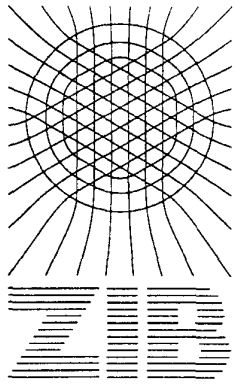
Konrad-Zuse-Zentrum für Informationstechnik Berlin



Folkmar A. Bornemann

An Adaptive Multilevel Approach to Parabolic Equations III.

2D Error Estimation and Multilevel Preconditioning

Preprint SC 91-1 (September 1991)

An Adaptive Multilevel Approach to Parabolic Equations

III. 2D Error Estimation and Multilevel Preconditioning *

FOLKMAR A. BORNEMANN

Konrad–Zuse–Zentrum für Informationstechnik Berlin, Heilbronner Strasse 10, D-1000 Berlin 31, Federal Republic of Germany

September 18, 1991

ABSTRACT

Part III of the paper is devoted to the construction of an adaptive FEM solver in two spatial dimensions, which is able to handle the singularly perturbed elliptic problems arising from discretization in time. The problems of error estimation and multilevel iterative solution of the linear systems — both uniformly well behaved with respect to the time step — can be solved simultaneously within the framework of preconditioning. A multilevel nodal basis preconditioner able to handle highly nonuniform meshes is derived. As a numerical example an application of the method to the bioheat–transfer equation is included.

AMS CLASSIFICATION: 65F10, 65F35, 65M50, 65M60, 65N30.

*This is Part III to the articles [6,7] with the same main title.

CONTENTS

INTRODUCTION

This paper condenses material of the author's thesis [8] and constitutes a direct continuation of Parts I [6] and II [7] [1], in which one can find detailed explanations of the proposed algorithm.

The objective of this part III is to present a 2D version of the algorithm. We will consider an adaptive finite element solver for the arising singularly perturbed elliptic subproblems

$$u + \tau Au = f.$$

The singular perturbation results from the time step τ of the discretization in time; standard adaptive finite element solvers like PLTMG [4] or KASKADE [12, 16, 21] run into difficulties for small time steps, which occur in transient phases.

In view of Section I.4 we have to construct two devices:

- Error estimator
- Linear solver

Both devices have to behave well – *uniformly* in the time step $\tau \geq 0$. Using a multilevel iteration as linear solver, the question of a *proper preconditioner* arises. As it turns out this preconditioner is the key to the error estimator as well.

Because of its use of orthogonal projections a recently presented preconditioner for elliptic equations due to BRAMBLE/PASCIAK/XU [9] — extended to the case of highly nonuniform meshes by YSERENTANT [30] — is ideally suited as conceptual base for our purposes. Moreover this concept is not restricted to certain space dimensions, like hierarchical basis preconditioners, but is easily extended to higher dimensions.

In Section 1 the singularly perturbed problem is stated. Details of triangulations and FEM approximations, which will be used later on, are given. Furthermore we state a list of requirements which a preconditioner has to obey.

Section 2 is devoted to the construction of a preconditioner on the base of the elliptic preconditioner of BRAMBLE/PASCIAK/XU [9]. We first deal with the case of an elliptic operator with no Helmholtz term and natural boundary conditions outside the Dirichlet boundary piece. The thus developed preconditioner, which gives a smooth transient from diagonal preconditioning of

[1] They will be cited throughout this paper with leading roman numbers I resp. II.

the mass matrix to a preconditioner of the stiffness matrix, is thereafter extended to the presence of a Helmholtz term and general Cauchy boundary conditions. For the need of error estimation we present the preconditioning of quadratic elements. We close the section with a discussion of error estimation.

In Section 3 some algorithmic details are given. They include such important issues as the optimal choice of certain parameters, the discussion of possible orders for the time discretization in dependence of the imposed accuracy, a stop criterion for the time error iteration, a stabilization of orthogonal projections and the direct solver on the coarsest triangulation in 2D. The latter becomes important when the starting grid already consists of "many" nodes.

Section 4 finally gives a real life application of our method to the bioheat–transfer equation. This equation plays a prominent role in planning *hyperthermia*, a recent clinical method for the treatment of malignancies (cancer), which at this time is in an experimental status. This shows the full applicability of our method to the given problem class.

Addendum to Part II

The author has recently proven that the conjecture of Remark II.2.9 is true. In fact, he proved [8, Lemma 2.1] for the Laguerre polynomials:

Lemma.

a) $|L_n(x)| \leq 1$ *for* $x \in [0,1]$, $n \geq 0$.

b) $L_n(1) \neq 0$ *for* $n > 1$.

Thus Theorem II.2.8 is true *for all* $j \geq 1$.

1. Triangulations and the Finite Element Discretization

In this section we slightly generalize the parabolic problem considered in Part II[1]. Furthermore the discretization of the 2D elliptic subproblems — which arise by using the time-stepping procedure of Part II — is discussed in detail.

1.1. The Parabolic Problem

We are concerned with linear scalar selfadjoint parabolic initial-boundary value problems:

Given a domain $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary $\partial\Omega = \Gamma_D \dot{\cup} \Gamma_C$, a time $T_{\text{fin}} > 0$, solve for $x \in \Omega$, $t \in]0, T_{\text{fin}}]$

$$
\begin{aligned}
&\text{i)} \quad \phi(x)\frac{\partial u(t,x)}{\partial t} + A(x,\partial)u(t,x) = f(x), \\
&\text{ii)} \quad u(t,\cdot)|_{\Gamma_D} = g(\cdot), \\
&\text{iii)} \quad \left(\frac{\partial}{\partial \nu_A} + \zeta(\cdot)\right) u(t,\cdot)\Big|_{\Gamma_C} = \xi(\cdot), \\
&\text{iv)} \quad u(0,\cdot) = u_0.
\end{aligned}
\tag{1.1}
$$

Here $A(x,\partial)$ denotes a *formally selfadjoint* elliptic operator of second order, which has a principal part in divergence form:

$$
A(x,\partial)u(x) = -\sum_{i,k=1}^{d} \partial_k \left(a_{ik}(x)\partial_i u(x)\right) + q(x)u(x),
$$

where $a_{ik} = a_{ki}$. The associated *conormal* derivative is defined as

$$
\frac{\partial}{\partial \nu_A} = \sum_{i,k=1}^{d} n_k a_{ik} \partial_i,
$$

where $n = (n_1, \ldots, n_d)^T$ denotes the outer unit normal on $\partial\Omega$. Boundary condition iii) is sometimes called *Cauchy boundary condition.*

Notation. The norms of the Sobolev spaces $H^s(\Omega)$ will be denoted by $\|\cdot\|_s$, their seminorms by $|\cdot|_s$, the norms of the spaces $W^{s,p}(\Omega)$ by $\|\cdot\|_{s,p}$

[1]Throughout this paper we use the *notation* introduced in Part II.

and the inner product of $L^2(\Omega)$ will be denoted by $(\cdot,\cdot)$. For a function $\psi \in L^\infty(\Omega) = W^{0,\infty}(\Omega)$ with $\psi \geq 0$ a.e. we abbreviate

$$\psi_{\max} = \|\psi\|_{0,\infty} \text{ and } \psi_{\min} = 1/\|1/\psi\|_{0,\infty}.$$

We make the following *assumptions*:

1. Ω has Lipschitz boundary, i.e., $\Omega \in C^{0,1}$. Furthermore Γ_D is a closed subset of $\partial\Omega$.
2. $\phi, q, \zeta, a_{ik} \in L^\infty(\Omega)$.
3. $\phi, q, \zeta \geq 0$ a.e., moreover $\phi_{\min} > 0$.
4. $A(x,\partial)$ is strongly elliptic, such that there are constants $0 < \delta \leq 1 \leq \Delta$

$$\delta \sum_{i=1}^{d} \xi_i^2 \leq \sum_{i,k=1}^{d} a_{ik}(x)\xi_i\xi_k \leq \Delta \sum_{i=1}^{d} \xi_i^2$$

   for all $\xi \in \mathbb{R}^d$ and almost all $x \in \Omega$.
5. $f, u_0 \in L^2(\Omega)$.
6. $g, \xi \in H^{1/2}(\partial\Omega)$.

By means of assumption 6 and the known properties of the trace operator, we can take by a simple transformation the case that

$$g, \xi \equiv 0.$$

For ease of representation we will assume mostly in this paper

$$\phi \equiv 1.$$

The *extension* to the case $\phi \not\equiv 1$ will be discussed in Section 4.3.1.

We introduce the space of weak solutions

$$H_D^1(\Omega) = \left\{ u \in H^1(\Omega) \mid u|_{\Gamma_D} = 0 \right\},$$

(the restriction is understood in the sense of traces) and consider the following continuous symmetric bilinear form $a(\cdot,\cdot)$ on $H_D^1(\Omega) \times H_D^1(\Omega)$:

$$a(u,v) = \sum_{i,k=1}^{d} \int_\Omega a_{ik}\partial_i u \partial_k v \, dx + \int_\Omega q\, uv \, dx + \int_{\Gamma_C} \zeta\, uv \, d\sigma,$$

$u, v \in H_D^1(\Omega)$. The $H_D^1(\Omega)$-*ellipticity* of the form $a(\cdot,\cdot)$, i.e. the existence of a constant $c_1 > 0$ such that

$$a(u,u) \geq c_1 \|u\|_1^2 \quad \text{for all } u \in H_D^1(\Omega),$$

follows from well known conditions given in the next Lemma, a proof may be found in [8, Lemma 1.1].

LEMMA 1.1. *Each of the following cases guarantees the $H_D^1(\Omega)$-ellipticity of the form $a(\cdot,\cdot)$:*

i) $\operatorname{mes}(\Gamma_C) = 0$. *In this case we estimate for $u \in H_D^1(\Omega)$*

$$a(u,u) \geq \frac{\delta}{1 + d_\Omega^2/2} \|u\|_1^2$$

and

$$a(u,u) \geq \frac{2\delta}{d_\Omega^2} \|u\|_0^2.$$

Here d_Ω denotes the band width of a strip containing Ω.

ii) $q_{\min} > 0$. *In this case we estimate for $u \in H_D^1(\Omega)$*

$$a(u,u) \geq \min(\delta, q_{\min}) \|u\|_1^2$$

and

$$a(u,u) \geq q_{\min} \|u\|_0^2.$$

iii) $\operatorname{mes}(\Gamma_D) > 0$.

iv) $\operatorname{mes}(\Gamma_C) > 0$ *and* $\zeta_{\min} > 0$.

The *weak representation* A of the differential operator $A(x,\partial)$ with the given boundary conditions and the spaces $\dot{H}^{2\alpha} = D_{A^\alpha}$ can now be introduced like in Theorem II.1.1.

1.2. THE SINGULARLY PERTURBED ELLIPTIC PROBLEMS

As we have seen in Section II.2.2 the elliptic problems resulting from discretization in time of the parabolic problem can always be given in the following variational form: Find $u \in H_D^1(\Omega)$ such that

$$(u,v) + \tau a(u,v) = \theta_0^*(v) + \tau \theta_1^*(v) \quad \text{for all } v \in H_D^1(\Omega).$$

Here $\theta_0^*, \theta_1^* \in L^2(\Omega) \subset H_D^{-1}(\Omega) = (H_D^1(\Omega))^*$. In order to bound the energy norm for large values of τ, we scale with $1+\tau$ to get the equivalent problem: Find $u \in H_D^1(\Omega)$ such that

$$a_\tau(u,v) = \theta_\tau^*(v) \quad \text{for all } v \in H_D^1(\Omega), \tag{1.2}$$

where we made use of the following notation:

$$\text{i)} \quad a_\tau(u,v) = \frac{1}{1+\tau}(u,v) + \frac{\tau}{1+\tau}a(u,v), \quad u,v \in H_D^1(\Omega),$$

$$\text{ii)} \quad \theta_\tau^*(v) = \frac{1}{1+\tau}\theta_0^*(v) + \frac{\tau}{1+\tau}\theta_1^*(v), \quad v \in H_D^1(\Omega).$$

REMARK 1.1. Note that this scaling is *not invariant* to a linear scale of the time variable. Thus the question of an appropriate scaling of given physical examples arises. This can be answered in a satisfying way as discussed in Section 4.3.2.

Furthermore we get as the corresponding weak representation (cf. Section II.1.1) the positive selfadjoint operator

$$\Lambda = \frac{1}{1+\tau}I + \frac{\tau}{1+\tau}A, \tag{1.3}$$

with the domain of definition

$$D(\Lambda) = \begin{cases} \dot{H}^2, & \tau > 0 \\ L^2(\Omega), & \tau = 0 \end{cases}.$$

The *energy norm* $\|\Lambda^{1/2} \cdot \|_0$ will be denoted by $\|\cdot\|_\Lambda$.

Case $\tau = 0$. In this case problem (1.2) reads as

$$(u,v) = \theta_0^*(v) \quad \text{for all } v \in H_D^1(\Omega).$$

The corresponding energy norm is the L^2–norm:

$$\|u\|_\Lambda = \|u\|_0, \quad u \in L^2(\Omega).$$

Case $\tau = \infty$. In this case problem (1.2) reduces to the stationary problem

$$a(u,v) = \theta_1^*(v), \quad v \in H_D^1(\Omega).$$

The corresponding energy norm is the energy norm of the elliptic operator A:

$$\|u\|_\Lambda = \|u\|_{\dot{H}^1}, \quad u \in H_D^1(\Omega).$$

1.3. Triangulations and the Finite Element Spaces

1.3.1. Triangulations

Let $\Omega \subset \mathbb{R}^2$ be a bounded simply connected polygonal domain. This implies that

$$\Omega \in C^{0,1},$$

i.e., assumption 1 of Section 1.1 is fulfilled.

A *triangulation* $\mathcal{T}$ of the polygonal domain Ω is given as the set of triangles resulting from a simplicial partition of Ω.

We start with a coarse triangulation $\mathcal{T}_0$ of Ω with the property that the Dirichlet boundary piece Γ_D is composed of edges of triangles $T \in \mathcal{T}_0$. The triangulation $\mathcal{T}_0$ is refined several times, giving a family of *nested* triangulations $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_j$. A triangle of $\mathcal{T}_{k+1}$ is either a triangle of $\mathcal{T}_k$ or is generated by subdividing a triangle of $\mathcal{T}_k$ into four congruent triangles or into two triangles by connecting one of its vertices with the midpoint of the opposite side. The first case is called a regular or *red* refinement and the resulting triangles as well as the triangles of the initial triangulation are called regular triangles. The second case is an irregular or *green* refinement and results in two so–called irregular triangles.

However, the irregular refinement has the character of a *closure* which we force by the following *rule*:

(T1) Each new vertex of $\mathcal{T}_k$, i.e., a vertex which does not belong to $\mathcal{T}_{k-1}$, is a vertex of a triangle which was generated by regular refinement.

The irregular refinement is potentially dangerous because interior angles are reduced. Therefore, we add the following rule:

(T2) Irregular triangles may not be further refined.

This rule insures that every triangle of any triangulation $\mathcal{T}_k$ is geometrically similar to a triangle of the initial triangulation $\mathcal{T}_0$ or to a green refinement of a triangle in $\mathcal{T}_0$. These triangulations are meanwhile standard and have been introduced by Bank et al. in [3, 4, 5].

The index of the final triangulation will always be denoted by j and will be fixed in most of the following considerations.

By the *depth* of a triangle

$$T \in \bigcup_{k=0}^{j} \mathcal{T}_k$$

we mean the number of successive ancestors in the family of triangulations. If we add the rule

(T3) Only triangles of depth $k-1$ are refined for the construction of $\mathcal{T}_k$,

we get the following expression for the depth of a triangle $T \in \bigcup_{k=0}^{j} \mathcal{T}_k$

$$\text{depth}(T) = \min\{0 \le k \le j \mid T \in \mathcal{T}_k\}.$$

Equipped with rule (T3) we can uniquely reconstruct the sequence $\mathcal{T}_1, \ldots, \mathcal{T}_{j-1}$ from the knowledge of the initial triangulation $\mathcal{T}_0$ and the final triangulation $\mathcal{T}_j$ *alone*, without knowing the actual dynamic refinement process leading to $\mathcal{T}_j$ in an adaptive algorithm, see [12]. However, if we choose the data-structures representing the triangulation cleverly, the sequence $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_j$ is *implicitly* given. For example this is the case in the adaptive FEM code KASKADE, cf. ROITZSCH [21, 22] or LEINEN [16].

1.3.2. Notation for Finite Element Spaces

Corresponding to the triangulations $\mathcal{T}_k$ we have *finite element* spaces $\mathcal{S}_k$. $\mathcal{S}_k$ consists of all functions which are linear on each triangle $T \in \mathcal{T}_k$ and continuous on Ω. Furthermore they vanish on the Dirichlet boundary piece Γ_D. Because the triangulations are nested we have

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \ldots \subset \mathcal{S}_j \subset H^1_D(\Omega).$$

Let $\mathcal{N}_k = \{x_1^{(k)}, \ldots, x_{n_k}^{(k)}\}$ be the set of vertices of triangles in $\mathcal{T}_k$, which do not lie on the Dirichlet boundary piece Γ_D.

The nodal basis. The set $\Gamma_k = \{\psi_1^{(k)}, \ldots, \psi_{n_k}^{(k)}\}$ of nodal basis functions, where

$$\psi_i^{(k)}(x_l^{(k)}) = \delta_{il} \quad \text{for } 1 \le i, l \le n_k,$$

forms a basis of $\mathcal{S}_k$. For $\psi \in \Gamma_k$ we denote by $x_\psi \in \mathcal{N}_k$ the *supporting point* of ψ, i.e.

$$\psi(x_\psi) = 1.$$

Structuring of the nodal bases of varying index k. We set

$$\begin{aligned} \text{i)} \quad \Psi &= \bigcup_{k=0}^{j} \Gamma_k, \\ \text{ii)} \quad \Psi_0 &= \Gamma_0, \\ \text{iii)} \quad \Psi_k &= \Gamma_k \setminus \Gamma_{k-1}, \qquad \text{whenever } 1 \le k \le j. \end{aligned}$$

It should be stressed that we split the set of nodal basis functions rather than the set of nodal points as done in hierarchical basis approaches. For $\psi \in \Psi$ we denote the set of indices, for which a nodal basis function ψ occurs, by

$$K_\psi = \{k \mid \psi \in \Gamma_k\}.$$

Here we abbreviate the first resp. the last occurrence of ψ in a set Γ_k by

$$\begin{aligned} \text{i)} \quad k_\psi^0 &= \min K_\psi, \\ \text{ii)} \quad k_\psi^1 &= \max K_\psi. \end{aligned}$$

The duality map. According to the Theorem of Fréchet–Riesz the duality map

$$\begin{aligned} \mathcal{I}_k : \mathcal{S}_k &\to \mathcal{S}_k^* \\ u &\mapsto u^* = (u, \cdot) \end{aligned}$$

is an isometrical isomorphism.
The dual basis. On $\mathcal{S}_k^*$ a natural basis is given by the canonical dual basis $\Gamma_k^* = \{\psi_* |\ \psi \in \Gamma_k\}$ to the basis Γ_k of $\mathcal{S}_k$. As usual ψ_* is defined as the *evaluation functional* at x_ψ:

$$\begin{aligned} \psi_* : \mathcal{S}_k &\to \mathbb{R} \\ u &\mapsto u(x_\psi), \end{aligned}$$

such that $\psi_*(\varphi) = \delta_{\psi\varphi}$ for all $\psi, \varphi \in \mathcal{S}_k$. The choice of these bases will be called the *natural representation* of the spaces $\mathcal{S}_k$ and $\mathcal{S}_k^*$.
The orthogonal L^2–projections. The orthogonal L^2 projections $\pi_k : L^2(\Omega) \to \mathcal{S}_k$, for $0 \le k \le j$ are given for $u \in L^2(\Omega)$ as

$$(\pi_k u, v) = (u, v) \qquad \text{for all } \ v \in \mathcal{S}_k.$$

1.4. The Finite Element Discretization

The finite element (FEM) discrete solution $u_k \in \mathcal{S}_k$ is given as the Galerkin approximation to the variational problem (1.2)

$$a_\tau(u_k, v_k) = f^*(v_k) \qquad \text{for all } \ v_k \in \mathcal{S}_k. \tag{1.4}$$

Here $f^* \in \mathcal{S}_j^*$ denotes an approximation of θ_τ^* on $\mathcal{S}_j$. Due to the Theorem of Fréchet–Riesz there are *symmetric positive definite* linear operators $A_k, \Lambda_k : \mathcal{S}_k \to \mathcal{S}_k$, such that for given $u, v \in \mathcal{S}_k$

$$(A_k u, v) = a(u, v)$$

resp.

$$(\Lambda_k u, v) = a_\tau(u, v).$$

For $0 \le k \le l$ we obtain the relations

$$A_k = \pi_k A_l|_{\mathcal{S}_k}$$

and

$$\Lambda_k = \pi_k \Lambda_l|_{\mathcal{S}_k} = \frac{1}{1+\tau} I_k + \frac{\tau}{1+\tau} A_k,$$

where I_k denotes the identity on the space $\mathcal{S}_k$. Problem (1.4) is now given as

$$\mathcal{I}_k \Lambda_k u_k = f^*|_{\mathcal{S}_k}. \tag{1.5}$$

1.5. The Solution Process and Requirements for a Preconditioner

Computationally problem (1.4) is realized for $k = j$ as follows: We have

$$u_j = \sum_{i=1}^{n_j} u_j(x_{\psi_i^{(j)}}) \psi_i^{(j)},$$

which implies the equivalence of (1.4) and

$$\sum_{i=1}^{n_j} u_j(x_{\psi_i^{(j)}})\ a_\tau(\psi_i^{(j)}, \psi_l^{(j)}) = f^*(\psi_l^{(j)}) \qquad \text{for all } 1 \le l \le n_j.$$

By introducing the *mass matrix* $\mathrm{M} = (m_{il})_{il}$ with

$$m_{il} = (\psi_i^{(j)}, \psi_l^{(j)})$$

and the *stiffness matrix* $\mathrm{A} = (a_{il})_{il}$ with

$$a_{il} = a(\psi_i^{(j)}, \psi_l^{(j)})$$

for $1 \le i, l \le n_j$, we gain as problem matrix A_τ the following convex combination of M and A:

$$\mathrm{A}_\tau = \frac{1}{1+\tau}\mathrm{M} + \frac{\tau}{1+\tau}\mathrm{A}.$$

Introducing the vectors $\vec{u} = \left(u_j(x_{\psi_i^{(j)}})\right)_i$ and $\vec{f} = \left(f^*(\psi_l^{(j)})\right)_l$ we obtain the computational problem

$$\mathrm{A}_\tau \vec{u} = \vec{f}. \tag{1.6}$$

However, this linear equation on $\mathbb{R}^{n_j}$ is just the natural matrix representation of the linear problem (1.5) in the case $k = j$

$$\mathcal{I}_j \Lambda_j u_j = f^*. \tag{1.7}$$

This fact is the reason why we have stressed the importance of the natural representation of the dual pair $(\mathcal{S}_j, \mathcal{S}_j^*)$, which will serve as a rather elegant method to describe the computational problem.

The large linear system (1.6) has to be solved iteratively. Since the involved matrices are symmetric positive definite a preconditioned conjugate gradient (CG) method is the method of choice.

We require several features for a preconditioning matrix B_τ:

(P1) The spectral condition number $\kappa = \kappa(B_\tau A_\tau)$ should only grow in j like $j^{2\nu}$, where $0 \leq \nu \leq 1$. Further it should remain bounded *independently* of the time step $\tau \geq 0$. These properties should neither depend (severely) on the shape of the domain under consideration nor on any quasi–uniformity of the triangulations.

(P2) The cost of computing $B_\tau \vec{r}$ should be proportional to the dimension n_j.

By requirement (P1) the number $\ell(\epsilon)$ of iterations necessary to reduce the error in the energy norm of A_τ by the factor ϵ is bounded by

$$\ell(\epsilon) \leq \frac{1}{2}\sqrt{\kappa}\left|\log\frac{\epsilon}{2}\right| = \mathcal{O}(j^\nu) \quad , \ 0 \leq \nu \leq 1,$$

independently of τ. If we solve each of the linear problems only as accurate as the discretization on the corresponding triangulation is expected to be, we end up with an *overall complexity* of

$$\mathcal{O}(j^{\nu+\sigma} n_j), \qquad 0 \leq \sigma \leq 1,$$

in view of requirement (P2) — an idea due to DEUFLHARD et al. [12] and implemented in the adaptive FEM solver KASKADE. The exponent σ is connected with the progression of unknowns during refinement: $\sigma = 0$ in the case of geometrical progression, whereas $\sigma = 1$ in the case of pure arithmetical progression. *Note that we do not propose to force the number of unknowns to progress geometrically* — for a reason discussed in [8, Example 8.3].

Reliable time–step control requires that the locally arising systems of ordinary differential equations, as which our algorithm can be viewed in each time–layer, are smooth, thus leading to

(P3) The matrix B_τ should depend *smoothly* on $\tau \geq 0$.

Finally we do not want to analyze the problem in matrix notation but in the corresponding operator version (1.7). If we introduce the operator $\Theta_j^* : \mathcal{S}_j^* \to \mathcal{S}_j$, whose matrix in the natural representation of $(\mathcal{S}_j, \mathcal{S}_j^*)$ is given by B_τ, the preconditioned CG–method can be written in its untransformed fashion as follows:

1.5.1. Preconditioned CG–method on $(\mathcal{S}_j, \mathcal{S}_j^*)$.

We want to solve

$$\mathcal{I}_j \Lambda_j u = f^*.$$

Given a start iteration $\tilde{u}_0$ we set

$$p_0 = r_0 = \Theta_j^* \left(f^* - \mathcal{I}_j \Lambda_j \tilde{u}_0 \right).$$

For $k = 0, 1, \ldots$ we iterate

$$\begin{aligned} \tilde{u}_{k+1} &= \tilde{u}_k + \alpha_k p_k \\ p_{k+1} &= r_{k+1} - \beta_k p_k \end{aligned}$$

with

$$\begin{aligned} r_{k+1} &= \Theta_j^* \left(f^* - \mathcal{I}_j \Lambda_j \tilde{u}_{k+1} \right) \\ \alpha_k &= \frac{(\mathcal{I}_j \Lambda_j r_k)(p_k)}{(\mathcal{I}_j \Lambda_j p_k) \left(\Theta_j^* \mathcal{I}_j \Lambda_j p_k \right)} \\ \beta_k &= \frac{(\mathcal{I}_j \Lambda_j r_{k+1}) \left(\Theta_j^* \mathcal{I}_j \Lambda_j p_k \right)}{(\mathcal{I}_j \Lambda_j p_k) \left(\Theta_j^* \mathcal{I}_j \Lambda_j p_k \right)}. \end{aligned}$$

In the natural representation of $(\mathcal{S}_j, \mathcal{S}_j^*)$ we directly get the computationally available version of the preconditioned CG–method. We thus have to require that the operator Θ_j^* is in fact already given in that representation, that means

(P4) The operator Θ_j^* should be given in such a form that directly allows to reconstruct the matrix B_τ without any further effort.

1.6. Quadratic Elements

For the use of error estimation the space of piecewise quadratic elements on $\mathcal{T}_j$ will be needed later on. Here we introduce the corresponding *notation:*

The space $\mathcal{S}_Q$ consists of all functions which are a quadratic polynomial on each triangle $T \in \mathcal{T}_j$ and which are continuous on Ω. Furthermore they vanish on the Dirichlet boundary piece Γ_D, such that

$$\mathcal{S}_Q \subset H_D^1(\Omega).$$

Let $\mathcal{N}_Q$ be the set of *midpoints* of edges belonging to $\mathcal{T}_j$ but not to the Dirichlet boundary piece Γ_D. Take the (quadratic) *hierarchical* basis Γ_Q, which consists of those $\psi \in \mathcal{S}_Q$, for which

$$\psi(x) = 0$$

for all $x \in \mathcal{N}_j$ and

$$\psi(x_\psi) = 1$$

for *exactly one* $x_\psi \in \mathcal{N}_Q$, the *supporting point* of ψ. With $\mathcal{V}_Q = \operatorname{span} \Gamma_Q$ we gain the direct composition

$$\mathcal{S}_Q = \mathcal{S}_j \oplus \mathcal{V}_Q.$$

The operators $I_Q, A_Q, \Lambda_Q : \mathcal{S}_Q \to \mathcal{S}_Q$ and $\mathcal{I}_Q : \mathcal{S}_Q \to \mathcal{S}_Q^*$ have the analogous meaning to I_j, A_j, Λ_j and $\mathcal{I}_j$.

2. The Multilevel Preconditioner

The objective of this section is the construction of a preconditioner, which obeys the requirements (P1)–(P4) of Section 1.5. This will be done for piecewise linear as well as for piecewise quadratic elements, for which a preconditioner is needed for the purpose of error estimation as we will see in the discussion of Section 2.4.

2.1. A Preconditioner for Piecewise Linear Elements

We first restrict the discussion to forms $a(\cdot,\cdot)$ which consists only of the principal part, i.e., $q \equiv 0$ and $\zeta \equiv 0$. Thus there is no Helmholtz term present and the boundary conditions on Γ_C are natural boundary conditions.

However, we do not exclude $\operatorname{mes}(\Gamma_D) = 0$ *in this section.* This will be important for the discussion of the next Section 2.2. Thus the form $a(\cdot,\cdot)$ might even be *not* $H^1_D(\Omega)$–elliptic.

For the finite element discretization of the *purely elliptic* problem

$$A_j u_j = f \quad \text{on } \mathcal{S}_j$$

two good preconditioners B_j are known:

- the *hierarchical basis preconditioner* due to Yserentant [28],
- the *multilevel nodal basis preconditioner* due to Bramble/Pasciak/Xu [9], also Xu [26, 27].

They are both based on a subspace decomposition of $\mathcal{S}_j$, which means

$$\mathcal{S}_j = \mathcal{S}_0 \oplus \mathcal{V}_1 \oplus \ldots \oplus \mathcal{V}_j, \qquad \mathcal{V}_k \subset \mathcal{S}_k \quad \text{for } 1 \le k \le j. \tag{2.1}$$

Both preconditioners lead to condition numbers $\kappa(B_j A_j) = \mathcal{O}(j^2)$. However, if we handle instead the problem resulting from time discretization of

a parabolic problem, we end up with the finite element equation (1.5) which is for $k = j$ equivalent to

$$\Lambda_j u_j = f \quad \text{on } \mathcal{S}_j. \tag{2.2}$$

A straightforward generalization of the preconditioners by just taking Λ_j instead of A_j is not possible since for $\tau \downarrow 0$ the ellipticity constant of the problem, which seriously enters $\kappa(B_j A_j)$, vanishes. On the other hand for $\tau = 0$ there is no need of preconditioning at all, since then $\Lambda_j = I_j$.

YSERENTANT suggested in [29] a τ–dependent version of his hierarchical basis preconditioner using local *Courant–numbers*, which allow locally to switch between the nodal and the hierarchical basis. However, this yields to a *non–smooth* dependence of the preconditioner on τ which is not desirable in the context of time–discretization, compare the requirement (P3) for a preconditioner as discussed in the last section. The same is true for the corresponding modification due to OSWALD [19] of the multilevel nodal basis preconditioner.

XU suggested in [26] a natural τ–dependent version of the multilevel nodal basis preconditioner depending smoothly on τ. However, he considers only the case of *quasi uniform* triangulations and moreover it is not at all clear whether multiplying this τ–dependent preconditioner by a vector can be realized within $\mathcal{O}(n_j)$ operations, as was required in (P2). This is also true if one uses XU's ideas together with the version of the multilevel nodal basis preconditioner for highly nonuniform triangulations by YSERENTANT [30].

However, by some modifications of XU's and YSERENTANT's constructions it is possible to overcome the above mentioned difficulties as will be shown in this section. Besides that we intend to clarify some aspects of their original constructions.

2.1.1. *A Preconditioner Based on an Orthogonal Splitting of the Finite Element Spaces*

BRAMBLE/PASCIAK/XU specified the subspace decomposition (2.1) as follows

$$\mathcal{V}_k = (\pi_k - \pi_{k-1})\mathcal{S}_j \quad \text{for } 1 \le k \le j,$$

and considered the symmetric positive definite operator

$$B_j^{-1} = A_0\pi_0 + \sum_{k=1}^{j} 4^k(\pi_k - \pi_{k-1}).$$

LEMMA 2.1. (YSERENTANT [30, Theorem 4.6]) *There are positive constants K_0 and K_1 with*

i) $\frac{\mu_0}{j+1}(B_j^{-1}u,u) \leq (A_j u,u) \leq \mu_1(j+1)(B_j^{-1}u,u)$, *where*

ii) $\mu_0 = \frac{\delta}{\Delta}K_0$,

iii) $\mu_1 = \Delta K_1$,

for all functions $u \in \mathcal{S}_j$. Furthermore $\mu_0 \leq 1 \leq \mu_1$ holds. The constants δ, Δ as introduced in assumption 4 of Section 1.1 describe the coefficient matrix a_{ik} of the elliptic operator $A(x,\partial)$, whereas the constants K_0, K_1 depend only on the geometry of the initial triangulation $\mathcal{T}_0$ and they are independent of the maximal depth j of the final triangulation.

REMARK 2.1. This Lemma is also valid for the case $\mathrm{mes}(\Gamma_D) = 0$. YSERENTANT assumes in his paper $\mathrm{mes}(\Gamma_D) > 0$, but this assumption nowhere enters the proof of Lemma 2.1. Surely then A_j and B_j^{-1} will only be positive *semi*–definite and

$$\|\cdot\|_{A_j} = (A_j\cdot,\cdot)^{1/2}, \|\cdot\|_{B_j^{-1}} = (B_j^{-1}\cdot,\cdot)^{1/2}$$

will only be seminorms then. According to the above Lemma 2.1 these seminorms must have the same null–spaces, a fact which can also be seen by the Poincaré inequality. This inequality specifies the null–space as $\mathrm{span}\{1\} \subset \mathcal{S}_j$.

By the symmetry of B_j^{-1} and A_j the symmetric positive *definite* operator

$$\Theta_j^{-1} = \frac{1}{1+\tau}I_j + \frac{\tau}{1+\tau}B_j^{-1}$$

should be spectrally close to Λ_j. Since $I_j = \pi_0 + \sum_{k=1}^{j}(\pi_k - \pi_{k-1})$ the representation

$$\Theta_j^{-1} = \Lambda_0\pi_0 + \sum_{k=1}^{j}\lambda_k(\pi_k - \pi_{k-1}) \tag{2.3}$$

holds, where

$$\lambda_k = \frac{1+\tau 4^k}{1+\tau}. \tag{2.4}$$

COROLLARY 2.1. *The following inequalities hold for all functions $u \in \mathcal{S}_j$:*

$$\frac{\mu_0}{j+1}(\Theta_j^{-1}u,u) \le (\Lambda_j u,u) \le \mu_1(j+1)(\Theta_j^{-1}u,u).$$

The constants μ_0, μ_1 are taken from Lemma 2.1.

Due to the properties of orthogonal projections

$$\Theta_j = \Lambda_0^{-1}\pi_0 + \sum_{k=1}^{j} \lambda_k^{-1}(\pi_k - \pi_{k-1})$$

holds on $\mathcal{S}_j$. Replacement by the operator

$$\begin{aligned} \bar{\Theta}_j &= \Theta_j + \frac{\tau}{1+\tau} A_0 \Lambda_0^{-1}\pi_0 \\ &= \frac{\tau}{1+\tau}\Lambda_0^{-1}\pi_0 + \sum_{k=0}^{j} \vartheta_k \pi_k, \end{aligned}$$

with

$$0 \le \vartheta_k = \begin{cases} \lambda_k^{-1} - \lambda_{k+1}^{-1} & \text{if } k < j \\ \lambda_j^{-1} & \text{if } k = j, \end{cases} \tag{2.5}$$

where $0 \le k \le j$, changes only the constants.

LEMMA 2.2. *The following inequalities hold for all $u \in \mathcal{S}_j$*

$$(\Theta_j u,u) \le (\bar{\Theta}_j u,u) \le (1+\Delta K_1)(\Theta_j u,u).$$

Here K_1 denotes the constant of Lemma 2.1.

Proof. The left inequality is obvious since the added operator is symmetric positive semi–definite. By using an inverse inequality like Lemma 3.3 of YSERENTANT [30] we get for $u \in \mathcal{S}_j$

$$\begin{aligned} \|\pi_0 u\|_0^2 &= \|\Lambda_0^{-1/2}\pi_0 u\|_{\Lambda_0}^2 \\ &= \frac{1}{1+\tau}\left(\|\Lambda_0^{-1/2}\pi_0 u\|_0^2 + \tau\|\Lambda_0^{-1/2}\pi_0 u\|_{H^1}^2\right) \\ &\le \frac{1}{1+\tau}\left(1+\tau\Delta K_1\right)\|\Lambda_0^{-1/2}\pi_0 u\|_0^2 \\ &\le \Delta K_1 \|\Lambda_0^{-1/2}\pi_0 u\|_0^2, \end{aligned}$$

since $\Delta K_1 \ge 1$. Here the the constant K_1 of Lemma 2.1 is exactly the constant of the inverse inequality which only depends on the local shape geometry of the initial triangulation $\mathcal{T}_0$. Hence

$$(\pi_0 u,u) \le \Delta K_1 (\Lambda_0^{-1}\pi_0 u,u).$$

Therefore we get

$$\left(\frac{\tau}{1+\tau}\Lambda_0^{-1}\pi_0 u, u\right) + (\pi_0 u, u) \leq (1 + \Delta K_1)(\Lambda_0^{-1}\pi_0 u, u),$$

which in turn implies the right inequality. ∎

2.1.2. *A Computationally Available Spectrally Equivalent Preconditioner*

However, our computational problem *is not* problem (2.2), but as we have seen in Section 1.5 problem (1.7). Identification of problems (2.2) and (1.7) would mean to compute $\mathcal{I}_j^{-1}$, i.e in the natural representation of $(\mathcal{S}_j, \mathcal{S}_j^*)$ to invert the *mass matrix*, a problem of the same complexity as (2.2) itself. Hence it would be far more desirable to get a cheap and easy representable expression for $\bar{\Theta}_j\mathcal{I}_j^{-1}$ to fulfill requirement (P4). However, the representation

$$\pi_k\mathcal{I}_j^{-1}u^* = \mathcal{I}_k^{-1}(u^*|_{\mathcal{S}_k}) \tag{2.6}$$

for $u^* \in \mathcal{S}_j$, $k \leq j$, shows that the computation of $\pi_k\mathcal{I}_j^{-1}$ can only be achieved by inverting $\mathcal{I}_k$, i.e., a mass matrix of dimension n_k. Thus we are led to replace $\mathcal{I}_k$ by an easily invertible $\hat{\mathcal{I}}_k$, the duality map with respect to a new inner product $(\cdot,\cdot)_k$ on $\mathcal{S}_k$. A rather simple possibility of inverting $\hat{\mathcal{I}}_k$ is given when $\hat{\mathcal{I}}_k$ is a diagonal matrix in the natural representation. This requires that the nodal basis functions of Γ_k are mutually *orthogonal* with respect to the new inner product, i.e.,

$$(u,\psi)_k = (\psi,\psi)_k u(x_\psi) = (\psi,\psi)_k\psi_*(u) \tag{2.7}$$

for $\psi \in \Gamma_k$, $u \in \mathcal{S}_k$. Thus the new inner product $(\cdot,\cdot)_k$ has to be a weighted Euclidian product in the basis Γ_k. We now exploit the advantage of the usage of $\hat{\mathcal{I}}_k$ by defining an operator $\hat{\pi}_k : \mathcal{S}_j \to \mathcal{S}_k$ through

$$\hat{\pi}_k\mathcal{I}_j^{-1}u^* = \hat{\mathcal{I}}_k^{-1}(u^*|_{\mathcal{S}_k}), \tag{2.8}$$

in analogy to relation (2.6). Replacement of π_k by $\hat{\pi}_k$ in the preconditioner would only be reasonable if these operators are spectrally equivalent. Hence the remaining degrees of freedom in the new inner product, the weights, are chosen in a way that the new inner product resembles the L^2 inner product as much as possible, which yields us to the construction of a *discrete* L^2 *inner product* by using a quadrature rule with nodes in the vertices of $\mathcal{T}_j$, i.e., for $u, v \in \mathcal{S}_k$

$$(u,v)_k = \frac{1}{3}\sum_{T\in\mathcal{T}_k}|T|\sum_{x\in\mathcal{N}_k\cap T}(uv)(x).$$

This discrete L^2 product satisfies a stability property yielding the spectral equivalence mentioned above. One can prove (cf. [8, Lemma 6.3]) using standard techniques the following estimates.

LEMMA 2.3. *The inequalities*

$$(u,u) \leq (u,u)_k \leq 4(u,u)$$

hold for all $u \in \mathcal{S}_k$.

The next definition is just an equivalent formulation of (2.8) and is due to XU.

DEFINITION 2.1. *The* L^2 *quasi-projection with respect to* $(u,v)_k$ *is given by the operator*

$$\hat{\pi}_k : L^2(\Omega) \to \mathcal{S}_k$$

for which

$$(\hat{\pi}_k u, v)_k = (u,v) \quad \text{for } u \in L^2(\Omega), v \in \mathcal{S}_k.$$

LEMMA 2.4. *The* L^2 *quasi-projection* $\hat{\pi}_k$ *is explicitly given as*

$$\hat{\pi}_k u = \sum_{\psi \in \Gamma_k} \frac{(u,\psi)}{(1,\psi)} \psi \tag{2.9}$$

for $u \in L^2(\Omega)$. *We further have for all* $u \in \mathcal{S}_j$

$$(\hat{\pi}_k u, u) \leq (\pi_k u, u) \leq 4(\hat{\pi}_k u, u). \tag{2.10}$$

Thus the L^2 *quasi-projections* $\hat{\pi}_k$ *and the* L^2 *projections* π_k *are spectrally equivalent, uniformly with respect to* k. *Finally the operator* $\pi_k^* = \hat{\pi}_k \mathcal{I}_j^{-1} : \mathcal{S}_j^* \to \mathcal{S}_k$ *has the representation*

$$\pi_k^* u^* = \sum_{\psi \in \Gamma_k} \frac{u^*(\psi)}{(1,\psi)} \psi \tag{2.11}$$

for $u^* \in \mathcal{S}_j^*$.

Proof. Equation (2.7) states for $u \in L^2(\Omega)$ and $\psi \in \Gamma_k$ that

$$(\hat{\pi}_k u, \psi)_k = (\psi, \psi)_k \, \hat{\pi}_k u(x_\psi).$$

The definition of $(\cdot,\cdot)_k$ implies $(\psi,\psi)_k = \frac{1}{3}\sum_{x_\psi \in T \in \mathcal{T}_k} |T| = (1,\psi)$, which yields

$$\hat{\pi}_k u(x_\psi) = \frac{(\hat{\pi}_k u, \psi)_k}{(\psi,\psi)_k} = \frac{(u,\psi)}{(1,\psi)},$$

i.e., (2.9). Next we define the operator $\sigma_k : \mathcal{S}_k \to \mathcal{S}_k$ such that

$$(\sigma_k u, v) = (u,v)_k$$

for all $u, v \in \mathcal{S}_k$. It is straightforward to check that $\hat{\pi}_k|_{\mathcal{S}_k} = \sigma_k^{-1}$. Lemma 2.3 states

$$(u,u) \le (\sigma_k u, u) \le 4(u,u)$$

for all $u \in \mathcal{S}_k$, which implies relation (2.10) for $u \in \mathcal{S}_k$. Replacing u by $\pi_k u$ we get (2.10) because of $\hat{\pi}_k \pi_k = \hat{\pi}_k$. ∎

We replace π_k by $2\hat{\pi}_k$ and end up with the following preconditioner of Λ_j

$$\hat{\Theta}_j = \frac{\tau}{1+\tau}\Lambda_0^{-1}\pi_0 + 2\sum_{k=0}^{j} \vartheta_k \hat{\pi}_k, \tag{2.12}$$

which is spectrally equivalent to $\bar{\Theta}_j$ and for which $\hat{\Theta}_j \mathcal{I}_j^{-1}$ is computational available without inversion of the mass matrix.

COROLLARY 2.2. *The following inequalities hold for all functions $u \in \mathcal{S}_j$:*

$$\frac{1}{2}(\hat{\Theta}_j u, u) \le (\bar{\Theta}_j u, u) \le 2(\hat{\Theta}_j u, u).$$

2.1.3. *Reduction of the Number of Terms*

However, the realization of $\hat{\Theta}_j \mathcal{I}_j^{-1} u^*$ for $u^* \in \mathcal{S}_j^*$ as suggested by the representation (2.12) would need at least

$$\mathcal{O}(\sum_{k=1}^{j} n_k)$$

operations since every $\hat{\pi}_k$ needs $\#\Gamma_k = n_k$ summations. For non-geometrical progression of the n_k, which we did not exclude, $\sum_{k=1}^{j} n_j$ will not be $\mathcal{O}(n_j)$ as desired, in the case of pure arithmetical progression it is even $\mathcal{O}(n_j^2)$.

For this reason we finally look whether the number of operations can be reduced to an effort of $\mathcal{O}(n_j)$. This objective can be achieved by a proper

rearrangement of the terms in (2.12). For $u \in \mathcal{S}_j$

$$\begin{aligned}\hat{\Theta}_j u &= \frac{\tau}{1+\tau}\Lambda_0^{-1}\pi_0 u + 2\sum_{k=0}^{j}\vartheta_k \sum_{\psi\in\Gamma_k}\frac{(u,\psi)}{(1,\psi)}\psi \\ &= \frac{\tau}{1+\tau}\Lambda_0^{-1}\pi_0 u + 2\sum_{\psi\in\Psi}\vartheta(\psi)\frac{(u,\psi)}{(1,\psi)}\psi\end{aligned}$$

holds with

$$\vartheta(\psi) = \sum_{k\in K_\psi}\vartheta_k.$$

These numbers have a simple expression as the following Lemma shows.

LEMMA 2.5. *For $\psi \in \Psi$ we have*

$$\vartheta(\psi) = \begin{cases} \lambda_{k_\psi^0}^{-1} - \lambda_{k_\psi^1+1}^{-1}, & \text{whenever } 0 \le k_\psi^0 \le k_\psi^1 < j \\ \lambda_{k_\psi^0}^{-1}, & \text{whenever } 0 \le k_\psi^0,\ k_\psi^1 = j. \end{cases} \tag{2.13}$$

Proof. The set K_ψ of those depths that ψ occurs in the corresponding nodal basis is according to [8, Lemma 4.1.iii] given by

$$K_\psi = \{k|\ k_\psi^0 \le k \le k_\psi^1\}.$$

Thus the sum

$$\vartheta(\psi) = \sum_{k=k_\psi^0}^{k_\psi^1}\vartheta_k,$$

reduces to (2.13) due to its telescope nature. ∎

Thus $\Theta_j^* = \hat{\Theta}_j\mathcal{I}_j^{-1}$ is given by

$$\Theta_j^* u^* = \frac{\tau}{1+\tau}(\mathcal{I}_0\Lambda_0)^{-1}(u^*|_{\mathcal{S}_0}) + 2\sum_{\psi\in\Psi}\vartheta(\psi)\frac{u^*(\psi)}{(1,\psi)}, \tag{2.14}$$

$u^* \in \mathcal{S}_j^*$.

LEMMA 2.6. (BORNEMANN [8, Lemma 4.1.iv]). *Suppose the family $\{\mathcal{T}_k\}_k$ obeys the rules (T1), (T2) and (T3) of Section 1. Then the estimate*

$$\#\Psi \le 2n_j - n_0$$

holds no matter how the sequence $\{n_k\}_k$ actually progresses.

This estimate shows that the sum in (2.14) consists of $\mathcal{O}(n_j)$ terms as desired, which in turn suggests that $\Theta_j^* u^*$ can be computed in $\mathcal{O}(n_j)$ operations. We constructed in [8] an implementation within the existing data structures of KASKADE [22], which really needs $\mathcal{O}(n_j)$ operations using an additional amount of storage of less than $4n_j$ reals and $3n_j$ integers – no matter how the n_k actually progress.

Summarizing our results so far we can state the main Theorem of this section.

THEOREM 2.1. *For all functions $u \in \mathcal{S}_j$ and for all numbers $\tau \geq 0$ the following inequalities hold*

$$\frac{\hat{\mu}_0}{j+1}(\Lambda_j^{-1}u, u) \leq (\hat{\Theta}_j u, u) \leq \hat{\mu}_1(j+1)(\Lambda_j^{-1}u, u)$$

where

$$\hat{\mu}_0 = \frac{\delta}{2\Delta}K_0 \quad \textit{and} \quad \hat{\mu}_1 = 2\Delta K_1(1+\Delta K_1).$$

The constants δ, Δ as introduced in assumption 4 of Section 1.1 describe the coefficient matrix a_{ik} of the elliptic operator $A(x,\partial)$, whereas the constants K_0, K_1 depend only on the geometry of the initial triangulation $\mathcal{T}_0$ and they are independent of the maximal depth j of the final triangulation and of the parameter τ.

REMARK 2.2. The case $\tau = 0$ is the preconditioning of $\mathcal{I}_j$, the operator which is represented by the *mass matrix* in the natural bases Γ_j, Γ_j^*. Here Theorem 2.1 states that $\Theta_j^* = 2\pi_j^*$ is a natural choice of a preconditioner. The operator $2\pi_j^*$ is now given by

$$(2\pi_j^*)u^* = \sum_{\psi \in \Gamma_j} \frac{u^*(\psi)}{(\psi,\psi)}\psi$$

since $(\psi,\psi) = \frac{1}{6}|\text{supp } \psi| = \frac{1}{2}(1,\psi)$ for $\psi \in \Gamma_j$. Hence $2\pi_j^*$ is represented in the natural bases by the matrix D^{-1} where $\mathrm{D} = \text{ diag}(m_{11}, \ldots, m_{n_j n_j})$ denotes the diagonal of the mass matrix M. Thus Lemma 2.4 gives in passing the result

COROLLARY 2.3. *The following holds for the diagonal preconditioner of the mass matrix:*

$$\kappa(\mathrm{D}^{-1}\mathrm{M}) \leq 4$$

and

$$\sigma(\mathrm{D}^{-1}\mathrm{M}) \subset [\frac{1}{2}, 2].$$

WATHEN [24] proved these results with a different technique.

REMARK 2.3. Since $\Lambda_j \to A_j$ for $\tau \to \infty$ and in turn, assuming now $\mathrm{mes}(\Gamma_D) > 0$,

$$\hat{\Theta}_j \to \hat{B}_j = A_0^{-1}\pi_0 + 2\sum_{k=0}^{j}\left(4^{-k} - 4^{-(k+1)}\right)\hat{\pi}_k,$$

Theorem 2.1 states that $\hat{B}_j$ is a good preconditioner for A_j. In fact the preconditioner

$$C_j = A_0^{-1}\pi_0 + \sum_{k=1}^{j} 4^{-k}\hat{\pi}_k$$

advocated by YSERENTANT in [29] is spectrally equivalent to $\hat{B}_j$:

$$(C_j u, u) \leq (\hat{B}_j u, u) \leq \frac{3}{2}(1 + \Delta K_1)(C_j u, u).$$

Comparison with Remark 2.2 shows that Θ_j^* *provides a continuous transition from the diagonal preconditioner of the mass matrix to the multilevel nodal basis preconditioner of the stiffness matrix.*

REMARK 2.4. Using Theorem 2 of OSWALD [18] we can prove for the case of *quasi–uniform* triangulations that

$$\kappa\left(\hat{\Theta}_j \Lambda_j\right) = \mathcal{O}(1),$$

with constants *independent* of $\tau \geq 0$ *and* j.

2.2. EXTENSION TO HELMHOLTZ TERMS AND GENERAL CAUCHY BOUNDARY CONDITIONS

Until now we considered the bilinear form $a(\cdot,\cdot)$ to consist only of the principal part, i.e., $q \equiv 0$ and $\zeta \equiv 0$. In this section we get rid of this restriction.

Denote by

$$\Lambda_P = \frac{1}{1+\tau} I + \frac{\tau}{1+\tau} A_P$$

the operator belonging to the principal part $a_P(\cdot,\cdot)$ of $a(\cdot,\cdot)$, i.e., exactly the operator which we considered in Section 2.1.

First assume $q \not\equiv 0$ but $\zeta \equiv 0$, which means that a Helmholtz term is present, but the Cauchy boundary conditions are still natural boundary conditions.

We have to consider the two cases

Case I. $\mathrm{mes}(\Gamma_D) > 0$

Case II. $q_{\min} > 0$,

either of which makes the form $a(\cdot,\cdot)$ $H^1_D(\Omega)$–elliptic by Lemma 1.1.

Case I and II will give two different versions of preconditioning, which will be discussed next. For the discussion of Case II it is essential that we did not exclude $\mathrm{mes}(\Gamma_D) = 0$ in Section 2.1.

2.2.1. *Version I of a Helmholtz Preconditioner*

By Lemma 1.1.iii Case I implies the $H^1_D(\Omega)$–ellipticity of the form $a_P(\cdot,\cdot)$, i.e., there is a constant $c_{1P} > 0$ such that

$$a_P(u,u) \geq c_{1P}\|u\|_1^2$$

for $u \in H^1_D(\Omega)$. Thus we can estimate for $u \in H^1_D(\Omega)$

$$a_P(u,u) \leq a(u,u) \leq a_P(u,u) + q_{\max}\|u\|_0^2 \leq \left(1 + \frac{q_{\max}}{c_{1P}}\right) a_P(u,u).$$

This inequalities yield

$$((\Lambda_P)_j u, u) \leq (\Lambda_j u, u) \leq \left(1 + \frac{q_{\max}}{c_{1P}}\right) ((\Lambda_P)_j u, u)$$

for $u \in \mathcal{S}_j$. Hence using the preconditioner $\hat{\Theta}_j$ of Section 2.1 we obtain

$$\kappa\left(\hat{\Theta}_j \Lambda_j\right) \leq \left(1 + \frac{q_{\max}}{c_{1P}}\right) \kappa\left(\hat{\Theta}_j (\Lambda_P)_j\right),$$

which has the desired form.

REMARK 2.5. For $\Gamma_D = \partial\Omega$, $\Gamma_C = \emptyset$ we can estimate by Lemma 1.1.i as follows: For $u \in H^1_0(\Omega)$ we obtain

$$a(u,u) \leq \left(1 + \frac{q_{\max} d_\Omega^2}{2\delta}\right) a_P(u,u),$$

which yields

$$\kappa\left(\hat{\Theta}_j \Lambda_j\right) \leq \left(1 + \frac{q_{\max} d_\Omega^2}{2\delta}\right) \kappa\left(\hat{\Theta}_j (\Lambda_P)_j\right).$$

2.2.2. *Version II of a Helmholtz Preconditioner*

In Case II we proceed as follows: Defining for some

$$0 < q_{\min} \leq \bar{q} \leq q_{\max}$$

the operator

$$\tilde{\Lambda} = \frac{1+\bar{q}\tau}{1+\tau} I + \frac{\tau}{1+\tau} A_P,$$

yields the estimate

$$\frac{q_{\min}}{\bar{q}}(\tilde{\Lambda}_j u, u) \leq (\Lambda_j u, u) \leq \frac{q_{\max}}{\bar{q}}(\tilde{\Lambda}_j u, u) \tag{2.15}$$

for $u \in \mathcal{S}_j$. Putting

$$\tilde{\lambda}_k = \frac{1+\tau(\bar{q}+4^k)}{1+\tau}$$

and replacing the λ_k of (2.4) in the derivation of $\hat{\Theta}_j$ by this $\tilde{\lambda}_k$ we obtain a preconditioner $\tilde{\Theta}_j$ of $\tilde{\Lambda}_j$ with

$$\kappa\left(\tilde{\Theta}_j \tilde{\Lambda}_j\right) \leq \kappa\left(\hat{\Theta}_j (\Lambda_P)_j\right)$$

since we never used in Section 2.1 the specific form of the λ_k, but only the fact that

$$\lambda_{k+1} \geq \lambda_k > 0.$$

By the above inequality we can estimate

$$\kappa\left(\tilde{\Theta}_j \Lambda_j\right) \leq \frac{q_{\max}}{q_{\min}} \kappa\left(\hat{\Theta}_j (\Lambda_P)_j\right),$$

which distinguishes $\tilde{\Theta}_j$ as preconditioner of the required form. The actual choice of $\bar{q}$ should be made in order to gauge the estimate (2.15) as $q_{\min}/\bar{q} = \bar{q}/q_{\max}$, i.e., $\bar{q} = \sqrt{q_{\min} q_{\max}}$, the geometric mean of the two bounds.

2.2.3. *A Case of Doubt: Case I and Case II*

If both cases, I and II, are present the question arises which version should be taken? Algorithmically both versions are quite related since Version I can be interpreted as the case $\bar{q} = 0$ of Version II, which gives $\tilde{\Theta}_j = \hat{\Theta}_j$. So, what value of $\bar{q}$ should be taken?

For $\Gamma_C = \emptyset$ we can answer this question definitely by means of Remark 2.5: Version II should be taken if

$$q_{\min} > \frac{2\delta}{d_\Omega^2}. \tag{2.16}$$

An example of such a decision will be given in Section 4.3.2.

2.2.4. *The Case of General Cauchy Boundary Conditions*

Now we consider $\zeta \not\equiv 0$ and $\Gamma_C \neq \emptyset$, assuming one of the above cases: I or II.

Hence we have

$$a(u,v) = a_H(u,v) + \int_{\Gamma_C} \zeta\, uv\, d\sigma$$

for all $u, v \in H^1_D(\Omega)$. The — due to Case I or II $H^1_D(\Omega)$-elliptic — form $a_H(\cdot,\cdot)$ with

$$a_H(u,u) \geq c_{1H} \|u\|_1^2$$

for $u \in H^1_D(\Omega)$ induces an operator Λ_H for which we know by Sections 2.1 and 2.2.1/2.2.2 a preconditioner $(\Theta_H)_j$. Since we may estimate for $u \in H^1_D(\Omega)$ by the continuity of the trace operator $H^1(\Omega) \to H^{1/2}(\partial\Omega)$

$$\begin{aligned} a_H(u,u) &\leq a(u,u) = a_H(u,u) + \int_{\Gamma_C} \zeta\, u^2\, d\sigma \\ &\leq a_H(u,u) + \zeta_{\max} \|u\|^2_{H^{1/2}(\partial\Omega)} \\ &\leq a_H(u,u) + \zeta_{\max} K_{\text{trace}} \|u\|_1^2 \\ &\leq \left(1 + \frac{\zeta_{\max} K_{\text{trace}}}{c_{1H}}\right) a_H(u,u), \end{aligned}$$

we obtain the estimate

$$\kappa\left((\Theta_H)_j \Lambda_j\right) \leq \left(1 + \frac{\zeta_{\max} K_{\text{trace}}}{c_{1H}}\right) \kappa\left((\Theta_H)_j (\Lambda_H)_j\right).$$

This tells us that $(\Theta_H)_j$ should be an as good preconditioner as in the case $\zeta \equiv 0$.

2.3. A Preconditioner for Piecewise Quadratic Elements

As introduced in Section 1.6 we consider the space of piecewise quadratic elements $\mathcal{S}_Q$ on the triangulation $\mathcal{T}_j$. The hierarchical splitting

$$\mathcal{S}_Q = \mathcal{S}_j \oplus \mathcal{V}_Q$$

shows that there is a *unique* decomposition of each $u \in \mathcal{S}_Q$ into

$$u = u_L + u_Q,$$

where $u_L \in \mathcal{S}_j$ and $u_Q \in \mathcal{V}_Q$. Based on that decomposition we introduce a partly discrete norm corresponding to an inner product on $\mathcal{S}_Q$: For $u \in \mathcal{S}_Q$ we define

$$\|u\|^2_{Q;\Lambda} = \|u_L\|^2_\Lambda + \sum_{\psi \in \Gamma_Q} (\Lambda_Q \psi, \psi)\, |u_Q(x_\psi)|^2. \tag{2.17}$$

This norm is in fact equivalent to the energy norm $\|\cdot\|_\Lambda$ on $\mathcal{S}_Q$.

LEMMA 2.7. (BORNEMANN [8, Corollary 6.4]). *There exists a positive constant γ_0 such that*

$$\gamma_0 \|u\|^2_{Q;\Lambda} \le \|u\|^2_\Lambda \le 4\|u\|^2_{Q;\Lambda},$$

for $u \in \mathcal{S}_Q$. The constant γ_0 depends only on the geometry of the initial triangulation $\mathcal{T}_0$.

Due to the Theorem of Fréchet–Riesz there is a symmetric positive definite operator $H_\Lambda : \mathcal{S}_Q \to \mathcal{S}_Q$ such that

$$(H_\Lambda u, v) = (\Lambda_j u_L, v_L) + \sum_{\psi \in \Gamma_Q} (\Lambda_Q \psi, \psi) u_Q(x_\psi) v_Q(x_\psi),$$

for all $u, v \in \mathcal{S}_Q$, which implies that

$$(H_\Lambda u, u) = \|u\|^2_{Q;\Lambda}.$$

LEMMA 2.8. (BORNEMANN [8, Lemma 6.8]). *The inverse of H_Λ is given by*

$$H_\Lambda^{-1} = \Theta_Q,$$

where

$$\Theta_Q u = \Lambda_j^{-1} \pi_j u + \sum_{\psi \in \Gamma_Q} \frac{(u, \psi)}{(\Lambda_Q \psi, \psi)} \psi, \tag{2.18}$$

for all $u \in \mathcal{S}_Q$.

Thus the operator Θ_Q is a preconditioner of Λ_Q. For ease of representation we restrict ourselves for the rest of this section to the case of $q \equiv 0$, $\zeta \equiv 0$, which was discussed in Section 2.1. The extension to the general case is obvious by means of Section 2.2.

In view of Section 2.1.2 the representation (2.18) suggests to use the operator $\hat{\Theta}_Q : \mathcal{S}_Q \to \mathcal{S}_Q$, defined by

$$\hat{\Theta}_Q u = \hat{\Theta}_j \pi_j u + \sum_{\psi \in \Gamma_Q} \frac{(u, \psi)}{(\Lambda_Q \psi, \psi)} \psi,$$

as computationally available preconditioner. Note that the operator $\hat{\Theta}_j$ can be defined *on all* $L^2(\Omega)$ and that

$$\hat{\Theta}_Q u = \hat{\Theta}_j u + \sum_{\psi \in \Gamma_Q} \frac{(u,\psi)}{(\Lambda_Q \psi, \psi)} \psi \tag{2.19}$$

due to $\pi_0 = \pi_0 \pi_j$ and $\hat{\pi}_k = \hat{\pi}_k \pi_j$ for $k \leq j$.

Therefore Theorem 2.1 and Lemma 2.7 immediately yield the following result.

THEOREM 2.2. *For all functions* $u \in \mathcal{S}_Q$ *the following inequalities hold*

$$\frac{\mu_0^Q}{j+1}(\Lambda_Q^{-1}u, u) \leq (\hat{\Theta}_Q u, u) \leq \mu_1^Q (j+1)(\Lambda_Q^{-1}u, u)$$

where $\mu_0^Q = \dfrac{\gamma_0 \delta K_0}{2\Delta}$ *and* $\mu_1^Q = 8\Delta K_1(1 + \Delta K_1)$. *The constants* δ, Δ *as introduced in assumption 4 of Section 1.1 describe the coefficient matrix* a_{ik} *of the elliptic operator* $A(x, \partial)$, *whereas the constants* γ_0, K_0, K_1 *depend only on the geometry of the initial triangulation* $\mathcal{T}_0$ *and they are independent of the maximal depth* j *of the final triangulation and of the parameter* τ.

2.4. Error Estimation

In this section we explain our concept of deriving error estimates for the elliptic subproblems. It clearly splits into two independent parts. First properties of the Galerkin approximation play a prominent role, while in the second part only finite dimensional linear problems are involved, where preconditioning very naturally comes into play. We will see that a good preconditioner of the linear systems provides us with a good error estimation.

2.4.1. Deviation Estimates Imply Error Estimates

Let $u_Q \in \mathcal{S}_Q$ be the solution of

$$\Lambda_Q u_Q = f_Q$$

and $u_j \in \mathcal{S}_j$ be the solution of

$$\Lambda_j u_j = f_j.$$

Let $u \in H_D^1(\Omega)$ be the solution of problem (1.2).

A detailed discussion in [8, Section 6.4.3] shows that it is reasonable to *assume* the existence of a $\beta \in]0,1[$ such that

$$\|u - u_Q\|_A \leq \beta \|u - u_j\|_A. \tag{2.20}$$

If the initial triangulation $\mathcal{T}_0$ reflects enough structure of the problem, such that the adaptive refinement process produces *adequate* triangulations, the above inequality should hold at least for $j \geq j_0$ with some j_0.

THEOREM 2.3. *Whenever the above assumption holds, we obtain for any* $\hat{u} \in \mathcal{S}_j$ *the error estimate*

$$\|u_Q - \hat{u}\|_A \leq \|u - \hat{u}\|_A \leq \gamma \|u_Q - \hat{u}\|_A,$$

where

$$\gamma = \sqrt{1 + \frac{1}{1 - \beta^2}}.$$

Proof. Orthogonality with respect to the inner product $a_\tau(\cdot,\cdot)$ yields

$$\|u - \hat{u}\|_A^2 = \|u - u_Q\|_A^2 + \|u_Q - \hat{u}\|_A^2, \tag{*}$$

since $\hat{u} \in \mathcal{S}_Q$. Thus the left inequality is proven. Similarly we get

$$\|u - u_j\|_A^2 \leq \|u - \hat{u}\|_A^2 = \|u - u_j\|_A^2 + \|u_j - \hat{u}\|_A^2. \tag{**}$$

Thus by (*) and the assumption (2.20)

$$\|u - u_j\|_A^2 \leq \frac{1}{1 - \beta^2} \|u_Q - \hat{u}\|_A^2.$$

Now (**) implies

$$\begin{aligned} \|u - \hat{u}\|_A^2 &\leq \|u - u_j\|_A^2 + \|u_Q - \hat{u}\|_A^2 \\ &\leq \left(1 + \frac{1}{1 - \beta^2}\right) \|u_Q - \hat{u}\|_A^2, \end{aligned}$$

which yields the right inequality. ∎

REMARK 2.6. The values of γ behave quite moderately: For example

$$\gamma \leq 2 \quad \text{for} \quad \beta \leq \frac{1}{3}\sqrt{6} \doteq 0.816.$$

Thus the energy norm of the *quadratic-deviation* $u_Q - \hat{u}$ is a good estimator for the energy norm of the error $u - \hat{u}$. However, we do not actually want to compute u_Q. Here preconditioning comes into play.

2.4.2. The Linear and the Quadratic Deviation Estimate

Consider the *linear residual*

$$r_j = f_j - \Lambda_j \hat{u}$$

and the *quadratic residual*

$$r_Q = f_Q - \Lambda_Q \hat{u}.$$

Because of the structure of Galerkin approximations and $\mathcal{S}_j \subset \mathcal{S}_Q$ we gain

$$r_j = \pi_j r_Q. \tag{2.21}$$

The linear–deviation estimate will be $\|r_j\|_{\hat{\Theta}_j}$, where $\|r_j\|^2_{\hat{\Theta}_j} = (\hat{\Theta}_j r_j, r_j)$ and the quadratic–deviation estimate will be $\|r_Q\|_{\hat{\Theta}_Q}$, where

$$\|r_Q\|^2_{\hat{\Theta}_Q} = (\hat{\Theta}_Q r_Q, r_Q) = (\hat{\Theta}_j \pi_j r_Q, \pi_j r_Q) + \sum_{\psi \in \Gamma_Q} \frac{(r_Q, \psi)^2}{(\Lambda_Q \psi, \psi)}.$$

By the projection property (2.21) we gain

$$\|r_Q\|^2_{\hat{\Theta}_Q} = \|r_j\|^2_{\hat{\Theta}_j} + \sum_{\psi \in \Gamma_Q} \eta^2_\psi,$$

where we define

$$\eta_\psi = \frac{|(r_Q, \psi)|}{\|\psi\|_\Lambda}$$

for every $\psi \in \Gamma_Q$.

THEOREM 2.4. *For any $\hat{u} \in \mathcal{S}_j$ we have*

$$\frac{1}{(j+1)\hat{\mu}_1} \|r_j\|^2_{\hat{\Theta}_j} \le \|u_j - \hat{u}\|^2_\Lambda \le \frac{j+1}{\hat{\mu}_0} \|r_j\|^2_{\hat{\Theta}_j}$$

and

$$\frac{1}{(j+1)\mu^Q_1} \|r_Q\|^2_{\hat{\Theta}_Q} \le \|u_Q - \hat{u}\|^2_\Lambda \le \frac{j+1}{\mu^Q_0} \|r_Q\|^2_{\hat{\Theta}_Q}.$$

The constants $\hat{\mu}_0$, $\hat{\mu}_1$ are from Theorem 2.1 and μ^Q_0, μ^Q_1 are from Theorem 2.2. These constants are independent of τ and j.

Proof. We only show the inequalities for $\|u_Q - \hat{u}\|_\Lambda$. This norm looks in terms of the residual as

$$\|u_Q - \hat{u}\|_\Lambda^2 = (\Lambda_Q^{-1} r_Q, r_Q).$$

By Theorem 2.2 this can be estimated from above and below with the given constants by

$$(\hat{\Theta}_Q r_Q, r_Q) = \|r_Q\|_{\hat{\Theta}_Q}^2.$$

■

REMARK 2.7. We want to measure the *quality* of an estimate $[\epsilon]$ for a quantity ϵ. Suppose we can prove $[\epsilon] = \zeta\epsilon$ with $\zeta \in [\zeta_{\min}, \zeta_{\max}]$. Now we introduce the *quality indicator* as the ratio of upper and lower bound,

$$\kappa = \frac{\zeta_{\max}}{\zeta_{\min}}.$$

A good estimator is therefore characterized by $\kappa \approx 1$, since $\kappa = 1$ would mean that we have computed the size exactly — with the exception of *gauging*. The above theorem now states that in the case of our deviation estimates

$$\kappa = \sqrt{\kappa(\hat{\Theta}\Lambda)}.$$

Thus we gain the same number which governs the number of PCG–iterations necessary for diminishing the error by a given factor.

2.4.3. *Refinement–Strategy*

The values η_ψ may serve as *indicators* for an edge–oriented refinement–strategy, since there is a one-to-one correspondence between Γ_Q and the edges of $\mathcal{T}_j$ which does not belong to the Dirichlet boundary piece Γ_D. The indicators η_ψ are in fact exactly the same as in [12], in the notation of [12]

$$\eta_\psi = \left(D_{QQ}^{-1/2} r_Q\right)\Big|_{\text{edge containing } x_\psi}.$$

Now, an edge containing x_ψ is marked for local refinement if

$$\eta_\psi \geq \eta_{\text{thresh}}.$$

We favor for the computation of η_{thresh} a procedure due to [2]:

It uses a simple heuristic prediction scheme to forecast what may happen to η_ψ if the edge containing x_ψ is subdivided. This forecast relies on the assumption that *locally*

$$\eta_\psi = c_\psi h_\psi^{\lambda_\psi} \text{ as } h_\psi \to 0.$$

Here h_ψ denotes the length of the edge containing x_ψ. Suppose this edge was generated by subdividing an edge with local error η_ψ^{old}. A simple extrapolation yields

$$\eta_\psi^{\text{new}} = \frac{\eta_\psi^2}{\eta_\psi^{\text{old}}}$$

as prediction for the error after a new subdivision of the edge. Clearly now, we should — in order of equidistributing the error — refine only those edges which have an η_ψ–value above the *largest* predicted *new* η^{new}–value of the virtual next triangulation, hence

$$\eta_{\text{thresh}} = \max_\psi \eta_\psi^{\text{new}}.$$

To avoid a refinement of too many triangles when the estimated error is near the given elliptic tolerance eps, we actually take with

$$\text{cut} = \max_\psi \eta_\psi^{\text{new}}$$

the value

$$\eta_{\text{thresh}} = \max\left(\text{cut}, \frac{\text{eps}}{\epsilon}\sqrt{\eta_{\max}\,\text{cut}}\right),$$

where $\eta_{\max} = \max_\psi \eta_\psi$ and ϵ is the actually estimated error.

This procedure of computing η_{thresh} yields triangulations with far fewer nodal points than the procedure originally proposed in [12]. For detailed comparisons see [13].

3. Algorithmic Details

Here we discuss in detail the consequences of the results of Section 2 for the time–stepping algorithm of Sections II.1.3 and II.3.

However, we have so far constructed an error estimator and linear solver belonging to the $\|\cdot\|_\Lambda$–norm and *not to the* L^2*–norm*, at least for $\tau \gg 0$. This relies on the fact that there is no iterative linear solver, at least to the knowledge of the author, comparable to the CG–method, which reduces the error with respect to the L^2–inner product. Since $\|u\|_\Lambda \geq \|u\|_0$ we can use

the $\|\cdot\|_A$–norm as upper bound estimate for the L^2–norm and control the time–error nevertheless in the L^2–norm.

Usage of the $\|\cdot\|_A$–norm for the stationary elliptic subproblems instead of the L^2–norm has certain disadvantages with respect to the amount of work, but is the best we can do yet. The impact for the time–stepping procedure is discussed below.

3.1. The Amount of Work and Qualitative Study of the Order Control

The measures for the amount of work, which has been introduced in Section II.1.2, should be chosen as

$$A_j = \frac{j+3}{\chi(j-1)^2}, \qquad j = 2, 3, \ldots .$$

Here we made use of [17, Theorem 8], which states that n_j grows for a correct sequence of triangulations like eps^{-2}, where eps denotes the accuracy for the elliptic $\|\cdot\|_A$–norm error. Usage of the L^2–norm would give instead

$$A_j^{L^2} = \frac{j+3}{\chi(j-1)},$$

compare with the 1D result given in (II.3.9).

Because of the use of the stationary $\|\cdot\|_A$–norm we have to minimize instead of the term given in Section II.3.3.3 the term

$$\frac{1}{(1-\varrho)^2\sqrt{\varrho}}$$

in view of the discussion in Section II.3.3.3 and the amount of work given above. This gives

$$\tilde{\varrho}_2 = \frac{1}{5}$$

instead of $\varrho_2 = \frac{1}{3}$ as in (II.3.8).

As in Section II.3.3.4 for the 1D case we can study the order control mechanism in dependence of the imposed accuracy TOL. For the choice $\sigma = 0.9$ as in Section II.3.3.4 we get the dependence of the maximal suggested order from TOL as listed in Table I.

It should be noted that $k_{\max} = 1$ means that we compute solutions u^1, u^2 of order 1 and 2 resp. at any time step. *Thus the time–step control is available anyway.*

TABLE I.
MAXIMAL ORDER $k_{\max}$ IN DEPENDENCE OF THE IMPOSED ACCURACY TOL; $\|\cdot\|_\Lambda$–NORM STATIONARY

TOL	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
$k_{\max}$	1	1	1	2	3	6	7

RESULT 3.1. For tolerances $\mathrm{TOL} \geq 10^{-3}$ the order control mechanism chooses the lowest possible order, since order switches do not pay off in terms of efficiency. For these tolerances we can restrict ourselves to the computation of u^1, u^2 at each time step because the order control would never decide to compute u^3. This result is still true if we choose $\sigma = 0.99$.

Quite a different result is obtained for the usage of the L^2–norm, as shown in Table II.

TABLE II.
MAXIMAL ORDER $k_{\max}$ IN DEPENDENCE OF THE IMPOSED ACCURACY TOL; $\|\cdot\|_0$–NORM STATIONARY

TOL	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
$k_{\max}$	1	2	3	5	6	7	7

3.2. MISCELLANEOUS ALGORITHMIC DETAILS

3.2.1. Stop Criterion for the Linear Solver for the Time Error $\hat{\eta}_l$

Approximation of the time error function η_l on the triangulation $\mathcal{T}_j$ gives the perturbed function $\hat{\eta}_l \in \mathcal{S}_j$ as discussed in Section II.3.1. However, in the 2D case we compute $\hat{\eta}_l$ only with an iterative solver, which has to be stopped efficiently. If we choose the starting value

$$\hat{\eta}_{l,0} = 0,$$

which gives a residual r_0, the value

$$\|r_0\|_{\hat{\Theta}_j}$$

is a reasonable measure for the size of the error $\|\hat{\eta}_l\|_\Lambda = \hat{\epsilon}_l$, due to Theorem 2.4. In view of the control criterion (II.1.11(ii)) we iterate

$$\hat{\eta}_{l,1}, \ldots, \hat{\eta}_{l,\iota}$$

until the following *stop criterion* is fulfilled:

$$\|r_\iota\|_{\hat{\Theta}_j} \leq \frac{1}{10}\|r_0\|_{\hat{\Theta}_j}.$$

We have

$$\begin{aligned}[\theta_l] &= \|r_{\iota,Q}\|_{\hat{\Theta}_Q} \\ &= \|r_\iota\|_{\hat{\Theta}_j} + [\tilde{\theta}_l],\end{aligned}$$

where $[\theta_l]$ is the estimate of the spatial perturbation of the time error as introduced in (II.3.2) and $[\tilde{\theta}_l]$ is the estimate without the part due to the linear solver. In view of the stop criterion we replace (II.1.11(ii)) by the computationally available

$$[\tilde{\theta}_l] \leq \frac{3}{20}\|\hat{\eta}_{l,\iota}\|_0.$$

3.2.2. Stabilization of the L^2–Projection

In the case of an inconsistent start function u_0 the L^2–projection into $\mathcal{S}_j$ may be unstable, hence our whole stationary problem becomes unstable for τ small. Here we replace the L^2–inner product by the discrete L^2–inner product $(\cdot,\cdot)_k$ of Section 2.1.2 and assemble the mass matrix, stiffness matrix and the right–hand side with respect to that discrete inner product, which means the usage of the corresponding quadrature rule. Now the case $\tau = 0$ reduces to a simple stable interpolation. Moreover the local order of approximation is not touched for $\tau > 0$ since the quadrature rule is exact for piecewise linear functions on $\mathcal{T}_j$, cf. Theorem 4.1.6 of CIARLET [10].

Because of the construction of our preconditioner no property of it is lost by usage of this quadrature rule.

3.2.3. The Direct Solver

The iterative solution process described in Section 1.5 as well as the preconditioner itself requires a direct solution on the coarsest triangulation $\mathcal{T}_0$. Due to the complex geometries in applications, e.g., the one given in Section 4, the number of nodal points in $\mathcal{T}_0$ may be quite big, about 200–1000. In 3D the number would be even more big. Thus a sophisticated direct solver is indispensable.

We choose a Cholesky decomposition solver, which exploits the envelope structure of its L–factor. In order to make this nearly optimal, it is necessary

to order the points in such a way, that the envelope is nearly minimal. The so–called *reverse Cuthill–McKee ordering* accomplishes that in a rather efficient way, cf. e.g., the textbook of GEORGE/LIU [15].

For a number of nodal points in $\mathcal{T}_0$ below about 1000 this choice turns out to be superior to the use of a fully sparse solver together with the nested dissection ordering.

4. A REAL LIFE APPLICATION: HYPERTHERMIA

Numerous numerical experiments on model problems have proved the algorithm to be very robust, reliable and efficient. Some of these computations including model problems of [1, 14] can be found in [8, Section 8.2]. These examples significantly back the developed theory. However, model problems tend to isolate the different kind of difficulties or to test for difficulties other than those arising in real applications. Thus we decided to include in this paper a real life application rather than a model problem.

4.1. HYPERTHERMIA

In order to prove the applicability of our method to real life problems, which *combine* the difficulties of complex problem geometry, discontinuous coefficients etc., we will present the solution of the so–called *bioheat–transfer equation* (BHT equation) in the framework of *hyperthermia.*

Hyperthermia, i.e., the heating of tissue to temperatures approximately above 42 °C, is a recently developed clinical method for cancer therapy. It allows *in combination with radiotherapy* an improvement of the local control of the tumor. The deep heating of tissue is obtained by an electric field (E–field), which is generated by the radio waves of four antenna pairs. Their parameters (frequency 60 – 120 MHz, phase and amplitude) have to be selected appropriately. It is essential to solve effectively and robust the BHT equation, which models the temperature distribution for a given E–field.

We will show for a set of real life data, that our method would in principle allow such an on line computation on a workstation. We will present computations for 2D cross sections generated by computer tomography (CT) data.

4.2. THE BIOHEAT–TRANSFER EQUATION

The BHT equation was developed 1948 by PENNES [20] to model the heat transport in live tissue. A characteristic is a local, isotropic blood flow term.

In the case of hyperthermia the BHT equation reads as (cf. [23]):

$$\text{i)} \qquad \varrho(x)c(x)\frac{\partial T(t,x)}{\partial t} = \operatorname{div}\left(\kappa(x)\ \operatorname{grad} T(t,x)\right) - \\ - \varrho_b c_b \varrho(x)\omega(x)\left(T(t,x) - T_a\right) + \\ + \frac{1}{2}\sigma(x)|E(t,x)|^2$$

(BHT)

$$\text{where } t \in [0, T_{\text{fin}}], \quad x \in \Omega$$

$$\text{ii)} \quad -\kappa(x)\frac{\partial}{\partial n}T(t,x)\Big|_{x\in\partial\Omega} = h\left(T(t,x) - T_{\text{bolus}}\right)\Big|_{x\in\partial\Omega}$$

$$\text{iii)} \qquad T(0,x) = T_a.$$

Here we denote, using SI–units:

$\varrho(x)$, ϱ_b	$\left[\frac{\text{kg}}{\text{m}^3}\right]$	:	density of tissue, resp. blood.
$c(x)$, c_b	$\left[\frac{\text{J}}{\text{kg °C}}\right]$	:	specific heat of tissue, resp. blood.
$T(x,t)$, T_a, T_{bolus}	[°C]	:	temperature of tissue, blood, bolus.
$\kappa(x)$	$\left[\frac{\text{W}}{\text{m °C}}\right]$	:	thermal conductivity of tissue.
$\omega(x)$	$\left[\frac{\text{m}^3}{\text{kg s}}\right]$	:	blood flow in tissue (perfusion).
$\sigma(x)$	$\left[\frac{1}{\text{m}\,\Omega}\right]$	:	electric conductivity of tissue.
$\lvert E(t,x)\rvert$	$\left[\frac{\text{V}}{\text{m}}\right]$	:	magnitude of the electric field.
h	$\left[\frac{\text{W}}{\text{m}^2\ \text{°C}}\right]$	:	heat flow at the boundary (body surface).

Figure 1 shows the CT–data of a rectum malignancy of a patient, who has been treated at the Klinikum Rudolf Virchow, Freie Universität Berlin.

The data of the involved tissues are given in Table III. The heat flow h for the Cauchy boundary condition (BHT.ii) is assumed to be

$$h = 45\frac{\text{W}}{\text{m}^2\ \text{°C}}$$

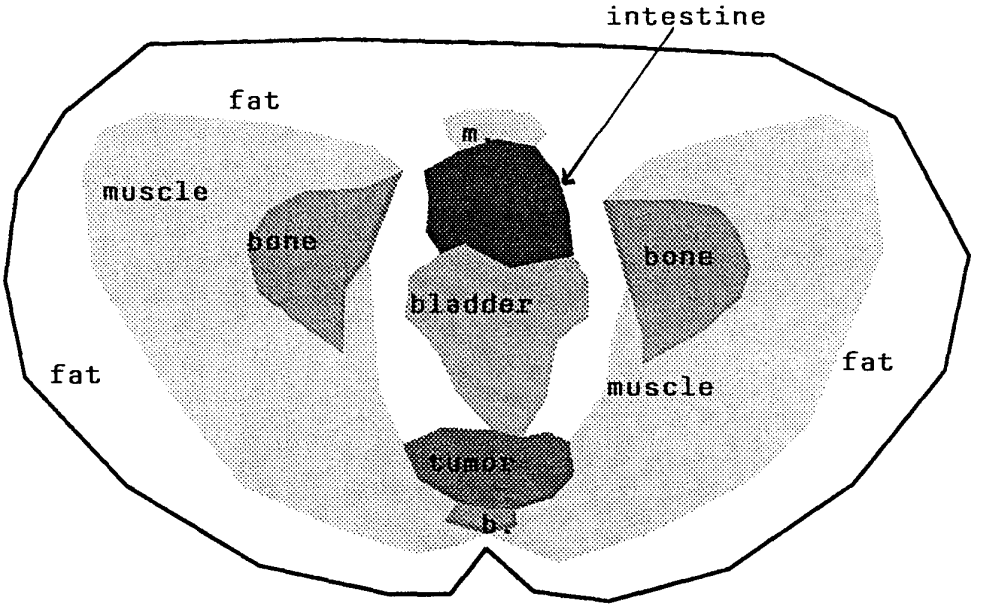


FIG. 1. CT-data of a rectum malignancy.

TABLE III.
DATA OF TISSUES

tissue	$\varrho[10^3\,\mathrm{kg/m^3}]$	$c[10^3\,\mathrm{J/kg\,°C}]$	$\kappa[\mathrm{W/m\,°C}]$	$\omega[\mathrm{ml/100g\ per\ min}]$	$\sigma[1/\mathrm{m\,\Omega}]$
blood	1.0	3.72	—	—	—
fat	0.9	2.36	0.210	5	0.21
muscle	1.0	3.72	0.642	20	0.80
bone	1.6	1.41	0.436	5	0.02
intestine	1.0	3.81	0.550	30	0.60
bladder	1.0	3.98	0.561	30	0.20
tumor	1.0	3.72	0.642	20	0.80

and the bolus temperature (i.e. the temperature of the water bolus, which is cooling the patient) is assumed to be

$$T_{\text{bolus}} = 25\ ^\circ\text{C}.$$

As temperature of the arterial blood we take

$$T_a = 37\ ^\circ\text{C}.$$

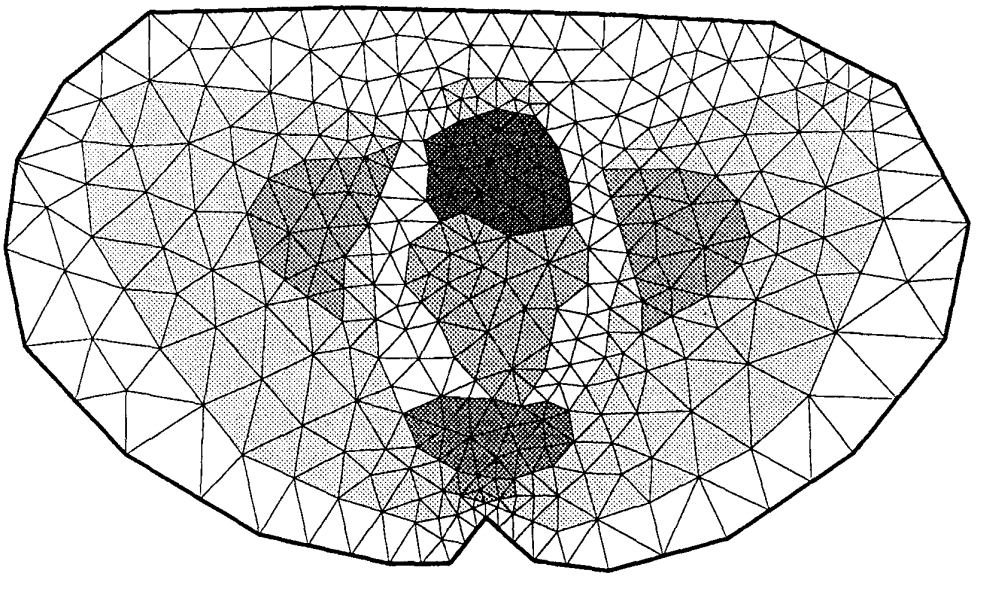


FIG. 2. Initial triangulation $\mathcal{T}_0$ of the CT cross section.

The initial triangulation $\mathcal{T}_0$ (Figure 2) (351 nodal points, 642 triangles) of the CT cross section was created by TRIGEN from the PLTMG–package of BANK [3]. On this triangulation an optimal E–field was computed by the second author of [25] which we used for our example.

4.3. COMPUTATIONAL DETAILS FOR THE BHT EQUATION

4.3.1. Time Discretization and Preconditioning

The abstract setting of the BHT equation is

$$\phi(x)\frac{\partial u(t,x)}{\partial t} + A(x,\partial)u(t,x) = f(x). \tag{4.1}$$

We briefly discuss the effect of $\phi \not\equiv 1$. Since assumptions 2 & 3 of Section 1.1 are fulfilled the semigroup setting of (4.1) is given by

$$\Phi u' + Au = f \tag{4.2}$$

with the bounded positive selfadjoint operator

$$\begin{aligned} \Phi : L^2(\Omega) &\to L^2(\Omega) \\ u &\mapsto \phi u. \end{aligned}$$

Since Φ^{-1} is bounded positive selfadjoint as well we obtain the equivalence of (4.2) and

$$\hat{u}' + \Phi^{-1/2} A \Phi^{-1/2} \hat{u} = \Phi^{-1/2} f,$$

with the transformed $\hat{u} = \Phi^{1/2} u$. Now the operator $\Phi^{-1/2} A \Phi^{-1/2}$ has the same properties as A. Note that a likely transformation by dividing equation (4.1) through $\sqrt{\phi}$ is impossible, since the principal part of $A(x, \partial)$ would loose its divergence form. Taking the time discretization (II.2.11) for the transformed $\hat{u}$ we get after back-transformation:

$$\begin{aligned} \text{i)} \quad u^1 &= (\Phi + \tau A)^{-1} (\Phi u_0 + \tau f), \\ \text{ii)} \quad \eta_1 &= \frac{1}{2} \tau A (\Phi + \tau A)^{-2} (u^1 - u^0), \\ \text{iii)} \quad u^2 &= u^1 + \eta_1. \end{aligned}$$

All results are valid as if $\phi \equiv 1$.

We now have to find a preconditioner for the operator

$$\Lambda_\Phi = \frac{1}{1+\tau} \Phi + \frac{\tau}{1+\tau} A.$$

We estimate with Λ of Section 2.2.4

$$\min(1, \phi_{\min}) (\Lambda u, u) \le (\Lambda_\Phi u, u) \le \max(1, \phi_{\max}) (\Lambda u, u)$$

for $u \in \dot{H}^2$. Thus we can take the same preconditioner $(\Theta_H)_j$ for $(\Lambda_\Phi)_j$ as for Λ_j. The conditioner number will grow at most like

$$\kappa((\Theta_H)_j (\Lambda_\Phi)_j) \le \frac{\max(1, \phi_{\max})}{\min(1, \phi_{\min})} \kappa((\Theta_H)_j \Lambda_j). \tag{4.3}$$

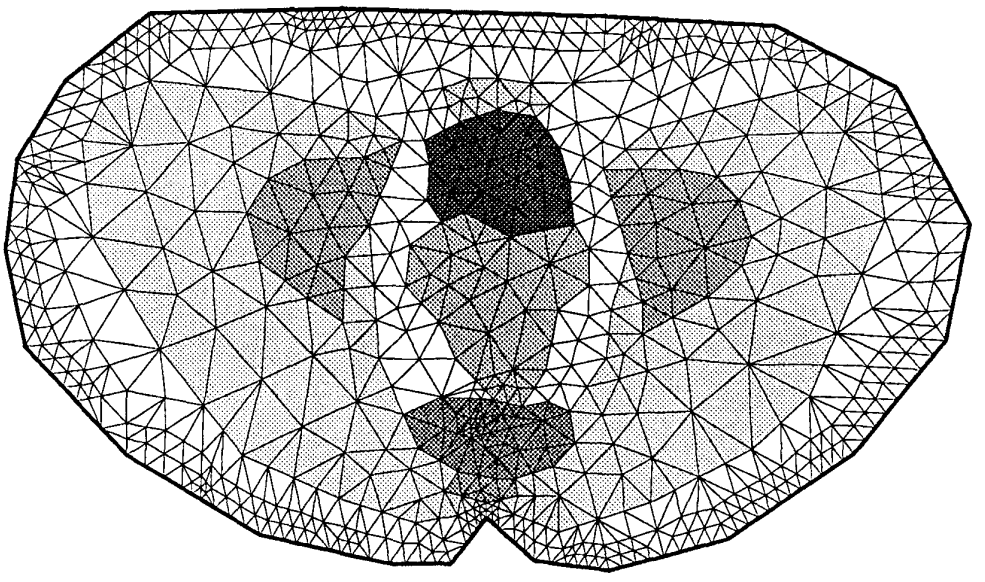


FIG. 3. Triangulation at time step 10, t = 13 min 11 s.

4.3.2. Time Scaling and Choice of $\bar{q}$

The grow factor of the condition number in (4.3) has for the above example the value

$$\frac{\max(1, 3.98 \cdot 10^6)}{\min(1, 2.12 \cdot 10^6)} = 3.98 \cdot 10^6,$$

which is *not feasible*. The reason for this big value is the comparison with the value 1 in the denominator. Now we make use of the possibility of a *time-scaling*: Introducing $t = t_{\text{scal}} \cdot \hat{t}$ we get

$$\varrho(x)c(x)\frac{\partial T(t,x)}{\partial t} = \frac{\varrho(x)c(x)}{t_{\text{scal}}}\frac{\partial T(\hat{t},x)}{\partial \hat{t}}.$$

In our example we choose

$$t_{\text{scal}} = 3.0 \cdot 10^6$$

and obtain the grow factor

$$\frac{\max(1, 1.33)}{\min(1, 0.71)} = 1.88.$$

Physically this scaling means that we take $3.0 \cdot 10^6$ s as time unit.

For the correct choice of $\bar{q}$, i.e., the version of the Helmholtz preconditioner of Section 2.2, we observe that exactly Case II of Section 2.2 is the case, because of $\Gamma_D = \emptyset$ and $q_{\min} > 0$. We thus have to use Version II of Section 2.2. In our example

$$q_{\min} = 2790$$

and

$$q_{\max} = 18600,$$

hence $\bar{q} = \sqrt{q_{\min} q_{\max}} = 7200$ is the correct value.

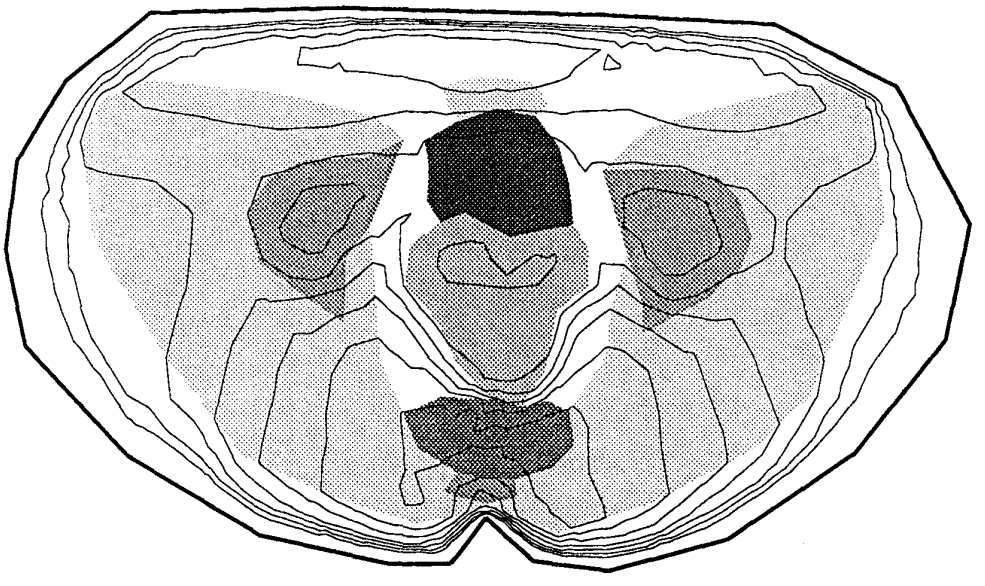


FIG. 4. Isothermals (37 - 43°C) at time step 10, t = 13 min 11 s.

REMARK 4.1. Even in the case of Dirichlet boundary conditions this choice of $\bar{q}$ would be preferable compared to the choice $\bar{q} = 0$. In this case $\Gamma_D = \partial\Omega$, $\Gamma_C = \emptyset$ would yield just the case of doubt mentioned in Section 2.2.3. Hence we would have to refer to the decision criterion (2.16) of Section 2.2.3: With

$$d_\Omega = 0.24[\mathrm{m}]$$

for the width of the vertical strip, i.e., the depth of the body, and

$$\delta = 0.21 \left[\frac{\mathrm{W}}{\mathrm{m}\ ^\circ\mathrm{C}} \right],$$

we would get

$$q_{\min} = 2790 \left[\frac{\mathrm{W}}{\mathrm{m}^3\ ^\circ\mathrm{C}}\right] \gg 7.29 \left[\frac{\mathrm{W}}{\mathrm{m}^3\ ^\circ\mathrm{C}}\right] = \frac{2\delta}{d_\Omega^2}.$$

Hence we would obtain a reduction of the condition number by a factor of $3.8 \cdot 10^2$ by using $\bar{q} = 7200$ instead of $\bar{q} = 0$. This reduction can be observed in numerical examples.

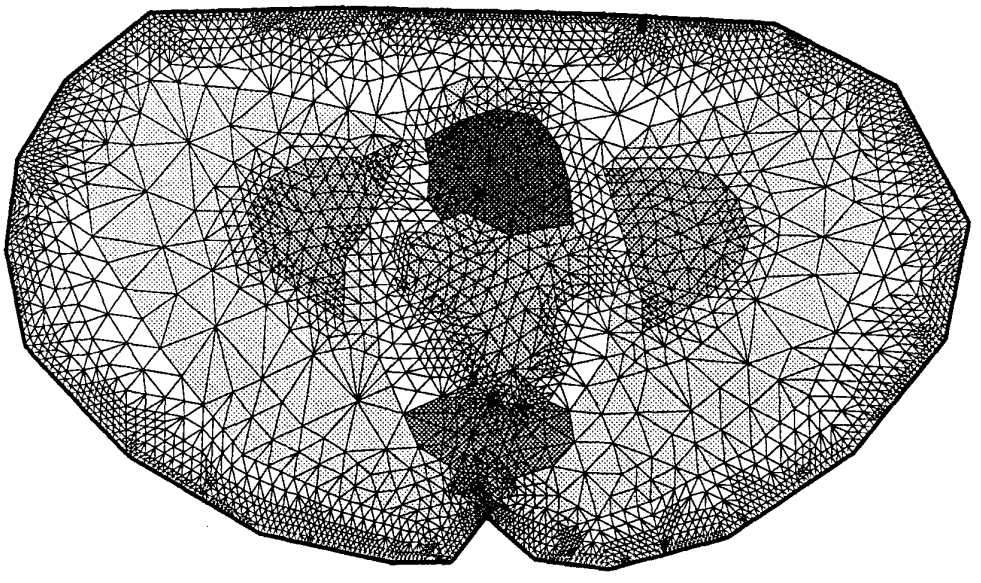


FIG. 5. Triangulation at time step 16, t = 58 min 41 s.

4.4. COMPUTATIONAL RESULTS

Here we present the computational results of the program KASTIO2 for the BHT equation with the above described data. We have chosen the accuracy

$$\mathrm{TOL} = 7.5 \cdot 10^{-2},$$

which corresponds to an accuracy of the temperature of $\pm 0.25\ ^\circ\mathrm{C}$, *assuming an equidistributed error.*

We did the computations until a treatment time of 1 h = 3600 s.

The triangulation computed at the problem time t = 13 min 11 s (time step 10) is show in Figure 3. This triangulation contains 724 points and

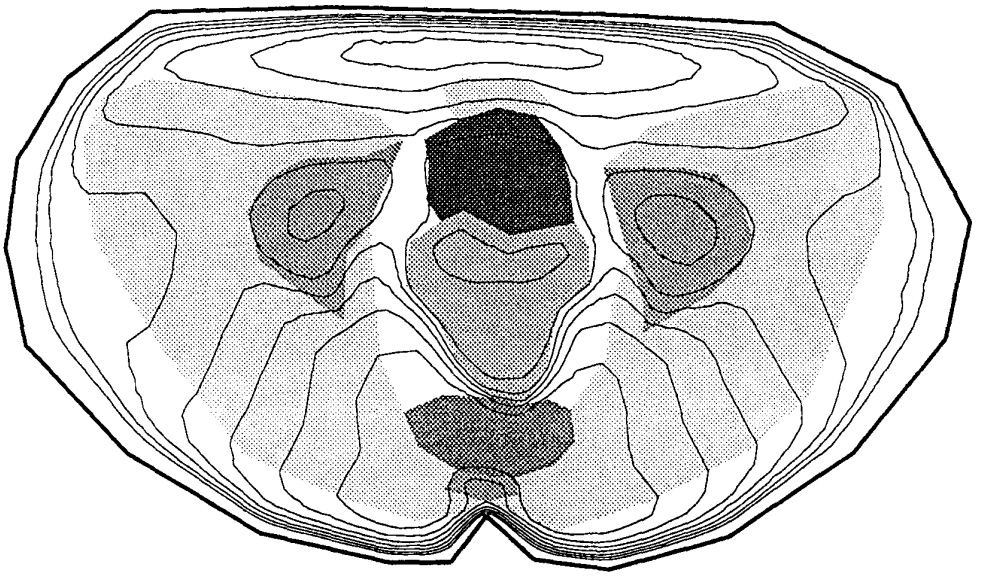


FIG. 6. Isothermals (37 - 43°C) at time step 16, t = 58 min 41 s.

1304 triangles. The refinement occurred mainly at the boundary where the steepest temperature differences can be found. The corresponding solution is show in Figure 4, where the isothermals are plotted for 37 - 43 °C in 1 °C steps.

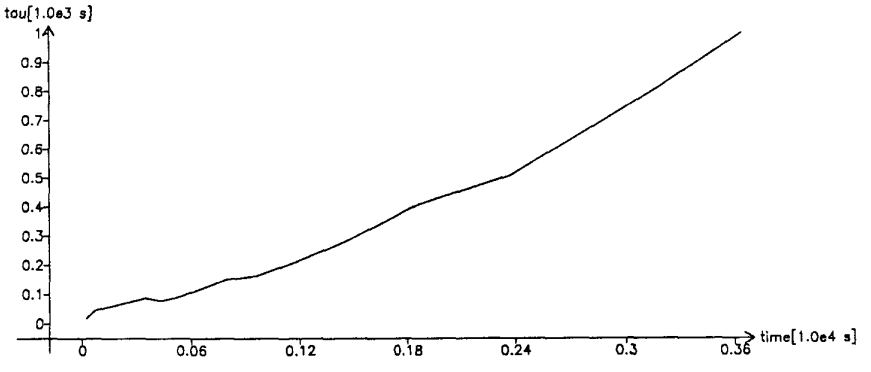


FIG. 7. Development of time step.

Figure 7 shows the development of the time step τ during the 16 time steps, Figure 8 the increase of the number of nodal points.

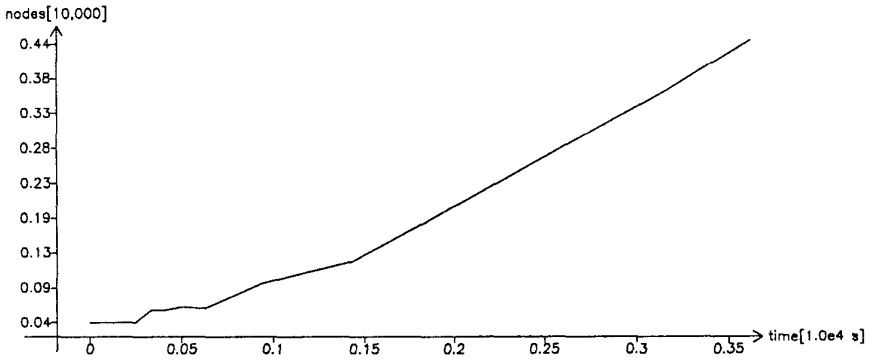


FIG. 8. Increase of number of nodal points.

The triangulation computed at the problem time $t = 58$ min 41 s (time step 16) is show in Figure 5. This triangulation contains 3688 points and 6922 triangles. The refinement occurred here also at the critical tissue boundaries like muscle/bone, fat/bladder and bladder/tumor. The corresponding solu-

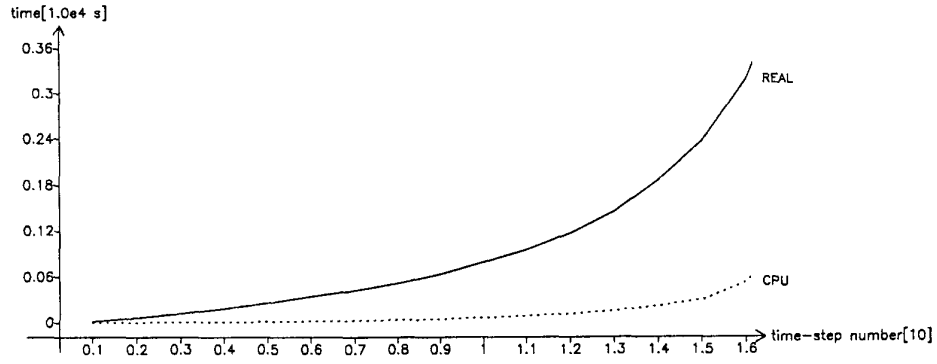


FIG. 9. Development of real treatment time (—) and cpu-time (···).

tion is show in Figure 6, where the isothermals are plotted for 37 – 43 °C in 1 °C steps. We observe that the region above 43 °C at the tumor has spread

out, now surrounding it nearly. Also the top fat region, which is a region of high electric field, did get an temperature increase of about 2 °C.

Figure 9 contains the treatment time and the cpu time versus the time step number, we observe an increasing gain of time to react: For 58 min 41 s treatment time the computation on the workstation (SPARC-station 1+) ended after 5 min 18.2 s.

The distribution of the computational time to the different tasks is shown in Table IV. It shows that the bulk of the computational time is spend in the local integrations for the mass matrix and the stiffness matrix. Of course, this would change if we would make use of the local constancy of the coefficients. However, more general problems would show up just the indicated behavior.

TABLE IV.
AMOUNT OF CPU TIME FOR THE COMPUTATIONAL TASKS

task	amount of CPU-time
direct solver on coarse grid	1.0%
PCG iterations for implicit Euler u^1	6.2 %
PCG iter. time error corrections η_1	4.8 %
refinement	2.1 %
local error indicators	4.3%
local integration	81.6%

ACKNOWLEDGMENTS

The author thanks his colleagues R. Kornhuber and M. Wulkow for a careful reading of a primary draft of this paper. Many thanks to J. Nadobny for making available the computer tomography data and for computing the E-field of the example of Section 4.

REFERENCES

[1] Adjerid, S., Flaherty, J.E.: *A Local Refinement Finite-Element Method for Two-Dimensional Parabolic Systems.* SIAM J. Sci. Stat. Comput. 9, 792–811 (1988).

[2] Babuška, I., Rheinboldt, W.C.: *Error Estimates for Adaptive Finite Element Computations.* SIAM J. Numer. Anal. 15, 736–754 (1978).

[3] Bank, R.E.: *PLTMG User's Guide.* Edition 5.0, Tech. Rep., Department of Mathematics, University of California at San Diego (1988).

[4] Bank, R.E.: *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations.* Philadelphia, SIAM (1990).

[5] Bank, R.E., Sherman, A.H., Weiser, A.: *Refinement Algorithms and Data Structures for Regular Local Mesh Refinement.* In: Scientific Computing (eds.: R. Stepleman et al.), Amsterdam: IMACS/North Holland, 3–17 (1983).

[6] Bornemann, F.A.: *An Adaptive Multilevel Approach to Parabolic Equations I. General Theory and 1D-Implementation.* IMPACT Comput. Sci. Engrg. 2, 279–317 (1990).

[7] Bornemann, F.A.: *An Adaptive Multilevel Approach to Parabolic Equations II. Variable-Order Time Discretization Based on a Multiplicative Error Correction.* IMPACT Comput. Sci. Engrg. 3, 93–122 (1991).

[8] Bornemann, F.A.: *An Adaptive Multilevel Approach to Parabolic Equations in Two Space Dimensions.* Technical Report TR 91-7, ZIB (1991).

[9] Bramble, J.H., Pasciak, J.E., Xu, J.: *Parallel Multilevel Preconditioners.* Math. Comp. 55, 1–22 (1990).

[10] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems.* North–Holland Publ. Comp., Amsterdam–New York–Oxford (1978).

[11] Deuflhard, P.: *Order and Stepsize Control in Extrapolation Methods.* Numer. Math. 41, 399–422 (1983).

[12] Deuflhard, P., Leinen, P., Yserentant, H.: *Concepts of an Adaptive Hierarchical Finite Element Code.* IMPACT Comput. Sci. Engrg. 1, 3–35 (1989).

[13] Erdmann, B., Roitzsch, R., Bornemann, F.A.: *KASKADE, Numerical Experiments.* Technical Report TR 91–1, ZIB (1991).

[14] Eriksson, K., Johnson, C.: *Adaptive Finite Element Method for Parabolic Problems I: A Linear Model Problem.* SIAM j. Numer. Anal. 28, 43–77 (1991).

[15] George, A., Liu, J.W.H.: *Computer Solution of Large Sparse Positive Definite Systems.* Prentice Hall, New Jersey (1981).

[16] Leinen, P.: *Ein schneller adaptiver Löser für elliptische Randwertprobleme auf Seriell- und Parallelrechnern.* Thesis, University of Dortmund (1990).

[17] Oswald, P.: *On Function Spaces Related to Finite Element Approximation Theory.* Z. Anal. Anwendungen 9, 43–64 (1990).

[18] Oswald, P.: *On Discrete Norm Estimates Related to Multilevel Preconditioners in the Finite Element Method.* In: Proc. Int. Conf. Constr. Theory of Functions Varna 1991 (to appear).

[19] Oswald, P.: *Remarks on Multilevel Preconditioners.* Manuscript June (1991).

[20] Pennes, H.H.: *Analysis of Tissue and Arterial Blood Temperatures in the Resting Human Forearm.* J. Appl. Physiol. 1, 93–122 (1948).

[21] Roitzsch, R.: *KASKADE User's Manual.* Technical Report TR 89–4, ZIB (1989).

[22] Roitzsch, R.: *KASKADE Programmer's Manual.* Technical Report TR 89–5, ZIB (1989).

[23] Seebaß, M.: *3D–Computersimulation der interstitiellen Mikrowellen–Hyperthermie von Hirntumoren.* Bericht Nr. CVR 1/90, Deutsches Krebsforschungszentrum Heidelberg (1990).

[24] Wathen, A.J.: *Realistic Eigenvalue Bounds for the Galerkin Mass Matrix.* IMA J. Numer. Anal. 7, 449–457 (1987).

[25] Wust, P., Nadobny, J., Felix, R., Deuflhard, P., Louis, A., John, W.: *Strategies for Optimized Application of Annular–Phased–Array Systems in Clinical Hyperthermia.* Int. J. Hyperthermia 7, 157–173 (1991).

[26] Xu, J.: *Theory of Multilevel Methods.* Report No. AM 48, Department of Mathematics, Pennsylvania State University (1989).

[27] Xu, J.: *Iterative Methods by Space Decomposition and Subspace Correction: A Unifying Approach.* Report No. AM 67, Department of Mathematics, Pennsylvania State University (1990).

[28] Yserentant, H.: *On the Multi–Level Splitting of Finite Element Spaces.* Numer. Math. 49, 379–412 (1986).

[29] Yserentant, H.: *Hierarchical Bases in the Numerical Solution of Parabolic Problems.* Large Scale Scientific Computing, Deuflhard, P., Engquist, B. (eds.), Birkhäuser, Boston, 22–37 (1987).

[30] Yserentant, H.: *Two Preconditioners Based on the Multi–Level Splitting of Finite Element Spaces.* Numer. Math. 58, 163–184 (1990).