\|\psi\|_1 \tag{2.3}$$

for any $\chi, \psi \in V_h$ and $0 \leq s \leq 1$ ($\|\cdot\|_s$ stands for the norm in H^s).

In what follows we make use of the L^2 projection operator Π_h onto V_h, which, for any $\theta \in L^2(\Omega)$, is defined by

$$(\Pi_h\theta, \psi)_h = (\theta, \psi) \tag{2.4}$$

for all ψ in V_h. Starting from the definition of the norm in H^{-1}, using the estimates in (2.3), since $\|\Pi_h\theta\| \leq C\|\theta\|$, it is easy to show that Π_h satisfies

$$\|\theta - \Pi_h\theta\|_{-s} \leq Ch^{r+s}\|\theta\|_r, \tag{2.5}$$

where $0 \leq r, s \leq 1$.

In our approach the maximum principle plays a crucial role. Hence the (weak) acuteness of S_h is assumed. In its stronger sense, this property means that the projection of the vertices of any d-simplex onto the hyperplane containing the opposite face lies in the closure of this face. This preserves the maximum principle for the Laplacian ([21]), since the following holds

$$\int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \leq 0, \quad \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_i > 0,$$

where $\{\varphi_i, i = \overline{1, N_h}\}$ is the corresponding basis for V_h and $i \neq j$. In this case, the stiffness matrix for the Laplacian is an M-matrix ([93]).

In fact, if S_h is acute, the contribution of each element to the global stiffness matrix has nonnegative diagonal entries and non-positive off diagonal ones. This is not a necessary condition but certainly a sufficient one. In two dimensions, acuteness can be relaxed to a weaker one, namely the sum of the opposite angles with respect to the common side of any pair of triangles does not exceed π (see, e.g., [91]). The major disadvantage with this assumption appears in connection with an adaptive mesh refinement. In general, local refinements are not allowed at least theoretically. If the weaker acuteness is accepted, an unstructured mesh following the free boundaries can be generated ([68], [69]), or, if the triangularization consists only of right triangles, the bisection method can be applied for a local refinement. Here we consider the (weakly) acute case and the mesh is quasi-uniform.

In the analysis of the semi-discrete approximation the Green operator G defined in (2.34) is used. For convenience, we assume G is regular ([19], p. 138), namely for any $\psi \in L^2(\Omega)$ we have $G\psi \in H^2(\Omega)$ and $\|G\psi\|_2 \leq C\|\psi\|_0$. This property holds, for instance, whenever mixed boundary conditions are avoided. Analogous, the corresponding discrete operator $G_h : H^{-1} \to V_h(0)$ (the finite element approximation to G) is defined by

$$(\nabla G_h\psi, \nabla\varphi) = (\psi, \varphi), \tag{2.6}$$

for all $\varphi \in V_h(0) \subset H_0^1(\Omega)$, where ψ is taken in $H^{-1}(\Omega)$. Since G and S_h are regular, standard error analysis for finite element approximations of elliptic equations give the following property

$$\|(G - G_h)\psi\|_s \leq Ch^{2-s-r}\|\psi\|_{-r}, \tag{2.7}$$

with $0 \leq s, r \leq 1$.

G_h generates on $V_h(0)$ an inner product and a norm related to the one in H^{-1}

$$(\psi, \varphi)_{-1,h} = (G_h\psi, \varphi) \quad \text{and} \quad \|\psi\|_{-1,h}^2 = (G_h\psi, \psi). \tag{2.8}$$

Similar to the continuous case, if $\psi \in V_h(0)$ we have

$$\|\nabla G_h\psi\| = \|\psi\|_{-1,h}^2, \quad \|\psi\|_{-1,h} \leq C\|\psi\|. \tag{2.9}$$

Using the approximation error in (2.7) it is easy to get the following

$$\|\psi\|_{-1}^2 \leq \|\psi\|_{-1,h}^2 + Ch^2\|\psi\| \quad \text{and} \quad \|\psi\|_{-1,h}^2 \leq \|\psi\|_{-1}^2 + Ch^2\|\psi\|. \tag{2.10}$$

For the sake of simplicity we consider the dual mesh given by the Donald diagram, which is based on the barycentres of the simplices building the initial triangularization

Figure 2.1: Dual box centered in P.

([3]). The nodal basis for B_h - the piecewise constant test function space - consists of the functions $\{\phi_i, i = \overline{1, N_h}\}$ which are equal to 1 inside the box around the node i and 0 outside of it. Figure 2.1 presents the two dimensional dual box B_P centered in the vertex P, its boundary being denoted by ∂B_P. We do not give here the definition of the fully discrete counterparts to the schemes MTI, MTC or MTL since, as mentioned above, the spatial discretization uses a divergence form of the equations. Moreover, the error estimates are obtained only in the absence of F. This lack in the analysis is due to the upwinding of the convection term, which does not allow us to write the fully discrete equations in a variational form.

2.3.2 The fully discrete problems

As seen before, the time discretization leads to a sequence of elliptic problems. The solutions are sought in an infinite dimensional Sobolev space. Here we consider the problems resulting after performing the spatial discretization. This makes the dimension of each problem finite, hence the solution can be represented as a finite array (containing the nodal values of the approximation) and the elliptic problems are transformed in finite systems. But these systems have to maintain the main features of the semi-discrete problems, like existence, uniqueness and - in our case - the maximum principle (which is synonym below with the stability in the maximum norm). Since this plays a crucial role here, we describe first an upwind method for the convection terms. Afterwards the convergence of the fully discrete counterpart of Iteration IJK without convection is analysed.

An upwind method

In the setting above, an appropriate discretization of the convection term is requested in order to obtain a stable scheme. If $\{\varphi_i, i = \overline{1, N_h}\}$ is the nodal basis for the piecewise linear finite element space V_h and $\{\phi_i, i = \overline{1, N_h}\}$ the one for B_h - the dual piecewise constant finite volumes, the lumped mass box method for the implicit scheme is obtained by integrating in (2.1) over an arbitrary box B_P, where P is an interior vertex - since we deal here with Dirichlet boundary conditions - ([40]). This can be seen as a weak form of the problem, but the test functions are the elements of B_h. Recalling (2.1) and the definition of the nodal basis of B_h, if i_P stands for the index of the vertex P, the first term (where no derivatives are involved) yields

$$\begin{aligned}\int_{B_P} I_h((\beta^{-1}(\theta^k)-\beta^{-1}(\theta^{k-1}))\phi_{i_P})dx &= meas(B_P)(\beta^{-1}(\theta^k(P))-\beta^{-1}(\theta^{k-1}(P))) \\ &= (\beta^{-1}(\theta^k(P))-\beta^{-1}(\theta^{k-1}(P)))\sum_{T\in T_P}\frac{meas(T)}{d+1},\end{aligned} \tag{2.11}$$

where T_P contains the simplices in S_h having P as a vertex. The reaction term gives analogously

$$\tau\int_{B_P} I_h(r(\beta^{-1}(\theta^k))\phi_{i_P})dx = r(\beta^{-1}(\theta^k(P)))\tau\sum_{T\in T_P}\frac{meas(T)}{d+1}. \tag{2.12}$$

Next, for the second order term we make use of Greens's identity

$$\int_{B_P}\phi\Delta\varphi dx = -\int_{B_P}\nabla\phi\cdot\nabla\varphi dx + \int_{\partial B_P}\phi\partial_n\varphi d\omega,$$

where n is the outward normal for the box B_P, defined edgewise and $\partial_n\varphi$ the normal derivative of φ. Applied to the second order term in (2.1) it leads to

$$-\tau\int_{B_P}\Delta\theta^k\phi_{i_P}dx = -\int_{\partial B_P}\partial_n\theta^k d\omega, \tag{2.13}$$

the first integral in the previous formula being 0 because ϕ_{i_P} is constant on B_P.

The remaining part is the convection term, for which we get

$$-\tau\int_{B_P}\nabla F(\beta^{-1}(\theta^k))\phi_{i_P}dx = -\tau\int_{\partial B_P}(n\cdot F(\beta^{-1}(\theta^k))d\omega.$$

This is rewritten in a form more convenient for a linearization procedure. First, $F(\beta^{-1}(\theta^k))$ is identical to $\dfrac{F(\beta^{-1}(\theta^k))}{\theta^k}\theta^k$. Next, for any simplex T, the ratio above is replaced by its mean value in T

$$\bar{v}_T^k := \frac{1}{meas(T)}\int_T\frac{F(\beta^{-1}(\theta^k))}{\theta^k}dx.$$

In practical computations a quadrature formula on T substitutes the integral. Denoting by $\bar{v}^k$ the piecewise constant function taking on each simplex T the value $\bar{v}_T^k$ yields

$$-\tau\int_{B_P}\nabla F(\beta^{-1}(\theta^k))\phi_{i_P}dx \approx -\tau\int_{\partial B_P}(n\cdot\bar{v}^k)\theta^k d\omega. \tag{2.14}$$

Here we have to make a brief comment. The degeneracy appears in 0. Since $\beta(0) = 0$, some problems may appear in the previous step. This situation is avoided because of the shift of the data and of the maximum principle.

Up to now we have projected the equation in (2.1) onto B_h. In order to obtain a complete discretization, a finite dimensional approximation of the solution θ^k has to be found. This is denoted by θ_h^k and is sought in $V_h(\beta(\varepsilon))$. Therefore θ_h^k can be decomposed w.r.t. the nodal basis $\{\varphi_j, j = \overline{1, N_h}\}$ as follows

$$\theta_h^k = \sum_{j=1}^{N_h} \Theta_j^k \varphi_j \equiv \Theta^k \cdot \Phi, \tag{2.15}$$

where Θ^k and Φ are N_h-dimensional vectors. The elements of Θ^k are the nodal values of θ_h^k (so $\theta_h^k(A_j) = \Theta_j^k$), while Φ contains the elements of the nodal basis of V_h.

Since the boundary value problem P is of Dirichlet type, some of the elements of Θ^k are already known (those corresponding to the vertices lying on the boundary of the domain Ω), while the others are the unknowns. Testing with the piecewise constant functions centered in the interior vertices a (nonlinear) system of equations is obtained. The steps described up to now (especially the one in (2.14)) make sense only in connection with the maximum principle. Therefore this property should be maintained also at the fully discrete level.

Now the we can proceed with the discrete counterparts of the relations in (2.11) - (2.14). Testing with ϕ_i, the first ones become

$$\begin{aligned} \textstyle\int_{B_i} I_h(\phi_i(\beta^{-1}(\theta_h^k) - \beta^{-1}(\theta_h^{k-1})))dx = &\quad meas(B_i)(\beta^{-1}(\Theta_i^k) - \beta^{-1}(\Theta_i^{k-1})), \\ \tau \textstyle\int_{B_i} I_h(\phi_i r(\beta^{-1}(\theta_h^k)))dx = &\quad \tau\ meas(B_i) r(\beta^{-1}(\Theta_i^k)), \end{aligned}$$

where B_i is the control volume centered in the vertex A_i. In both cases the contribution to the global stiffness matrix is diagonal and identical to the one obtained when applying the lumped mass finite element method. Its entries are $M_{ij} = \delta_{ij} meas(B_i)$, δ_{ij} standing for Kronecker's symbol.

The diffusion term gets

$$-\tau \int_{B_i} \Delta\theta_h^k \phi_i dx = -\tau \int_{\partial B_i} \partial_n \theta_h^k d\omega = -\sum_{j=1}^{N_h} \tau \int_{\partial B_i} \partial_n \varphi_j d\omega \ \ \Theta_j^k.$$

In the setting mentioned previously, the stiffness matrix due to the diffusion is identical to the one resulting in a classical Galerkin finite element formulation. In two dimensions this statement was proven in [40], while the general case is considered in [15]. Therefore, the second order term adds an M-matrix to the stiffness matrix, namely $D_{ij} = -\tau \int_{\partial B_i} \partial_n \varphi_j d\omega = \tau(\nabla\varphi_i, \nabla\varphi_j)$.

Finally we have to consider the convection term. Doing as before gives

$$-\tau \int_{B_i} \nabla F(\beta^{-1}(\theta^k))\phi_i dx \approx -\tau \int_{\partial B_i} (n \cdot \bar{v}^k)\theta_h^k d\omega = -\sum_{j=1}^{N_h} \tau \int_{\partial B_i} (n \cdot \bar{v}^k)\varphi_j d\omega \quad \Theta_j^k,$$

but θ_h^k replaces θ^k when computing $\bar{v}_T^k$. This can be seen as an approximation of the flux of θ_h^k over the boundary of the control volume B_i. The convective term contributes to the stiffness matrix with the entries $C_{ij} = -\tau \int_{\partial B_i} (n \cdot \bar{v}^k)\varphi_j d\omega$. Generally, nothing can be said about the sign of these values. Since a stable discretization is desired, we have to continue with an upwinding procedure. To do so we notice that a given simplex $T \in S_h$ intersecting B_i contains d (linear) portions of the box boundaries, denoted by $\partial B_{i,l}^T$ ($l = \overline{1,d}$). By renaming the vertices of T we can say that $\partial B_{i,l}^T$ separates the vertex $A_0^T = A_i$ from the vertex A_l^T. Because we work here with d-simplices, for any vertex A of T an index l can be found such that $A = A_l^T$. Thus we get

$$-\tau \int_{\partial B_i} (n \cdot \bar{v}^k)\varphi_j d\omega = -\tau \sum_{T \in S_h,\ T \cap B_i \neq \phi} \sum_{l=1}^{d} \int_{\partial B_{i,l}^T} (n_{i,l}^T \cdot \bar{v}_T^k)\varphi_j d\omega,$$

with $n_{i,l}^T$ being the outer normal to $\partial B_{i,l}^T$ (pointing from A_0^T to A_l^T). Therefore the following holds

$$C_{ij} = \sum_{T \in S_h, T \cap B_i \neq \phi} C_{ij}^T = \sum_{T \in S_h, T \cap B_i \neq \phi} \sum_{l=1}^{d} \Phi_{i,l}^{T,j},$$

where $\Phi_{i,l}^{T,j}$ is given by $\Phi_{i,l}^{T,j} := -\tau \int_{\partial B_{i,l}^T} (n_{i,l}^T \cdot \bar{v}_T^k)\varphi_j d\omega$. This means that the values of θ_h^k in $\bar{A}_0$ and $\bar{A}_l$ have the same weight in the approximation of the flux over $\partial B_{i,l}^T$, which contradicts the physical point of view requesting that the flux should be approximated essentially on the basis of the vertices lying against sense of the flow. Therefore in what follows an upwinding procedure is considered. This is slightly different from the usual approach in the box method ([15], [9]) and is inspired by the upwinding methods for finite differences.

For getting a stable discretization we have to consider the sign of the term $-\tau(n_{i,l}^T \cdot \bar{v}_T^k)$ (denoted by γ_T^l), which is constant inside T. In case this is negative, the flow over $\partial B_{i,l}^T$ is inside the box B_i, while in the opposite situation the flow over $\partial B_{i,l}^T$ is outside B_i. Since $\partial B_{i,l}^T$ lies between $A_i \equiv A_0^T$ and A_l^T (vertex indexed in the initial numbering by i_l, so

$A_{i_l} \equiv A_l^T$), it is natural to assume that the flow over $\partial B_{i,l}^T$ outside B_i is inside B_{i_l}. As a consequence, in building the upwind discretization matrix $\Phi_{i,l}^{T,j} \leq 0$ is added to the entry corresponding to the ansatz-function φ_j when testing with ϕ_i

$$U_{i,j}^{new} = U_{i,j}^{old} + \Phi_{i,l}^{T,j}.$$

If γ_T^l is positive, $\Phi_{i,l}^{T,j} \geq 0$ contributes to the diagonal entry regarding ϕ_{i_l}. In fact, in this situation we have

$$U_{i_l,i_l}^{new} = U_{i_l,i_l}^{old} + \sum_{j,\ supp(\varphi_j)\cap T\neq\phi} \Phi_{i,l}^{T,j},$$

since the sign of $\Phi_{i,l}^{T,j}$ is the same for all j. Here the following remark is necessary. If γ_T^l is negative, when testing with ϕ_{i_l} the same flux over $\partial B_{i_l,p}^T$ - p being now the index of A_i when the numbering inside T starts at A_{i_l} - has the opposite sign because of the change in the orientation of the normal vector. As a consequence, $\Phi_{i_l,p}^{T,j} = -\Phi_{i,l}^{T,j} \geq 0$ is added to $U_{i,i}$ for all j, thus

$$U_{i,i}^{new} \; = \; U_{i,i}^{old} + \sum_{j,\ supp(\varphi_j)\cap T\neq\phi} \Phi_{i_l,p}^{T,j} \; = \; U_{i,i}^{old} - \sum_{j,\ supp(\varphi_j)\cap T\neq\phi} \Phi_{i,l}^{T,j}.$$

In this way, the contribution of φ_i to the upwind discretization matrix is lost since the term $\Phi_{i,l}^{T,i}$ is once added to $U_{up,up}$ and afterwards subtracted from the same element, with $up = i$ in case γ_T^l is negative or $up = i_l$ otherwise. To avoid this we make a further modification. If $\gamma_T^l < 0$ (hence the flow over $\partial B_{i,l}^T$ is inside B_i), $\Phi_{i,l}^{T,i} \leq 0$ adds to U_{i,i_l}, otherwise to U_{i_l,i_l}. In fact, if $\Phi_{i,l}^{T,i} \leq 0$ is added to U_{i,i_l}, the opposite will be added to $U_{i,i}$ because of the change of sign of the flux corresponding to the neighbouring box B_{i_l}. This means that if φ_i contributes to the flow over $\partial B_{i,l}^T$ outside B_i, the effect is inside B_{i_l} and reciprocally.

Now a more precise formula for the local upwind matrix (denoted by U_T) can be given

$$U_{i,j}^T = \begin{cases} -\displaystyle\sum_{k,\ supp(\varphi_k)\cap T\neq\phi} \sum_{l=1}^{d} \min\{\Phi_{i,l}^{T,k}, 0\}, \ \text{ if } i = j \\ \min\{\Phi_{i,l_j}^{T,i}, 0\} + \displaystyle\sum_{l=1}^{d} \min\{\Phi_{i,l}^{T,j}, 0\}, \ \text{ if } i \neq j \end{cases}, \tag{2.16}$$

where l_j stands for the (local) index l in $1, \ldots, d$ for which $A_j = A_{l_j}^T$. To be more rigorous we should make a distinction between the finite elements φ_j corresponding to the vertices of T (for which an l_j exists) and the other ones. But since only piecewise linear finite elements on d-simplices are considered, if A_j is not a vertex of T, the support of φ_j does not intersect this simplex and therefore the corresponding entry in U^T is 0. Adding the local upwind matrices U^T for all simplices T, the global upwind matrix (U) is obtained.

It is easy to see that U^T has zero row sums, with nonnegative diagonal and nonpositive off diagonal entries. For elements with vertices on the boundary, the rows and columns corresponding to Dirichlet vertices are ignored in computing the global upwind matrix. Thus U will be an irreducible, diagonally dominant M-matrix w.r.t. its rows, leading to the discrete maximum principle.

Generally, such upwinding procedures generate schemes of lower order. But we need this approach only in that regions where the convection dominates the problem. It is possible to combine the upwind strategy with the original one. If we consider the local stiffness matrices corresponding to the simplex T, this is obtained from the local diffusion matrix D^T, the original convection matrix C^T and the upwind one U^T. As mentioned above, A^T and U^T possess the M-property. A possible strategy is to consider a convex combination of the two convection matrices such that, with the aid of D^T, the resulting local stiffness matrix is still an M-matrix. Thus,

$$A^T = D^T + s_T C^T + (1 - s_T) U^T,$$

where $0 \leq s_T \leq 1$ is a local parameter. By requesting for A^T to have positive diagonal entries and negative off diagonal ones we obtain local limitations for s_T. Since we want to obtain an approximation having as much as possible a higher order, s_T should be taken equal to the local upper limit above (but not greater than 1). However, if the convection dominates the problem - this happens at least around the free boundary - the weight of the upwind discretization matrix is increasing so s_T goes to 0. Therefore the approximation becomes worser, but still stable.

The discrete maximum principle

In what follows we consider the problem without convection. Since the dual box mesh is the Donald diagram and mass lumping is used, the box scheme for the semi-discrete Problem WMTI is equivalent to a modified finite element formulation ([40], [15]).

Problem WMDI. For any $1 \leq k \leq n$, find $\theta_h^k \in V_h(\beta(\varepsilon))$ such that for all $\varphi \in V_h(0)$ the following holds true

$$(\beta^{-1}(\theta_h^k) - \beta^{-1}(\theta_h^{k-1}), \varphi)_h + \tau(\nabla\theta_h^k, \nabla\varphi) = \tau(I_h r(\beta^{-1}(\theta_h^k)), \varphi)_h, \tag{2.17}$$

where $\underline{k}$ is either k or $k-1$, depending on the discretization of the reaction term.

Remark 2.3.1 *For each $k \geq 0$, let u_h^k be the piecewise linear interpolant of $\beta^{-1}(\theta_h^k)$. Hence we have*

$$\theta_h^k \equiv I_h(\beta((u_h^k)) \quad \textit{and} \quad (u_h^k, \varphi)_h = (\beta^{-1}(\theta_h^k), \varphi)_h$$

for all $\varphi \in V_h$.

In the above definition the initial data is not given. This is done now, recalling the perturbed initial data described in (2.1) and (2.3)

$$u_h^0 = \Pi_h(\beta^{-1}(\theta^0)), \quad \theta_h^0 = I_h\beta(u_h^0) \tag{2.18}$$

It is easy to show that u_h^0 is bounded by ε and M (the essential supremum of u_0). In fact, replacing ψ with an element of the basis of V_h in (2.4), an explicit formula can be given for the nodal values of u_h^0

$$u_{hi}^0 \equiv u_h^0(A_i) = \frac{d+1}{meas\{\varphi_i > 0\}} \int_\Omega \beta^{-1}(\theta^0)\varphi_i dx$$

for all $i = \overline{1, N_h}$. Therefore we have $\theta_h^0 \in V_h(\beta(\varepsilon)) \cap V_0$, where V_0 is defined in (2.9).

Remark 2.3.2 *This particular choice for the discrete initial data requests some quadrature formula for computing the above integrals. We do not consider the errors induced by this procedure since they are mostly of the same order as the global one. Moreover, in the computations we have used the linear interpolation of θ_0 instead of the one defined in (2.18) and the results were almost the same.*

Now we are able to show the discrete maximum principle.

Lemma 2.3.1 *Assume (A1), (A3), $\theta_h^{k-1} \in V_h \cap V_{k-1}$ and $r(u) \geq 0$ for all u. If S_h is of acute type and a solution of Problem WMDI exists, then it belongs to $V_h \cap V_k$.*

Proof. The proof makes use of the properties of the discretization matrices. Since S_h is acute, The stiffness matrix A of the Laplace operator has the properties

$$A_{ii} > 0 \geq A_{ij}, \quad \sum_{j=1}^{N_h} A_{ij} \geq 0$$

for all $i, j = \overline{1, N_h}$, $i \neq j$. Let $v_h^k := u_h^k - \varepsilon$, $\psi_h^k := \theta_h^k - \beta(\varepsilon)$, then the equation in (2.17) becomes

$$(v_h^k, \varphi)_h + \tau(\nabla\psi_h^k, \nabla\varphi) = (v_h^{k-1}, \varphi)_h + \tau(P_h r(\beta^{-1}(\theta_h^k)), \varphi)_h \tag{2.19}$$

In order to get the lower bounds for θ_h^k it is enough if we obtain the positiveness of v_h^k or ψ_h^k. To do so, let i_m be the index of the vertex where v_h^k attains its minimum (since v_h^k is piecewise linear, its minimum is reached in a vertex). Assuming now $v_{h\,i_m}^k < 0$, since β is monotone and $v_{h\,i_m}^k \leq v_{hj}^k$, we get $\psi_{h\,i_m}^k \leq 0$ and $\psi_{h\,i_m}^k \leq \psi_{hj}^k$ for all $j = \overline{1, N_h}$. Taking $\varphi = \varphi_{i_m}$ in (2.19) yields

$$\tau \sum_{j=1}^{N_h} A_{i_m j}\psi_{hj}^k = (meas\{\varphi_{i_m} > 0\})(v_h^{k-1}{}_{i_m} - v_{h\,i_m}^k) + \tau(P_h r(\beta^{-1}(\theta_h^k)), \varphi_{i_m})_h > 0,$$

where for the last inequality the positiveness of v_h^{k-1} and r have been used. Recalling the properties of A and the minimality of $\psi_{h\,i_m}^k$ we get

$$\sum_{j=1}^{N_h} A_{i_m j}\psi_{hj}^k = \psi_{h\,i_m}^k \sum_{j=1}^{N_h} A_{i_m j} + \sum_{j=1, j\neq i_m}^{N_h} A_{i_m j}(\psi_{hj}^k - \psi_{h\,i_m}^k) \leq 0.$$

This contradicts the previous result and therefore the assumption on $v_{h\,i_m}^k$ has to be false.

The upper bounds for θ_h^k can be obtained in the same manner. Then we arrive in a situation analogous to the one in the semi-discrete case, resulting similar restrictions on the time step.

Remark 2.3.3 *Since the initial data described in (2.18) lies in $V_h \cap V_0$, a mathematical induction argument shows that the conclusion of the above lemma holds true for the whole sequence of discrete solutions.*

Remark 2.3.4 *As in the semi-discrete case, the global positivity of r is not necessary if r is discretized explicitly. Moreover, the restrictions on the time step dissapear here too.*

Remark 2.3.5 *Lemma 2.3.1 also holds true if the convection term is present, but it has to be discretized in an upwinding manner.*

An iterative method for the nonlinear discrete equations

Here we consider an iterative method for solving the nonlinear discrete system in (2.17). In fact this is the fully discrete counterpart of Iteration IJK, which is an alternative to Iteration ISm, having better convergence properties from practical point of view. The reaction term is discretized explicitly, while convection is not present here.

For each $i > 0$, the discrete formulation of the problems associated to Iteration IJK reads

Problem WIJK. Find $\bar{\theta}_h^i \in V_h(\beta(\varepsilon))$ such that

$$(\sigma(\bar{\theta}_h^{i-1}, \theta_h^{k-1})(\bar{\theta}_h^i - \theta_h^{k-1}), \varphi)_h + \tau(\nabla\bar{\theta}_h^i, \nabla\varphi) = \tau(I_h r(\beta^{-1}(\theta_h^{k-1})), \varphi)_h \tag{2.20}$$

holds true for all $\varphi \in V_h(0)$, where $\bar{\theta}_h^0 = \theta_h^{k-1}$, $\sigma(\bar{\theta}_h^0, \theta_h^{k-1}) = (\beta^{-1})'(\theta_h^{k-1})$ and θ_h^{k-1} belongs to $V_h(\beta(\varepsilon))$.

Taking the elements of the nodal basis as test functions in (2.20) gives the linear system

$$(M_h^{i-1} + \tau A_h)\bar{\Theta}^i = M_h^{i-1}\Theta^{k-1} + \tau R_h^{k-1}, \tag{2.21}$$

where $\bar{\Theta}^i$ and Θ^{k-1} are vectors containing the nodal values of $\bar{\theta}_h^i, \theta_h^k$. A_h represents the stiffness matrix, while M_h^i is a diagonal matrix obtained from the lumped mass one by multiplication with the corresponding nodal values of $\sigma(\bar{\theta}_h^i, \theta_h^k)$. R_h^{k-1} is obtained similarly, the lumped mass matrix being multiplied by the nodal values of the reaction term. Because of the assumptions we have made on the decomposition S_h, $M_h^i + \tau A_h$ is irreducible and diagonal dominant ([41], p. 51), hence an M-matrix. Therefore, if we have $\theta_h^k \in V_h(\beta(\varepsilon))$, the same holds for the entire array $\{\bar{\theta}_h^i, i \geq 0\}$ and Problem WIJK makes sense for any i. Moreover, a discrete maximum principle analogous to the one stated in Lemma 2.3.1 is valid also in this case.

Now we want to obtain the convergence of the sequence $\{\bar{\Theta}^i, i \geq 0\}$. To this aim the Lipschitz continuity of β' is needed supplementary. Then, the limit

$$\Theta^k = \lim_{i\to\infty} \bar{\Theta}^i$$

gives the nodal values of θ_h^k.

Let $i > 1$. Multiplying for each i the system in (2.21) by $(M_h^{i-1})^{-1}$ and subtracting the result obtained for $i-1$ from the one corresponding to i yields

$$(I_{N_h} + \tau D_h^{i-1} A_h) \cdot (\bar{\Theta}^i - \bar{\Theta}^{i-1}) = \tau(D_h^{i-1} - D_h^{i-2}) \cdot A_h \cdot \bar{\Theta}^{i-1}, \tag{2.22}$$

where D_h^i stands for the inverse of M_h^i (which is diagonal) and I_{N_h} is the identity matrix. Denoting by

$$\bar{e}^i \equiv \bar{\Theta}^i - \bar{\Theta}^{i-1}$$

the difference between two successive approximations, recalling the assumptions on β and β', it is easy to see that for all $i \geq 2$, the following holds true (elementwise)

$$|(D_h^{i-1} - D_h^{i-2})_{jj}| \leq L_\beta C_2(\varepsilon) \frac{d+1}{meas(\Delta_j)} |(\bar{e}^{i-1})_j|,$$

where Δ_j is the support of φ_j, $j = \overline{1, N_h}$. This is true for any j, hence

$$\|D_h^{i-1} - D_h^{i-2}\|_\infty \leq C h^{-d} L_\beta C_2(\varepsilon) \|\bar{e}^{i-1}\|_\infty. \tag{2.23}$$

Multiplying (2.22) by $(I_{N_h} + \tau D_h^{i-1} A_h)^{-1}$ and using the matrix norm associated to the l_∞ vectorial one we obtain

$$\|\bar{e}^i\|_\infty \leq \tau \|(I_{N_h} + \tau D_h^{i-1} A_h)^{-1}\|_\infty \|D_h^{i-1} - D_h^{i-2}\|_\infty \|A_h\|_\infty \|\bar{\Theta}^{i-1}\|_\infty. \tag{2.24}$$

Because S_h is regular, A_h satisfies

$$\|A_h\|_\infty \leq C h^{d-2}. \tag{2.25}$$

For estimating $\|(I_{N_h} + \tau D_h^{i-1} A_h)^{-1}\|_\infty$ we can proceed as in the proof of Lemma 2.3.1 and obtain

$$\|(I_{N_h} + \tau D_h^{i-1} A_h)^{-1}\|_\infty \leq 1. \tag{2.26}$$

Replacing the inequalities in (2.23), (2.25) and (2.26) in (2.24) and recalling the maximum principle for $\|\bar{\Theta}^{i-1}\|_\infty$ yields

$$\|\bar{e}^i\|_\infty \leq C \tau h^{-2} C_2(\varepsilon) \|\bar{e}^{i-1}\|_\infty. \tag{2.27}$$

Hence we have proven the convergence of the iterations

Lemma 2.3.2 *If $\tau C_2(\varepsilon) \leq C h^2$, the sequence of solutions of the problems in (2.20) converges to the solution of Problem WMDI.*

Remark 2.3.6 *In the above lemma a severe restriction is imposed to the time step τ. Practical computations show that this appears only theoretically.*

Remark 2.3.7 *In proving the convergence of the iterations we have used only the fact that A_h is irreducible and diagonal dominant. Therefore, if the convection is discretized using an upwind procedure, the conclusion of the Lemma 2.3.2 still holds true.*

2.3.3 Error estimates for the complete discretization

Finally we consider the convergence of the fully discrete problems. For proving this, the following lemma will be useful

Lemma 2.3.3 *Let $u_h \in V_h$ and $\theta_h = I_h\beta(u_h)$, where β satisfies the assumption in (A1). Then, if r is Lipschitz w.r.t. $\beta(u)$, we have*

$$\begin{aligned} \|\nabla\theta_h\|^2 &\leq C(\nabla u_h, \nabla\theta_h), \\ \|I_h\beta(u_h) - \beta(u_h)\| &\leq Ch\|\nabla I_h\beta(u_h)\|, \\ \|I_h r(u_h) - r(u_h)\| &\leq Ch\|\nabla\theta_h\|, \end{aligned} \tag{2.28}$$

where $C > 0$ is a generic constant independent on h and u_h.

Proof. The first two inequalities are proven in [22] and [29]. The assumption on r is needed only for the third one. Here we notice that the interpolation makes sense since u_h and r are continuous.

Consider a single element $T \in S_h$. A linear function on a simplex attains its extreme values at the vertices. Let A_m and A_M be these vertices, hence, for every $x \in \bar{T}$,

$$u_h(A_m) \leq u_h(x) \leq u_h(A_M) \quad \text{and} \quad \beta(u_h(A_m)) \leq \beta(u_h(x)) \leq \beta(u_h(A_M)),$$

the last part is due to the monotonicity of β. Similarly, there are two vertices B_m and B_M in which the extremal values of $I_h r(u_h)$ are attained,

$$r(u_h(B_m)) = I_h r(u_h(B_m)) \leq I_h r(u_h(x)) \leq I_h r(u_h(B_M)) = r(u_h(B_M)) \quad \forall x \in \bar{T}.$$

Moreover, the continuity of $r(u_h)$ on the compact set $\bar{T}$ yields the existence of two points C_m and C_M in $\bar{T}$ (not necessary vertices) such that the following holds true

$$r(u_h(C_m)) \leq r(u_h(x)) \leq r(u_h(C_M)) \quad \forall x \in \bar{T}.$$

The above inequalities together with the assumption (A3) on r show that for any $x \in T$ we have

$$\begin{aligned} |r(u_h(x)) - I_h r(u_h(x))|^2 \leq\ & (\max\{r(u_h(C_M)) - r(u_h(B_m)), r(u_h(B_M)) - r(u_h(C_m))\})^2 \\ \leq\ & C_r(\max\{\beta(u_h(C_M)) - \beta(u_h(B_m)), \beta(u_h(B_M)) - \beta(u_h(C_m))\})^2 \\ \leq\ & C_r(\beta(u_h(A_M)) - \beta(u_h(A_m)))^2 \\ =\ & C_r((A_M - A_m)^t(\nabla\theta_h)|_T)^2 \\ \leq\ & Ch^2|(\nabla\theta_h)|_T|^2. \end{aligned}$$

Integrating over T yields the inequality for the simplex T. Summing up for all simplices in S_h completes the proof.

Remark 2.3.8 *The last inequality in the lemma above is proven only if the reaction term is Lipschitz continuous in $\beta(u)$. We were not able to get a proof for the more general situation assumed in (A3). As it will be seen below, the existence of an estimate of the type above is essential in obtaining bounds for the fully discrete scheme.*

Error estimates for the fully discrete nonlinear scheme

Now we need some stability results for the fully discrete implicit scheme.

Theorem 2.3.4 *Assume (A1), (A2), (A3) and $F \equiv 0$. Then, for $k \leq n$, if θ_h^k solves Problem WMDI and $u_h^k = I_h\beta^{-1}(\theta_h^k)$, we have*

$$\sum_{k=1}^{p} \|u_h^k - u_h^{k-1}\|^2 + \tau \sum_{k=1}^{p} \|\nabla\theta_h^k\|^2 \leq C. \tag{2.29}$$

Proof. Since $u_h^k - \varepsilon \in V_h(0)$, (2.17) implies

$$\begin{aligned} & \frac{1}{2}\left(\|u_h^k\|_h^2 - \|u_h^{k-1}\|_h^2 + \|u_h^k - u_h^{k-1}\|_h^2\right) + \tau(\nabla\theta_h^k, \nabla u_h^k) \\ = \; & \tau(I_h r(\beta^{-1}(\theta_h^k)), u_h^k - \varepsilon)_h + (u_h^k - u_h^{k-1}, \varepsilon)_h. \end{aligned}$$

Summing up over k, recalling the inequalities in (2.28) and the maximum principle for u_h^k (hence the boundedness of $I_h r(\beta^{-1}(\theta_h^k))$ leads to the inequality in (2.29).

The error estimates in the fully discrete case are using the results in the semi-discrete approximation. The following notations are related to the ones proposed there

$$e_u^{k,h} := \beta^{-1}(\theta^k) - I_h\beta^{-1}(\theta_h^k) \equiv \beta^{-1}(\theta^k) - u_h^k, \quad e_\theta^{k,h} := \theta^k - \theta_h^k,$$

where $k \geq 0$. Remembering the definition of θ_h^0 in (2.18), the initial error satisfies the following inequality

$$\|e_u^{0,h}\|_{-1} \leq Ch. \tag{2.30}$$

The analysis below follows the ideas in [29] and [73] and continues the one for the time discretization. Having already the result in Theorem 2.2.11 for the semi-discrete approximation, it is enough if we estimate the error due to the spatial discretization of the elliptic problems. This is done in the following theorem.

Theorem 2.3.5 *Under the assumptions (A1), (A2), (A3), if $F \equiv 0$, r is Lipschitz continuous in $\beta(u)$ and θ^k, θ_h^k solve, for each $k > 0$ Problem WMTI, respectively Problem WMDI, then*

$$\sup_{k=\overline{1,n}} \|e_u^{k,h}\|_{-1}^2 + C\tau \textstyle\sum_{k=1}^n \|e_\theta^{k,h}\|^2 \le C\left(h + \dfrac{h^2}{\tau}\right),$$

provided τ is reasonably small.

Proof. Taking $\varphi = Ge_u^{j,h} \in H_0^1$ in Problem WMTI and $\varphi = G_h e_u^{j,h} \in V_h(0)$ in (2.17), subtracting the last equality from the first one summing up for $j = \overline{1,k}$ yields

$$\begin{aligned}
(I_1) + (I_2) := \quad & \sum_{j=1}^k (e_u^{j,h} - e_u^{j-1,h}, Ge_u^{j,h}) \\
& + \tau \sum_{j=1}^k \left\{(\nabla\theta^j, \nabla Ge_u^{j,h}) - (\nabla\theta_h^j, \nabla G_h e_u^{j,h})\right\} \\
= & - \sum_{j=1}^k (u_h^j - u_h^{j-1}, (G - G_h)e_u^{j,h}) \\
& + \sum_{j=1}^k \left\{(u_h^j - u_h^{j-1}, G_h e_u^{j,h})_h - (u_h^j - u_h^{j-1}, G_h e_u^{j,h})\right\} \\
& + \tau \sum_{j=1}^k \left\{(r(\beta^{-1}(\theta^j)), Ge_u^{j,h}) - (I_h r(\beta^{-1}(\theta_h^j)), G_h e_u^{j,h})\right\} \\
& + \tau \sum_{j=1}^k \left\{(I_h r(\beta^{-1}(\theta_h^j)), G_h e_u^{j,h}) - (I_h r(\beta^{-1}(\theta_h^j)), G_h e_u^{j,h})_h\right\} \\
=: & (I_3) + (I_4) + (I_5) + (I_6).
\end{aligned} \tag{2.31}$$

We go on with the estimates for each term appearing above. Using the identity in (2.18) and the properties of G, (I_1) becomes

$$(I_1) = \frac{1}{2}\left(\|e_u^{k,h}\|_{-1}^2 - \|e_u^{0,h}\|_{-1}^2 + \sum_{j=1}^k \|e_u^{j,h} - e_u^{j-1,h}\|_{-1}^2\right). \tag{2.32}$$

The second term in (2.31) is decomposed into two sums

$$\begin{aligned}
(I_2) = \;& \tau \textstyle\sum_{j=1}^k (\theta^j, e_u^{j,h}) - (\theta_h^j, e_u^{j,h}) = \tau \sum_{j=1}^k (e_\theta^{j,h}, e_u^{j,h}) \\
= \;& \tau \textstyle\sum_{j=1}^k \left\{(\theta^j - \beta(u_h^j), \beta^{-1}(\theta^j) - u_h^j) + (\beta(u_h^j) - I_h\beta(u_h^j), \beta^{-1}(\theta^j) - u_h^j)\right\} \\
=: \;& (I_{21}) + (I_{22}).
\end{aligned}$$

For (I_{21}) we recall the relation in (2.2), Lemma 2.3.3 and the apriori estimates in order to obtain

$$\begin{aligned}(I_{21}) \geq & C_1\tau\sum_{j=1}^{k}\|\theta^j-\beta(u_h^j)\|^2 \geq C_1\tau\sum_{j=1}^{k}(\|\theta^j-\theta_h^j\|-\|\theta_h^j-\beta(u_h^j)\|)^2\\ \geq & \frac{C_1\tau}{2}\sum_{j=1}^{k}\|\theta^j-\theta_h^j\|^2-C_1\tau\sum_{j=1}^{k}\|\theta_h^j-\beta(u_h^j)\|^2\\ \geq & \frac{C_1\tau}{2}\sum_{j=1}^{k}\|\theta^j-\theta_h^j\|^2-C_1\tau h^2\sum_{j=1}^{k}\|\nabla\theta_h^j\|^2\\ \geq & \frac{C_1\tau}{2}\sum_{j=1}^{k}\|\theta^j-\theta_h^j\|^2-C_1h^2.\end{aligned} \tag{2.33}$$

Similarly, (I_{22}) yields

$$\begin{aligned}|(I_{22})| \leq & \tau\sum_{j=1}^{k}\|\theta^j-\beta(u_h^j)\|\|\beta^{-1}(\theta^j)-u_h^j\| \leq C\tau h\sum_{j=1}^{k}\|\nabla\theta_h^j\|\|\beta^{-1}(\theta^j)-u_h^j\|\\ \leq & Ch\tau\sum_{j=1}^{k}\|\nabla\theta^j\|^2+Ch \leq Ch.\end{aligned} \tag{2.34}$$

Using (2.20), (I_3) can be rewritten as

$$(I_3)=(u_h^k,(G-G_h)e_u^{k,h})-(u_h^0,(G-G_h)e_u^{0,h})+\sum_{j=1}^{k}(u_h^{j-1},(G-G_h)(e_u^{j,h}-e_u^{j-1,h})).$$

This implies the following estimates

$$\begin{aligned}|(I_3)|= & \quad\|u_h^k\|\|(G-G_h)e_u^{k,h}\|+\|u_h^0\|\|(G-G_h)e_u^{0,h}\|\\ & +\sum_{j=1}^{k}\|u_h^{j-1}\|\|(G-G_h)(e_u^{j,h}-e_u^{j-1,h})\|\\ \leq & Ch^2+Ch\sum_{j=1}^{k}\|e_u^{j,h}-e_u^{j-1,h}\|_{-1}\\ \leq & Ch^2+C\frac{h^2}{\tau\delta_3}+C\delta_3\sum_{j=1}^{k}\|e_u^{j,h}-e_u^{j-1,h}\|_{-1}^2,\end{aligned} \tag{2.35}$$

where the approximation property in (2.7) and the maximum bounds for u_h^j have been used. Analogous, (I_4) gives

$$
\begin{aligned}
|(I_4)| \leq \quad & |(u_h^k, G_h e_u^{k,h}) - (u_h^k, G_h e_u^{k,h})_h| + |(u_h^0, G_h e_u^{0,h}) - (u_h^0, G_h e_u^{0,h})_h| \\
& + \sum_{j=1}^{k} |(u_h^{j-1}, G_h(e_u^{j,h} - e_u^{j-1,h})) - (u_h^{j-1}, G_h(e_u^{j,h} - e_u^{j-1,h}))_h| \\
\leq & Ch(\|u_h^k\| \|\nabla G_h e_u^{k,h}\| + \|u_h^0\| \|\nabla G_h e_u^{0,h}\|) \\
& + Ch \sum_{j=1}^{k} \|u_h^{j-1}\| \|\nabla G_h(e_u^{j,h} - e_u^{j-1,h})\| \\
\leq & Ch \left(\|e_u^{k,h}\|_{-1} + \|e_u^{0,h}\|_{-1} + \sum_{j=1}^{k} \|e_u^{j,h} - e_u^{j-1,h}\|_{-1} \right) \\
\leq & Ch + Ch \sum_{j=1}^{k} \|e_u^{j,h} - e_u^{j-1,h}\|_{-1} \\
\leq & Ch + C\frac{h^2}{\tau\delta_4} + C\delta_4 \sum_{j=1}^{k} \|e_u^{j,h} - e_u^{j-1,h}\|_{-1}^2.
\end{aligned}
\tag{2.36}
$$

(I_5) can be splitted into three sums

$$
\begin{aligned}
(I_5) = \quad & \tau \textstyle\sum_{j=1}^{k} (r(\beta^{-1}(\theta^j)) - r(\beta^{-1}(\theta_h^j)), G e_u^{j,h}) + \tau \textstyle\sum_{j=1}^{k} ((I - I_h) r(\beta^{-1}(\theta_h^j)), G e_u^{j,h}) \\
& + \tau \textstyle\sum_{j=1}^{k} (I_h r(\beta^{-1}(\theta_h^j)), (G - G_h) e_u^{j,h}) =: (I_{51}) + (I_{52}) + (I_{53}).
\end{aligned}
$$

Recalling (A3), for (I_{51}) we get

$$
\begin{aligned}
|(I_{51})| \leq & C\tau \sum_{j=1}^{k} \|\theta^j - \theta_h^j\| \|G e_u^{j,h}\| \\
\leq & C\tau\delta_{51} \sum_{j=1}^{k} \|\theta^j - \theta_h^j\|^2 + \frac{C\tau}{\delta_{51}} \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1}^2.
\end{aligned}
\tag{2.37}
$$

Now, since r is Lipschitz continuous in $\beta(u)$, Lemma 2.3.3 and the apriori estimates in Theorem 2.3.4 yield

$$
\begin{aligned}
|(I_{52})| \leq & C\tau h \sum_{j=1}^{k} \|\nabla\theta_h^j\| \|G e_u^{j,h}\| \leq C\tau h^2 \delta_{52} \sum_{j=1}^{k} \|\nabla\theta_h^j\|^2 + \frac{C\tau}{\delta_{52}} \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1}^2 \\
\leq & Ch^2\delta_{52} + \frac{C\tau}{\delta_{52}} \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1}^2.
\end{aligned}
\tag{2.38}
$$

Recalling again the inequality in (2.7), (I_{53}) leads to

$$\begin{aligned}|(I_{53})| \leq & C\tau \sum_{j=1}^{k} \|(G - G_h)e_u^{j,h}\| \leq C\tau h \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1} \\ \leq & Ch^2\delta_{53} + \frac{C\tau}{\delta_{53}} \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1}^2.\end{aligned} \tag{2.39}$$

Finally, for (I_6) the relations in (2.3) and (2.8) give

$$\begin{aligned}|(I_6)| \leq & C\tau h \sum_{j=1}^{k} \|I_h r(\beta^{-1}(\theta_h^j))\| \|\nabla G_h e_u^{j,h}\| \leq C\tau h \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1,h} \\ \leq & C\tau h \sum_{j=1}^{k} (\|e_u^{j,h}\|_{-1} + h^2 \|e_u^{j,h}\|) \leq Ch^2\delta_6 + \frac{C\tau}{\delta_6} \sum_{j=1}^{k} \|e_u^{j,h}\|_{-1}^2.\end{aligned} \tag{2.40}$$

Inserting all the inequalities in (2.32) - (2.40) in (2.31), recalling (2.33) and choosing the δ's properly, we get

$$\begin{aligned} & \|e_u^{k,h}\|_{-1}^2 + \sum_{j=1}^k \|e_u^{j,h} - e_u^{j-1,h}\|_{-1}^2 + \sum_{j=1}^k \|e_\theta^{j,h}\|^2 \\ \leq \;& C \left(h + \frac{h^2}{\tau} \right) + C\tau \sum_{j=1}^k \|e_u^{j,h}\|_{-1}^2. \end{aligned}$$

The inequality above can be obtained provided τ is not bigger then C, a constant depending on the problem but not on the discretization parameters. The discrete Gronwall lemma concludes the proof.

Remark 2.3.9 *As in the semi-discrete case, an explicit treatment of the reaction term does not affect the results above.*

The error estimates for Scheme WMDI are a direct consequence of the Theorems 2.2.11 and 2.3.5

Theorem 2.3.6 *In the setting of Theorem 2.3.5, if u is the weak solution of Problem WP and θ_h^k solves for each $k > 0$ Problem WMDI, then*

$$\begin{aligned} & \sup_{k=\overline{1,n}} \|u(t_k) - u^{k,h}\|_{-1}^2 + C \sum_{k=1}^n \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta^{k,h}\|^2 dt \\ & \qquad \leq C \left(\tau + h + \frac{h^2}{\tau} + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2 \right). \end{aligned}$$

Remark 2.3.10 *Taking above $h = O(\tau)$ we can notice that the order of convergence for the fully discrete approximation is the same as in the semi-discrete case. But this choice may not be always convenient, especially when a nonuniform mesh or a locally refined one is considered. Moreover, it contradicts the convergence condition for the iterations defined through Problem WIJK. But the restriction on h appears only theoretically. Different approaches in proving the convergence of the fully discrete scheme (see, e.g. [88]) do not request any relation between the discretization parameters, but the estimates depend additionally on $C_2(\varepsilon)$ resulting a lower order of convergence compared with the semi-discrete case. This was avoided in the previous result.*

Error estimates for the fully discrete linear scheme

Here we extend the analysis for the semi-discrete linear scheme MTL to the fully discrete case. Analogous to the nonlinear one, if the convection term is not present, the linear scheme can be brought to a finite element formulation

Problem WMDL. For any $1 \leq k \leq n$, find $\theta_h^k \in V_h(\beta(\varepsilon))$ such that for all $\varphi \in V_h(0)$ the following holds true

$$\begin{aligned}(\sigma_{k-1,h}(\theta_h^k - \theta_h^{k-1}), \varphi)_h + \tau(\nabla\theta_h^k, \nabla\varphi) = \tau(I_h r(\beta^{-1}(\theta_h^{k-1})), \varphi)_h,\\ \sigma_{k,h} = (\beta^{-1})'(\theta_h^k).\end{aligned} \tag{2.41}$$

with $\sigma_{0,h} = (\beta^{-1})'(\theta_h^0)$.

For the implicit scheme, the choice of the initial data in the form proposed in (2.18) was accepted in order to avoid a worsening of the convergence order. But other possibilities may be considered. For example, if θ^0 is continuous,

$$u_h^0 = I_h(\beta^{-1}(\theta^0)), \;\; \theta_h^0 = I_h\beta(u_h^0)$$

is more convenient here. Again, u_h^0 is bounded by ε and M (the essential supremum of u_0), but now we have also $\theta_h^0 = I_h\theta^0$ and consequently $\|\nabla\theta_h^0\| \leq \|\nabla\theta^0\| \leq C$, not depending on h. This choice is better for the apriori estimates, but the bounds for the initial error $e_u^{0,h}$ depend on ε as follows

$$\|e_u^{0,h}\|_{-1} \leq C\|e_u^{0,h}\| \leq Ch\|\nabla\beta^{-1}(\theta^0)\| \leq ChC_2(\varepsilon).$$

Due to this the estimates in Theorem 2.3.5 become

$$\sup_{k=\overline{1,n}} \|e_u^{k,h}\|_{-1}^2 + C\tau \textstyle\sum_{k=1}^n \|e_\theta^{k,h}\|^2 \leq C\left(h + \dfrac{h^2}{\tau} + h^2 C_2(\varepsilon)^2\right),$$

and the same change appears in Theorem 2.3.6.

The maximum principle stated in Lemma 2.3.1 remains valid in this case too, therefore the scheme defined above makes sense. Since convection is not present here, stability results can be obtained as in Theorem 2.2.9.

Theorem 2.3.7 *Assume (A1), (A2), (A3) and $F \equiv 0$. Then, for $p < n$, if θ_h^k solves Problem WMDL, the following holds*

$$\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1,h}}(\theta_h^k - \theta_h^{k-1})\|_h^2 + \tau\|\nabla\theta_h^p\|^2 + \tau\sum_{k=1}^{p}\|\nabla(\theta_h^k - \theta_h^{k-1})\|^2 \leq C\tau. \tag{2.42}$$

Using this result, the error estimates for the fully discrete linear scheme are given in the theorem below.

Theorem 2.3.8 *In the setting of Theorem 2.3.7, if u is the weak solution of the Problem WP and θ_h^k solves for each $k > 0$ Problem WMDL, then*

$$\begin{aligned} &\sup_{k=\overline{1,n}} \|u(t_k) - u^{k,h}\|_{-1}^2 + C\tau \textstyle\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta^{k,h}\|^2 dt \\ &\leq C\left(\tau C_2(\varepsilon)^2 + hC_2(\varepsilon)^2 + \frac{h^2}{\tau} + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2\right). \end{aligned}$$

Proof. The proof is identical to the one for the nonlinear scheme, but because of the apriori estimates above the estimation of the terms (I_{21}), (I_{22}) and (I_{52}) yield $ChC_2(\varepsilon)^2$ instead of Ch. Moreover, an additional term has to be considered, namely

$$T \equiv \sum_{j=1}^{k}(I_h\beta^{-1}(\theta_h^j) - I_h\beta^{-1}(\theta_h^{j-1}) - \sigma_{j-1,h}(\theta_h^j - \theta_h^{j-1}), G_h e_u^{j,h})_h.$$

As in the semi-discrete case, the Lipschitz continuity of β', the apriori estimates and the maximum principle gives

$$|T| \leq C\tau C_2(\varepsilon)^2,$$

which concludes the proof.

Remark 2.3.11 *Taking again $h = O(\tau)$, the estimates in the semi-discrete case are recovered.*

Chapter 3

Regularization by modifying the nonlinearity function

In this chapter we proceed with the analysis of a different regularization method which can be applied to more general situations, including also two phase Stefan problems. This relies on the modification of the nonlinearity function such that its derivative is uniformly bounded by a (small) nonzero constant. As mentioned at the beginning of the chapter before, this is the underlying idea of most of the regularization procedures for degenerate equations. Again, the convergence of the algorithm is shown by proving error estimates.

3.1 Basic setting

In order to avoid additional complications and to keep the proofs in the same framework as before, the basic setting defined in the previous chapter is maintained here. Hence, Ω is a bounded domain in $\mathbb{R}^d (d \geq 1)$ with a Lipschitz continuous boundary and $Q_T \equiv (0,T)\times\Omega$, where $0 < T < \infty$ is fixed. Problem P defined Chapter 2 is also considered here

Problem P:

$$\begin{aligned} \partial_t u - \nabla\cdot(\nabla\beta(u) + F(u)) = r(u), \qquad & \text{in } \; Q_T \equiv (0,T)\times\Omega, \\ u(0,x) = u_0(x), \qquad & \text{in } \; \Omega, \\ \beta(u) = 0, \qquad & \text{on } \; \partial\Omega. \end{aligned} \tag{3.1}$$

The function $\beta : \mathbb{R} \to \mathbb{R}$ is monotone increasing and smooth, but may now vanish on sets of nonzero measure. Again, degeneracy means no diffusion, namely $\beta'(u) = 0$ for some u. It is worth to notice that in this case there is no need to have an unique degeneracy point. Moreover, the approach works also if the diffusion vanishes on intervals, therefore it applies, e.g., to two-phase Stefan problems (in the enthalpy formulation).

Before writing a numerical method for solving Problem P we consider a regularization procedure for it. As mentioned in the previous chapter, the resulting problem becomes non-degenerate and therefore any discretization method for nonlinear parabolic problems can be applied for solving it numerically. Roughly speaking, the approach considered here relies on the perturbation of the nonlinearity β. This idea has been widely used in the analysis of such kind of problems and became an important source for numerical algorithms.

There are several possibilities to do this step. The simplest way is to construct a strictly increasing function β_ε which approximates β, but its derivative is bounded from below by the small parameter $\varepsilon > 0$, (as done, e.g., in [84], [50], [31] or [88]). Another possibility consists in the regularization of the inverse of the nonlinearity, $\gamma = \beta^{-1}$. A justification for this approach appears when the numerical method is written in terms of the more regular unknown, $\theta \equiv \beta(u)$, since at least the time derivative for $u \equiv \beta^{-1}(\theta)$ has to be considered. In this case, because the derivative of γ is infinite in the points of degeneracy, γ'_ε is bounded by a large constant (like $1/\varepsilon$). Numerical methods based on this procedure are proposed in [63], [66], [67] or [73]. The last category of regularization algorithms is akin to the nonlinear Chernoff formula ([17]) and gives rise to some relaxation schemes. Again, the more regular unknown is considered, but the modification appears only in the derivative of the nonlinearity function, not in the function itself. This possibility was first observed by Berger, Brézis & Rogers ([14]) and improved in [60] and [70]. The resulting schemes are linear, but the accuracy is affected especially around the free boundaries. This drawback is avoided through a nonlinear version of the scheme in [46], [47] or [54].

The assumptions on β, F, r and u_0 are similar to the ones in the previous chapter.

(A1) β is Lipschitz and differentiable, $\beta(0) = 0$, $\beta'(u) \geq 0$.

(A2) $u_0 \in L^\infty(\Omega)$.

(A3) $r : \mathbb{R} \to \mathbb{R}$ and $F : \mathbb{R} \to \mathbb{R}^d$ are continuous in u and satisfy the condition

$$|r(u) - r(v)|^2 + |F(u) - F(v)|^2 \leq C(u - v)(\beta(u) - \beta(v))$$

for any $u, v \in \mathbb{R}$, where $C > 0$ does not depend on x, t, u and v. Moreover, it is assumed here that r is positive for all positive arguments and the graph of both functions contain the origin, hence $r(0) = 0$ and $F(0) = \bar{0}$.

Again, Lipschitz continuity for β' or the existence of a function $f = F'$ are requested in some cases additionally to (A1) or (A3). Moreover, then $\beta(u_0)$ should be more regular, belonging to $H_0^1(\Omega)$. Sometimes the constants appearing above are mentioned in a distinct way. The Lipschitz constant of both β and β' (where this will be needed) is denoted by L_β, while the growth of F and r is controlled by C_F, respectively C_r.

Remark 3.1.1 *The positivity of the initial data requested in the maximum principle algorithm is abandoned here, thus the solution may become negative. However, we still consider only essentially bounded solutions, therefore a maximum principle for Problem P is assumed. This allows us to avoid some technical difficulties, but may not be necessary. For the sake of simplicity, β, F and r do not depend on the variables x and t, but this would not affect the results significantly (see again [53]). Moreover, this approach can be extended also for nonlinearities β which are not Lipschitz continuous.*

Remark 3.1.2 *The assumptions on F and r are - as in the previous chapter - weaker than the usual ones, namely a Lipschitz continuity w.r.t. $\beta(u)$*

$$|r(u) - r(v)| + |F(u) - F(v)| \leq C|\beta(u) - \beta(v)|.$$

Remark 3.1.3 *Non-homogeneous Dirichlet or natural boundary conditions may be considered without any problem here.*

The schemes we want to consider here rely on the first regularization approach described above. The nonlinearity is approximated by a function β_ε satisfying $\beta' \geq \varepsilon$. There are several possibilities to obtain the approximation. The simplest one is given by

$$\beta_\varepsilon(u) \equiv \beta(u) + \varepsilon u, \tag{3.2}$$

hence a global perturbation is added to the derivative of the original function,

$$\beta_\varepsilon'(u) \equiv \beta'(u) + \varepsilon$$

and thus we have

$$\beta_\varepsilon(u) \equiv \int_0^u \beta_\varepsilon'(s)ds. \tag{3.3}$$

Other possibility is given by a local perturbation of β', namely

$$\beta_\varepsilon'(u) \equiv \max\{\beta'(u), \varepsilon\}.$$

In any case, two main features should be fulfilled, namely

$$\beta_\varepsilon'(u) \geq \varepsilon \quad \text{and} \quad 0 \leq \beta_\varepsilon'(u) - \beta'(u) \leq \varepsilon \tag{3.4}$$

for any real u. Clearly, these are carried out by any of the two constructions proposed above. In this framework, the elementary lemma below can be proven easily.

Lemma 3.1.1 *If β satisfies the assumption (A1) and its approximation β_ε fulfills the relations in (3.4), the following inequalities hold true for any reals u and v.*

$$\begin{aligned} C \leq \left(\beta_\varepsilon^{-1}\right)' (\beta_\varepsilon(u)) \leq \frac{1}{\varepsilon}, \\ sgn(\beta_\varepsilon(u) - \beta(u)) = sgn(u), \\ |\beta(u) - \beta(v)| \leq |\beta_\varepsilon(u) - \beta_\varepsilon(v)| \leq |\beta(u) - \beta(v)| + \varepsilon|u - v|. \end{aligned} \tag{3.5}$$

Proof. Since $\beta_\varepsilon(u) \geq \varepsilon$, the first inequality is trivial. For the second one the identity in (3.3) can be used in order to get

$$\beta_\varepsilon(u) - \beta(u) = \int_0^u \beta_\varepsilon'(s) - \beta'(s) ds,$$

so the identity of the signs follows since the integrand is positive. For the last part we proceed in a similar manner and obtain

$$\beta(u) - \beta(v) = \int_v^u \beta'(s) ds \quad \text{and} \quad \beta_\varepsilon(u) - \beta_\varepsilon(v) = \int_v^u \beta_\varepsilon'(s) ds.$$

Applying now the inequalities in (3.4) the conclusion is immediate.

Clearly, β_ε is a strictly increasing function and admits a differentiable inverse. The next result is a direct consequence of the previous lemma.

Corollary 3.1.2 *In the setting above, if the assumption (A3) is satisfied by F and r, an inequality of the same type holds for β_ε,*

$$|r(u) - r(v)|^2 + |F(u) - F(v)|^2 \leq C(u - v)(\beta_\varepsilon(u) - \beta_\varepsilon(v))$$

for all $u, v \in \mathbb{R}$. Moreover, the Lipschitz continuity in $\beta(u)$ implies the same in terms of β_ε, namely

$$|r(u) - r(v)| + |F(u) - F(v)| \leq C|\beta_\varepsilon(u) - \beta_\varepsilon(v)|$$

for any real numbers u and v.

As done in the previous chapter, we deal here with the variational formulation of Problem P since a solution in the classical sense does not generally exist.

Problem WP. u is called a solution of the problem in (3.1) iff

$$u \in H^1(0,T;H^{-1}(\Omega)), \quad \beta(u) \in L^2(0,T;H^1_0(\Omega)), \quad u(0) = u_0 \text{ (in } H^{-1})$$

and for all $\varphi \in L^2(0,T;H^1_0(\Omega))$ the equation holds true

$$\int_0^T (\partial_t u(t), \varphi(t))\, dt + \int_0^T (\nabla\beta(u(t)) + F(u(t)), \nabla\varphi(t))dt = \int_0^T (r(u(t)), \varphi(t))dt. \quad (3.6)$$

Beyond the frame stated in the chapter before, existence, uniqueness and regularity of the solution for this problem without convection and reaction has been established in several papers (see, e.g. [49] or [36] and the references therein). Moreover, a maximum principle can be proven. There are no significant changes if a reaction term is added, provided this is Lipschitz continuous w.r.t. $\beta(u)$. If F has this property too and its graph passes through the origin, the existence of a weak solution has been shown in [18]. Under similar growth conditions for the convective and reactive terms, an existence result can be found in [1]. However, there are less uniqueness results. It is easy to imagine situations when the problem becomes purely hyperbolic, therefore the existence of a unique solution cannot be expected. Since our aim here is the analysis of some schemes based on regularization procedure described above, we assume Problem WP admits an unique solution which is uniformly bounded (a.e.) in the whole cylinder Q_T. As mentioned in the chapter before, in this case u belongs to $C(0,T;H^{-1}(\Omega))$. Moreover, the notations used there are maintained here too.

Remark 3.1.4 *The existence of an unique solution makes the error estimates possible. If uniqueness does not hold true, convergence can be obtained proceeding from compactness arguments (the basic ideas can be found in [46], [47]; see also [35]).*

3.2 Time discretization

The main goal here is the investigation of some approximation schemes for Problem P. We go on with the equation written in the more regular unknown, $\beta(u)$. Since in the regularization step a perturbation of the nonlinearity function has been introduced, the numerical schemes are given w.r.t. β_ε instead of β itself. Therefore, after computing the unknown θ, u can be obtained by inverting β_ε in θ.

As done in the previous chapter, due to the lack of regularity, the time discretization relies on first order methods. The time step is denoted again by τ and we have $\tau = T/n$ for some integer n. Correspondingly, an Euler implicit scheme reads

Scheme RTI:

$$\begin{aligned} \beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}) =& \tau\nabla \cdot (\nabla\theta^k + F(\beta_\varepsilon^{-1}(\theta^k)) + \tau r(\beta_\varepsilon^{-1}(\theta^k)), \\ \theta^k|_{\partial\Omega} =& 0 \end{aligned} \tag{3.1}$$

for $k = \overline{1, n}$ where θ^0 will be given below.

Remark 3.2.1 *Looking at the relations in (3.2) or (3.3), an explicit formula for the inverse of β_ε may not be available. In fact, even for β itself, computing the value in some point u can be a quite tedious (this problem appears also in case β has a complicated form). Moreover, any function call increase the computing time significantly. Therefore in the implementation of the schemes the values of β_ε in some points are inserted into a look-up table, with an additional memory requirement. Surely, this involves also a simple (linear) interpolation step for obtaining values which are not included in the table, but now the computing time is significantly reduced and the errors can be controlled through an appropriate choice of the interpolation knots. Because of the monotonicity of the function, searching in this table is fast and therefore the values of β_ε or its inverse can be obtained efficiently.*

To simplify the nonlinear scheme RTI, the convection term $\nabla F(\beta_\varepsilon^{-1}(\theta^k))$ can be linearized by $(\beta_\varepsilon^{-1})'(\theta^{k-1})f(\beta_\varepsilon^{-1}(\theta^{k-1})) \cdot \nabla\theta^k$, with $f(u) \equiv F'(u)$. A fully explicit discretization of this term can be also considered (as done, e.g., in [73], [47], [51]), but then some stability problems may appear especially near the free boundaries, where convection becomes dominant. Since L^∞ bounds are used in the forthcoming estimates, we have not considered this approach. The scheme becomes

Scheme RTC:

$$\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}) = \tau\Delta\theta^k + \tau(\beta_\varepsilon^{-1})'(\theta^{k-1})f(\beta_\varepsilon^{-1}(\theta^{k-1})) \cdot \nabla\theta^k + \tau r(\beta_\varepsilon^{-1}(\theta^k)), \tag{3.2}$$

together with the initial and boundary data. Replacing $r(\beta_\varepsilon^{-1}(\theta^k))$ by $r(\beta_\varepsilon^{-1}(\theta^{k-1}))$, an explicit discretization of the reaction term is obtained.

Now a linear approximation scheme for Problem P can be given,

Scheme RTL:

$$\begin{aligned}\sigma_{k-1}(\theta^k - \theta^{k-1}) =& \tau\Delta\theta^k + \tau(\beta_\varepsilon^{-1})'(\theta^{k-1})f(\beta_\varepsilon^{-1}(\theta^{k-1}))\cdot\nabla\theta^k + \tau r(\beta_\varepsilon^{-1}(\theta^{k-1})),\\ \theta^k|_{\partial\Omega} =& 0,\\ \sigma_k =& (\beta_\varepsilon^{-1})'(\theta^k)\end{aligned} \tag{3.3}$$

for $k = \overline{1,n}$, where $\sigma_0 = (\beta_\varepsilon^{-1})'(\theta^0)$.

Remark 3.2.2 *In the linearization of the convection term, $(\beta_\varepsilon^{-1})'(\theta)f(\beta_\varepsilon^{-1}(\theta))$ multiplies the gradient of θ^k. Lemma 3.1.1 and the assumption (A3) on F assure the boundedness of this 'speed', namely*

$$|(\beta_\varepsilon^{-1})'(\theta)f(\beta_\varepsilon^{-1}(\theta))| \leq \sqrt{\frac{C_F}{\varepsilon}}.$$

Moreover, if F is Lipschitz continuous in $\beta(u)$, the bound above becomes $\sqrt{C_F}$.

Up to now we did not gave the initial data. For stability reasons, if the convective term is linearized, θ_0 should belong to $H_0^1(\Omega)$. In any case $\beta_\varepsilon^{-1}(\theta_0)$ has to approximate u_0 in some sense. Concretely, in obtaining the error estimates, these depend on the initial approximation error in H^{-1}, namely $\|u_0 - \beta_\varepsilon^{-1}(\theta_0)\|_{-1}$. If u_0 belongs to $H^1(\Omega)$, the choice $\theta_0 \equiv \beta_\varepsilon(u_0)$ fulfills both requirements since there is no initial error for u and H^1 boundedness for θ_0 follows from the one for u_0 and the upper limits of β_ε'. Moreover, these bounds do not depend on ε.

The previous choice works in the implicit case even if u_0 is not a H^1 function since θ^0 needs not to be in $H^1(\Omega)$. In fact, the assumption $u_0 \in H^1(\Omega)$ is quite restrictive and fails to hold true in important cases. But $\theta_0 \in H^1$ is essential for obtaining useful apriori estimates for the schemes RTC and RTL. Therefore, if the regularization given in (3.2) is considered, we can proceed as follows. Let ρ be a $C_0^\infty(\mathbb{R}^d)$ positive function defining a mollifier sequence $\{\rho_\mu\}_{1>>\mu>0}$,

$$supp\ \rho \subset B(0,1), \quad \int_{\mathbb{R}^d}\rho(x)dx = 1 \quad \text{and} \quad \rho_\mu(x) \equiv \frac{1}{\mu^d}\,\rho\left(\frac{x}{\mu}\right).$$

Now $\theta_0 \equiv \beta(u_0) + \varepsilon\rho_\mu * u_0$ is a H^1 function, where $*$ stands for the convolution operator. From practical point of view there is no need to compute the convolution since this can be simulated by taking the solution of the heat equation after one (small) time step with the initial data u_0. We have

$$\|\nabla\theta_0\| \leq \|\nabla\beta(u_0)\| + \varepsilon\|\nabla(\rho_\mu * u_0)\| = \|\nabla\beta(u_0)\| + \varepsilon\|(\nabla\rho_\mu) * u_0\|.$$

The first term above is bounded since $\beta(u_0) \in H^1$. For the second one we have

$$\|(\nabla \rho_\mu) * u_0\| \le \|\nabla \rho_\mu\|_{0,1} \|u_0\| = \frac{1}{\mu} \|\nabla \rho\|_{0,1} \|u_0\| \le \frac{C}{\mu}.$$

Therefore, if μ is of the same order as ε, the H^1 norm of θ_0 is uniformly bounded w.r.t. ε.

The initial error due to this step goes to zero as follows from

$$\|u_0 - \beta_\varepsilon^{-1}(\theta_0)\| \le \frac{1}{\varepsilon} \|\theta_0 - \beta_\varepsilon(u_0)\| = \|\rho_\mu * u_0 - u_0\|.$$

Since u_0 is a L^2 function, the last term goes to zero together with μ. It is worth noticing here that convergence to 0 takes place in a stronger norm as necessary, but unfortunately we did not succeed in getting a convergence order.

In particular cases it is possible to find other perturbations of the initial data for which estimates of the above error are possible. For the porous medium equation, the nonlinearity $\beta(u) = u^m$ yields for any real u,

$$|\beta_\varepsilon^{-1}(\beta(u)) - u| \le C \varepsilon^{\frac{1}{m-1}},$$

where β_ε is the regularization defined in (3.3). This shows that the choice $\theta_0 \equiv \beta(u_0)$ may be considered, but theoretically the global error estimates are affected by this step.

Thinking at the Stefan problem, if an initial non-degeneracy property for the temperature holds true ($meas\{x \in \Omega \ / \ 0 < \theta_0(x) < \varepsilon\} \le C\varepsilon^{\frac{1}{2}}$), it is possible to give initial data satisfying

$$\|\beta_\varepsilon^{-1}(\theta_0) - u_0\| \le C\varepsilon,$$

as proposed by R. Nochetto in [63].

Before giving the weak forms of the schemes defined above, some remarks can be formulated. The regularization is necessary when simplified versions of the implicit scheme are considered. This step is useful also when some linear iterations are applied for solving the nonlinear problems. As already mentioned in the previous chapter, up to the explicit treatment of the non-diffusive terms, Scheme RTI is akin to Scheme JK ([46], [47]) in a particular form, for which the proof there does not apply. The difference appears due to the fact that in the Jäger-Kačur method only the derivative of the nonlinearity function is modified, while in the schemes above this applies to the function itself. Moreover, if the spatial discretization in [32] is completed by a backward Euler method in time, the resulting scheme is a fully discrete counterpart for RTI. Regarding the linear scheme RTL, this was proposed in [88] again in a simplified framework (Lipschitz continuous reaction terms and no convection), while the nonlinearity is perturbed globally.

Similar to the continuous case, a rigorous formulation for the schemes RTI, RTC and RTL can be given only in a weak sense.

Problem WT. For any $1 \le k \le n$, find $\theta^k \in H_0^1(\Omega)$ such that for all $\varphi \in H_0^1(\Omega)$ one of the equations below (each corresponding to one of the schemes mentioned before) hold true.

Problem WRTI.

$$(\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}), \varphi) + \tau(\nabla\theta^k + F(\beta_\varepsilon^{-1}(\theta^k)), \nabla\varphi) = \tau(r(\beta_\varepsilon^{-1}(\theta^{\underline{k}})), \varphi) \tag{3.4}$$

for Scheme RTI, or

Problem WRTC.

$$\begin{aligned} &(\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}), \varphi) + \tau(\nabla\theta^k, \nabla\varphi) \\ = \;& \tau((\beta_\varepsilon^{-1})'(\theta^{k-1}) f(\beta_\varepsilon^{-1}(\theta^{k-1})) \cdot \nabla\theta^k + r(\beta_\varepsilon^{-1}(\theta^{\underline{k}})), \varphi) \end{aligned} \tag{3.5}$$

if the convection is linearized, respectively

Problem WRTL.

$$\begin{aligned} &(\sigma_k(\theta^k - \theta^{k-1}), \varphi) + \tau(\nabla\theta^k, \nabla\varphi) \\ = \;& \tau((\beta_\varepsilon^{-1})'(\theta^{k-1}) f(\beta_\varepsilon^{-1}(\theta^{k-1})) \cdot \nabla\theta^k + r(\beta_\varepsilon^{-1}(\theta^{k-1})), \varphi) \end{aligned} \tag{3.6}$$

for the linear scheme.

The initial data - θ^0 - has been already chosen in Scheme RTI, while Scheme RTL contains the definition of σ_k. $\underline{k}$ stands either for k or for $k-1$, depending on the way the reaction term is treated.

3.2.1 The elliptic problems

Now we proceed with the analysis of the elliptic problems arising in the time discretization process. This follows essentially the lines mapped out in the chapter before, therefore the proofs are limited to the cases where some differences occur.

Recalling the setting for the continuous problem, we deal here with essentially bounded solutions. Therefore a similar property is interesting also for the numerical approximation of the solution. Hence, for each k less then n, solutions of the semi-discrete problems are sought in the space

$$V_k = \left\{ \varphi \in H_0^1(\Omega) : \|\varphi\|_\infty \le \beta_\varepsilon\left(Me^{\bar{C}k\tau}\right), \quad a.e. \right\}, \tag{3.7}$$

where $\|\cdot\|_\infty$ stands for the $L^\infty(\Omega)$ norm and $\bar{C} > 0$ is a constant appropriately chosen. Clearly, for any $k > 0$, V_k includes all the previous sets and is convex and closed. It is worth noticing that also negative solutions are admitted here. But if the solution of Problem WP is positive, the sets above can be restricted to their positive halves.

As done in the chapter before, a maximum principle for the semi-discrete approximations can be proven. To do so we make the following notations

$$\begin{aligned} a(\theta, \varphi; \psi) = &\ (\beta_\varepsilon^{-1}(\theta), \varphi) + \tau(\nabla\theta, \nabla\varphi) - \tau(((\beta_\varepsilon^{-1})'(\psi) f(\beta_\varepsilon^{-1}(\psi)) \cdot \nabla\theta + r(\beta_\varepsilon^{-1}(\theta)), \varphi), \\ l(\varphi; \chi) = &\ (\beta_\varepsilon^{-1}(\chi), \varphi). \end{aligned}$$

Both a and l are linear w.r.t. $\varphi \in H_0^1$. If ψ and χ are essentially bounded, the form a is bounded on $H_0^1 \times H_0^1$, while l becomes continuous on H^1. If the perturbation parameter ε satisfies the inequality $\varepsilon > C\tau$ for an appropriate constant C depending on F, r and the L^∞ norm for ψ and χ, coercivity can be easily proven for a. Based on these notations an auxiliary problem can be defined,

Problem AUX. Find $\theta \in H_0^1(\Omega)$ such that

$$a(\theta, \varphi; \psi) = l(\varphi; \chi) \tag{3.8}$$

holds for all $\varphi \in H_0^1(\Omega)$.

If ψ and χ are essentially bounded, existence and uniqueness of the solution for the above problem is guaranteed either by the nonlinear Lax-Milgram lemma or by the theory of monotone operators (see, e.g., [56]).

Clearly, Problem AUX is related to the elliptic problems produced by the nonlinear schemes RTI and RTC. It is easy to write the corresponding for Scheme RTL. The maximum principle is stated in the following lemma.

Lemma 3.2.1 *Assume (A1), (A3), $\psi \in V_k$ and $\chi \in V_{k-1}$. Then, for reasonably small τ, the solution of Problem AUX belongs to V_k.*

It is enough here to repeat the proof for the upper bounds in Lemma 2.2.1. Moreover, if χ is positive a.e. in Ω and $r(u) \geq 0$ for any u, the solution θ is bounded from below by 0. In fact, the additional assumption on r (global positivity) is necessary only when the reaction term is discretized implicitly and the same can be said regarding the restriction imposed to the time step.

Remark 3.2.3 *The above lemma establishes a maximum principle for the schemes RTI or RTC. Similarly, essential boundedness can be obtained for the linear scheme.*

In order to solve the nonlinear problems arising in the schemes RTI or RTC, the iterative procedures defined in the previous chapter can be applied here too. However, the regularization step - modifying the original nonlinearity function β - implies little modifications. For defining a correspondent to Iteration ISm let K be a constant above $1/\varepsilon$ and define, for $\psi, \varphi, \underline{\psi} \in H_0^1(\Omega)$

$$\begin{aligned} a_K(\psi, \varphi; \underline{\psi}) = &\quad K(\psi, \varphi) + \tau(\nabla\psi, \nabla\varphi) - \tau((\beta_\varepsilon^{-1})'(\underline{\psi}) f(\beta_\varepsilon^{-1}(\underline{\psi})) \cdot \nabla\psi, \varphi), \\ l_K(\underline{\psi}; \varphi) = &\quad K(\underline{\psi}, \varphi) + (\beta_\varepsilon^{-1}(\theta^{k-1}) - \beta_\varepsilon^{-1}(\underline{\psi}), \varphi) + \tau(r(\beta_\varepsilon^{-1}(\underline{\psi})), \varphi), \end{aligned}$$

which are linear and bounded. The iterative scheme relies on the operator $T : H_0^1 \to H_0^1$ giving the solution of the following problem

Problem PISm: Let $\psi \in H_0^1$. Find $T\psi \in H_0^1(\Omega)$ such that

$$a_K(T\psi, \varphi; \theta^{k-1}) = l_K(\psi; \varphi) \tag{3.9}$$

for all $\varphi \in H_0^1(\Omega)$.

Now the first iteration can be defined as

Iteration ISm:

$$\psi^{i+1} = T\psi^i \tag{3.10}$$

for $i \geq 0$ and $\psi^0 = \theta^{k-1} \in H_0^1$.

For τ reasonably small, yielding $K \geq C\tau/\varepsilon$, the Lax-Milgram lemma ensures the existence and uniqueness of a solution for Problem PISm. Moreover, if θ^{k-1} is taken in V_{k-1} defined above, the set V_k is invariant w.r.t. the operator T, the proof for this being a simple reproduction of the one for Lemma 2.2.1. Moreover, Lemmas 2.2.2 and 2.2.3 have a correspondent here, namely

Lemma 3.2.2 *Assume (A1), (A3) and $\theta^{k-1} \in V_{k-1}$. If K is greater than C/ε and $\tau \leq C'\varepsilon$ for appropriately chosen constants C and C', then there is a norm on $H_0^1(\Omega)$ equivalent to the usual one, such that T maps V_k in itself and is contractive on the same set.*

Proof. The proof is identical to the ones for the two lemmas mentioned above.

Remark 3.2.4 *Problem PISm is related to Scheme WRTC. The same holds also if θ^{k-1} in (3.9) is replaced by ψ, in order to obtain an iterative scheme for the implicit time discretization method. In this case we can obtain only weak convergence in H^1, but without imposing any relation between τ and ε.*

Remark 3.2.5 *The above lemma states a convergence result (in H^1) for the sequence $\{\psi^i\}_{i\geq 0} \subset V_k$. It is easy to see that its limit solves the problem WRTC, therefore we have*

$$\theta^k = \lim_{i\to\infty} \psi^i.$$

As already noticed in the previous chapter, Iteration ISm has a bad convergence rate. For practical computations we have used the following

Iteration IJK:

$$\begin{aligned}
&\bar{\theta}^i \in H_0^1(\Omega),\\
&(\sigma(\bar{\theta}^{i-1}, \theta^{k-1})(\bar{\theta}^i - \theta^{k-1}), \varphi) + \tau(\nabla\bar{\theta}^i, \nabla\varphi) - \tau((\beta_\varepsilon^{-1})'(\underline{\theta}) f(\beta_\varepsilon^{-1}(\underline{\theta})) \cdot \nabla\bar{\theta}^i, \varphi)\\
= \;& \tau(r(\beta_\varepsilon^{-1}(\underline{\theta})), \varphi),\\
&\sigma(\bar{\theta}^i, \theta^k) = \int_0^1 \left(\beta_\varepsilon^{-1}\right)' (s\bar{\theta}^i + (1-s)\theta^k) ds
\end{aligned} \tag{3.11}$$

for all $\varphi \in H_0^1(\Omega)$ and $i \geq 1$, where $\bar{\theta}^0 = \theta^{k-1}, \sigma(\bar{\theta}^0, \theta^{k-1}) = \left(\beta_\varepsilon^{-1}\right)' (\theta^{k-1})$.

Iteration JK is more appropriate for practical purposes (see [46], [47] or [53]). Without giving the proof - which is similar to the one for Lemma 3.2.1 - if $\theta^{k-1} \in V_{k-1}$, then all the elements of the sequence of solutions $\{\bar{\theta}^i\}_{i=0}^{\infty}$ given by this iteration are in V_k. Since the problems WRTC and WRTI have unique solutions, compactness arguments can be considered for showing that $\bar{\theta}^i$ converges weakly in H^1 to the semi-discrete solution θ_k, a function in V_k. Particularly, taking $\underline{\theta} = \theta^{k-1}$ above leads to an iterative method for Problem WRTC, while the choice $\underline{\theta} = \bar{\theta}^{i-1}$ gives the alternative for Problem WRTI. Because we deal here with a perturbation of the nonlinearity β_ε, there is no need to consider a a cut-off approach for the regularization, as done in [46] and [47].

3.2.2 Error estimates for the semi-discrete approximation

In order to show the convergence of the schemes considered here it is enough to obtain some error estimates. Thinking at the maximum principle based algorithms, there are no essential changes in the proofs, the basic ideas being the same.

The apriori estimates

The first step in getting some bounds for the errors consists in proving apriori estimates for the semi-discrete solutions. For the implicit scheme, this is done in the following theorem.

Theorem 3.2.3 *Assume (A1), (A2) and (A3). Then, for $p < n$, if θ^k solves Problem WRTI, we have*

$$\tau \sum_{k=1}^{p} \|\nabla\theta^k\|^2 \leq C, \tag{3.12}$$

$$\sum_{k=1}^{p}(\beta^{-1}(\theta^k) - \beta^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}) + \sum_{k=1}^{p} \|\theta^k - \theta^{k-1}\|^2 \leq C. \tag{3.13}$$

Proof. The first inequality can be proven in the same manner as done for the implicit scheme MTI. Here β should be replaced by its perturbation β_ε, having similar properties. Now $\varphi = \theta^k \in H_0^1(\Omega)$ can be taken as a test function in Problem WRTI. Summing up for $k = \overline{0,p}$ we get

$$\begin{aligned} &\sum_{k=1}^{p}(\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}), \theta^k) + \tau \sum_{k=1}^{p} \|\nabla\theta^k\|^2 \\ = \quad & -\tau \sum_{k=1}^{p}(F(\beta_\varepsilon^{-1}(\theta^k)), \nabla\theta^k) + \tau \sum_{k=1}^{p}(r(\beta_\varepsilon^{-1}(\theta^k)), \theta^k). \end{aligned} \tag{3.14}$$

For estimating the first term we proceed as in Theorem 2.2.6 with β_ε instead of β and obtain

$$\sum_{k=1}^{p}(\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}), \theta^k) \geq -C.$$

By taking a primitive function for $F(\beta_\varepsilon^{-1}(\cdot))$ it can be shown that the first term in the right hand side disappears (as done in (2.12)). The last sum is also uniformly bounded w.r.t. k since $\|\theta_k\| \leq C$ for any k. This gives the first part of the theorem.

For the second estimate, if $k \geq 2$, φ in (3.4) can be replaced by $\theta^k - \theta^{k-1} \in H_0^1(\Omega)$ for any k greater than 2. Summing up again over k from 2 to p and applying the elementary identity in (2.18) yields

$$\begin{aligned} (I) + (II) :=& \sum_{k=2}^{p}(\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}) \\ &+ \frac{\tau}{2}\left(\|\nabla\theta^p\| - \|\nabla\theta^1\| + \sum_{k=2}^{p}\|\nabla(\theta^k - \theta^{k-1})\|^2\right) \\ =& -\tau \sum_{k=2}^{p}(F(\beta_\varepsilon^{-1}(\theta^k)), \nabla(\theta^k - \theta^{k-1})) \\ &+ \tau \sum_{k=2}^{p}(r(\beta_\varepsilon^{-1}(\theta^k)), \theta^k - \theta^{k-1}) =: (III) + (IV). \end{aligned} \tag{3.15}$$

The first inequality in (3.5) gives

$$(I) \geq C \sum_{k=2}^{p} \|\theta^k - \theta^{k-1}\|^2 + \frac{1}{2} \sum_{k=2}^{p} (\beta_\varepsilon^{-1}(\theta^k) - \beta_\varepsilon^{-1}(\theta^{k-1}), \theta^k - \theta^{k-1}).$$

As a consequence of the first part of this theorem we get

$$(II) \geq \frac{\tau}{2} \sum_{k=2}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2 - C$$

for a generic constant C. For (III) we go on as in Remark 2.2.13 in the chapter before and obtain

$$|(III)| \leq C + \frac{\tau}{2} \sum_{k=2}^{p} \|\nabla(\theta^k - \theta^{k-1})\|^2,$$

Based on the essential boundedness of the semi-discrete solutions, the estimates for the last term are trivial,

$$|(IV)| \leq C.$$

The above inequalities applied into (3.15) show the remaining part of the theorem. Here the sum can be taken from 1 since both θ^0 and θ^1 are bounded in the L^2 norm.

Remark 3.2.6 *It does not make any difference for the apriori estimates if the reaction term is treated explicitly.*

Remark 3.2.7 *The estimates obtained here are worser than those for the implicit scheme in the previous chapter. But now the assumptions on the initial data are weaker. However, if θ^0 is taken in H^1 (as mentioned in the previous section), a similar approach lead to estimates similar to those for the maximum principle based algorithm (Theorem 2.2.6). The bounds in the second inequality there become $C\tau/\varepsilon$. Moreover, if F is Lipschitz continuous in $\beta(u)$, the apriori estimates are optimal, namely $C\tau$.*

For the remaining two schemes proposed in the beginning we have to take the initial approximation θ^0 in $H^1(\Omega)$. Now the methods are completely the same as in the previous chapter. Analogous to Scheme MTC, Scheme RTC yields

Theorem 3.2.4 *Assume (A1), (A2), (A3) and θ^0 in H^1. Then, for $p < n$, if θ^k solves Problem WRTC, there are constants C independent on p, τ and ε such that*

$$\tau \sum_{k=1}^{p} \|\nabla \theta^k\|^2 \leq \frac{C}{\varepsilon}, \tag{3.16}$$

$$\sum_{k=1}^{p}(\beta_\varepsilon^{-1}(\theta^k)-\beta_\varepsilon^{-1}(\theta^{k-1}),\theta^k-\theta^{k-1})+\sum_{k=1}^{p}\|\theta^k-\theta^{k-1}\|^2$$
$$+\tau\sum_{k=1}^{p}\|\nabla(\theta^k-\theta^{k-1})\|^2\leq C\frac{\tau}{\varepsilon^2}. \tag{3.17}$$

Proof. The proof is almost identical to the one for Theorem 2.2.7. The only difference appears in the estimates, where $1/\varepsilon$ replaces $C_2(\varepsilon)$, because of the bounds in Remark 3.2.2 for the term $(\beta_\varepsilon^{-1})'(\theta)f(\beta_\varepsilon^{-1}(\theta))$.

Remark 3.2.8 *As for the implicit case, optimal estimates are obtained if F is Lipschitz continuous in $\beta(u)$. Then ε disappears in the above inequalities. Moreover, an explicit discretization of the reaction term gives the same results.*

Now our attention turns to the linear scheme. Yet the Lipschitz continuity of β' is necessary. The estimates in the previous chapter were obtained either assuming F to be Lipschitz w.r.t. $\beta(u)$ or for convex nonlinearities. The last alternative works only if positive solutions are sought, and positivity is needed for the approximate solutions too (this property is assured by the maximum principle for the elliptic problems). In this case the perturbation β_ε should be also convex, which is guaranteed only in the construction proposed in (3.2).

Theorem 3.2.5 *Assume (A1), (A2), (A3) and one of the alternatives mentioned above. Then, for $p<n$, if θ^k solves Problem WRTL, we have*

$$\sum_{k=1}^{p}\|\sqrt{\sigma_{k-1}}(\theta^k-\theta^{k-1})\|^2+\tau\|\nabla\theta^p\|^2+\tau\sum_{k=1}^{p}\|\nabla(\theta^k-\theta^{k-1})\|^2\leq C\tau \tag{3.18}$$

if F is Lipschitz w.r.t. $\beta(u)$, or

$$\tau\sum_{k=1}^{p}\|\nabla\theta^k\|^2\leq\frac{C}{\varepsilon}, \tag{3.19}$$

$$\sum_{k=1}^{p}\|\sqrt{\sigma_{k-1}}(\theta^k-\theta^{k-1})\|^2+\tau\|\nabla\theta^p\|^2+\tau\sum_{k=1}^{p}\|\nabla(\theta^k-\theta^{k-1})\|^2\leq C\frac{\tau}{\varepsilon^{\frac{3}{2}}} \tag{3.20}$$

for convex β and positive initial and boundary data in Problem P.

Proof. Here we can follow the proofs for Theorems 2.2.8 and 2.2.9. In the first case, Corollary 3.1.2 shows that F is Lipschitz w.r.t. $\beta_\varepsilon(u)$ too, hence the constant C_F is an upper bound for $(\beta^{-1})'(\theta)f(\beta^{-1}(\theta))$. As mentioned in Remarks 3.2.7 and 3.2.8, the estimates become optimal.

Working with the global perturbation defined in (3.2), if β is convex, the same holds for β_ε. If the solution of Problem WP is positive, its semi-discrete approximation have similar properties (as stated in Lemma 3.2.1). In this case $\varphi = \theta^k \in H_0^1(\Omega)$ is a positive test function in (3.6). Replacing β by β_ε in the proof of Theorem 2.2.8 for the maximum principle based approach we get the desired results.

Error estimates for the implicit method

As mentioned before, the convergence of the schemes is shown by obtaining error estimates. This is done in the framework proposed in the previous chapter. Therefore most of the notations are the same, e.g.

$$\bar{f}^k := \frac{1}{\tau}\int_{(k-1)\tau}^{k\tau} f(s,\cdot)ds,$$

for any function f integrable in time and defined in Q_T - if $k \geq 1$ - and $\bar{f}^0 := f(0,\cdot)$. The errors are obtained in terms of e_u^k and e_θ^k given below

$$e_u^k := \bar{u}^k - \beta_\varepsilon^{-1}(\theta^k), \quad e_\theta^k := \overline{\beta(u)}^k - \theta^k,$$

where $k \geq 0$. Again, $G : H^{-1}(\Omega) \to H_0^1(\Omega)$ denotes the Green operator defined by

$$(\nabla G\psi, \nabla\varphi) = (\psi, \varphi), \tag{3.21}$$

for all $\varphi \in H_0^1(\Omega)$, where ψ is taken in $H^{-1}(\Omega)$, so

$$\|\nabla G\psi\| = \|\psi\|_{-1}, \quad \|\psi\|_{-1} \leq C\|\psi\| \tag{3.22}$$

(where the last inequality applies only if $\psi \in L^2(\Omega)$).

The error estimates for the implicit scheme are given in the following theorem.

Theorem 3.2.6 *Assuming (A1), (A2) and (A3), if u is the weak solution of Problem WP and θ^k solves for each $k > 0$ Problem WRTI, then*

$$\sup_{k=\overline{1,n}} \|e_u^k\|_{-1}^2 + \int_0^T (\beta_\varepsilon(u(t)) - \theta_\Delta(t), u(t) - \beta_\varepsilon^{-1}(\theta_\Delta(t)))dt + \|\beta(u) - \theta_\Delta\|_{L^2(Q)}^2$$
$$\leq C\left\{\tau + \varepsilon\right\},$$

where $\theta_\Delta(t) = \theta^k$ for $t \in (t_{k-1}, t_k]$ and $k = \overline{1,n}$.

Proof. We follow here essentially the steps done when proving Theorem 2.2.11. If $\chi|_I$ is the characteristic function of the time interval $I \subset [0,T]$, choosing $\varphi\chi|_{[t_{j-1},t_j]}$ for an arbitrary $\varphi \in H_0^1(\Omega)$ as test function in (3.6) gives

$$\begin{aligned}(u(t_j)-u(t_{j-1}),\varphi)+&\left(\nabla\int_{t_{j-1}}^{t_j}\beta(u(t))dt+\int_{t_{j-1}}^{t_j}F(u(t))dt,\nabla\varphi\right)\\ &=\left(\int_{t_{j-1}}^{t_j}r(u(t))dt,\varphi\right),\end{aligned}\tag{3.23}$$

for any j between 1 and n. Taking $\varphi = Ge_u^j \in H_0^1$ in (3.4) and (3.23), subtracting the first from the second one and summing up for $j=\overline{1,k}$ yields

$$\begin{aligned}(I_1)+(I_2) :=\quad &\sum_{j=1}^{k}(u(t_j)-u(t_{j-1})-\beta_\varepsilon^{-1}(\theta^j)+\beta_\varepsilon^{-1}(\theta^{j-1}),Ge_u^j)\\ &+\tau\sum_{j=1}^{k}(\nabla e_\theta^j,\nabla Ge_u^j)\\ &=-\sum_{j=1}^{k}\left(\int_{t_{j-1}}^{t_j}F(u(t))-F(\beta_\varepsilon^{-1}(\theta^j))dt,\nabla Ge_u^j\right)\\ &+\sum_{j=1}^{k}\left(\int_{t_{j-1}}^{t_j}r(u(t))-r(\beta_\varepsilon^{-1}(\theta^j))dt,Ge_u^j\right)=:(I_3)+(I_4).\end{aligned}\tag{3.24}$$

The terms in (3.24) are estimated as done in the proof of the corresponding theorem for Scheme MTI. (I_1) can be decomposed as follows.

$$\begin{aligned}(I_1) =\quad &\textstyle\sum_{j=1}^{k}(u(t_j)-\bar u^j-u(t_{j-1})+\bar u^{j-1},Ge_u^j)+\sum_{j=1}^{k}(e_u^j-e_u^{j-1},Ge_u^j)\\ =:\ &(I_{11})+(I_{12}).\end{aligned}$$

Recalling the steps performed in (2.38) and (2.39), since $\partial_t u \in L^2(0,T;H^{-1}(\Omega))$, (I_{11}) gets

$$\begin{aligned}|(I_{11})| \le\ &\tau\delta_{111}\,\|\partial_t u\|^2_{L^2(t_{k-1},t_k;H^{-1}(\Omega))}+\frac{1}{4\delta_{111}}\|e_u^k\|_{-1}^2\\ &+\tau\delta_{112}\,\|\partial_t u\|^2_{L^2(0,t_{k-1};H^{-1}(\Omega))}+\frac{1}{4\delta_{112}}\sum_{j=2}^{k}\|e_u^j-e_u^{j-1}\|_{-1}^2,\end{aligned}\tag{3.25}$$

where δ_{111} and δ_{112} are positive constants chosen below.

For (I_{12}) we make use of the identities in (2.18) and (3.22) and obtain

$$\begin{aligned}(I_{12}) =& \sum_{j=1}^{k} (e_u^j - e_u^{j-1}, Ge_u^j) = \sum_{j=1}^{k} (\nabla Ge_u^j - \nabla Ge_u^{j-1}, \nabla Ge_u^j) \\ =& \frac{1}{2}\left(\|e_u^k\|_{-1}^2 - \|e_u^0\|_{-1}^2 + \sum_{j=1}^{k} \|e_u^j - e_u^{j-1}\|_{-1}^2 \right).\end{aligned} \tag{3.26}$$

Now we go on with the second term in (3.24).

$$\begin{aligned}(I_2) =& \ \tau \textstyle\sum_{j=1}^{k} (e_\theta^j, e_u^j) \\ =& \ \textstyle\sum_{j=1}^{k} \left(\int_{t_{j-1}}^{t_j} (\beta(u(t)) - \theta^j) dt, \frac{1}{\tau} \int_{t_{j-1}}^{t_j} (u(s) - \beta_\varepsilon^{-1}(\theta^j)) ds \right) \\ =& \ \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta(u(t)) - \theta^j, u(t) - \beta_\varepsilon^{-1}(\theta^j))\, dt \\ & + \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \left(\beta(u(t)) - \theta^j, \frac{1}{\tau} \int_{t_{j-1}}^{t_j} (u(s) - u(t)) ds \right) dt \\ =:& \ (I_{21}) + (I_{22}).\end{aligned}$$

(I_{21}) can be decomposed into two sums

$$\begin{aligned}(I_{21}) =& \ \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta_\varepsilon(u(t)) - \theta^j, u(t) - \beta_\varepsilon^{-1}(\theta^j))\, dt \\ & + \textstyle\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (\beta(u(t)) - \beta_\varepsilon(u(t)), u(t) - \beta_\varepsilon^{-1}(\theta^j))\, dt \\ =:& \ (I_{211}) + (I_{212}).\end{aligned}$$

The properties of β_ε stated in Lemma 3.1.1 are useful in the estimates for the two terms from above.

$$\begin{aligned}(I_{211}) \geq & \frac{1}{2} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \big(\beta_\varepsilon(u(t)) - \theta^j, u(t) - \beta_\varepsilon^{-1}(\theta^j)\big)\, dt \\ & + \frac{C}{2} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|\beta_\varepsilon(u(t)) - \theta^j\|^2 dt.\end{aligned} \tag{3.27}$$

For (I_{212}), since u is essentially bounded in the parabolic cylinder Q_T, it follows that u belongs to $L^2((t_{j-1}, t_j) \times \Omega)$ for any $j > 0$. Because of the maximum principle for the semi-discrete approximations, we can go on as follows

$$
\begin{aligned}
|(I_{212})| \leq & \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|\beta_\varepsilon(u(t)) - \beta(u(t))\| \|u(t) - \beta_\varepsilon^{-1}(\theta^j)\| dt \\
\leq & \sum_{j=1}^{k} \left(\int_{t_{j-1}}^{t_j} \|\beta_\varepsilon(u(t)) - \beta(u(t))\|^2 dt \right)^{\frac{1}{2}} \left(\int_{t_{j-1}}^{t_j} \|u(t) - \beta_\varepsilon^{-1}(\theta^j)\|^2 dt \right)^{\frac{1}{2}} \\
\leq & \frac{1}{2\delta_{212}^2} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|\beta_\varepsilon(u(t)) - \beta(u(t))\|^2 dt + \frac{\delta_{212}^2}{2} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|u(t) - \beta_\varepsilon^{-1}(\theta^j)\|^2 dt \\
\leq & \frac{C}{\delta_{212}^2} \varepsilon^2 \|u\|_{L^2(Q_T)}^2 + C\delta_{212}^2.
\end{aligned}
\tag{3.28}
$$

For $\delta_{212} = \varepsilon^{\frac{1}{2}}$ the bound above become $C\varepsilon$.

(I_{22}) can be treated as done in (2.44).

$$
\begin{aligned}
|(I_{22})| = & \frac{1}{\tau} \left| \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \left(\beta(u(t)) - \theta^j, \int_{t_{j-1}}^{t_j} \int_t^s \partial_r u(r) dr ds \right) dt \right| \\
\leq & \frac{1}{\tau} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \int_{t_{j-1}}^{t_j} \|\nabla(\beta(u(t)) - \theta^j)\| \int_t^s \|\partial_r u(r)\|_{-1} dr ds dt \\
\leq & \tau \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|\nabla(\beta(u(t)) - \theta^j)\| \|\partial_r u(t)\|_{-1} dt \leq C\tau,
\end{aligned}
\tag{3.29}
$$

where the regularity properties of the solution u and the apriori estimates in Theorem 3.2.3 have been used.

Considering now the right hand side in (3.24), because of the assumption (A3) on F, Corollary 3.1.2 can be used again in order to get

$$
\begin{aligned}
|(I_3)| \leq & \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} \|F(u(t)) - F(\beta_\varepsilon^{-1}(\theta^j))\| \|\nabla G e_u^j\| dt \\
\leq & C \sum_{j=1}^{k} \|e_u^j\|_{-1} \int_{t_{j-1}}^{t_j} (u(t) - \beta_\varepsilon^{-1}(\theta^j), \beta_\varepsilon(u(t)) - \theta^j)^{\frac{1}{2}} dt \\
\leq & C\tau^{\frac{1}{2}} \sum_{j=1}^{k} \|e_u^j\|_{-1} \left(\int_{t_{j-1}}^{t_j} (u(t) - \beta_\varepsilon^{-1}(\theta^j), \beta_\varepsilon(u(t)) - \theta^j) dt \right)^{\frac{1}{2}} \\
\leq & \frac{C}{4\delta_3} \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} (u(t) - \beta_\varepsilon^{-1}(\theta^j), \beta_\varepsilon(u(t)) - \theta^j) dt + C\tau\delta_3 \sum_{j=1}^{k} \|e_u^j\|_{-1}^2.
\end{aligned}
\tag{3.30}
$$

Proceeding exactly in the same manner as before, the Poincaré inequality applied to Ge_u^j leads to the same estimates for (I_4). Inserting all the inequalities in (3.25) - (3.30) in (3.24) and choosing the δ's properly, we get

$$\begin{aligned} &\|e_u^k\|_{-1}^2 + \textstyle\sum_{j=1}^k \|e_u^j - e_u^{j-1}\|_{-1}^2 \\ &+ \textstyle\sum_{j=1}^k \int_{t_{j-1}}^{t_j} (\beta_\varepsilon(u(t)) - \theta^j, u(t) - \beta_\varepsilon^{-1}(\theta^j))dt + C \textstyle\sum_{j=1}^k \int_{t_{j-1}}^{t_j} \|\beta_\varepsilon(u(t)) - \theta^j\|^2 dt \\ &\leq C\tau \|\partial_t u\|_{L^2(0,t_k;H^{-1}(\Omega))}^2 + C(\tau + \varepsilon + \|e_u^0\|_{-1}^2) + C\tau \textstyle\sum_{j=1}^k \|e_u^j\|_{-1}^2. \end{aligned}$$

As noticed in the beginning of the chapter, the semi-discrete initial data can be chosen as $\theta^0 \equiv \beta_\varepsilon(u_0)$, therefore the initial error e_u^0 is zero. Since $u \in H^1(0,T;H^{-1}(\Omega))$, applying now the discrete Gronwall inequality yields

$$\|e_u^k\|_{-1}^2 + \int_0^{t_k} (\beta_\varepsilon(u) - \theta_\Delta, u - \beta_\varepsilon^{-1}(\theta_\Delta)) + \|\beta_\varepsilon(u) - \theta_\Delta\|_{L^2(0,t_k;L^2(\Omega))}^2 \leq C\,(\tau + \varepsilon) \quad (3.31)$$

for any $k \geq 0$.

Now we observe that

$$\|\beta_\varepsilon(u) - \theta^j\| \geq |\|\beta(u) - \theta^j\| - \|\beta_\varepsilon(u) - \beta(u)\|| \geq |\|\beta(u) - \theta^j\| - C\varepsilon|\,,$$

and so

$$2\|\beta_\varepsilon(u) - \theta^j\|^2 \geq \|\beta(u) - \theta^j\|^2 - C\varepsilon^2.$$

This, together with (3.31), gives the desired result.

Remark 3.2.9 *As for the implicit scheme MTI in the chapter before, the above error estimates, $O(\tau^{\frac{1}{2}})$, are at least as good as those for the algorithms in [60], [73], [46], [4] or [28]. However, in particular cases, it is possible to get better estimates which are discussed later.*

Remark 3.2.10 *An explicit discretization of the reaction term r does not affect the estimates. Then, the last term in (3.24) becomes*

$$\begin{aligned} (\bar{I}_4) = &\ \textstyle\sum_{j=1}^k \left(\int_{t_{j-1}}^{t_j} r(u(t)) - r(\beta_\varepsilon^{-1}(\theta^{j-1}))dt, Ge_u^j \right) \\ = &\ (I_4) + \tau \textstyle\sum_{j=1}^k (r(\beta_\varepsilon^{-1}(\theta^{j-1})) - r(\beta_\varepsilon^{-1}(\theta^j)), Ge_u^j). \end{aligned}$$

Recalling the assumption on r in (A3), the last sum yields

$$|(\bar{I}_4)| \leq |(I_4)| + C\tau \sum_{j=1}^k \|e_u^j\|_{-1} (\beta_\varepsilon^{-1}(\theta^j) - \beta_\varepsilon^{-1}(\theta^{j-1}), \theta^j - \theta^{j-1})^{\frac{1}{2}}.$$

The apriori estimates in (3.13) give

$$|(\bar{I}_4)| \leq |(I_4)| + C\tau + C\tau \sum_{j=1}^{k} \|e_u^j\|_{-1}^2$$

and the rest follows as before.

Remark 3.2.11 *The error estimates for u are obtained in the $H^1(0,T;H^{-1}(\Omega))$ norm. If an inequality of the form*

$$(\beta^{-1}(\theta) - \beta^{-1}(\psi))(\theta - \psi) \geq C(\beta^{-1}(\theta) - \beta^{-1}(\psi))^p$$

holds true for a positive constant C and an exponent $p > 1$. Lemma 3.1.1 allows us to rewrite the above inequality in terms of β_ε^{-1}. Then, using the estimate for the scalar product in (3.31), an error estimate in the better $L^p(Q)$ norm can be obtained. For example, if $\beta(u) = u^m$, the estimate becomes

$$\|u - \beta_\varepsilon^{-1}(\theta_\Delta)\|_{L^{m+1}(Q_T)}^{m+1} \leq C(\tau + \varepsilon)$$

(see, e.g. [73], [28]).

As we have already mentioned, the estimates obtained here are not optimal. Based on the semigroup theory, if the generator is a subgradient in a Hilbert space, the order of convergence of the implicit Euler method becomes $O(\tau)$ ([85], see also Theorem 2.2.14). An example in this sense is Problem P without convection or reaction, if β is maximal monotone supposing its range is $\mathbb{R}$. This result may be used here in order to obtain optimal estimates for the implicit scheme RTI. The approach is identical to the one for the maximum principle based algorithm and relies on the comparison between Scheme EI defined in (2.50) and RTI, in a variational formulation.

Theorem 3.2.7 *In the setting of Theorem 3.2.6, if θ^k solves for each $k > 0$ Problem WRTI without convection or reaction, then*

$$\|\beta(u) - \theta_\Delta\|_{L^2(Q)}^2 + \tau\|\beta(u) - \theta_\Delta\|_{L^2(0,T;H_0^1(\Omega))}^2 \leq C\left\{\tau^2 + \varepsilon^2\right\}.$$

Proof. Recalling the estimates in for the Euler implicit scheme EI (Theorem 2.2.14), it is enough to compare the semi-discrete approximation generated by Scheme RTI with the one corresponding to the above mentioned. To do so, Scheme EI is rewritten in a weak form

$$(u^j - u^{j-1}, \varphi) + \tau(\nabla\beta(u^j), \nabla\varphi) = 0,$$

for all $\varphi \in H_0^1(\Omega)$, with $j = \overline{1,n}$. Subtracting (3.4) from this equality, summing up for $j = \overline{1,k}$, taking $\varphi = \beta(u^k) - \theta^k \in H_0^1$ in the resulting difference and summing up again for $k = \overline{1,n}$ gives

$$
\begin{aligned}
(I_1) + (I_2) := &\sum_{k=1}^{n} (u^k - \beta_\varepsilon^{-1}(\theta^k), \beta(u^k) - \theta^k) \\
&+ \tau \sum_{k=1}^{n} \sum_{j=1}^{k} (\nabla(\beta(u^j) - \theta^j), \nabla(\beta(u^k) - \theta^k)) \\
&= \sum_{k=1}^{n} (u^0 - \beta_\varepsilon^{-1}(\theta^0), \beta(u^k) - \theta^k) =: (I_3).
\end{aligned}
\tag{3.32}
$$

If we take $\theta^0 \equiv \beta_\varepsilon(u_0)$ as initial data in Scheme RTI, the right hand side above vanishes. Hence it is enough to estimate the first two terms. First, (I_1) can be decomposed in

$$
(I_1) = \sum_{k=1}^{n} (u^k - \beta_\varepsilon^{-1}(\theta^k), \beta_\varepsilon(u^k) - \theta^k) + \sum_{k=1}^{n} (u^k - \beta_\varepsilon^{-1}(\theta^k), \beta(u^k) - \beta_\varepsilon(u^k)) =: (I_{11}) + (I_{12}).
$$

Recalling Lemma 3.1.1, (I_{11}) gives

$$
(I_{11}) \geq C \sum_{k=1}^{n} \|\beta_\varepsilon(u^k) - \theta^k\|^2 \geq \frac{C}{2} \sum_{k=1}^{n} \|\beta(u^k) - \theta^k\|^2 - \frac{C}{\tau}\varepsilon^2. \tag{3.33}
$$

As a consequence of the maximum principle for the solution of the implicit scheme, (I_{12}) gives

$$
\begin{aligned}
|(I_{12})| \leq &C\varepsilon \sum_{k=1}^{n} \|u^k - \beta_\varepsilon^{-1}(\theta^k)\| \leq C\varepsilon \sum_{k=1}^{n} \|\beta_\varepsilon(u^k) - \theta^k\| \\
\leq &\frac{C}{\tau}\varepsilon^2 + \frac{C}{4} \sum_{k=1}^{n} \|\beta_\varepsilon(u^k) - \theta^k\|^2.
\end{aligned}
\tag{3.34}
$$

For (I_2), the elementary identity in (2.19) yields

$$
(I_2) = \frac{\tau}{2} \sum_{k=1}^{n} \|\nabla(\beta(u^k) - \theta^k)\|^2 + \frac{\tau}{2} \|\nabla \sum_{k=1}^{n} (\beta(u^k) - \theta^k)\|^2. \tag{3.35}
$$

Inserting all the inequalities in (3.33) - (3.35) in (3.32) and multiplying everything with τ we arrive at

$$
\frac{C}{4} \sum_{k=1}^{n} \tau \|\beta(u^k) - \theta^k\|^2 + \frac{\tau^2}{2} \sum_{k=1}^{n} \|\nabla(\beta(u^k) - \theta^k)\|^2 \leq C\varepsilon^2. \tag{3.36}
$$

The rest of the proof results from Theorem 2.2.14.

Remark 3.2.12 *This result shows that if the implicit Euler discretization has a linear order of convergence, the same holds also for Scheme RTI. Therefore, from theoretical point of view, the implicit scheme RTI behaves at least as good as other schemes.*

Error estimates for the simplified schemes

Here we consider the simplified schemes RTC and RTL. Since the apriori estimates were obtained under the assumption $\theta^0 \in H^1$, this is assumed to hold true here. Hence the initial data should be taken in a special way, as mentioned, e.g., in the first section of this chapter. Moreover, the initial error $\|e_u^0\|_{-1}$ plays also an essential role in the error estimates. For the implicit scheme RTI, its effect was neglected because of the choice for θ^0, namely $\beta_\varepsilon(u_0)$. But in general this function does not belong to $H^1(\Omega)$ and therefore we had to consider other alternatives, as mentioned before. In the forthcoming, the estimates depend on the initial error $\|e_u^0\|_{-1}$. For any of the situations discussed in this chapter, this error vanishes as ε goes to 0, but theoretically the convergence order may not be so good. The next result applies to Scheme RTC.

Theorem 3.2.8 *In the setting of Theorem 3.2.6, if $\theta^0 \in H^1$, the following estimates can be obtained for Scheme RTC*

$$\sup_{k=\overline{1,n}} \|e_u^k\|_{-1}^2 + \int_0^T (\beta(u(t)) - \theta_\Delta(t), u(t) - \beta_\varepsilon^{-1}(\theta_\Delta(t)))dt + \|\beta(u) - \theta_\Delta\|_{L^2(Q)}^2$$
$$\leq C \left\{ \frac{\tau}{\varepsilon^3} + \varepsilon + \|e_u^0\|_{-1}^2 \right\}.$$

Proof. The steps in proving Theorem 3.2.6 can be repeated here, but some estimates are not the same. First, because of the apriori estimates in (3.16), $|(I_{22})|$ is bounded from above by $C\tau/\varepsilon$ in (3.29). But the main difference appears when dealing with the convection part, where we get

$$\begin{aligned}(\bar{I}_3) = (I_3) \quad & + \sum_{j=1}^k \int_{t_{j-1}}^{t_j} (F(\beta_\varepsilon^{-1}(\theta^{j-1})) - F(\beta_\varepsilon^{-1}(\theta^j)), \nabla G e_u^j) dt \\ & + \sum_{j=1}^k \int_{t_{j-1}}^{t_j} ((\beta_\varepsilon^{-1})'(\theta^{j-1}) f(\beta_\varepsilon^{-1}(\theta^{j-1})) \nabla(\theta^{j-1} - \theta^j), G e_u^j) dt,\end{aligned}$$

the last terms being denoted by $(\bar{I}_{31})$ and $(\bar{I}_{32})$. The estimates for (I_3) in Theorem 3.2.6 are valid in this case too, while $(\bar{I}_{31})$ gives

$$\begin{aligned}|(\bar{I}_{31})| \leq \;& C\tau \sum_{j=1}^k \|e_u^j\|_{-1} (\beta_\varepsilon^{-1}(\theta^j) - \beta_\varepsilon^{-1}(\theta^{j-1}), \theta^j - \theta^{j-1})^{\frac{1}{2}} \\ \leq \;& C\tau \sum_{j=1}^k (\beta_\varepsilon^{-1}(\theta^j) - \beta_\varepsilon^{-1}(\theta^{j-1}), \theta^j - \theta^{j-1}) + C\tau \sum_{j=1}^k \|e_u^j\|_{-1}^2.\end{aligned}$$

Recalling Remark 3.2.2 and the apriori estimates in Theorem 3.2.4, $(\bar{I}_{32})$ can be bounded as follows

$$\begin{aligned}|(\bar{I}_{32})| \leq\ & C\frac{\tau}{\sqrt{\varepsilon}}\sum_{j=1}^{k}\|\nabla(\theta^j-\theta^{j-1})\|\|Ge_u^j\| \\ \leq\ & C\tau\sum_{j=1}^{k}\|\nabla Ge_u^j\|^2 + C\frac{\tau}{\varepsilon}\sum_{j=1}^{k}\|\nabla(\theta^j-\theta^{j-1})\|^2 \\ \leq\ & C\tau\sum_{j=1}^{k}\|e_u^j\|_{-1}^2 + C\frac{\tau}{\varepsilon^3}.\end{aligned}$$

Now, the proof continues as the one for Theorem 2.2.11.

Remark 3.2.13 *As before, there is no difference when r is discretized explicitly. Moreover, if F is Lipschitz continuous w.r.t. $\beta(u)$, the error estimates are identical to the ones for the implicit scheme.*

Yet our attention turns to the linear scheme. The following theorem give the error bounds for the semi-discrete approximation provided by Scheme RTL.

Theorem 3.2.9 *In the setting of Theorem 3.2.8, assuming additionally the Lipschitz continuity of F w.r.t. $\beta(u)$ and the same property (w.r.t. u) for β', if θ^k solves for each $k > 0$ Problem WRTL, then*

$$\sup_{k=\overline{1,n}}\|e_u^k\|_{-1}^2 + \int_0^T(\beta(u(t))-\theta_\Delta(t), u(t)-\beta_\varepsilon^{-1}(\theta_\Delta(t)))dt + \|\beta(u)-\theta_\Delta\|_{L^2(Q)}^2$$
$$\leq C\left\{\frac{\tau}{\varepsilon^2}+\varepsilon+\|e_u^0\|_{-1}^2\right\},$$

with θ_Δ being defined in Theorem 2.2.11.

Proof. The proof is similar to the ones above. The same steps lead to

$$\sum_{j=1}^{k}(u(t_j)-u(t_{j-1})-\sigma_{j-1}(\theta^j-\theta^{j-1}),Ge_u^j)+\tau\sum_{j=1}^{k}(\nabla e_\theta^j,\nabla Ge_u^j)$$
$$=-\sum_{j=1}^{k}\left(\int_{t_{j-1}}^{t_j}F(u(t))-F(\beta_\varepsilon^{-1}(\theta^j))dt,\nabla Ge_u^j\right)$$
$$-\tau\sum_{j=1}^{k}\left(F(\beta_\varepsilon^{-1}(\theta^j))-F(\beta_\varepsilon^{-1}(\theta^{j-1})),\nabla Ge_u^j\right) \tag{3.37}$$
$$+\tau\sum_{j=1}^{k}\left((\beta_\varepsilon^{-1})'(\theta^{j-1})f(\beta_\varepsilon^{-1}(\theta^{j-1}))\nabla(\theta^j-\theta^{j-1}),Ge_u^j\right)$$
$$+\sum_{j=1}^{k}\left(\int_{t_{j-1}}^{t_j}r(u(t))-r(\beta_\varepsilon^{-1}(\theta^{j-1}))dt,Ge_u^j\right).$$

Following the proof for Theorem 2.2.13 - where β is replaced everywhere by β_ε - we get

$$|\sum_{j=1}^{k}(\beta_\varepsilon^{-1}(\theta^j)-\beta_\varepsilon^{-1}(\theta^{j-1})-\sigma_{j-1}(\theta^j-\theta^{j-1}),Ge_u^j)|\le C\frac{\tau}{\varepsilon^2},$$

and anything else is as before.

Remark 3.2.14 *Similar to the apriori estimates in Theorem 3.2.5, if F does not satisfy the stronger assumption, error bounds can be obtained for a convex nonlinearity β, but only if the solution is positive a.e.. In this case the result is worser, namely*

$$\sup_{k=\overline{1,n}}\|e_u^k\|_{-1}^2+\int_0^T(\beta(u(t))-\theta_\Delta(t),u(t)-\beta_\varepsilon^{-1}(\theta_\Delta(t)))dt+\|\beta(u)-\theta_\Delta\|_{L^2(Q)}^2$$
$$\le C\left\{\frac{\tau}{\varepsilon^{\frac{7}{2}}}+\varepsilon+\|e_u^0\|_{-1}^2\right\},$$

Remark 3.2.15 *Even though the theoretical results are not so good for Scheme RTL, its importance appears in practical computations. The results obtained are comparable to the ones provided by the implicit scheme RTI, but require less computational effort.*

Remark 3.2.16 *As mentioned before, Scheme RTL was proposed by M. Slodička ([88]) in a more restrictive framework. By our analysis we have obtained the same order of convergence up to the presence of the initial error - $\|e_u^0\|_{-1}$, which is due to the convective term.*

3.3 Full discretization

For solving Problem P numerically, the semi-discrete approximations provided by the schemes RTI, RTC and RTL considered here should be completed by the spatial discretization. As already being discussed in the previous chapter, there are several possibilities to do this step. Even though the maximum principle is not essential here (from theoretical point of view), in practical computations a stable discretization is useful. Therefore we maintain the frame defined in the chapter before, resulting an upwind box method. Moreover, this helps us to keep the proofs here similar to the ones for the maximum principle based algorithms. Again, convergence of the fully discrete schemes is obtained only in the absence of the convective terms. Consequently, this part is not of interest for Scheme RTC.

In the sequel, $\Omega \subset \mathbb{R}^d$ is decomposed into closed d-simplices (this set is denoted by S_h, where h stands for the mesh-size). S_h contains N_h vertices and is assumed regular and weakly acute. This guarantees a discrete maximum principle for the Laplace operator ([21]), since its finite element discretization matrix is irreducible and diagonal dominant. Again, Ω is assumed polygonal, hence the decomposition is exact. V_h includes the piecewise linear finite element space defined for S_h, while $V_h(g) = \{\varphi \in V_h : \varphi|_{\partial\Omega} = g|_{\partial\Omega}\}$ contains those elements of V_h satisfying a Dirichlet boundary condition. $\{\varphi_i, i = \overline{1, N_h}\}$ is the usual finite element basis for V_h. We recall here the definition of the discrete semi-inner product in (2.1) together with its properties mentioned in (2.2) and (2.3). As before, I_h denotes the local linear interpolation operator, while Π_h stands for the L^2 projection operator defined in (2.4).

For obtaining the error estimates for the fully discrete case in the manner proceeded up to now, regularity of the Green operator G - defined in (3.21) - is assumed ([19], p. 138). Correspondingly, the discrete Green operator $G_h : H^{-1} \to V_h(0)$ is defined by

$$(\nabla G_h \psi, \nabla \varphi) = (\psi, \varphi), \tag{3.1}$$

for all $\varphi \in V_h(0) \subset H_0^1(\Omega)$, where ψ is taken in $H^{-1}(\Omega)$. The error done when approximating G by G_h is given in (2.7). Other properties related to G_h have already been mentioned in (2.8), (2.9) and (2.10).

3.3.1 The fully discrete problems

After performing the time discretization, the algebraic systems corresponding to the fully discrete problems are obtained by applying the upwind box method described in the previous chapter. The dual mesh is generated by Donald diagrams. The nodal basis for B_h - the piecewise constant test function space - is denoted by $\{\phi_i, i = \overline{1, N_h}\}$, and its elements are 1 inside the box around the node i and 0 outside of it.

As mentioned before, for obtaining error estimates we consider Problem P without convection. In this case, because the dual box mesh is the Donald diagram and mass lumping is used, the box scheme for the semi-discrete Problem WRTI can be written in a finite element formulation ([40], [15]).

Problem WRDI. For any $1 \leq k \leq n$, find $\theta_h^k \in V_h(0)$ such that for all $\varphi \in V_h(0)$ the following holds true

$$(\beta_\varepsilon^{-1}(\theta_h^k) - \beta_\varepsilon^{-1}(\theta_h^{k-1}), \varphi)_h + \tau(\nabla\theta_h^k, \nabla\varphi) = \tau(I_h r(\beta_\varepsilon^{-1}(\theta_h^{\underline{k}})), \varphi)_h, \tag{3.2}$$

where $\underline{k}$ is either k or $k-1$, depending on the discretization of the reaction term.

Remark 3.3.1 *For each $k \geq 0$, if u_h^k is the piecewise linear interpolant of $\beta_\varepsilon^{-1}(\theta_h^k)$, we have*

$$\theta_h^k \equiv I_h(\beta_\varepsilon((u_h^k)) \quad \textit{and} \quad (u_h^k, \varphi)_h = (\beta_\varepsilon^{-1}(\theta_h^k), \varphi)_h$$

for all $\varphi \in V_h$.

The above definition has to be completed by the initial data. If θ_0 is the one involved in the time discretization (with the particular choices mentioned in the beginning of this chapter), the following construction can be considered

$$u_h^0 = \Pi_h(\beta_\varepsilon^{-1}(\theta^0)), \;\; \theta_h^0 = I_h\beta_\varepsilon(u_h^0). \tag{3.3}$$

Since u_0 is essentially bounded, the same holds for u_h^0 (and the bounds are the same).

Remark 3.3.2 *This particular choice for the discrete initial data involves a quadrature formula for computing the above integrals, inducing some additional errors which have the same order as the global ones. Therefore they are not taken into consideration here.*

Because of the assumptions on the triangularization, the fully discrete approximations satisfy a discrete maximum principle.

Lemma 3.3.1 *Assume (A1), (A3), $\theta_h^{k-1} \in V_h \cap V_{k-1}$ and $r(u) \geq 0$ for all u. If a solution of Problem WRDI exists, then it belongs to $V_h \cap V_k$ (where V_k has been defined in (3.7)).*

Proof. The proof is identical to the one for Theorem 2.3.1, where β has to be replaced by its approximation β_ε.

Remark 3.3.3 *As in the semi-discrete case, the global positivity of r is not necessary if r is discretized explicitly.*

Remark 3.3.4 *Lemma 2.3.1 also holds true if the convection term is present, but it has to be discretized in an upwinding manner.*

The algebraic nonlinear system arising in (3.2) can be solved by a fully discrete counterpart of Iteration IJK, defined in (3.11). Having k fixed, for any $i > 0$, the iteration can be formulated as follows

Problem WIJK. Find $\bar{\theta}_h^i \in V_h(0)$ such that

$$(\sigma(\bar{\theta}_h^{i-1}, \theta_h^{k-1})(\bar{\theta}_h^i - \theta_h^{k-1}), \varphi)_h + \tau(\nabla\bar{\theta}_h^i, \nabla\varphi) = \tau(I_h r(\beta_\varepsilon^{-1}(\theta_h^{k-1})), \varphi)_h \tag{3.4}$$

holds true for all $\varphi \in V_h(0)$, where $\bar{\theta}_h^0 = \theta_h^{k-1}$, $\sigma(\bar{\theta}_h^0, \theta_h^{k-1}) = (\beta_\varepsilon^{-1})'\,(\theta_h^{k-1})$ and θ_h^{k-1} belongs to $V_h(0)$.

Proceeding in the same manner as for the maximum principle based approach, assuming the discretization parameters satisfy the relation $\tau \leq C\varepsilon h^2$, we can show that the sequence of solutions of the above problems converges to the solution of Problem WRDI. The restriction on the time step τ is severe, but appears only theoretically. Moreover, since the proof relies essentially on the properties of the discretization matrix, the same holds true also if a convective term is present, but it should be discretized using an upwind procedure.

3.3.2 Error estimates for the complete discretization

In the remaining part of the chapter we seek for some error estimates for the fully discrete approximation of the solution of Problem P. Because β_ε has the same properties as β (stated in the assumption (A1)), the results in Lemma 2.3.3 are still valid here

Lemma 3.3.2 *Let $u_h \in V_h$ and $\theta_h = I_h\beta_\varepsilon(u_h)$. Then, if r is Lipschitz w.r.t. $\beta(u)$, we have*

$$\begin{aligned} \|\nabla\theta_h\|^2 \leq & C(\nabla u_h, \nabla\theta_h), \\ \|I_h\beta_\varepsilon(u_h) - \beta_\varepsilon(u_h)\| \leq & Ch\|\nabla I_h\beta_\varepsilon(u_h)\|, \\ \|I_h r(u_h) - r(u_h)\| \leq & Ch\|\nabla\theta_h\|, \end{aligned} \tag{3.5}$$

where $C > 0$ is a generic constant independent on h and u_h.

Error estimates for the fully discrete nonlinear scheme

We start with the implicit scheme WRDI, for which some stability results can be obtained.

Theorem 3.3.3 *Assume (A1), (A2), (A3) and $F \equiv 0$. Then, for $k \leq n$, if θ_h^k solves Problem WRDI and $u_h^k = I_h\beta_\varepsilon^{-1}(\theta_h^k)$, we have*

$$\sum_{k=1}^{p} \|u_h^k - u_h^{k-1}\|^2 + \tau\sum_{k=1}^{p} \|\nabla\theta_h^k\|^2 \leq C. \tag{3.6}$$

Proof. The result can be easily obtained taking $\varphi = u_h^k \in V_h(0)$ in (3.2) and using the first inequality in Lemma 3.3.2.

Now the errors due to the space discretization are estimated. To do so, we consider the following notations

$$e_u^{k,h} := \beta_\varepsilon^{-1}(\theta^k) - I_h\beta_\varepsilon^{-1}(\theta_h^k) \equiv \beta_\varepsilon^{-1}(\theta^k) - u_h^k, \quad e_\theta^{k,h} := \theta^k - \theta_h^k,$$

where $k \geq 0$. Because of the definition of θ_h^0 in (3.3), the initial error fulfills

$$\|e_u^{0,h}\|_{-1} \leq Ch. \tag{3.7}$$

The lemma below gives the estimates for the spatial discretization errors.

Theorem 3.3.4 *Under the assumptions (A1), (A2), (A3), if $F \equiv 0$, r is Lipschitz continuous w.r.t. $\beta(u)$ and θ^k, θ_h^k solve, for each $k > 0$, Problem WRTI respectively WRDI, then*

$$\sup_{k=\overline{1,n}} \|e_u^{k,h}\|_{-1}^2 + C\tau\textstyle\sum_{k=1}^{n} \|e_\theta^{k,h}\|^2 \leq C\left(h + \dfrac{h^2}{\tau}\right),$$

provided τ is reasonably small.

Proof. Again replacing β by β_ε, using Lemma 3.3.2, the proof is exactly the same with the one for Theorem 2.3.5. Moreover, an explicit treatment of the reaction term does not affect the result.

Yet the error estimates for Scheme WRDI are a direct consequence of the Theorems 3.2.6 and 3.3.4.

Theorem 3.3.5 *In the setting of Theorem 3.3.4, if u is the weak solution of Problem WP and θ_h^k solves for each $k > 0$ Problem WRDI, then*

$$\sup_{k=\overline{1,n}} \|u(t_k) - u^{k,h}\|_{-1}^2 + C\sum_{k=1}^n \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta^{k,h}\|^2 dt \leq C\left(\tau + \varepsilon + h + \frac{h^2}{\tau}\right).$$

Remark 3.3.5 *The same result has been obtained also for Scheme MDI, provided τ, ε and h are of the same order. The above estimates hold also for the schemes in [29] or [73], while in [86] they become optimal.*

Error estimates for the fully discrete linear scheme

We continue now by extending the above analysis to the fully discrete counterpart of the linear scheme RTL. As for the nonlinear one, if the convection term is not present, the spatial discretization can be brought into a finite element formulation

Problem WRDL. For any $1 \leq k \leq n$, find $\theta_h^k \in V_h(0)$ such that for all $\varphi \in V_h(0)$ the following holds true

$$\begin{aligned}(\sigma_{k-1,h}(\theta_h^k - \theta_h^{k-1}), \varphi)_h + \tau(\nabla\theta_h^k, \nabla\varphi) &= \tau(I_h r(\beta_\varepsilon^{-1}(\theta_h^{k-1})), \varphi)_h,\\ \sigma_{k,h} &= (\beta_\varepsilon^{-1})'(\theta_h^k).\end{aligned} \tag{3.8}$$

with $\sigma_{0,h} = (\beta_\varepsilon^{-1})'(\theta_h^0)$.

As done in the previous chapter, the initial data are given by

$$u_h^0 = I_h(\beta_\varepsilon^{-1}(\theta^0)), \quad \theta_h^0 = I_h\beta_\varepsilon(u_h^0).$$

Again, θ_h^0 is uniformly bounded (w.r.t. h) in the H^1 norm and the initial error becomes

$$\|e_u^{0,h}\|_{-1} \leq C\|e_u^{0,h}\| \leq C\frac{h}{\varepsilon}.$$

Now the estimates in Theorem 3.3.5 change to

$$\sup_{k=\overline{1,n}} \|u(t_k) - u^{k,h}\|_{-1}^2 + C \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta^{k,h}\|^2 dt$$
$$\leq C \left(\tau + \varepsilon + h + \frac{h^2}{\tau} + \frac{h^2}{\varepsilon^2} \right).$$

Similarly to Lemma 3.3.1, a maximum principle can be proven in this case too. Moreover, proceeding as for Theorem 3.2.5 we get

Theorem 3.3.6 *Assume (A1), (A2), (A3) and $F \equiv 0$. Then, for $p < n$, if θ_h^k solves Problem WRDL, the following holds true*

$$\sum_{k=1}^{p} \|\sqrt{\sigma_{k-1,h}}(\theta_h^k - \theta_h^{k-1})\|_h^2 + \tau\|\nabla\theta_h^p\|^2 + \tau \sum_{k=1}^{p} \|\nabla(\theta_h^k - \theta_h^{k-1})\|^2 \leq C\tau. \tag{3.9}$$

Now the error estimates for the fully discrete linear scheme become

Theorem 3.3.7 *In the setting of Theorem 3.3.6, if u is the weak solution of the Problem WP and θ_h^k solves for each $k > 0$ Problem WRDL, then*

$$\sup_{k=\overline{1,n}} \|u(t_k) - u^{k,h}\|_{-1}^2 + C \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta^{k,h}\|^2 dt$$
$$\leq C \left(\frac{\tau}{\varepsilon^2} + \frac{h}{\varepsilon^2} + \frac{h^2}{\tau} + \varepsilon + \|e_u^0\|_{-1}^2 \right).$$

Proof. The result follows directly from Theorem 3.2.9 if we can estimate again the errors due to the spatial discretization. The proof is identical to the one for Theorem 2.3.8, but $1/\varepsilon$ should replace $C_2(\varepsilon)$ in the final result.

Remark 3.3.6 *The error estimates here are similar to the ones for the semi-discrete linear scheme RTL. Taking above $h = C\tau$ and $\varepsilon = C\tau^{\frac{1}{3}}$, the order of convergence for the linear scheme WRDL becomes $\tau^{\frac{1}{6}}$, up to the error due to the approximation of the initial data. This improves the result obtained in [89], where the spatial discretization brings forth a loss in the power of τ, namely $\tau^{\frac{1}{8}}$. However, for particular nonlinearities modelling the Stefan problem, the linear scheme proposed in [67] behaves better - $O(\tau^{\frac{1}{3}})$.*

Chapter 4

Numerical examples

In this chapter we give some examples appropriate to the algorithms analysed up to now. The applications have been implemented in UG ([11], see also *http://www.ica3.uni-stuttgart.de*), a software toolbox providing tools for the generation and manipulation of unstructured meshes in two and three space dimensions and for the implementation of different algebraic solvers (including parallel adaptive multigrid methods) on the resulting grids. The fully discrete linear problems are solved by multigrid procedures provided by the above mentioned package. All calculations have been carried out on a $SGI\ O2$ computer with a MIPS R5000 processor. AVS ([100]) was used for the representation of the two-dimensional data, while in the three-dimensional case this is done by a volume rendering visualization program ([24]).

The first two examples are considered for testing purposes. The problems are simpler and do not include convection or reaction. Therefore the linear schemes have provided approximations of similar quality as the ones produced by the implicit methods. In this cases the resulting discretization matrices are symmetric and a Gauß-Seidel V-Cycle with two pre- and post-smoothing steps are enough for obtaining good approximations for the solutions of the discrete linear systems. Since exact solutions are available, the convergence order of the methods can be evaluated. To this aim we approximate the $L^2(Q_T)$ errors in terms of u and $\beta(u)$ by

$$E_u := \left[\sum_{i=1}^{N_h} \tau \left(|u^{k,h}(A_i) - u(k\tau, A_i)| \;\; meas(B_{A_i}) \right)^2 \right]^{\frac{1}{2}},$$

respectively

$$E_\theta := \left[\sum_{i=1}^{N_h} \tau \left(|\theta^{k,h}(A_i) - \beta(u)(k\tau, A_i)| \;\; meas(B_{A_i}) \right)^2 \right]^{\frac{1}{2}},$$

where $\{A_i,\ i = \overline{1, N_h}\}$ are the nodes of the triangularization and B_A stands for the dual box centered in A. Assuming the above errors are of order τ^α - more precisely $E_u = C_u \tau^{\alpha_u}$ and $E_\theta = C_\theta \tau^{\alpha_\theta}$ for some exponents α_u, α_θ and constants C_u, C_θ - a reasonable evaluation is offered by the computation of approximations corresponding to different sets of discretization parameters τ_j, h_j and ε_j. Now the convergence order can be estimated from the relations

$$\alpha_u = \frac{ln(E_{u_1}/E_{u_2})}{ln(\tau_1/\tau_2)} \quad \text{and} \quad \alpha_\theta = \frac{ln(E_{\theta_1}/E_{\theta_2})}{ln(\tau_1/\tau_2)}.$$

We have avoided a comparison with other discretization methods because this always depend on the particular choices of the parameters. However, tests in one spatial dimension show that for the porous medium equation, the results obtained with the methods considered here are similar to the ones produced by the Jäger-Kačur scheme (more details can be found in [98]).

The last example is a model for a one-phase flow in unsaturated porous media. The simulation is done both in two and three spatial dimensions. Because of the strongly oscillating diffusion coefficients, the simplest multigrid procedures are not efficient anymore. In this case an algebraic one-level procedure ([82]) has been applied with satisfactory results.

4.1 The porous medium equation

The first example considered here models the diffusion of a gas through a homogeneous porous medium. The process takes place in the time interval $[0, T]$ in a domain $\Omega \in \mathbb{R}^d$. The flow is governed by Darcy's law ([12]),

$$v = -\frac{\mu}{\nu} \nabla p,$$

where μ is the permeability of the porous medium, ν the viscosity of the gas and p its pressure. Denoting by ρ the density of the gas, the conservation of mass implies that

$$\kappa \partial_t \rho + \nabla \cdot (\rho v) = 0.$$

Here κ stands for the porosity of the medium (the fraction of the volume of the medium available for the flow). Because both the gas and the porous medium are homogeneous, ν, μ and κ are positive constants. For relating the density of the gas to the pressure, the following equation of state is assumed to hold true for some real constants $p_0 > 0$ and $\alpha \geq 1$

$$p = p_0 \rho^\alpha.$$

Eliminating p and v from the above equations and scaling ρ properly, the porous medium equation is obtained

$$\partial_t u = \Delta(u^m), \tag{4.1}$$

with $m = 1 + \alpha$. Together with suitable initial and boundary data, the above equation has been analysed in several papers (see, e.g., [5] for a detailed discussion). Whenever u vanishes, the above problem looses its parabolic character and as a consequence the free boundary - which separates the region occupied by the gas from the one where no gas is present - is propagating with finite speed.

Explicit solutions can be given in some particular cases. We have considered here a famous example given by G. I. Barenblatt ([10]), namely

$$u(t,x) = (t+1)^{-\frac{d}{(md+2-d)}} \left\{ \left[1 - \frac{m-1}{2m(md+2-d)} \left(\frac{x}{(t+1)^{\frac{1}{md+2-d}}} \right)^2 \right]_+ \right\}^{\frac{1}{m-1}}, \tag{4.2}$$

where d denotes the dimension of the space.

The domain Ω is taken sufficiently large in order to include the support of the above solution, namely $(-7,7)^2$. The boundary conditions are of Dirichlet type and homogeneous, while the initial data is exactly the value of u above at $t = 0$. All the results presented here are obtained at $t = 1.0s$. The semi-discrete nonlinear problems are solved by applying the iterative scheme IJK 2 to 6 times at any time step.

The largest time step is $\tau = 0.04$, the same as the perturbation parameter ε. We start with a uniform spatial grid which is refined three times first. A finer computation halves τ and ε and refines the mesh once more. The refinement level is denoted by i, in the first column of the tables. Table 4.1 presents the errors E_u and E_θ for the maximum principle based approach together with the estimated order of convergence for the fully discrete schemes - denoted by WMDI and WMDL in the second chapter. Here α_u and α_θ stand for the order of the errors E_u, respectively E_θ. The upper half corresponds to the

i	E_u	E_θ	α_u	α_θ
0	$5.04 \cdot 10^{-1}$	$3.06 \cdot 10^{-2}$	$*$	$*$
1	$2.46 \cdot 10^{-1}$	$9.31 \cdot 10^{-3}$	1.03	1.71
2	$1.23 \cdot 10^{-1}$	$3.93 \cdot 10^{-3}$	1.00	1.24
0	$5.03 \cdot 10^{-1}$	$3.04 \cdot 10^{-2}$	$*$	$*$
1	$2.45 \cdot 10^{-1}$	$8.40 \cdot 10^{-3}$	1.04	1.85
2	$1.22 \cdot 10^{-1}$	$2.86 \cdot 10^{-3}$	1.01	1.55

i	E_u	E_θ	α_u	α_θ
0	$7.47 \cdot 10^{-1}$	$7.89 \cdot 10^{-2}$	$*$	$*$
1	$5.38 \cdot 10^{-1}$	$4.33 \cdot 10^{-2}$	0.47	0.86
2	$3.10 \cdot 10^{-1}$	$1.47 \cdot 10^{-2}$	0.80	1.56
0	$7.52 \cdot 10^{-1}$	$7.24 \cdot 10^{-2}$	$*$	$*$
1	$5.46 \cdot 10^{-1}$	$4.06 \cdot 10^{-2}$	0.46	0.84
2	$3.27 \cdot 10^{-1}$	$1.57 \cdot 10^{-2}$	0.74	1.37

Table 4.1: L^2 errors for Schemes WMDI and WMDL, $m = 2$ and $m = 6$.

i	E_u	E_θ	α_u	α_θ
0	$1.00 \cdot 10^{-1}$	$2.92 \cdot 10^{-2}$	$*$	$*$
1	$4.23 \cdot 10^{-2}$	$1.15 \cdot 10^{-2}$	1.24	1.35
2	$1.57 \cdot 10^{-2}$	$5.39 \cdot 10^{-3}$	1.43	1.09
0	$1.05 \cdot 10^{-1}$	$2.96 \cdot 10^{-2}$	$*$	$*$
1	$5.16 \cdot 10^{-2}$	$1.27 \cdot 10^{-2}$	1.02	1.22
2	$2.24 \cdot 10^{-2}$	$5.78 \cdot 10^{-3}$	1.22	1.14

i	E_u	E_θ	α_u	α_θ
0	$6.17 \cdot 10^{-1}$	$1.72 \cdot 10^{-1}$	$*$	$*$
1	$4.87 \cdot 10^{-1}$	$8.36 \cdot 10^{-2}$	0.34	1.04
2	$2.76 \cdot 10^{-1}$	$3.39 \cdot 10^{-2}$	0.82	1.30
0	$6.19 \cdot 10^{-1}$	$1.62 \cdot 10^{-1}$	$*$	$*$
1	$4.86 \cdot 10^{-1}$	$7.84 \cdot 10^{-2}$	0.35	1.05
2	$2.80 \cdot 10^{-1}$	$3.14 \cdot 10^{-2}$	0.80	1.32

Table 4.2: L^2 errors for Schemes WRDI and WRDL, $m = 2$ and $m = 6$.

nonlinear scheme, while the lower one to the linear method. Similarly, Table 4.2 displays the results for the schemes WRDI and WRDL defined in the previous chapter, with the same discretization parameters. As we can see in the tables, the convergence rates for the more regular variable ($\beta(u)$) lie above the estimated one, 1. A better behaviour can be noticed if the nonlinearity is milder, namely $m = 2$. Then u itself is H^1 w.r.t. the spatial variable (see, e.g., [48]), implying better convergence orders also for E_u. The situation is different for $m = 6$, when only E_θ is of $O(\tau)$ order.

The exact solutions u and $\beta(u)$ are presented in Figure 4.1 for $m = 2$ and Figure 4.2 for $m = 6$. The pictures are zoomed 15 times in the z direction. We give here only the results on the finest mesh. For $m = 2$, the differences between the exact solution and its approximation provided by any of the schemes are not significant, at least graphically. It is worth here to notice the steep gradients of u appearing in the second case (for $m = 6$). Figure 4.3 contains the approximation of u in this last case, obtained with the implicit schemes WMDI and WRDI (the linear ones providing similar results). Due to the regularization step, Scheme WRDI smoothes the gradients of the numerical solution.

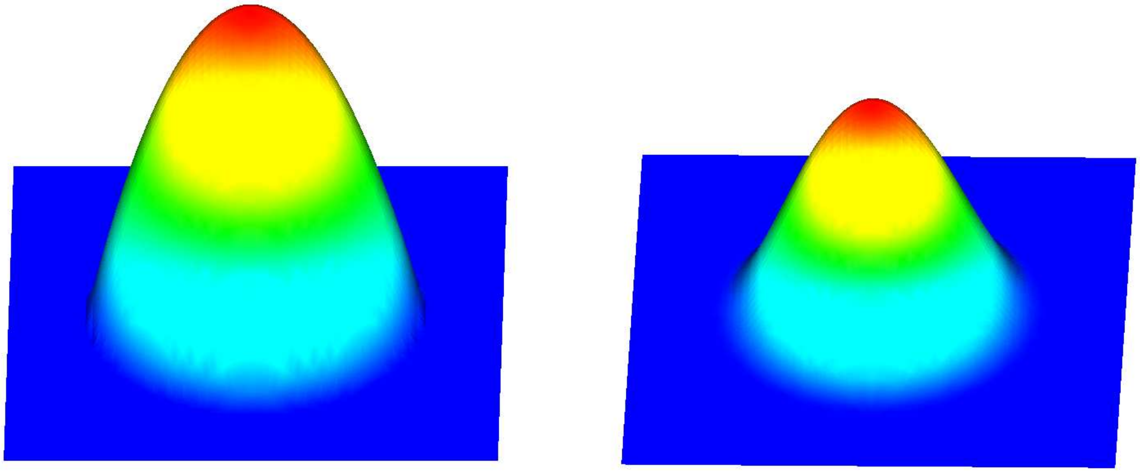

Figure 4.1: Exact u (left) and $\beta(u)$ (right) after $1.0s$, $m = 2$.

Figure 4.4 displays the absolute error (zoomed 30 times in the z-direction) for the fully discrete implicit scheme WMDI after $1.0s$, where the nonlinearity is $\beta(u) = u^2$. They range until $2 \cdot 10^{-2}$, respectively $1 \cdot 10^{-3}$. The results for $m = 6$ are shown in Figure 4.5. The errors are less than 0.13, respectively $2 \cdot 10^{-2}$. It becomes clear that in this case the errors are larger, because of the missing regularity of the solution u. It is worth here to notice that the errors appear mainly around the free boundary (namely inside a region covered by two, maximum three elements).

As it comes out from Figures 4.6 and 4.7, we have the same situation for the scheme WRDI (and also for WRDL). Now the maximal errors in u are $2.2 \cdot 10^{-2}$ and 0.35, while in $\beta(u)$ they become $1.5 \cdot 10^{-3}$, respectively $4.0 \cdot 10^{-3}$. Again, the errors are localized around the free boundary, but now, because the regularization parameter was chosen rather large, the spreading area is covered by 4 to 6 elements. A reason for this is the smoothing of the gradient of the solution u, as we have already noticed in Figure 4.3.

The maximal values of the errors are less relevant here, at least for u. Because this solution develops infinite gradients, errors having the same magnitude as u itself can appear around the free boundary and the only thing a finer approximation does is to make this region thinner.

Similar results are obtained for the linear schemes. It is difficult to distinguish them from the ones produced by the implicit methods (at least graphically), so we skip the corresponding pictures here.

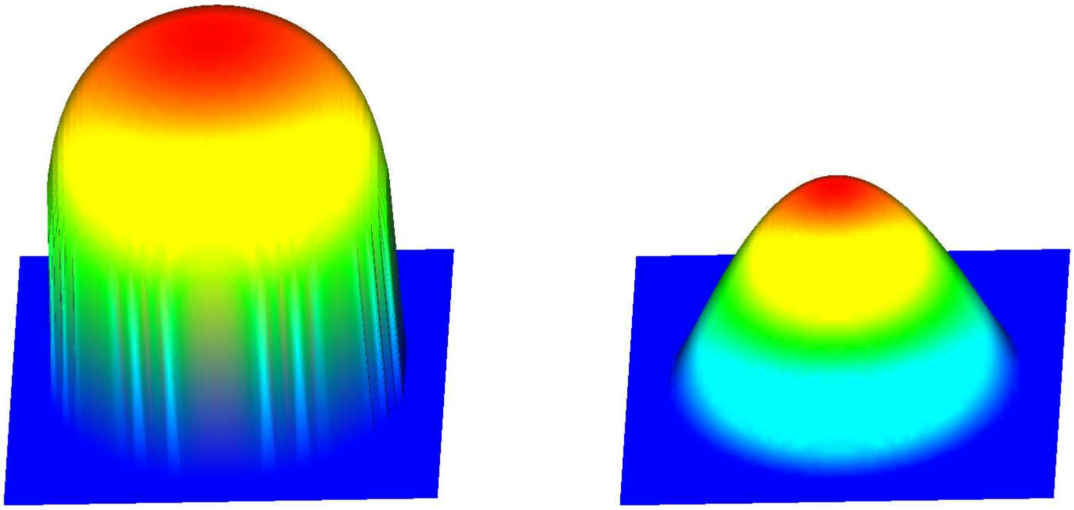

Figure 4.2: Exact u (left) and $\beta(u)$ (right) after $1.0s$, $m = 6$.

4.2 The Stefan problem

The second example is a simple model of a melting-solidification process ([36], [30]). A substance occupies a domain $\Omega \in \mathbb{R}^d$ during the time interval $[0, T]$. The phase change takes place at a fixed temperature, assumed 0. Ω_+ is the part of the domain where the temperature θ is positive (and consequently occupied by the liquid), while in Ω_- the solid phase - having a negative temperature - is present. The two sub-domains are separated by a free boundary Γ which is assumed smooth, with the normal n_Γ pointing in the outward direction of Ω_+. If the phase changes appear strictly due to the heat conduction (thus no heating or cooling sources are present and there is no transport of the substance), the conservation of energy implies the equation

$$\rho\partial_t H + \nabla \cdot v = 0 \quad \text{in } \Omega_- \cup \Omega_+,$$

where ρ is the density of the substance (assumed constant for simplicity), H denotes the heat content while v describes the heat flux. The Stefan condition relates the velocity V_Γ of the free boundary in the normal direction n_Γ with the jumps of H and v at the interface,

$$[v]_{sol}^{liq} n_\Gamma = \rho [H]_{sol}^{liq} V_\Gamma \quad \text{on } \Gamma.$$

Hence the phase change takes place at the temperature $\theta = 0$ with release or uptake of a latent heat $L = [H]_{sol}^{liq}$. Fourier's law describes the heat flux in terms of the temperature

$$v = -\kappa(\theta)\nabla\theta,$$

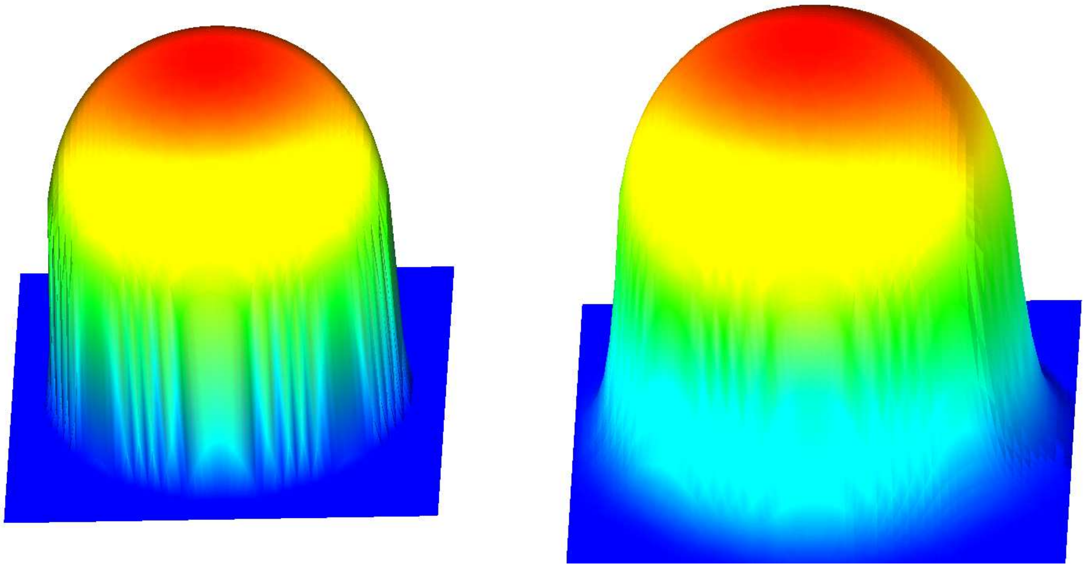

Figure 4.3: Approximation of u for $m = 6$, Scheme WMDI (left) and WRDI (right).

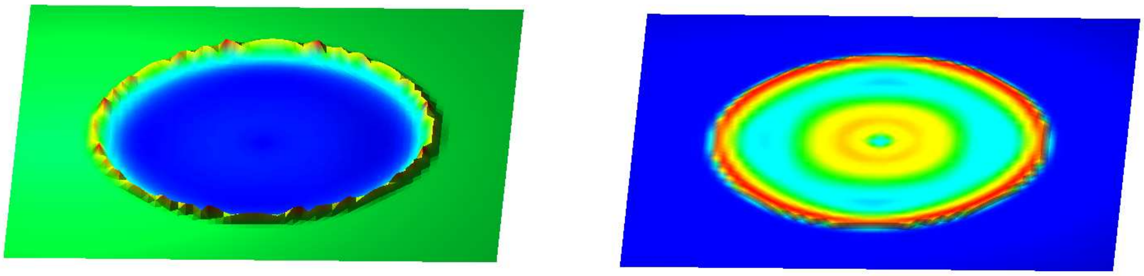

Figure 4.4: Errors in u (left) and θ (right) for Scheme WMDI, $m = 2$.

where $\kappa(\theta)$ denotes the thermal conductivity. Introducing the specific heat $c(\theta)$, the heat content can be expressed in terms of the temperature

$$H(\theta) = \int_0^\theta c(\theta) d\theta + \begin{cases} 0 & \text{if } \theta < 0, \\ [0, L] & \text{if } \theta = 0, \\ L & \text{if } \theta > 0, \end{cases}$$

thus the enthalpy function H is multi-valued.

For simplicity, the specific heat and the thermal conductivity are assumed constant over the two phases, namely c_-, c_+, κ_- and κ_+. Taking into account the above equations and applying a normalization to the enthalpy and the temperature, the two-phase Stefan

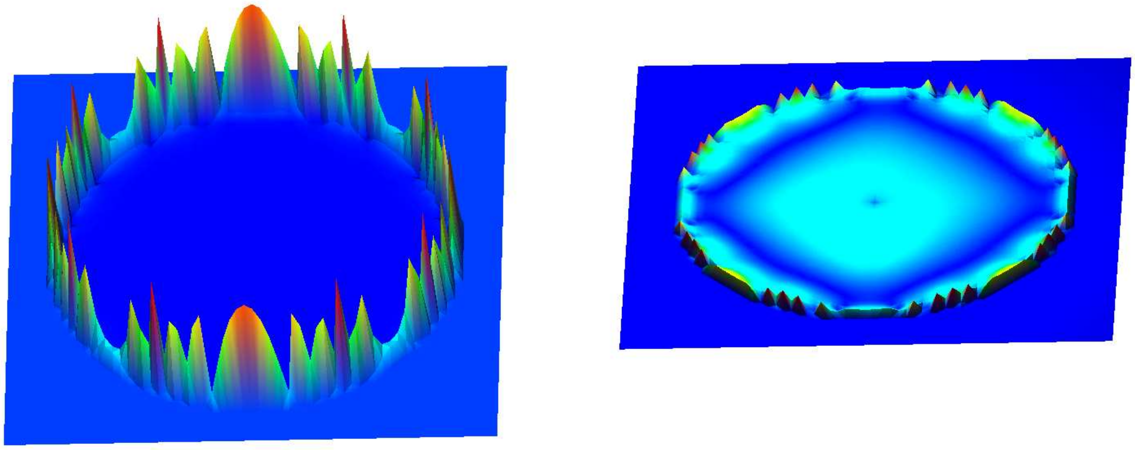

Figure 4.5: Errors in u (left) and θ (right) for Scheme WMDI, $m = 6$.

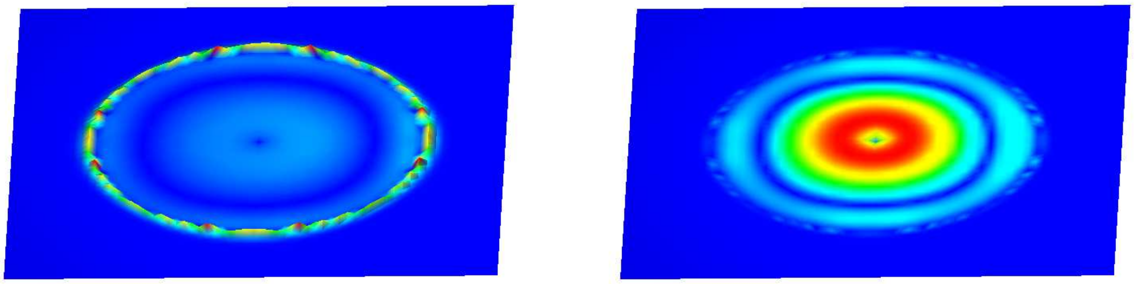

Figure 4.6: Errors in u (left) and θ (right) for Scheme WRDI, $m = 2$.

problem reads

$$\partial_t u = \Delta\beta(u), \tag{4.3}$$

with

$$\beta(u) = \begin{cases} \dfrac{\kappa_-}{\rho c_-}\, u, & \text{if } u < 0, \\ 0 & \text{if } u, \in [0, \rho L], \\ \dfrac{\kappa_+}{\rho c_+}\,(u - \rho L), & \text{if } u > \rho L. \end{cases}$$

This equation is completed by suitable initial and boundary conditions.

Here we have taken an example from [14], where $\Omega := (0, 0.5) \times (0, 0.25)$ and $T := 0.4$.

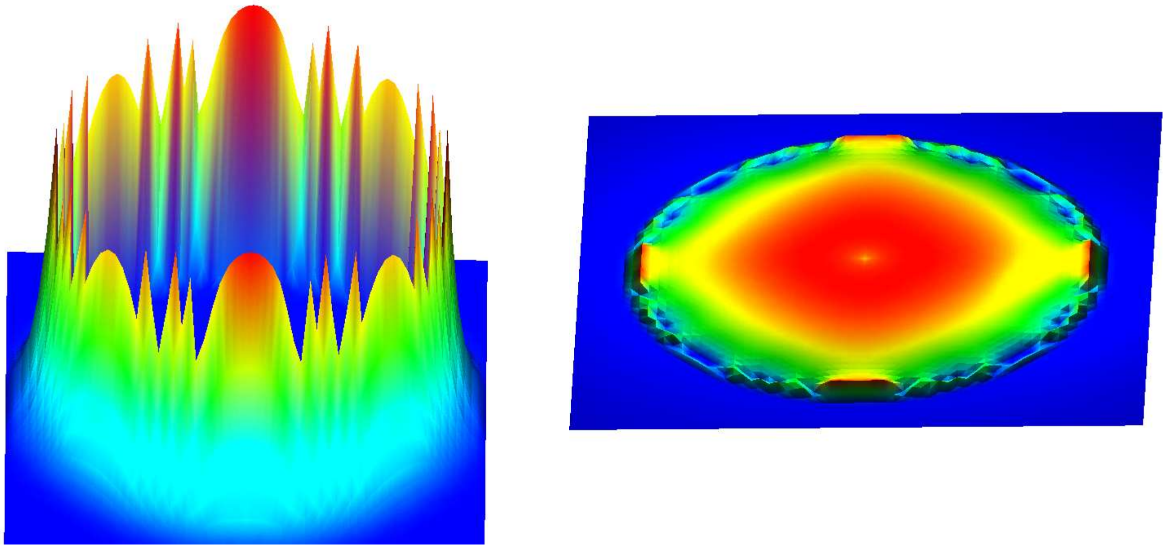

Figure 4.7: Errors in u (left) and θ (right) for Scheme WRDI, $m = 6$

i	E_u	E_θ	α_u	α_θ
0	$2.50 \cdot 10^{-2}$	$3.50 \cdot 10^{-2}$	$*$	$*$
1	$1.94 \cdot 10^{-2}$	$1.73 \cdot 10^{-2}$	0.36	1.02
2	$1.46 \cdot 10^{-2}$	$8.60 \cdot 10^{-3}$	0.41	1.01

i	E_u	E_θ	α_u	α_θ
0	$2.34 \cdot 10^{-2}$	$3.67 \cdot 10^{-3}$	$*$	$*$
1	$2.02 \cdot 10^{-2}$	$1.86 \cdot 10^{-3}$	0.21	0.98
2	$1.62 \cdot 10^{-2}$	$9.67 \cdot 10^{-4}$	0.32	0.94

Table 4.3: L^2 errors for Schemes WRDI (left) and WRDL (right).

The constants above are taken such that $\dfrac{\kappa_-}{\rho c_-} = \dfrac{\kappa_+}{\rho c_+} = \rho L = 1$. The exact enthalpy is

$$u(t,x,y) = \begin{cases} e^{\Phi(t,x,y)} - 1, & \text{if } \Phi(t,x,y) < 0, \\ 2[e^{\Phi(t,x,y)} - 1] + 1, & \text{if } \Phi(t,x,y) \geq 0, \end{cases}$$

where $\Phi(t,x,y) := 2t - x - y + 0.1$. Obviously, the free boundary is given by the points where Φ becomes 0. We have considered again Dirichlet boundary conditions, and the values are those provided by the exact solution.

For this problem, only the algorithms based on the second regularization approach can be applied (Schemes WRDI and WRDL). The nonlinearity is stronger here than the one in the previous case. At each time step, the first iteration step reduces the residuum significantly, while afterwards the convergence rate goes above 0.1. Moreover, this rate is affected by the choice of the discretization parameters. In order to make the iterative method effective, ε should be chosen in relation to the values of τ.

Table 4.3 contains the L^2 errors for the enthalpy (E_u) and the temperature (E_θ) for the parabolic cylinder mentioned above, together with the estimated order of convergence. As done in the previous section, we have started with a particular choice of parameters and a number of uniform refinements of the coarse grid (namely 3). Afterwards τ and ε are halved, while the grid is refined once more. At the beginning we have $\tau = 0.02$. As mentioned before, in order to make the iterations for the nonlinear scheme WRDI effective, we have chosen ε significantly larger (0.2), but the proportion to τ is maintained also on the finest level. However, some time steps have required more iterations (10 was the maximal number admitted here). For the linear scheme we have taken $\varepsilon = \tau$.

Since the enthalpy has a jump at the interface, we can expect that the errors are large there (attaining even the order of magnitude of the jump). This affects also the convergence order of both schemes w.r.t u, which is significantly lower than the one for the temperature.

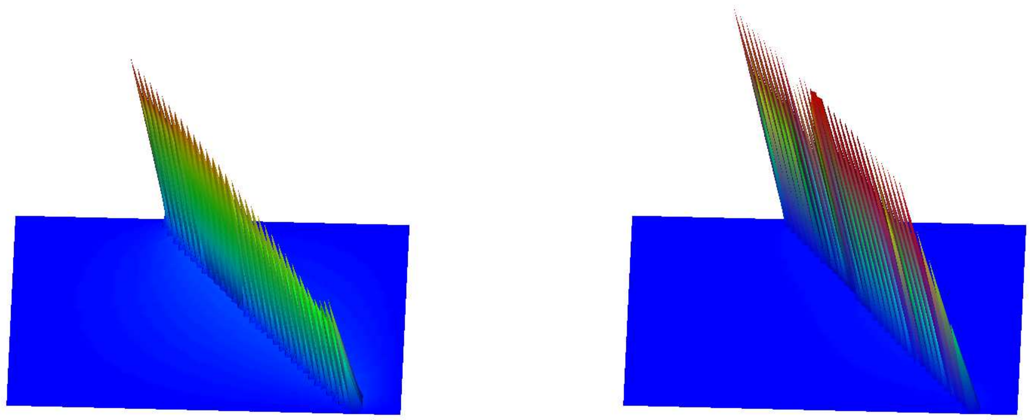

Figure 4.8: Enthalpy error for Scheme WRDI (left) and WRDL (right).

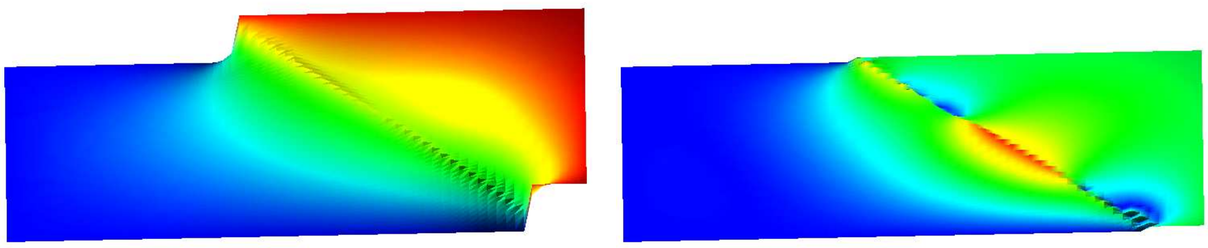

Figure 4.9: Temperature error for Scheme WRDI (left) and WRDL (right).

Figures 4.8 and 4.9 display the absolute errors for both schemes after 0.1s on the finest mesh. For a better visualization the domain is rotated around the z-axis with 180°. As mentioned before, due to the jumps of the enthalpy the corresponding errors are large (up to 1) at the interface, but restricted mainly to a region covered by two or three elements. The errors for the temperature appear mainly in the domain occupied by the liquid (where the temperature is positive), and this happens because of the regularization procedure. The range of values here are 0.05 (in fact ε) and 0.01.

4.3 The Richards equation

The last application we want to present here appears as a model of a one-phase flow in unsaturated porous media. Some basic ideas in modelling such kind of phenomena are given here briefly, for a a comprehensive presentation we mention books like [12], [13]. The same model has been considered in [95] or [94] (see also [78]). A unsaturated porous media wetted by a liquid (water) of density ρ is considered. The whole domain $\Omega \in \mathbb{R}^d$ is partitioned in representative elementary volumes (REV), each of them containing a volume of the wetting phase V_w. This fact is expressed through the saturation (fluid content) θ defined as

$$\theta = \frac{V_w}{V_{REV}},$$

where V_{REV} is the volume of the REV. The flow is governed by Darcy's law,

$$q = -K(\theta)\nabla\Phi,$$

where q stands for the specific flow rate and K for the hydraulic conductivity. The piezometric head Φ is given by the sum of the capillary pressure head Ψ and the vertical coordinate z,

$$\Phi = -\Psi + z.$$

The continuity condition

$$\partial_t(\rho\theta) + \nabla \cdot (\rho q) = 0$$

combined with Darcy's law leads to Richard's equation

$$\partial_t(\rho\theta) - \nabla \cdot (\rho K(\theta)\nabla\Phi) = 0.$$

Assuming the wetting phase has a constant density, this equation contains two unknowns, the saturation and the piezometric head. Moreover, a relation between the hydraulic conductivity and the saturation is necessary in order to solve the above equation.

To do so, the dimensionless fluid content (reduced saturation) is introduced,

$$S = \frac{\theta - \theta_r}{\theta_s - \theta_r},$$

where θ_r is the residual fluid content (the volume of the wetting phase always remaining in the porous media) and θ_s the saturated fluid content. A class of retention curves for describing the capillary pressure - saturation relation was suggested in [38]

$$S = \left(\frac{1}{1 + (\alpha \Psi)^n} \right)^m,$$

where $\alpha > 0$, $m \in (0, 1)$ and $n > 1$ are parameters depending on the media. As done in [38] (see also [94] and [95]), we assume here the relation $n = \frac{1}{m-1}$ holds true. One possibility for predicting the hydraulic conductivity knowing the saturation and the pressure head is given in [62]

$$K = K_s S^{\frac{1}{2}} \left[\int_0^S \frac{1}{\Psi(u)} du \Big/ \int_0^1 \frac{1}{\Psi(u)} du \right]^2,$$

where K_s denotes the hydraulic conductivity of the saturated porous medium. In this setting, the dependence of the hydraulic conductivity on the reduced saturation can be given explicitly

$$K(S) = K_s S^{\frac{1}{2}} \left[1 - (1 - S^{\frac{1}{m}})^m \right]^2.$$

Now the Richard's equation can be rewritten in terms of the reduced saturation, namely

$$\partial_t S - \nabla \cdot (D(S) \nabla S + \frac{K(S)}{\theta_s - \theta_r} \nabla z) = 0,$$

with the moisture diffusivity $D(S)$ defined by

$$D(S) = -K(S) \frac{\partial \Psi}{\partial \theta} = \frac{(1-m) K_s}{\alpha m (\theta_s - \theta_r)} S^{\frac{1}{2} - \frac{1}{m}} \left[\left(1 - S^{\frac{1}{m}}\right)^{-m} + \left(1 - S^{\frac{1}{m}}\right)^m - 2 \right].$$

In the simulations below we have considered a heterogeneous porous medium. This character is given by a variable hydraulic conductivity at saturation (K_s), which is generated randomly with a log-normal distribution ([83]; the implementation is based essentially on the procedures from [87]). The mean value of the decadic logarithm is $\overline{logK_s} = 0$, for the standard correlation we have $\sigma_{logK_s} = 0.4$, while the correlation length is $\Lambda = 0.02$. In the model we have taken $m = \frac{2}{3}$. Figure 4.10 displays the decadic logarithm of K_s both

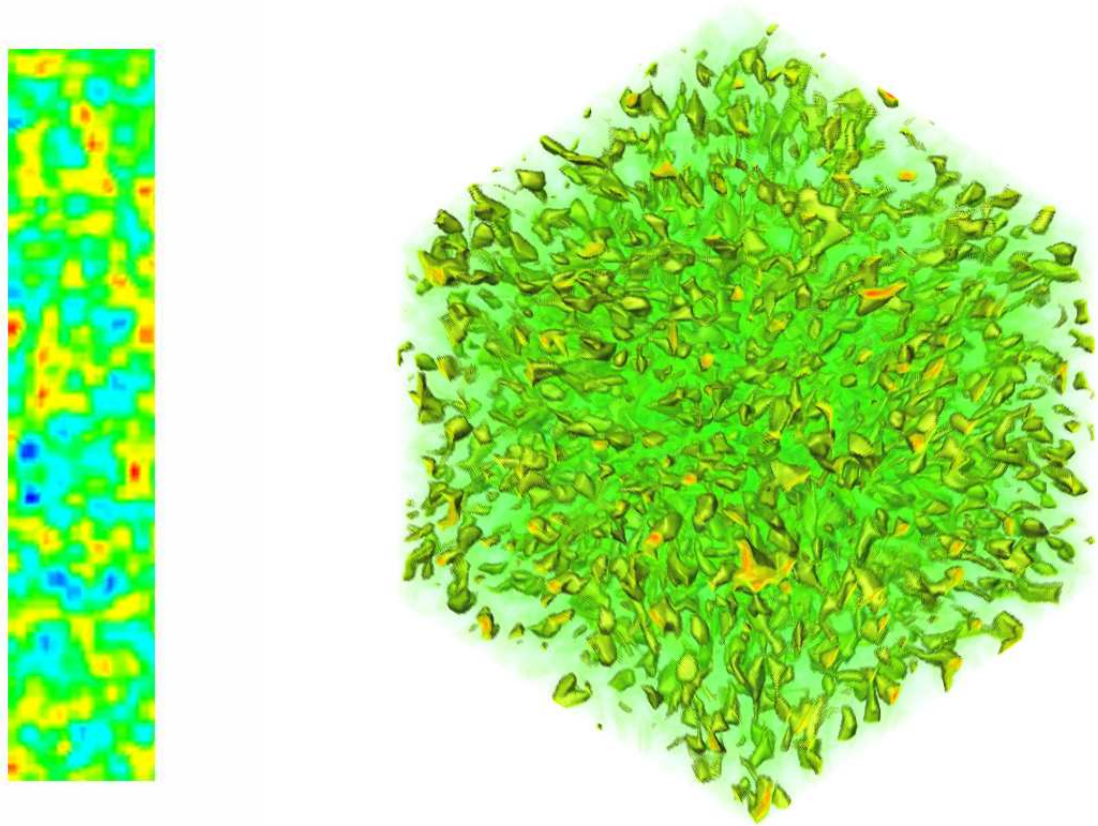

Figure 4.10: Decadic logarithm of K_s, 2 (left) and 3 (right) dimensions.

in two and three dimensions. These values start near -2 (in the blue colored regions) and go up to 2 (in the red part).

The above equation is completed with zero initial data (thus the medium is completely unsaturated). The boundary data are homogeneous and of Neumann type everywhere excepting the bottom part of the domain. There we assume the medium is in contact with the wetting phase, so the reduced saturation is maximal - 1 - in any moment. The two dimensional domain is a test stripe of length 5 and width 1, while in the three dimensional case we have considered the unit cube.

For the implementation of the numerical schemes we have approximated the primitive function of the moisture diffusivity divided by the hydraulic conductivity at saturation

$$\mathcal{D}(S) = \int_0^S \frac{D(u)}{K_s} du.$$

In fact we solve the problems with respect to the unknown $\Psi = \mathcal{D}(S)$ and recover the reduced saturation afterwards by inverting the function $\mathcal{D}$. Because of the form of D, its primitive (and the corresponding inverse) are approximated by a quadrature method.

An elementary calculus shows that $D(S)$ is positive (in fact it is also strictly increasing). We have now two degeneracy points, namely at 0, where D vanishes, and in 1, where the moisture diffusivity goes to infinity. Thus, for applying the maximum principle based algorithm, a shift of the initial and boundary data should be performed both from below and above. Even though a rigorous justification of the maximum principle is difficult because of the presence of the spatially dependent coefficient K_s in the convective term, we were able to apply Scheme WMDI without any problem.

In order to work also with Scheme WRDI we have approximated $\mathcal{D}(S)$ by $\mathcal{D}_\varepsilon(S)$, where

$$\mathcal{D}_\varepsilon(S) = \int_0^S \max\left\{\varepsilon, \min\left\{\frac{1}{\varepsilon}, \frac{D(u)}{K_s}\right\}\right\} du.$$

The two dimensional results (namely the reduced saturation) provided by both schemes are displayed in Figures 4.11 - 4.13 (where the pictures have been rotated with 90°), while the figures 4.14 - 4.16 contain the three dimensional results. The approximations provided by the two schemes are similar, but the effect of the strongly oscillating coefficients are milder in the case of the maximum principle based approach.

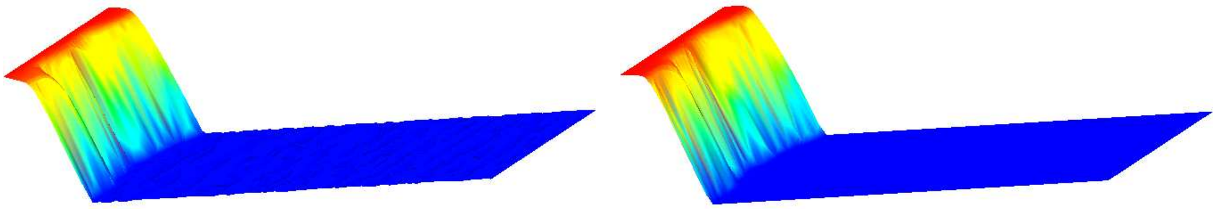

Figure 4.11: Reduced saturation, 20 time steps, Scheme WMDI (left) and WRDI (right).

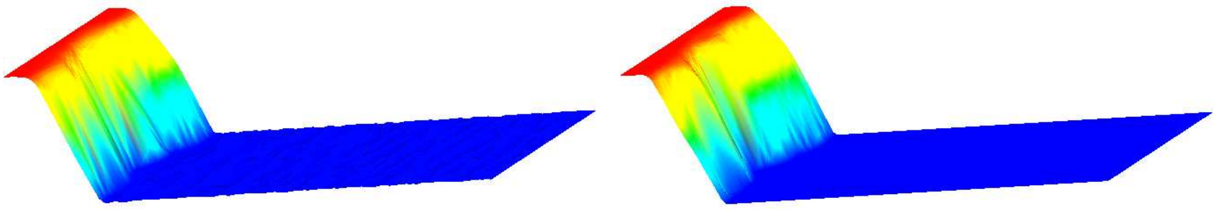

Figure 4.12: Reduced saturation, 60 time steps, Scheme WMDI (left) and WRDI (right).

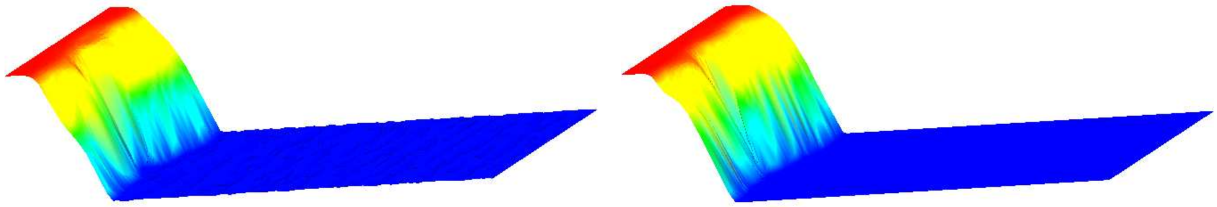

Figure 4.13: Reduced saturation, 100 time steps, Scheme WMDI (left) and WRDI (right).

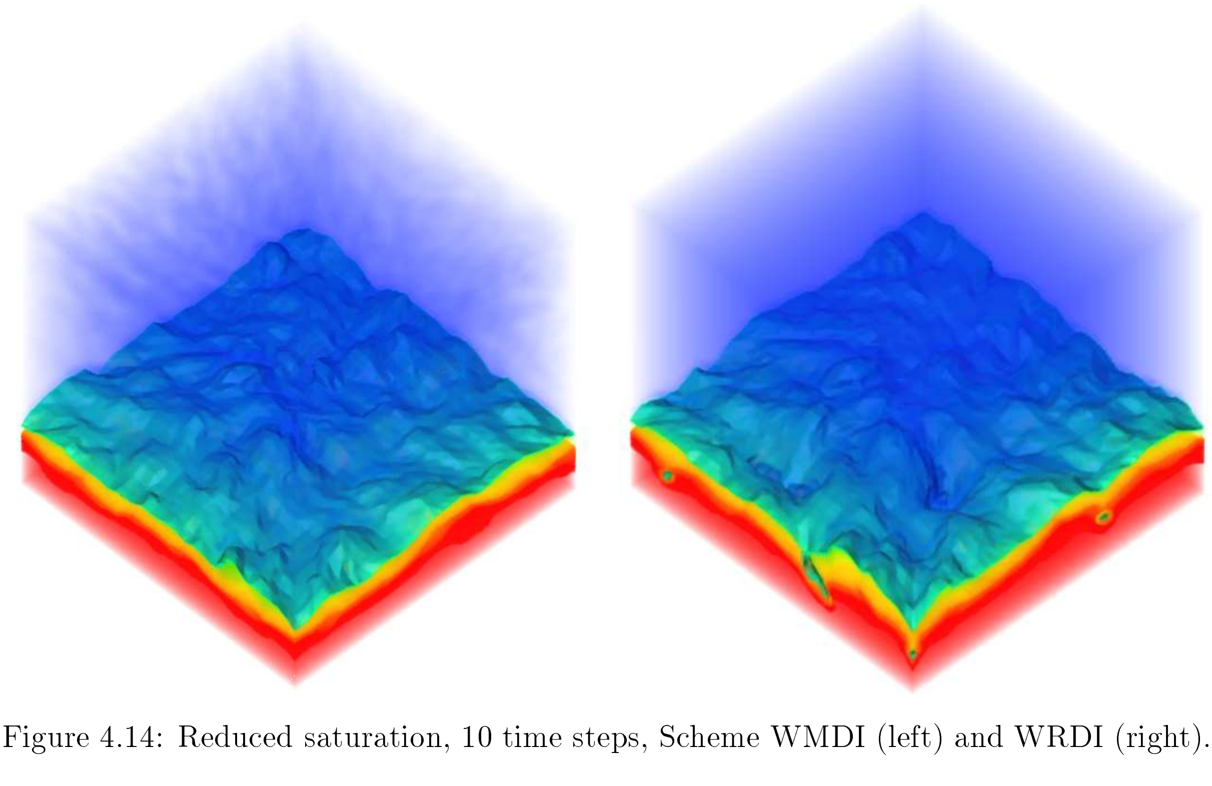

Figure 4.14: Reduced saturation, 10 time steps, Scheme WMDI (left) and WRDI (right).

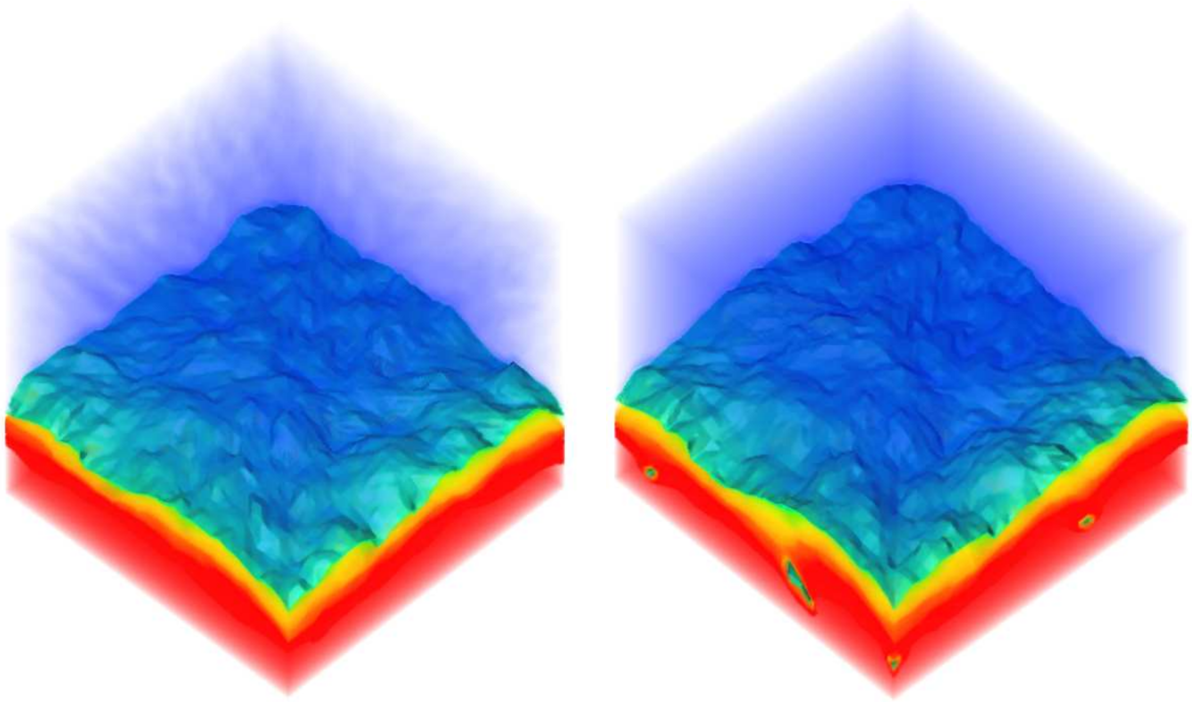

Figure 4.15: Reduced saturation, 20 time steps, Scheme WMDI (left) and WRDI (right).

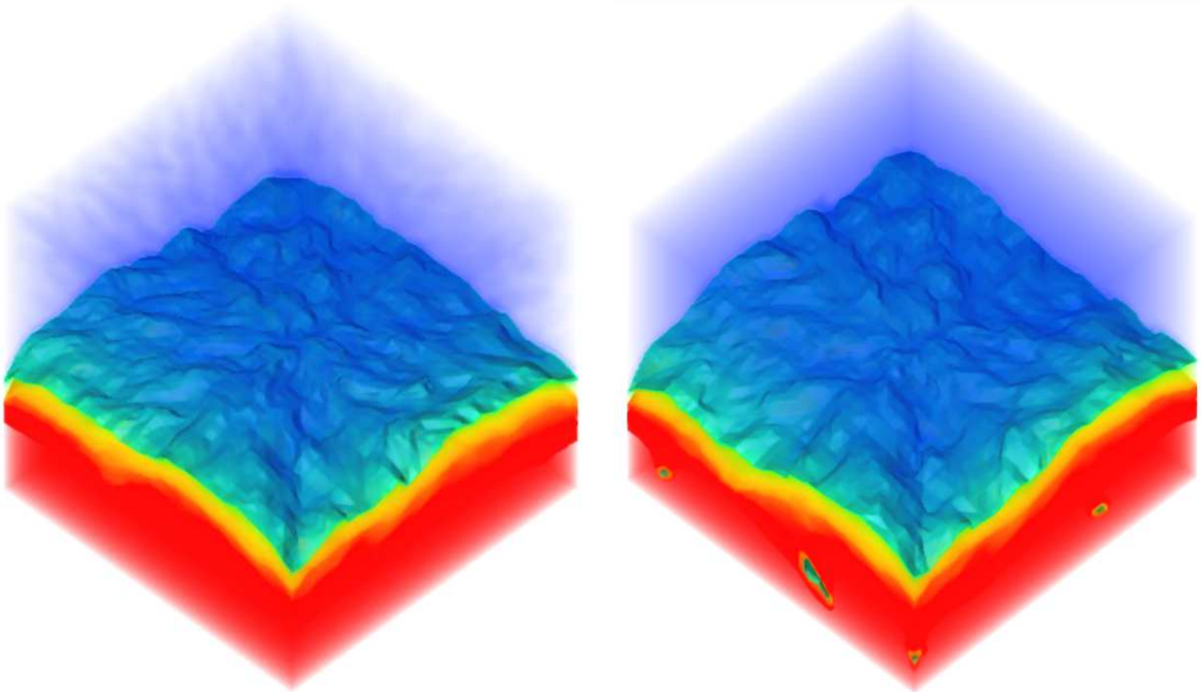

Figure 4.16: Reduced saturation, 30 time steps, Scheme WMDI (left) and WRDI (right).

Chapter 5

Conclusions and perspectives

In this work we have considered two regularization methods for some degenerate parabolic equations. We are interested here in the possibility to apply them for numerical purposes. While the perturbation of the nonlinearity was widely exploited for this purpose, we are not aware of the existence of some numerical schemes relying on the maximum principle based regularization. In both cases, the resulting algorithms are quite simple. We have proven the convergence of the approximation schemes by obtaining some error estimates. Due to them, the theoretical behaviour of the nonlinear schemes is at least as good as the one of other schemes appearing in the literature. We have done also an analysis of some linear algorithms, where the price for simplicity is a lower convergence order - at least theoretically. The analysis is completed by some examples of applications on which the methods are tested. On simpler problems - like those on which we have estimated the convergence order of the method - the linear schemes provide good approximations, so the iterations have not improved the results significantly. We do not expect the same situation on more complex problems, including reaction and - especially - transport.

The degenerate problems may have regions where the convection part dominates the flow. For stability purposes, an upwind method in connection with a box discretization method has been proposed. We were not able to show the convergence of this approach because of the missing regularity of the convective terms. Another possible candidate for a stable discretization is the method of characteristics (see, e.g., [96] or [34]), where a dominating transport influences the orientation of the time-stepping scheme. This is one of the directions we want to follow in the future.

The framework considered here is restricted to scalar equations where a maximum principle holds true (this property is not essential, but makes the analysis easier). However, there are several problems of practical interest which do not fit in this setting and therefore an extension to more general situations would be desirable. Among them we mention the systems of equations, arising, e.g., in chemistry or biology. At least one of the above methods (the one relying on the regularization of the nonlinear diffusion) could be applicable there if the system has a diagonal form (as shown in [46]).

Several free and moving boundary phenomena can be formulated as variational inequalities. We have not considered this type of problems here, but this may be another subject to be studied in the future. The applicability of the Jäger-Kačur time-discretization scheme to degenerate parabolic inequations is studied in [7]. For solving the resulting elliptic problems, monotone methods are essential (see [55], where an appropriate multigrid method is proposed). The situation becomes more complicated in the presence of some transport terms, which alter the symmetry of the problem.

Another important direction of improvement consists in the adaptive discretization. Our analysis considers only equal time steps and uniform refinement of the spatial mesh. Even though some difficulties in maintaining a discrete maximum principle can occur, it is worth considering this approach due to its efficiency.

Bibliography

[1] H. W. Alt & S. Luckhaus, Quasilinear Elliptic-Parabolic Differential Equations, *Math. Z.*, 183(1983), 311 - 341.

[2] G. Amiez & P. A. Gremaud, On a Numerical Approach to Stefan-like Problems, *Numer. Math.*, 59(1991), 71 - 89.

[3] L. Angermann, An Introduction to Finite Volume Methods for Linear Elliptic Equations of Second Order, *Script, University of Erlangen*, 1995.

[4] T. Arbogast, M. F. Wheeler & N.-Y. Zhang, A Nonlinear Mixed Finite Element Method for a Degenerate Parabolic Equation Arising in Flow in Porous Media, *SIAM J. Numer. Anal.*, 33(1996), 1669 - 1687.

[5] D. G. Aronson, The Porous Medium Equation, in Nonlinear Diffusion Problems, A. Fasano & M. Primicerio (eds.), *Lecture Notes in Mathematics 1224, Springer Verlag, Berlin*, 1985, 1 - 46.

[6] D. G. Aronson, L. A. Caffarelli & S. Kamin, How an Initially Stationary Interface Begins to Move in Porous Medium Flow, *SIAM J. Math. Anal.*, 14(1983), 639 - 658.

[7] J. Babušiková, Numerical Solution of Degenerate Variational Inequalities, *Ph.D. thesis, University of Bratislava*, 1997.

[8] E. Bänsch, Numerical Experiments with Adaptivity for the Porous Medium Equation, *Acta Math. Univ. Comen.*, 64(1995), 157 - 172.

[9] R. E. Bank, J. F. Bürgler, W. Fichtner & R. Kent Smith, Some Upwinding Techniques for Finite Element Approximations of Convection-Diffusion Equations, *Numer. Math.*, 58(1990), 185 - 202.

[10] G. I. Barenblatt, On Some Unsteady Motion of a Liquid or a Gas in a Porous Medium, *Prikl. Math. Meh.*, 16(1952), 67 - 78.

[11] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuss, H. Rentz-Reichert & C. Wieners, UG - A Flexible Software Toolbox for Solving Partial Differential Equations, *Computing and Visualization in Science*, 1(1997), 27 - 40.

[12] J. Bear, Dynamics of Fluids in Porous Media, *American Elsevier, New York*, 1972.

[13] J. Bear & Y. Bachmat, Introduction to Modelling of Transport Phenomena in Porous Media, *Kluwer Academic, Dordrecht*, 1991.

[14] A. E. Berger, H. Brezis & J. C. W. Rogers, A Numerical Method for Solving the Problem $u_t - \Delta f(u) = 0$, *R.A.I.R.O. Anal. Numer.*, 13(1979), 297 - 312.

[15] J. Bey, Finite- Volumen- und Mehrgitterverfahren für elliptische Randwertprobleme, *Ph.D. thesis, University of Tübingen*, 1997.

[16] H. Brezis, Monotonicity Methods in Hilbert Spaces and Some Applications to Nonlinear Partial Differential Equations, in Contributions to Nonlinear Functional Analysis, E. Zarantonello (ed.), *Academic Press, New York*, 1971, 101 - 156.

[17] H. Brezis & A. Pazy, Convergence and Approximation of Semigroups of Nonlinear Operators in Banach Spaces, *J. Funct. Anal.*, 9(1972), 63 - 74.

[18] J. R. Cannon & E. DiBenedetto, On the Existence of Weak-Solutions to an n-Dimensional Stefan Problem with Nonlinear Boundary Conditions, *SIAM J. Math. Anal.*, 11(1980), 632 - 645.

[19] P. G. Ciarlet, The Finite Element Method for Elliptic Problems, *North-Holland, Amsterdam*, 1978.

[20] P. G. Ciarlet, Basic Error Estimates for Elliptic Problems, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet & J. L. Lions (eds.), *North-Holland, Amsterdam*, 1991, 18 - 351.

[21] P. G. Ciarlet & P. A. Raviart, Maximum Principle and Uniform Convergence for the Finite Element Method, *Comput. Methods Appl. Mech. Engrg.*, 2(1973), 17 - 31.

[22] J. F. Ciavaldini, Analyse numérique d'un probléme de Stefan á deux phases par une méthode d'éléments finis, *SIAM J. Numer. Anal.*, 12(1975), 464 - 487.

[23] M. G. Crandall & T. M. Liggett, Generation of Semi-Groups of Nonlinear Transformation on General Banach Spaces, *Amer. J. Math.*, 93(1971), 265 - 298.

[24] C. Dârtu & W. Jäger, Volume Rendering and Applications in Scientific Visualization, *Preprint (SFB 359)* (1998), IWR, University of Heidelberg, in preparation.

[25] J. I. Diaz & J. de Thelin, On a Nonlinear Parabolic Problem Arising in Some Models Related to Turbulent Flows, *SIAM J. Math. Anal.*, 25(1994), 1085 - 1111.

[26] E. DiBenedetto & D. Hoff, An Interface Tracking Algorithm for the Porous Medium Equation, *Trans. Amer. Math. Soc.*, 284(1984), 463 - 500.

[27] C. Ebmeyer, A Non-Degeneracy Property for a Class of Degenerate Parabolic Equations, *Z. Anal. Anwend.* 15(1996), 637 - 650.

[28] C. Ebmeyer, Error Estimates for a Class of Degenerate Parabolic Equations, *SIAM J. Numer. Anal.*, 35(1998), 1095 - 1113.

[29] C. M. Elliott, Error Analysis of the Enthalpy Method for the Stefan Problem, *IMA J. Numer. Anal.*, 7(1987), 61 - 71.

[30] C. M. Elliott & J. R. Ockendon, Weak and Variational Methods for Moving Boundary Problems, *Pitman, London*, 1982.

[31] J. F. Epperson, An Error Estimate for Changing the Stefan Problem, *SIAM J. Numer. Anal.*, 19(1982), 114 - 120.

[32] J. F. Epperson, Finite Element Method for a Class of Nonlinear Evolution Equations, *SIAM J. Numer. Anal.*, 21(1984), 1066 - 1079.

[33] S. Evje & K. H. Karlsen, A Note on Viscous Splitting of Degenerate Convection-Diffusion Equations, to appear.

[34] R. E. Ewing & H. Wang, An Optimal-Order Estimate for Eulerian-Lagrangian Localized Adjoint Methods for Variable-Coefficient Advection-Reaction Problems, *SIAM J. Numer. Anal.*, 33(1996), 318 - 348.

[35] R. Eymard, T. Gallouët, D. Hilhorst & Y. Naït Slimane, Finite Volumes and Nonlinear Diffusion Equations, to appear.

[36] A. Friedman, Variational Principles and Free-Boundary Problems, *John Wiley & Sons, New York*, 1982.

[37] J. Fuhrmann, Numerical Solution Schemes for Nonlinear Diffusion Problems Based on Newton's Method, in ALGORITMY'97, Proceedings of the 14^{th} Conference on Scientific Computing, A. Handlovičová, M. Komorníková & K. Mikula (eds.), *Slovak Technical University, Bratislava*, 1997, 32 - 41.

[38] M. Th. van Genuchten, A Closed-Form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils, *Soil Sci. Soc. Am. J.*, 44(1980), 892 - 898.

[39] D. Gilbarg & N. S. Trudinger, Elliptic Partial Differential Equations of Second Order, Grundlehren der Mathematischen Wissenschaften, 224, *Springer Verlag, Berlin*, 2^{nd} ed., 1983.

[40] W. Hackbusch, On First and Second Order Box Schemes, *Computing*, 41(1989), 277 - 296.

[41] W. Hackbusch, Theorie und Numerik elliptischer Differentialgleichungen, *Teubner, Stuttgart*, 1986.

[42] A. Handlovičová, Error Estimates of a Linear Approximation Scheme for Nonlinear Diffusion Problems, *Acta Math. Univ. Comenianae*, 61(1992), 27 - 39.

[43] A. Handlovičová, Error Estimates of a Fully Discrete Linear Approximation Scheme for Stefan Problem, *Acta Math. Univ. Comenianae*, 65(1996), 65 - 85.

[44] T. Ikeda, Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena, Lecture Notes in Numerical and Applied Analysis Vol. 4, *North-Holland/Kinokuniya, Amsterdam/Tokyo*, 1983.

[45] A. V. Ivanov & W. Jäger, Existence and Uniqueness of a Regular Solution of Cauchy-Dirichlet Problem for Equation of Turbulent Filtration, *Preprint 96-14 (SFB 359)* (1996), IWR, University of Heidelberg.

[46] W. Jäger & J. Kačur, Solution of Porous Medium Type Systems by Linear Approximation Schemes, *Numer. Math.*, 60(1991), 407 - 427.

[47] W. Jäger & J. Kačur, Solution of Doubly Nonlinear and Degenerate Parabolic Problems by Relaxation Schemes, *M^2AN (Math. Model. Numer. Anal.)*, 29(1995), 605 - 627.

[48] W. Jäger & Y. G. Lu, Hölder Continuity of Solutions of Degenerate Parabolic Equations, to appear.

[49] J. W. Jerome, Approximation of Nonlinear Evolution Systems, *Academic Press, New York*, 1983.

[50] J. W. Jerome & M. E. Rose, Error Estimates for the Multidimensional Two-Phase Stefan Problems, *Math. Comp.*, 39(1982), 377 - 414.

[51] X. Jiang & R. H. Nochetto, Optimal Error Estimates for Semidiscrete Phase Relaxation Models, *M^2AN (Math. Model. Numer. Anal.)*, 31(1997), 91 - 120.

[52] J. Kačur, On a Solution of Degenerate Elliptic-Parabolic Systems in Orlicz-Sobolev Spaces I; II, *Math. Z.*, 203(1990), 153 - 171; 569 - 579.

[53] J. Kačur, Solution to Strongly Nonlinear Parabolic Problems by a Linear Approximation Scheme, *Preprint M1-96* (1996), Comenius University Bratislava, Faculty of Mathematics and Physics.

[54] J. Kačur, A. Handlovičová & M. Kačurová, Solution of Nonlinear Diffusion Problems by Linear Approximation Schemes, *SIAM J. Numer. Anal.*, 30(1993), 1703 - 1722.

[55] R. Kornhuber, Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems, *B. G. Teubner, Stuttgart*, 1997.

[56] A. Kufner, & S. Fučic, Nonlinear Differential Equations, *SNTL, Praha, Elsevier*, 1980.

[57] A. Kufner, O. John & S. Fučic, Function Spaces, *Academia CSAV, Prague*, 1967.

[58] O. A. Ladyzhenskaya & N. N. Ural'ceva, Linear and Quasilinear lliptic Equatins, *Academic Press, New York*, 1968.

[59] J. L. Lions & E. Magenes, Non Homogenous Boundary Value Problems and Applications, Vol. I, *Springer Verlag, Berlin*, 1972.

[60] E. Magenes, R. H. Nochetto & C. Verdi, Energy Error Estimates for a Linear Scheme to Approximate Nonlinear Parabolic Problems, *M^2AN (Math. Model. Numer. Anal.)*, 21(1987), 655 - 678.

[61] G. H. Meyer, Multidimensional Stefan Problems, *SIAM J. Numer. Anal.*, 10(1973), 522 - 538.

[62] Y. Mualem, A New Model for Predicting the Hydraulic Conductivity of Unsaturated Porous Media, *Water Resour. Res.*, 12(1976), 513 - 522.

[63] R. H. Nochetto, Error Estimates for Two-Phase Stefan Problems in Several Space Variables, I: Linear Boundary Conditions *Calcolo*, 22(1985), 501 - 534.

[64] R. H. Nochetto, A Note on the Approximation of Free Boundaries by Finite Element Method, *Math. Model. Numer. Anal.*, 20(1986), 355 - 368.

[65] R. H. Nochetto, A Class of Non-Degenerate Two-Phase Stefan Problems in Several Space Dimensions, *Comm. Partial Differential Equations*, 12(1987), 21 - 45.

[66] R. H. Nochetto, Error Estimates for Multidimensional Singular Parabolic Problems, *Japan J. Appl. Math.*, 4(1987), 111 - 138.

[67] R. H. Nochetto, A Stable Extrapolation Method for Multidimensional Degenerate Parabolic Problems, *Math. Comput.*, 53(1989), 73 - 108.

[68] R. H. Nochetto, M. Paolini & C. Verdi, An Adaptive Finite Element Method for Two-Phase Stefan Problems in Two Space Dimensions. Part I: Stability and Error Estimates, *Math. Comput.*, 57(1991), 73 - 108.

[69] R. H. Nochetto, M. Paolini & C. Verdi, An Adaptive Finite Element Method for Two-Phase Stefan Problems in Two Space Dimensions. Part II: Implementation and Numerical Experiments *SIAM J. Sci. Stat. Comput.*, 12(1991), 1207 - 1244.

[70] R. H. Nochetto, M. Paolini & C. Verdi, A Fully Discrete Adaptive Nonlinear Chernoff Formula, *SIAM J. Numer. Anal.*, 30(1993), 991 - 1914.

[71] R. H. Nochetto, A. Schimdt & C. Verdi, Adapting Meshes and Time-Steps for Phase Change Problems, submitted.

[72] R. H. Nochetto, A. Schimdt & C. Verdi, A Posteriori Error Estimation and Adaptivity for Degenerate Parabolic Problems, *Math. Comp.* to appear.

[73] R. H. Nochetto & C. Verdi, Approximation of Degenerate Parabolic Problems Using Numerical Integration, *SIAM J. Numer. Anal.*, 25(1988), 784 - 814.

[74] R. H. Nochetto & C. Verdi, An Efficient Linear Scheme to Approximate Parabolic Free Boundary Problems: Error Estimates and Implementation, *Math. Comp.*, 183(1988), 27 - 53.

[75] O. A. Oleinik, A. S. Kalashnikov & Y.-L. Zhou, The Cauchy Problem and Boundary Value Problems for Equations of the Type of Unsteady Filtration, *Izv. Akad. Nauk SSSR Ser. Mat.*, 22(1958), 667 - 704.

[76] F. Otto, L^1-Contraction and Uniqueness for Quasilinear Elliptic-Parabolic Equations, *J. Differ. Equations*, 131(1996), 20 - 38.

[77] I. S. Pop, Regularization Methods in the Numerical Analysis of Some Degenerate Parabolic Equations, *Preprint 98-43 (SFB 359)* (1998), IWR, University of Heidelberg.

[78] I. S. Pop, Numerical Simulation of Infiltration Phenomena in Unsaturated Porous Media, in preparation.

[79] I. S. Pop & W. A. Yong, A Maximum Principle Based Numerical Approach to Porous Medium Equations, in ALGORITMY'97, Proceedings of the 14^{th} Conference on Scientific Computing, A. Handlovičová, M. Komorníková & K. Mikula (eds.), *Slovak Technical University, Bratislava*, 1997, 207 - 218.

[80] P. A. Raviart, The Use of Numerical Integration in Finite Element Methods for Solving Parabolic Equations, in Topics in Numerical Analysis, J. J. H. Miller (ed.), *Academic Press, London*, 1973, 233 - 264.

[81] P. A. Raviart & J. M. Thomas, A Mixed Finite Element Method for 2^{nd} Order Elliptic Problems, in Mathematical Aspects of the Finite Element Method, J. J. H. Miller (ed.), *Lect. Notes Math. 606, Springer Verlag, New York*, 1977, 292 - 315.

[82] M. Raw, A Coupled Algebraic Multigrid Method for the 3D Navier-Stokes Equations, in Fast Solvers for Flow Problems, Proceedings of the tenth GAMM-Seminar Kiel,

Germany, January 14–16, 1994, W. Hackbusch et al. (eds.), *Notes Numer. Fluid Mech. 49, Vieweg, Wiesbaden*, 1995, 204-215.

[83] M. J. L. Robin, A. L. Gutjahr, E. A. Sudicky & J. L. Wilson, Cross-Corelated Random Field Generation with the Direct Fourier Transform Method, *Water Resour. Res.*, 29(1993), 2385 - 2397.

[84] M. E. Rose, Numerical Methods for Flows through Porous Media I, *Math. Comp.*, 40(1983), 435 - 467.

[85] J. Rulla, Error Analysis for Implicit Approximations to Solutions to Cauchy Problems, *SIAM J. Numer. Anal.*, 33(1996), 68 - 87.

[86] J. Rulla & N. J. Walkington, Optimal Rates of Convergence for Degenerate Parabolic Problems in Two Dimensions, *SIAM J. Numer. Anal.*, 33(1996), 56 - 67.

[87] C. Schwarz, Effective Parameter für adsorbtiven Transport im Grundwasser, *Diploma Thesis* (1995), University of Heidelberg.

[88] M. Slodička, Solution of Nonlinear Parabolic Problems by Linearization, *Preprint M3-92* (1992), Comenius University Bratislava, Faculty of Mathematics and Physics.

[89] M. Slodička, On a Numerical Approach to Nonlinear Degenerate Parabolic Problems, *Preprint M6-92* (1992), Comenius University Bratislava, Faculty of Mathematics and Physics.

[90] J. Smoller, Shock Waves and Reaction-Diffusion Equations, *Springer Verlag, New York*, 1983.

[91] G. Strang & G. Fix,, An Analysis of the Finite Element Method, *Prentice-Hall, Englewood Cliffs, New Jersey*, 1973.

[92] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems, Springer Series in Computational Mathematics. 25, *Springer Verlag, Berlin*, 1973.

[93] R. S. Varga, Matrix Iterative Analysis, *Prentice-Hall, Englewood Cliffs, New Jersey*, 1963.

[94] C. Wagner, Numerical Methods for Diffusion-Reaction-Transport Processes in Unsaturated Porous Media, *Computing and Visualisation in Science*, 1(1998), 97 - 105.

[95] C. Wagner, G. Wittum, R. Fritsche & H. P. Haar, Diffusions-Reaktionsprobleme in ungesättigten porösen Medien, in Mathematik-Schlüsseltechnologie für die Zukunft, K. H. Hoffmann, W. Jäger, T. Lohmann & H. Schunck (eds.), *Springer Verlag, Berlin*, 1997.

[96] H. Wang & R. E. Ewing, Optimal-Order Convergence Rates for Eulerian-Lagrangian Localized Adjoint Methods for Reactive Transport and Contamination in Groundwater, *Numer. Methods Partial Differ. Equations*, 11(1995), 1 - 31.

[97] M. Watanabe, An Approach by Difference to the Porous Medium Equation with Convection, *Hiroshima Math. J.*, 25(1995), 623 - 645.

[98] W. A. Yong & I. S. Pop, A Numerical Approach to Porous Medium Equations, *Preprint 96-50 (SFB 359)* (1996), IWR, University of Heidelberg, submitted.

[99] E. Zeidler, Applied Functional Analysis, Vols. I, II, Applied Mathematical Sciences 108, 109, *Springer Verlag, New York*, 1995.

[100] ***, AVS/Express User Guide, *AVS/Uniras*, 1997.