Mathematical Modelling and Numerical Analysis M2AN, Vol. 33, N° 5, 1999, p. 895–922
Modélisation Mathématique et Analyse Numérique

ORIGINS, ANALYSIS, NUMERICAL ANALYSIS, AND NUMERICAL APPROXIMATION OF A FORWARD-BACKWARD PARABOLIC PROBLEM

A. Kadir Aziz[1], Donald A. French[1, 2], Soren Jensen[1] and R. Bruce Kellogg[3]

Abstract. We consider the analysis and numerical solution of a forward-backward boundary value problem. We provide some motivation, prove existence and uniqueness in a function class especially geared to the problem at hand, provide various energy estimates, prove *a priori* error estimates for the Galerkin method, and show the results of some numerical computations.

AMS Subject Classification. 65M60, 65N30, 65N15, 76D15, 76M10.

Received: July 28, 1997. Revised: June 30, 1998.

1. Introduction

We study a class of forward-backward heat equations in this report. Problems such as these arise in a remarkable variety of physical applications which we will describe in the next section. It seems that this problem-type has been avoided to some degree due to the nontrivial task of finding a proper formulation.

Let $\Omega \subset \mathbb{R}^2$ be a rectangle $(0, L) \times (0, H)$. Let $\sigma(x, y)$ be a smooth function ($\in C^1(\overline{\Omega})$ with $|\sigma|_y \in L_\infty(\Omega)$) in Ω such that $\sigma = 0$ defines a curve $\mathcal{C}$ which divides Ω into two parts, $\Omega_\pm$. It will minimally suffice that σ is piece-wise smooth with $|\sigma|_y \in L_\infty(\Omega)$ on each of finitely many segments, these being regular domains. We are concerned with the problem

$$\sigma u_y - u_{xx} + \lambda u = f \ \text{ on } \ \Omega, \tag{1.1}$$

$$u = 0 \ \text{ on } \ \partial\Omega \cap \{(x, H) : \sigma(x, H) < 0\} =: \Gamma_-^H, \tag{1.2}$$

$$u = 0 \ \text{ on } \ \partial\Omega \cap \{(x, 0) : \sigma(x, 0) > 0\} =: \Gamma_+^0, \ \text{ and} \tag{1.3}$$

$$u = 0 \ \text{ on } \ \partial\Omega \cap (\{(0, y), 0 < y < H\} \cup \{(L, y), 0 < y < H\} =: \Gamma_0. \tag{1.4}$$

Some of these boundary conditions become vacuous if the sets on which they are posed become empty (or of zero measure). We assume Γ_-^H and Γ_+^0 are either empty or consist of finitely many open intervals. Note that in a region where $\sigma > 0$, the above equation resembles a (forward) heat equation for which one expects to prescribe initial and lateral boundary values. Conversely, in a region where $\sigma < 0$, the equation becomes a (backward) heat equation for which one expects to pose terminal as well as lateral boundary values. One easily imagines

Keywords and phrases. Forward-backward, heat equation, degenerate parabolic problem, Brownian motion, electron and neutron scattering, separated flow boundary layers.

[1] Department of Mathematics, UMBC, Baltimore, MD 21250, USA. e-mail: aziz@math.umbc.edu

[2] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA. e-mail: french@math.uc.edu

[3] Department of Mathematics, University of Maryland, College Park, MD 20742, USA. e-mail: kellogg@ipst.umd.edu

© EDP Sciences, SMAI 1999

an interesting problem as to the mathematical structure of the solution globally as well as along the level set $\sigma^{-1}(\{0\})$ and its intersection with the boundary Γ.

Because of the unusual nature of these forward-backward problems, it is important to describe their origins. This is done briefly in two sections: Section 2 discusses the mathematical origin of the problem as a singular perturbation limit, and gives an indication of the variety of problems that can arise in this way. Section 3 indicates briefly a number of physical situations where equations with a forward-backward character arise.

The first main results of this paper consist in establishing existence and uniqueness of a weak solution to the problem in a certain class of functions. The essential result, in Theorem 4.2, is the precise identification of the space of solutions. This identification requires a trace theorem, Theorem 4.1., that generalizes a corresponding theorem in [3]. A consequence of knowing the space of solutions is that the uniqueness of the solution follows readily, as is shown in Theorem 4.2.

We then formulate in Section 5 a Galerkin method for the solution of the problem. Error bounds are given for both parts of the norm (Th. 5.1 and 5.2). In Section 6 we do some numerical computations with a method that uses second degree polynomials. We furnish some experiments on the accuracy confirming our results in Section 5 and examine some cases where the true solution is not known. Other numerical work on similar equations abound; we can mention [1, 2, 13, 14, 23, 24, 26–28, 46, 47]. Regarding this literature, we mention that [28] is concerned with problems that do not necessarily satisfy a coercivity condition. Some careful eigenvalue estimates are used in the analysis, and several bilinear forms are used; a form that is the same one as used here, and a weighted form with the weight chosen to enhance the coercivity. Some numerical results are given. Among the other numerical methods proposed, we mention [13,14]. These deal with the special case $\sigma(x, y) = x$. Both [13, 14] have bilinear forms that involve integrations one slab at a time. In [14] both the solution and test functions are discontinuous in y, and while in [13] the solution functions are continuous in y. In very simple cases [14] is like the backward Euler method while [13] is like Crank-Nicolson. Finally, various members of the fluid dynamics community have produced numerical solutions of the reversed flow boundary layer problem closely associated herewith, for example [8, 40].

Our results are carried out in two dimensions although it is reasonable to suppose they can be extended to higher dimensions.

2. Problems

Problems of the form (1.1–1.4) occur in a variety of applications (to which we shall return in the next section) and in addition have an independent mathematical interest. One source for the problem (1.1–1.4) is the singular perturbation limit as $\epsilon \to 0$ of the elliptic boundary value problem

$$\begin{aligned} -u^\epsilon_{xx} - \epsilon u^\epsilon_{yy} + \sigma u^\epsilon_y + \lambda u^\epsilon &= f \quad \text{in } \Omega \\ u^\epsilon &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

If $\sigma_y < 2\lambda$, a simple integration by parts and use of the Lax-Milgram shows that (2.1) has a solution. It is shown in [35] that u^ϵ converges weakly to the solution u of (1.1–1.4) as $\epsilon \to 0$. (A singular perturbation result in a special case is also given in [12].) Different choices of the function σ give rise to interesting examples of (1.1–1.4). We cite some of these. If $\sigma(x, y) = x - 1/2$, we have (with a change of variable) the forward-backward parabolic equation described in earlier papers. In this example, depicted in Figure 1, we have a forward parabolic equation in the right half rectangle, and a backward heat equation in the left half rectangle. This problem arises in the theory of stochastic processes (see Sect. 3.1), in a simple model of neutron scattering (see Sect. 3.5), in the modeling of counter-current separators (see [18]), and also in some astronomical problems (see Sect. 3.3). It is, to our knowledge, the first example of (2.1) that has been studied. Some mathematical properties of the solution of (2.1) in this case are given in [12, 35], or [20]. A second choice of σ is given by the formula $\sigma(x, y) = (x - 1)^2 + (y - 1/2)^2 - 1/16$, see Figure 2. Here we have a situation reminiscent of fluid dynamics and the use of the "parabolized Navier-Stokes" (PNS) equations as a simplified version of the Navier Stokes equations to model fluid flow near a boundary. To obtain the parabolized Navier-Stokes equations, one

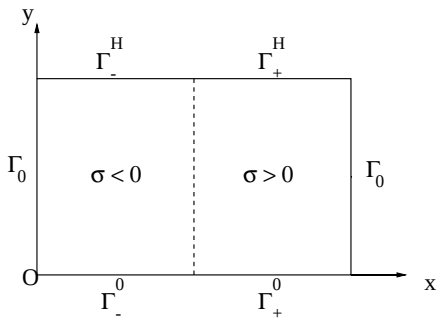

FIGURE 1. $\sigma(x, y) = x - 1/2$.

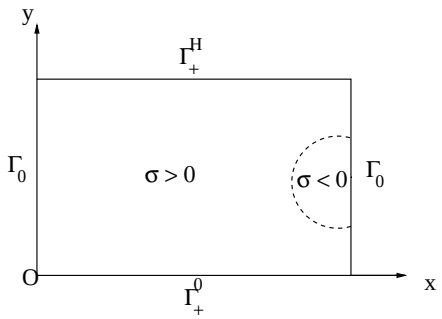

FIGURE 2. $\sigma(x, y) = (x - 1)^2 + (y - 1/2)^2 - 1/16$.

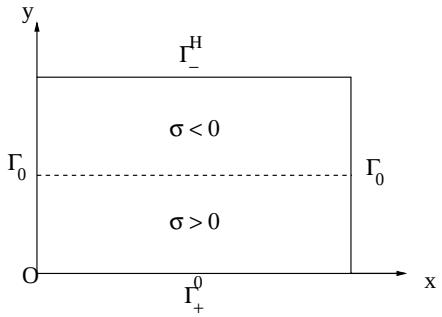

FIGURE 3. $\sigma(x, y) = 1 - 2y$.

simply omits the second derivatives in the direction of the boundary. The resulting system is parabolic in nature, and so can be solved numerically by marching forward in the time-like direction. If there are regions of separation and reverse flow, this marching procedure becomes unstable which is not surprising since the marching procedure is attempting to generate an approximate solution to a backward parabolic equation. This difficulty has been the object of several papers [7,40]. Our second example may be regarded as a simplified form of the PNS equations with a region of reverse flow. We cite some further choices of σ. If $\sigma(x, y) = 1 - 2y$ (see Fig. 3), we have a forward parabolic equation in the lower rectangle and a backward parabolic equation in the upper rectangle. The problems in the two rectangles are independent of each other, and it is not hard to see

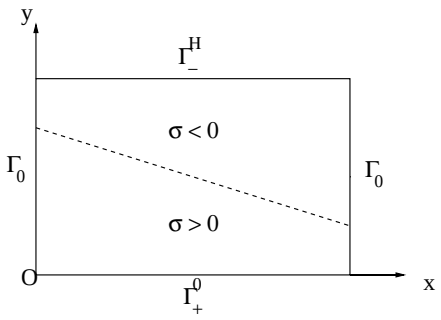

FIGURE 4. $\sigma(x, y) = 3 - x - 5y$.

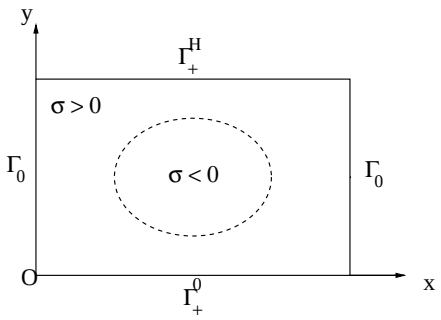

FIGURE 5. $\sigma(x, y) = (x - 1/2)^2 + (y - 1/2)^2 - 1/16$.

that the solution on the line $y = 1/2$ is given by the two point boundary value problem

$$-u_{xx}\left(x, \frac{1}{2}\right) + \lambda u = f\left(x, \frac{1}{2}\right), \quad u\left(0, \frac{1}{2}\right) = u\left(1, \frac{1}{2}\right) = 0.$$

It turns out that the solution u is continuous across the line $y = 1/2$. If $\sigma(x, y) = 3 - x - 5y$, see Figure 4, the zero line of σ is tilted. In this case, it would be interesting to show that u is continuous across the zero line of σ, and to identify the solution on the zero line. In the case $\sigma(x, y) = 2x - 1$, our solution is obtained without boundary conditions on the top or bottom of the square. As a final example, see Figure 5, we mention the choice $\sigma(x, y) = (x - 1/2)^2 + (y - 1/2)^2 - 1/16$. Here one has a backward region imbedded in the square. Our theory gives a solution to this problem. As with all these examples, it would be interesting to determine the regularity of the solution across the zero line of σ.

At times one considers the problem on a half-space or a semi-infinite strip. Recently, the problem arose in [10], as part of the solution of $(u + u^p)_t = u_{xx} + u_{yy} - u_x$ in case (c) where $1 < p < 3/2$. With ζ, η replacing x, y, their function is nontrivial:

$$\sigma(\zeta, \eta) = p v_0^{p-1}(\zeta, \eta) - \frac{3 - p}{2p}\eta,$$

where $v_0 \geq 0$ and $\int\int_{\mathbb{R}^2} v_0 \, d\eta d\zeta = M > 0$. By the properties (i) through (vii) of v_0 derived in [10], one sees that the zero set of σ must behave as depicted in Figure 6 below. There is an even symmetry about $\zeta = 0$. The fashion in which the forward-backward character arises here is similar to that in Section 3.2 in that one seeks similarity or traveling wave solutions. The interesting character under study arises in the moving frame.

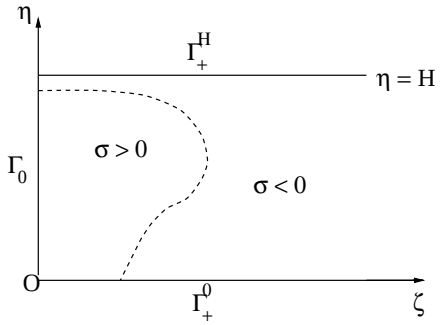

FIGURE 6. $\sigma(\zeta,\eta) = pv_0^{p-1}(\zeta,\eta) - \frac{3-p}{2p}\eta$.

Historically the problem goes back to [21, 32] – viewed as parabolic problems with degenerate coefficients, to [16] where the forward-backward problem was treated for the first time and to [15] where the results of the former were announced. Note also the paper [22], which treats a general class of degenerate parabolic equations.

3. Motivation for a forward-backward heat equation

In this section we present five different physical problems that are modeled by the forward-backward heat equation.

3.1. **Randomly accelerated particle**

Certain stochastic processes involving the motion of a particle undergoing random acceleration lead to degenerate type equations such as the one we are describing in this paper. A complete derivation and analysis as well as original references are given in Franklin and Rodemich [11].

Consider the problem of determining the time $T(x,y)$ that it takes a particle which is restricted to move on the line segment $[-1,1]$ with initial position x and initial velocity y undergoing random acceleration to reach either of the boundaries $x=-1$ or $x=1$. In the derivation in [11] it is assumed that the particle experiences an acceleration due to white noise so its velocity follows a Brownian motion. They work with the time probability density function $p(x,y,t)$ and an equation is derived for it by postulating that if a particle starting at position x with velocity y reaches a boundary in time t then a particle starting at $x-y\Delta t$ with velocity $y+\Delta y$ should reach the boundary at time $t-\Delta t$. They conclude after a formal probability argument with the following initial/boundary value problem for T:

$$y\frac{\partial T}{\partial x} + \frac{1}{2}\frac{\partial^2 T}{\partial y^2} = -1 \qquad \text{for} \quad -1 < x < 1, \quad -\infty < y < \infty, \tag{3.1}$$

$$T(x,y) \cong O(\frac{1}{|y|}) \qquad \text{for } |y| \text{ large and } -1 < x < 1,$$

$$T(1,y) = 0 \qquad \text{for } y > 0 \text{ and } T(-1,y) = 0 \text{ for } y < 0.$$

Equation (3.1) is an example of (1.1) if one switches x and y.

3.2. **LaRosa's electron beam model**

Solar type III radio bursts have since the 1950's been explained by non-thermal beams of electrons being accelerated in solar flares. The physics of electron beam propagation through the solar corona is not yet

completely settled, but in [23, 24], a nonlinear theory is proposed that seems to at least explain several of the phenomena observed.

Let the beam electron velocity distribution function be denoted by f, it specifies the measure of how large a population of electrons at location x and time t travel at velocity v. At the leading edge of the beam it is considered appropriate to model the diffusion of electrons by the quasi-linear plasma diffusion equations, which in [30] (pp. 131–133) have been derived for one dimension:

$$f_t = \frac{\pi\omega_{pe}}{m_e n_e}(vWf_v)_v. \tag{3.2}$$

It describes how electrons that are traveling sufficiently fast create Langmuir waves, hence plasma energy, and acts as a diffusion. W is the wave energy distribution in velocity space and f is integrated over all components of $\vec{v}$ normal to the beam direction. Here ω_{pe} denotes the wave frequency of the beam which is on the order of 100 Hz, n_b denote the beam density which we later will relate to the background electron density (n_e) and $n_b/n_e \in (10^{-7}, 10^{-4})$ (from what is termed a weak to a strong beam), and let v_b denote the average beam velocity which is on the order of 10^{10} cm s^{-1}, as well as the front velocity v_{front}. Adding the possibility of inhomogeneity of the beam, *i.e.*, the effect of drift within the distribution f due to varying velocity, the classical diffusion equation is modified by adding a convective derivative term to the homogeneous beam equation as in [30] (p. 135) to:

$$f_t + vf_x = \frac{\pi\omega_{pe}}{m_e n_e}(vWf_v)_v$$

for $x \geq x_{\text{front}}$ and with a constraint due to nonlinear[1] effects: $\int W(\xi = 0, v)\,\mathrm{d}v \leq W_{\text{thr}}$, expressing that there is an upper limit to the energy in the beam, *i.e.*, energy is absorbed into the plasma beyond a certain energy level. Some names of physicists involved in the theory behind these nonlinear effects are: Zakharov, Tsytovich, Rudakov, and Papadopoulos. In such a beam with a distribution of velocities, fast electrons will out-pace slower ones. At the front of the beam, the density of fast electrons increases in comparison to the slower electrons. This creates a finite positive slope ($f_v > 0$) at the beam front. This allows for energy generation there in the form of plasma waves. At the back of the beam, however, $f_v < 0$ and these slower electrons may re-absorb energy from the waves generated by the fast electrons at the head. Supposing this re-absorption is so efficient so as to allow almost all of the energy generated by the fast particles to be re-absorbed later by the slow ones, then the beam may maintain its energy and propagate large distances without losing energy to the background medium. The beam length, denoted by Δx, turns out to be on the order of 10^9 cm. The beam may then travel 10 to 100 times its length, *i.e.* as far as into interplanetary space. One may thus consider a steady-state approximation in which v_{front} and W_{thr} are considered to be constant throughout most of the life-span of the electron beam. Here we introduce a new variable moving with the front: $\xi = x - v_{\text{front}}t$. Rewriting the drift-diffusion equation in the new variable, seeking traveling wave solutions (or as LaRosa terms it, a steady-state solution), yields:

$$(v - v_{\text{front}})f_\xi = \frac{\pi\omega_{pe}}{m_e n_e}(vWf_v)_v, \tag{3.3}$$

which is a forward-backward heat equation identifiable with our own, when we set $x = v$ and $y = \xi$ and $\sigma = 1 - x_0/x$ and there is a first-order term $(W/x)f_x$ if we choose to differentiate through rather than keep the divergence form ($x_0 = v_{\text{front}}$). (3.3) is coupled with an energy equation, with $v_{\text{front}} > v_g$, one has

$$(v_g - v_{\text{front}})W_\xi = \frac{\pi\omega_{pe}}{n_e}v^2 Wf_v, \tag{3.4}$$

indicating exponential decrease of W with increasing ξ. See also [4, 6].

[1]The term here refers to strong turbulence modulational interactions, see [23, 24].

3.3. **Prandtl boundary layer equations**

A nonlinear forward-backward heat equation arises in two-dimensional fluid flow near a boundary when separation occurs. To derive this problem we start with the nondimensionalized Navier-Stokes equations for a viscous incompressible fluid,

$$\vec{u}_t + (\vec{u}\cdot\vec{\nabla})\vec{u} = -\vec{\nabla}p + Re^{-1}\Delta\vec{u} \quad \text{and} \quad \vec{\nabla}\cdot\vec{u} = 0$$

where $\vec{u} = (u, v)$ is the velocity and p is the pressure. We examine the flow near a boundary, assume it is primarily unidirectional, require no-slip boundary conditions so $\vec{u} = 0$ on the boundary, take the Reynolds number, Re, as large, and assume steady flow, $\vec{u}_t = 0$. Away from the boundary the flow is primarily inviscid. To focus our analysis on the boundary layer we make the variable change $\tilde{y} = \sqrt{Re}y$ and $\tilde{v} = \sqrt{Re}v$ which balances the important viscous and convection processes. We obtain for the x-velocity equation,

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{\partial p}{\partial x} + Re^{-1}\frac{\partial^2 u}{\partial x^2}, + \frac{\partial^2 u}{\partial y^2}$$

the y-velocity equation,

$$\frac{\partial p}{\partial y} = 0 \quad \text{and} \quad \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0.$$

Dropping the small terms, noting that p is independent of y so its behavior is completely determined in the inviscid region we have

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - \frac{\partial^2 u}{\partial y^2} = 0, \tag{3.5}$$

and

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0. \tag{3.6}$$

Typical boundary conditions associated with (1, 2) are

$$u(x, 0) = v(x, 0) = 0,\ u(0, y), u(x, 1) \text{ are given functions.} \tag{3.7}$$

In the mathematical analysis of (3.5, 3.6, 3.7), it is assumed that $u > 0$. (See [34] or [33].) In some important problems there are regions where the flow "separates" from the solid boundary, with a backflow region next to the boundary. A typical situation is shown in Figure 7. Here, there is depicted a "separation region" near the x-axis in which the flow moves to the left. The values x_S and x_R are respectively called the separation point and the re-attachment point. (Fig. 7 is essentially the same as Fig. 2, with the x and y axes interchanged and the fluid mechanics would be more reasonable if we had a wall rise along $y = x$, $x > 0$, say, rather than lie flat along the x-axis.)

Following (1.2–1.4), since $0 < x_S < x_R < 1$, the boundary conditions (3.7) are still appropriate for the problem with separation. However, the presence of a separation region complicates the solution process. In the case of no reverse flow, (3.5) is a forward parabolic equation, so there is a possibility that (3.5) can be solved by a marching procedure, moving in the positive x direction. With the presence of the reverse flow region, the marching becomes unstable because one is solving a parabolic equation in the unstable direction. A way around this difficulty has been proposed by Flügge-Lotz and Reyhner: in regions where $u < 0$, one simply drops the uu_x term and continues solving the system in a forward direction. The Flügge-Lotz and Reyhner technique is inconsistent with the equation (3.5), and results in an inaccurate solution. Modifications of the Flügge-Lotz

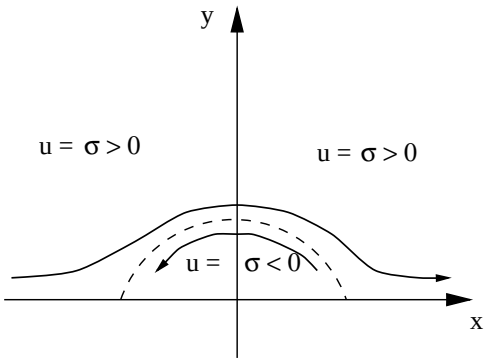

FIGURE 7. $\sigma(x, y) = u$, the first component of the velocity

and Reyhner method have been proposed in [7] that involve iterations in the separated region. These modifications are in the spirit of [47] and, if convergent, retain consistency with (3.5). It would be interesting to establish the convergence of the [7] iterations. See also [31, 41–45].

3.4. **Transport during flow reversal**

The forward-backward heat equation emerges in the modeling of the transport by convection dominated flow of temperature or a pollutant or salt in the the boundary layer of a fluid undergoing a flow separation or reversal (see [38, 39], for a specific examples).

Assume u represents temperature or the concentration of some other substance. Let (p, q) be the x and y components of the fluid velocity. Then if convection, diffusion, and some type of reaction are involved in the transport the equation is

$$p\frac{\partial u}{\partial x} + q\frac{\partial u}{\partial y} - \epsilon\Delta u + \lambda u = f$$

where f represents some heat or pollutant sources. Assume $0 < \epsilon \ll 1$. We again examine the boundary layer behavior. In the usual singular perturbation analysis one introduces the solution v of the reduced equation

$$p\frac{\partial v}{\partial x} + q\frac{\partial v}{\partial y} + \lambda v = f$$

and seeks an approximation to the difference $w = u - v$. One has

$$p\frac{\partial w}{\partial x} + q\frac{\partial w}{\partial y} - \epsilon\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) + \lambda w = 0.$$

Suppose $q(x, 0) = 0$ as is reasonable if there is no flow through the boundary. Write $q(x, y) = q_0(x)y + O(y^2)$. Introduce the stretched variable $\eta = y/\sqrt{\epsilon}$, and set $\hat{w}(x, \eta) = w(x, y)$. Then

$$p\frac{\partial \hat{w}}{\partial x} + (q_0 y + O(y^2))\epsilon^{-1/2}\frac{\partial \hat{w}}{\partial \eta} - \epsilon\frac{\partial^2 \hat{w}}{\partial x^2} - \frac{\partial^2 \hat{w}}{\partial \eta^2} + \lambda w = 0$$

so

$$p\frac{\partial \hat{w}}{\partial x} + q_0\eta\frac{\partial \hat{w}}{\partial x} - \epsilon\frac{\partial^2 \hat{w}}{\partial x^2} - \frac{\partial^2 \hat{w}}{\partial \eta^2} + \lambda w = O(\epsilon^{1/2}).$$

Replacing the right side by 0, one obtains a parabolic equation for $\hat{w}$, with x the time-like variable. If the function p changes sign in a region, the equation is forward-backward equation.

3.5. **Neutron scattering**

An example of the forward-backward heat equation occurs in the scattering of neutrons. In a simple case, the scattering medium is contained between two parallel planes perpendicular to the z-axis and placed at $z = 0$ and $z = 1$. The dependent variable of interest is the density $u(z, v)$ of neutrons at position z and velocity v. Suppose the the neutrons have constant kinetic energy, so v may be taken to be a unit vector. Suppose also that u depends only on $\mu = v \cdot e_z$, where e_z is the unit vector along the z-axis. The density function which we now write $u(z, \mu)$, is given by a linear integro-differential equation known as the linear Boltzmann equation:

$$\mu u_z + \sigma u = \sigma_s \int_{-1}^{1} p(\mu, \nu) u(z, \nu) \mathrm{d}\nu.$$

Here, $p(\mu, \nu)$ gives the probability that a neutron at position z and traveling in direction ν is scattered to the direction μ. The quantity σ_s, known as the scattering cross-section, gives the fraction of neutrons at z that encounter scattering, and the quantity σ, known as the total cross-section, gives the fraction of neutrons at z that are removed from the neutron population at (z, μ), either through scattering or absorption. This equation is considered in the infinite slab $0 < z < 1$ and for $-1 < \mu < 1$. For $\mu > 0$, $u(z, \mu)$ represents neutrons moving in a positive z-direction, whereas for $\mu < 0$, $u(z, \mu)$ represents neutrons moving in a negative z-direction. The equation is generally considered with boundary conditions that represent a specified source of neutrons entering the slab at the boundaries $z = 0$ and $z = 1$. So we have for boundary conditions

$$u(0, \mu) = g_0(\mu) \text{ for } 0 < \mu < 1, \quad u(1, \mu) = g_1(\mu) \text{ for } -1 < \mu < 0. \tag{3.8}$$

This integro-differential boundary value problem is discussed, for example, in [19, 48].

In a certain energy range, $p(\mu, \nu)$ has a significant maximum when $\nu = \mu$. (This is discussed in [5], where the following equation is derived.) In this case, we may approximate the integral by expanding $u(z, \nu)$ in a power series around $\nu = \mu$ and retaining only the first 3 terms. We get

$$\int_{-1}^{1} p(\mu, \nu) u(z, \nu) \mathrm{d}\nu \approx u(z, \mu) + u_\mu(z, \mu) \int_{-1}^{1} (\nu - \mu) p(\mu, \nu) \mathrm{d}\nu + \frac{1}{2} u_{\mu\mu}(z, \mu) \int_{-1}^{1} (\nu - \mu)^2 p(\mu, \nu) \mathrm{d}\nu,$$

which leads to the forward-backward equation

$$\mu u_z + \sigma u - a(\mu) u_{\mu\mu} - b(\mu) u_\mu = 0 \tag{3.9}$$

for appropriate functions $a(\mu) > 0$ and $b(\mu)$. In addition to the boundary conditions (1), one imposes $u(z, \pm\mu) = 0$, which corresponds to the requirement that there are no neutrons moving in a direction parallel to the slab.

4. Existence for the forward-backward heat equation

In the spirit of a classical paper by Baouendi and Grisvard [3], we aim at proving existence and uniqueness of a weak solution to forward-backward heat equations lying within a certain class. We start by assuming that $\sigma \in C^1(\overline{\Omega})$. Define the following spaces:

$$F = H^{(1,0)}(\Omega) = L_2(H_0^1),$$
$$\Phi = \{\phi \in F \cap C^1(\overline{\Omega}) : \phi(x, 0) = 0 \text{ if } \sigma(x, 0) < 0 \wedge \phi(x, H) = 0 \text{ if } \sigma(x, H) > 0\}$$

equipped with the norms:

$$\|u\|_F^2 = \|u\|_{L_2(\Omega)}^2 + \|u_x\|_{L_2(\Omega)}^2$$

and

$$\|\phi\|_\Phi^2 = \|\phi\|_F^2 + \frac{1}{2}\int_{\Gamma_+^0} \sigma(x,0)|\phi(x,0)|^2\,\mathrm{d}x - \frac{1}{2}\int_{\Gamma_-^H} \sigma(x,H)|\phi(x,H)|^2\,\mathrm{d}x.$$

Then define the following bilinear form:

$$E(u,\phi) = \iint_\Omega \left[-u(\sigma\phi)_y + \iint_\Omega u_x\phi_x + \lambda u\phi\right] \tag{4.1}$$

over the product space $F \times \Phi$. Suppose $\phi \in \Phi$, then

$$E(\phi,\phi) = -\iint_\Omega \phi(\sigma\phi)_y + \iint_\Omega \phi_x^2 + \lambda\phi^2.$$

Simple algebra yields that $\phi(\sigma\phi)_y = \frac{1}{2}(\sigma\phi^2)_y + \frac{1}{2}\sigma_y\phi^2$ and the following formula:

$$\begin{aligned} E(\phi,\phi) &= -\iint_\Omega \frac{1}{2}(\sigma\phi^2)_y + \iint_\Omega \phi_x^2 + (\lambda - \frac{1}{2}\sigma_y)\phi^2 \\ &= -\int_0^L \frac{1}{2}(\sigma\phi^2)(x,H)\,\mathrm{d}x + \int_0^L \frac{1}{2}(\sigma\phi^2)(x,0)\,\mathrm{d}x + \iint_\Omega \phi_x^2 + (\lambda - \frac{1}{2}\sigma_y)\phi^2 \\ &= \iint_\Omega \phi_x^2 + (\lambda - \frac{1}{2}\sigma_y)\phi^2 + \frac{1}{2}\int_{\Gamma_+^0} \sigma\phi^2 - \frac{1}{2}\int_{\Gamma_-^H} \sigma\phi^2. \end{aligned}$$

Applying Poincaré's inequality:

$$\iint_\Omega \phi_x^2 \geq c_\Omega \iint_\Omega \phi^2$$

where we explicitly may exhibit $c_\Omega = (\pi/L)^2$, we finally get, for some $\delta \in (0,1)$,

$$E(\phi,\phi) \geq (1-\delta)\iint_\Omega \phi_x^2 + \iint_\Omega (\delta c_\Omega + \lambda - \frac{1}{2}\sigma_y)\phi^2 + \frac{1}{2}\int_{\Gamma_+^0} \sigma\phi^2 - \frac{1}{2}\int_{\Gamma_-^H} \sigma\phi^2,$$

which coerces $\|\phi\|_\Phi^2$ provided there exists $\mu > 0$ such that $c_\Omega + \lambda - \frac{1}{2}\sigma_y \geq \mu$ on $\overline{\Omega}$. This holds if

$$\lambda - \frac{1}{2}\sigma_y > -(\pi/L)^2 \text{ on } \overline{\Omega}, \tag{4.2}$$

which is the same result obtained by [29] in their Corollary 3.2. using – and allowing – some more terms in the pde. In conclusion,

$$E(\phi,\phi) \geq \alpha\|\phi\|_\Phi^2,$$

with $\alpha = \min\{1-\delta, \mu\delta\}$. Next we show that $E(\cdot,\phi)$ at fixed $\phi \in \Phi$ belongs to F^*:

$$E(u,\phi) = -\iint_\Omega u(\sigma\phi_y) + \iint_\Omega u_x\phi_x + \lambda u\phi \leq C\|u\|_F\|\phi\|_{C^1}$$

which allows us to use Lions' projection theorem ([25]: Chap. 3, Sect. 1) to conclude that

Proposition 4.1. *Suppose $\sigma \in C^1(\overline{\Omega})$ and (4.2) holds. Then if $g \in L_2(H^{-1})$, there exists $u \in L_2(H_0^1)$ such that*

$$E(u, \phi) = (f, \phi), \qquad \forall \phi \in \Phi. \tag{4.3}$$

Taking the strong form of the equation in distributional sense, as the solution $u \in F, u_{xx} \in L_2(H^{-1})$ in which space also f lies. One would therefore naturally wish to seek u so that also $\sigma u_y \in L_2(H^{-1})$. If we are sufficiently lucky, this will also provide uniqueness.

4.1. **A generalized trace theorem**

The plan is to develop a tighter variational formulation (evidently, the use of Φ is a bit crude) employing a formula of Green's type which in turn necessitates a theory of traces with σ as a weight. For this we require a further assumption on σ. We suppose that $\sigma \in C^2(\overline{\Omega})$ and that

$$\sigma(x^*, 0) = 0 \Rightarrow \sigma_x(x^*, 0) \neq 0 \quad \text{and} \quad \sigma(x^*, H) = 0 \Rightarrow \sigma_x(x^*, H) \neq 0. \tag{4.4}$$

The assumption (4.4) implies that σ has at most a finite number of zeros on the lines $y = 0$ and $y = H$. Each of the 5 examples in Section 2 satisfies this assumption. The assumption can be relaxed considerably; in fact, if $\sigma(x, 0) = \operatorname{sgn}(x - x^*)$ in a neighborhood of a zero of σ, the proofs given below go through.

Let

$$\mathcal{B} = \{u \in L_2(H_0^1) : \sigma u_y \in L_2(H^{-1})\} \tag{4.5}$$

equipped with the norm:

$$\|u\|_{\mathcal{B}}^2 = \|u_x\|_0^2 + \int_0^H \|\sigma u_y\|_{H^{-1}}^2 \, \mathrm{d}y.$$

Let also

$$\mathcal{A} = \{u \in L_2(H_0^1) : u_y \in L_2\}.$$

Theorem 4.1. *Suppose $\sigma \in C^2(\overline{\Omega})$ and satisfies (4.4). Let $\mathcal{A}$ and $\mathcal{B}$ be defined as above. Then $\mathcal{A}$ is a dense subset of $\mathcal{B}$. Furthermore, the trace maps*

$$u \mapsto u(x, 0) \ \text{ and } \ u \mapsto u(x, H),$$

defined for all $u \in \mathcal{A}$, are extendable to $\mathcal{B}$ as bounded operators and the trace boundedness,

$$\int_0^L (|\sigma||u|^2)(x, 0) \, \mathrm{d}x + \int_0^L (|\sigma||u|^2)(x, H) \, \mathrm{d}x \leq \beta \|u\|_{\mathcal{B}}^2,$$

holds.

The proof of this theorem will come after some lemmas. We start by dealing with a half-line. Suppose for the moment that σ is defined and continuously differentiable in the closed upper half plane, and that $\sigma(x, 0) > 0$ for $x > x_0$ where $x_0 \in (0, L)$ and that $\sigma < 0$ to the left of x_0. Define $\mathbb{R}_+^0 = \{x \in \mathbb{R} : x > x_0\}$ and

$$B = \{u \in L_2(H^1(\mathbb{R}_+^0)) : \sigma u_y \in L_2(H^{-1}(\mathbb{R}_+^0))\} \tag{4.6}$$

equipped with the norm:

$$\|u\|_B^2 = \int_0^\infty \int_{x_0}^\infty (u^2 + u_x^2)\, \mathrm{d}x\mathrm{d}y + \int_0^\infty \|\sigma u_y\|^2_{H^{-1}(\mathbb{R}^0_+)}\, \mathrm{d}y.$$

We shall demonstrate that $u \in B$ has trace on $y = 0$ in a weighted L_2-space.

Proposition 4.2. *Let B be defined as above. Then for every $v \in B$, $v(x,0)$ is measurable in $\mathbb{R}_+$ and there is a constant k such that, for all $v \in B$:*

$$\int_{x_0}^\infty (\sigma|v|^2)(x,0)\, \mathrm{d}x \le k\|v\|_B^2. \tag{4.7}$$

This proposition will be proved *via* two lemmas, but first some more spaces are needed: Let us introduce

$$W = \{u \in L_2(H^1(\mathbb{R})) : \sigma u_y \in L_2(H^{-1}(\mathbb{R}))\}$$

equipped with the obvious norm (inherited from the B space) and

$$V = W \cap \mathcal{E}'(\overline{\mathbb{R}^2_+}) \cap \mathcal{C}^\infty(\overline{\mathbb{R}^0_+}, H^1(\mathbb{R}))$$

where $\mathcal{E}'(\overline{\mathbb{R}^2_+})$ denotes the set of distributions over $\mathbb{R}^2_+ = \mathbb{R} \times \mathbb{R}^0_+$ of bounded support.

Lemma 4.1. *(i) V is dense in W, and (ii) the trace map $u \mapsto |\sigma|^{1/2}u(x,0)$ defined for all $u \in V$ is extendable to W as a bounded, linear operator from W to $L_2(\mathbb{R})$.*

Proof. (i) It suffices to show that $W \cap \mathcal{E}'(\overline{\mathbb{R}^2_+})$ is dense in W. There exists a $\phi \in \mathcal{C}_0^\infty(\mathbb{R})$ which satisfies

$$0 \le \phi(x) \le 1, \quad \phi(x) = 1 \ \text{ for } \ -1 \le x \le 1, \ \text{ and } \ \phi(x) = 0 \ \text{ for } \ |x| \ge 2.$$

For a given $u \in W$, let

$$u_n(x,y) = \phi(\frac{y}{n})\phi(\frac{x - x_0}{n})u(x,y) \qquad \text{for } \ x \in \mathbb{R}, y \in \mathbb{R}_+.$$

Then $u_n \in \mathcal{E}'(\overline{\mathbb{R}^2_+}) \cap W$. Using Lebesgue's dominated convergence theorem, we may show that u_n tends to u in $L_2(H^1(\mathbb{R}))$. Let us now show that $|\sigma|(\partial u_n/\partial y)$ tends to $|\sigma|(\partial u/\partial y)$ in $L_2(H^{-1}(\mathbb{R}))$. We have that

$$|\sigma|\frac{\partial u_n}{\partial y}(x,y) = \phi(\frac{y}{n})\phi(\frac{x - x_0}{n})|\sigma|\frac{\partial u}{\partial y}(x,y) + \phi'(\frac{y}{n})\phi(\frac{x - x_0}{n})\frac{|\sigma|}{n}u(x,y).$$

The first term on the right-hand-side tends to $|\sigma|(\partial u/\partial y)(x,y)$ in $L_2(H^{-1}(\mathbb{R}))$. Now for the second term: as $\phi(x/n) = 0$ for $|x| \ge 2n$ we have $|x - x_0|\phi([x - x_0]/n)/n \le 2$ and – with σ Lipschitz – $\phi([x - x_0]/n)|\sigma|/n$ is also uniformly bounded. Lebesgue's theorem now shows that this term tends to zero in $L_2(H^0(\mathbb{R}))$ and hence in $L_2(H^{-1}(\mathbb{R}))$. The remaining part to show in (i) is done by smoothing, using well-known techniques with mollifiers.

(ii) Let $u \in V$. Since $\sigma \in C^1(\overline{\Omega})$, the weak derivative $|\sigma|_y$ is a bounded function given by the formula: $|\sigma|_y = \sigma_y$ if $\sigma > 0$, $|\sigma|_y = -\sigma_y$ if $\sigma < 0$, and $|\sigma|_y = 0$ if $\sigma = 0$. We have

$$\begin{aligned}\int_{-\infty}^{\infty} |\sigma(x,0)||u(x,0)|^2 \mathrm{d}x &= -\int_0^{\infty} \frac{\mathrm{d}}{\mathrm{d}y}\left(\int_{-\infty}^{\infty} |\sigma(x,y)u(x,y)^2|\mathrm{d}x\right)\mathrm{d}y \\ &= -2\int_0^{\infty}\int_{-\infty}^{\infty} |\sigma(x,y)|u(x,y)u_y(x,y)\mathrm{d}x\mathrm{d}y - \int_0^{\infty}\int_{-\infty}^{\infty} |\sigma(x,y)|_y|u(x,y)|^2\mathrm{d}x\mathrm{d}y \\ &\le \|\sigma u_y\|^2_{L_2(H^{-1}(\mathbb{R}))} + \|u\|^2_{L_2(H^1(\mathbb{R}))} + \||\sigma|_y\|_{L_\infty}\|u\|^2_{L_2} \qquad (4.8) \\ &\le C\|u\|^2_{\mathcal{B}}.\end{aligned}$$

This proves (ii) employing (i). □

Lemma 4.2. *There exists an extension operator P such that*

$$P \in \mathcal{L}(H^1(\mathbb{R}^0_+), H^1(\mathbb{R})), \qquad (4.9)$$

and $|\sigma(\cdot,0)|Pu = Q(\sigma(\cdot,0)u)$ where Q is another extension operator satisfying

$$Q \in \mathcal{L}(H^{-1}(\mathbb{R}^0_+), H^{-1}(\mathbb{R})). \qquad (4.10)$$

Proof. Explicitly, for any $u \in H^1(\mathbb{R}^0_+)$, let (in the style of Calderon)

$$(Pu)(x) = u(x), \quad \text{for } x \ge x_0, \quad (Pu)(x) = \sum_{k=1}^{2} \alpha_k u(x_0 + k(x_0 - x)), \quad \text{for } x < x_0.$$

In order for $Pu \in H^1(\mathbb{R})$ and hence also $\mathcal{C}^0$, we must have

$$\sum_{k=1}^{2} \alpha_k = 1. \qquad (4.11)$$

In order to create the commutativity, we see that

$$(Qv)(x) = v(x), \qquad \text{for } x \ge x_0, \quad (Qv)(x) = \sum_{k=1}^{2} \beta_k v(x_0 + k(x_0 - x)), \sigma \text{ for } x < x_0.$$

Matching $|\sigma|P$ with $Q \circ \sigma$ is achieved by $\beta_k \sigma(x_0 + k(x_0 - x), 0) = \alpha_k(-\sigma(x,0))$, so we set

$$\beta_k = -\frac{\sigma(x,0)}{\sigma(x_0 + k(x_0 - x), 0)}\alpha_k.$$

To have Q be continuous in the topology mentioned, we may show that the adjoint of Q satisfies $Q^* \in \mathcal{L}(H^1(\mathbb{R}), H^1_0(\mathbb{R}^0_+))$. By a simple integration-by-parts argument we see that

$$(Q^*u)(x) = u(x) + \sum_{k=1}^{2} \frac{\alpha_k}{k} \frac{\sigma(x_0 + (x_0 - x)/k, 0)}{\sigma(x,0)} u(x_0 + \frac{x_0 - x}{k})$$

and the new compatibility condition at x_0,

$$\lim_{x \to x_0+} \sum_{k=1}^{2} \frac{\alpha_k}{k} \frac{\sigma(x_0 + (x_0 - x)/k, 0)}{\sigma(x,0)} = \left(-\sum_{k=1}^{2} \frac{\alpha_k}{k^2} \frac{\sigma'_-(x_0,0)}{\sigma'_+(x_0,0)} \right) = -1, \qquad (4.12)$$

arises. By the hypothesis (4.4), $\sigma(x_0+(x_0-x)/k,0)/\sigma(x,0)\in W^{1,\infty}(0,x_0)$ for $k=1,2$. The two-by-two linear system for $(\alpha_k)_{k=1}^2$, (4.11–4.12), is well-posed as its determinant is nonzero. □

Proof of Proposition 4.2. Suppose $v\in B$. Using the extension P from Lemma 4.2, we have

$$Pv\in L_2(H^m(\mathbb{R})),\ \text{ and }\quad \frac{\partial}{\partial y}|\sigma|Pv=\frac{\partial}{\partial y}Q(\sigma v)=Q\left(\frac{\partial(\sigma v)}{\partial y}\right)=Q(\sigma v_y)+Q(\sigma_y v)\in L_2(H^{-1}(\mathbb{R})).$$

The mapping $B\ni v\mapsto Pv\in W$ is thus bounded. Now, to use Lemma 0.4, we get

$$\int_{x_0}^{\infty}\sigma(x,0)|v(x,0)|^2\,\mathrm{d}x\le\int_{-\infty}^{\infty}|\sigma(x,0)||(Pv)(x,0)|^2\,\mathrm{d}x\le\|Pv\|_W^2\le k\|v\|_B^2,$$

which ends the proof. □

Proof of Theorem 4.1. The density follows by well-known techniques. For the trace-boundedness, let us concentrate on verifying this property for the mapping:

$$u\mapsto u(x,0)$$

for $x_0<x<L$:

Pick ϕ and ψ to belong to $\mathcal{C}^1([x_0,\infty)\times[0,\infty))$ in such a way that

$$\begin{aligned}&\phi(x,y)=0, && \text{for } y\ge\frac{1}{2}H \text{ or } x\ge\frac{1}{2}(x_0+L),\\ &\psi(x,y)=0, && \text{for } x\le\frac{3}{4}x_0+\frac{1}{4}L, \text{ and}\\ &\phi(x,0)+\psi(x,0)=1, && \text{for } x_0\le x\le L.\end{aligned}$$

We now split u locally as follows:

$$\begin{aligned}&v(x,y)=\ \ \phi(x,y)u(x,y), && \text{for } 0\le y\le H \text{ and } x_0<x<L,\\ &v(x,y)=0, && \text{for } y\ge H \text{ or } x\ge L.\end{aligned}$$

We let $w(x,y)=\psi(x,y)u(x,y)$ for $0<y<H$ and $x_0<x<L$. Then $v\in B$ and

$$\|v\|_B\le k_1\|u\|_{\mathcal{B}},\quad\text{where}\quad\int_{-\infty}^{\infty}\sigma(x,0)|\phi(x,0)u(x,0)|^2\,\mathrm{d}x\le kk_1^2\|u\|_{\mathcal{B}}^2,$$

due to the Proposition. On the other hand, we have

$$w\in L_2(H_0^1(0,L)),\quad \partial w/\partial y\in L_2(H^{-1}(0,L)),$$

$$\int_0^H\int_0^L|\frac{\partial w}{\partial x}|^2\,\mathrm{d}x\mathrm{d}y+\int_0^H\|\frac{\partial w}{\partial y}\|_{H^{-1}(0,L)}^2\,\mathrm{d}y\le k_2\|u\|_{\mathcal{B}}^2.$$

Using an interpolation result by Lions and Peetre, the trace of w on $y=0$ is in L_2, and one obtains

$$\int_0^L|\psi(x,0)u(x,0)|^2\,\mathrm{d}x\le k_3\|u\|_{\mathcal{B}}^2.$$

A similar bound holds with σ present in the integrand. By stringing together the recent bounds, we see that

$$\int_{x_0}^{L} \sigma(x,0)|u(x,0)|^2 \,\mathrm{d}x \leq k_4 \|u\|_{\mathcal{B}}^2.$$

Applying this inequality the finite number of times it takes to account for changes of sign of σ on $y = 0$ or $y = H$ ends the proof of the theorem. □

4.2. **Uniqueness for the forward-backward heat equation**

Now we tighten up the variational formulation. We first form a Green type formula.

Corollary 4.1. *For $u, v \in \mathcal{B}$ the following formula holds*

$$\langle \sigma \frac{\partial u}{\partial y}, v\rangle + \langle u, \frac{\partial(\sigma v)}{\partial y}\rangle = \int_0^L \sigma(x,H)u(x,H)v(x,H)\,\mathrm{d}x - \int_0^L \sigma(x,0)u(x,0)v(x,0)\,\mathrm{d}x. \tag{4.13}$$

Proof. Both sides of the equation are well-defined due to the definition of $\mathcal{B}$, that $\sigma_y \in L^\infty$, and the trace Theorem. Both are continuous, linear functionals on $\mathcal{B} \times \mathcal{B}$. It thus suffices to verify this identity on the dense subset $\mathcal{A}$. Here

$$\begin{aligned} \iint_\Omega \sigma \frac{\partial u}{\partial y} v \,\mathrm{d}x\mathrm{d}y + \iint_\Omega u \frac{\partial(\sigma v)}{\partial y}\,\mathrm{d}x\mathrm{d}y &= \iint_\Omega \frac{\partial}{\partial y}(\sigma u v)\,\mathrm{d}x\mathrm{d}y \\ &= \int_0^L \sigma(x,H)u(x,H)v(x,H)\,\mathrm{d}x - \int_0^L \sigma(x,0)u(x,0)v(x,0)\,\mathrm{d}x, \end{aligned}$$

which ends the proof of the corollary. □

From the identity in the Existence Proposition, u being a distributional solution to the original forward-backward heat equation, we get for all $\phi \in \Phi$,

$$-\langle u, \frac{\partial(\sigma\phi)}{\partial y}\rangle + \langle \frac{\partial u}{\partial x}, \frac{\partial \phi}{\partial x}\rangle + \langle \lambda u, \phi\rangle = \langle \sigma \frac{\partial u}{\partial y} - \frac{\partial^2 u}{\partial x^2} + \lambda u, \phi\rangle\cdot$$

Since $\Phi \subseteq \mathcal{B}$ we may apply the Corollary and get

$$\int_0^L \sigma(x,0)u(x,0)\phi(x,0)\,\mathrm{d}x - \int_0^L \sigma(x,H)u(x,H)\phi(x,H)\mathrm{d}x + \langle \frac{\partial u}{\partial x}, \frac{\partial \phi}{\partial x}\rangle = -\langle \frac{\partial^2 u}{\partial x^2}, \phi\rangle$$

for all $\phi \in \Phi$. Integrating by parts yields

$$-\langle \frac{\partial^2 u}{\partial x^2}, \phi\rangle = \langle \frac{\partial u}{\partial x}, \frac{\partial \phi}{\partial x}\rangle$$

and taking into account the b.c. for ϕ, we see that

$$\int_{\Gamma_+^0} \sigma(x,0)u(x,0)\phi(x,0)\,\mathrm{d}x - \int_{\Gamma_-^H} \sigma(x,H)u(x,H)\phi(x,H)\,\mathrm{d}x = 0$$

for all $\phi \in \Phi$. This gives a weak imposition of the initial and terminal conditions on u since those traces are now well-defined by the trace theorem.

Theorem 4.2. *Suppose $\sigma \in C^2(\overline{\Omega})$ and satisfies (4.4). For every $g \in L_2(H^{-1})$ there exists a unique solution $u \in \mathcal{B}$ satisfying (1.1–1.4).*

Proof. Existence is already proved. Suppose $g = 0$. By the argument in the proof of the Corollary, specifically letting $u = v$ in (4.13),

$$2\langle \sigma \frac{\partial u}{\partial y}, u\rangle + \langle \sigma_y u, u\rangle = \int_{\Gamma_+^H} \sigma u^2 - \int_{\Gamma_-^0} \sigma u^2,$$

so that

$$\langle \sigma \frac{\partial u}{\partial y}, u\rangle = -\frac{1}{2}\langle \sigma_y u^2, 1\rangle + \frac{1}{2}\int_{\Gamma_+^H} \sigma u^2 - \frac{1}{2}\int_{\Gamma_-^0} \sigma u^2.$$

However, as we are dealing with a distributional solution to the homogeneous partial differential equation, we get

$$\langle \sigma \frac{\partial u}{\partial y}, u\rangle = \langle \frac{\partial^2 u}{\partial x^2} - \lambda u, u\rangle = -\langle \frac{\partial u}{\partial x}, \frac{\partial u}{\partial x}\rangle - \langle \lambda u, u\rangle\cdot$$

When we combine these two identities, we see that

$$\langle u_x, u_x\rangle + \langle (\lambda - \frac{1}{2}\sigma_y)u, u\rangle = -\frac{1}{2}\left(\int_{\Gamma_+^H} \sigma u^2 - \int_{\Gamma_-^0} \sigma u^2\right) \le 0,$$

whence – using once more the argument in the proof of Proposition 4.1 – , almost everywhere in Ω, $u = 0$ as required to complete the proof. □

It is easily seen that each of the 5 examples of Section 2 satisfies (4.2) and satisfies (4.4) for sufficiently large λ.

5. Galerkin method

In this section we introduce and analyze a higher order finite element method based on a cross-product space of continuous piecewise polynomials of possibly high degree. Higher order methods have not appeared in the earlier papers cited here.

5.1. **Energy properties of the continuous problem**

We shall gather a few identities and inequalities of energy type that will be useful for our discrete method to be introduced and analyzed in the next subsection.

Let us, in addition to our earlier hypotheses on the coefficients, assume, instead of (4.2), that

$$\lambda - \frac{1}{2}\sigma_y \ge 0 \quad \text{in } \Omega. \tag{5.1}$$

This assumption applies to all the examples of Section 2 for which λ is sufficiently large.

We now give some discussion of the boundary value problem (1.1–1.4). We start with a formal manipulation. Suppose u is a suitably smooth solution of (1.1–1.4). Let ϕ be a suitable function on Ω which vanishes on $x = 0, L$. Multiplying both sides of (1.1) by ϕ and integrating by parts, one obtains

$$\int_\Omega [\sigma(x,y)u_y\phi + u_x\phi_x + \lambda u\phi]\mathrm{d}x\mathrm{d}y = \int_\Omega f\phi \mathrm{d}x\mathrm{d}y. \tag{5.2}$$

Now set $\phi = u$ – which warrants the formal identity as true since we established earlier that $u \in L_2(H_0^1)$ and $\sigma u \in L_2(H^{-1})$ so that (5.2) could be considered as a consequence of Theorem 4.2 – and use the boundary condition to write

$$\int_\Omega [\sigma(x,y)uu_y + \lambda u^2]\,\mathrm{d}x\mathrm{d}y = \frac{1}{2}\int_\Omega (\sigma u^2)_y \mathrm{d}x\mathrm{d}y + \int_\Omega [\lambda - \frac{1}{2}\sigma_y]u^2\,\mathrm{d}x\mathrm{d}y \geq \frac{1}{2}\int_0^L \sigma(x,y)u(x,y)^2\big|_{y=0}^{y=H}\,\mathrm{d}x \geq 0.$$

Using these inequalities we obtain

$$\int_\Omega u_x^2 \mathrm{d}x\mathrm{d}y \leq \int_\Omega uf\mathrm{d}x\mathrm{d}y.$$

From this and the Poincaré inequality, it follows that a solution u of (1.1–1.4) satisfies

$$\int_\Omega u_x^2 \mathrm{d}x\mathrm{d}y \leq C\int_\Omega f^2 \mathrm{d}x\mathrm{d}y \tag{5.3}$$

which is our first energy inequality.

We now write another, related, energy inequality for (1.1–1.4). First, consider the two point boundary value problem

$$-u_{xx} + \lambda u = f,\ u(0) = u(L) = 0,$$

with $\lambda \geq 0$. Write the solution operator for this equation as $u = T_\lambda f$. The energy formula gives

$$(u_x, u_x) + \lambda(u,u) = (f,u) = (f, T_\lambda f) = \|T_\lambda^{1/2} f\|^2.$$

Hence, since $f = T_\lambda^{-1}u$,

$$\|u_x\|^2 + \lambda\|u\|^2 = \|T_\lambda^{1/2}(T_\lambda^{-1}u)\|^2 = \|T_\lambda^{-1/2}u\|^2. \tag{5.4}$$

Now we consider the problem (1.1–1.4). We write (5.1) as $-u_{xx} + \lambda u = f - \sigma u_y$, so $u = T_\lambda f - T_\lambda(\sigma u_y)$. Multiplying both sides by σu_y and integrating over $(0, L)$, we obtain

$$(u, \sigma u_y) + (\sigma u_y, T_\lambda(\sigma u_y)) = (T_\lambda f, \sigma u_y).$$

Integrating over $(0, H)$, we get the identity

$$\frac{1}{2}\int_0^L \left[\sigma(x,H)u^2(x,H) - \sigma(x,0)u^2(x,0)\right]\mathrm{d}x + \int_\Omega [\lambda - \frac{1}{2}\sigma_y]u^2 \mathrm{d}x\mathrm{d}y + \int_0^H \|T_\lambda^{1/2}(\sigma u_y)\|^2 \mathrm{d}y = \int_0^H (T_\lambda f, \sigma u_y)\mathrm{d}y.$$

One now gets the following stability inequality for the forward-backward equation,

$$\int_0^H \|T_\lambda^{1/2}(\sigma u_y)\|^2 \mathrm{d}y \leq \int_0^H \|T_\lambda^{1/2} f\|^2 \mathrm{d}y. \tag{5.5}$$

Now we return to the equation $u = T_\lambda f - T_\lambda(\sigma u_y)$. Multiplying by $T_\lambda^{-1/2}$ and using the triangle inequality, $\|T_\lambda^{-1/2}u\| \leq \|T_\lambda^{1/2} f\| + \|T_\lambda^{1/2}(\sigma u_y)\|$. We therefore obtain

$$\int_0^H [\|T_\lambda^{-1/2}u\|^2 + \|T_\lambda^{1/2}(\sigma u_y)\|^2]\mathrm{d}y \leq 2\int_0^H \|T_\lambda^{1/2} f\|^2 \mathrm{d}y.$$

Using (5.4), we get

$$\int_0^H [\|u_x\|^2 + \lambda\|u\|^2 + \|T_\lambda^{1/2}(\sigma u_y)\|^2]\mathrm{d}y \le 2\int_0^H \|T_\lambda^{1/2} f\|^2 \mathrm{d}y \tag{5.6}$$

which is our second energy identity. We shall next develop a similar structure on the discrete level.

5.2. **Discrete Galerkin method**

Now we give a precise description of our numerical method. For this, we use the subspace $\mathcal{S}$ of $S_h^p \otimes S_k^q$ and of $\{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_-^H \cup \Gamma_+^0 \cup \Gamma_0\}$ where S_h^p is the set of continuous piecewise polynomials on $(0, L)$ of degree p and S_k^q is the set on $(0, H)$. The numerical method takes the form: find $\tilde{u} \in \mathcal{S}$ such that

$$\int_0^H [(\sigma\tilde{u}_y, \tilde{w}) + (\tilde{u}_x, \tilde{w}_x) + \lambda(\tilde{u}, \tilde{w})]\mathrm{d}y = \int_0^H (f, \tilde{w})\mathrm{d}y, \quad \text{for all } \tilde{w} \in \mathcal{S}. \tag{5.7}$$

In order to study stability along the lines of earlier energy inequalities, we introduce a discrete variant of the operator T_λ: let

$$T_{\lambda,h} : \mathcal{S}^\star \to \mathcal{S} : \mathcal{S}^\star \ni g \mapsto T_{\lambda,h} g = \tilde{z} \in \mathcal{S}$$

be defined by

$$(\tilde{z}_x, \tilde{w}_x) + \lambda(\tilde{z}, \tilde{w}) = (g, \tilde{w}) \quad \forall \tilde{w} \in \mathcal{S}.$$

We also introduce a discrete L^2-projection operator P_h: let

$$P_h : \mathcal{S}^\star \to \mathcal{S} : \mathcal{S}^\star \ni g \mapsto P_h g = \tilde{g} \in \mathcal{S}$$

be defined by

$$(\tilde{g}, \tilde{w}) = (g, \tilde{w}) \quad \forall \tilde{w} \in \mathcal{S}.$$

We then claim the following commutative property.

Lemma 5.1. $T_{\lambda,h} P_h = P_h T_{\lambda,h}$.

Proof. It suffices to verify that $T_{\lambda,h} f = T_{\lambda,h} P_h f$ for all $f \in \mathcal{S}^\star$: if $T_{\lambda,h} f = \tilde{z} \in \mathcal{S}$, then $(\tilde{z}_x, \tilde{w}_x) = (f, \tilde{w})$ which in turn equals $(P_h f, \tilde{w})$ for all $\tilde{w} \in \mathcal{S}$. Hence $\tilde{z} = T_{\lambda,h} P_h f$. □

Lemma 5.2. *$T_{\lambda,h}$ is a symmetric, positive definite operator with a square root defined on $\mathcal{S}$.*

Proof. If $T_{\lambda,h} f = \tilde{z} \in \mathcal{S}$ and $T_{\lambda,h} g = \tilde{w} \in \mathcal{S}$, then

$$(T_{\lambda,h} f, g) = (\tilde{z}, g) = (\tilde{z}_x, \tilde{w}_x) + \lambda(\tilde{z}, \tilde{w}) = (f, \tilde{w}) = (f, T_{\lambda,h} g)$$

and if $f = g \in \mathcal{S}$, then $(T_{\lambda,h} f, f) = (\tilde{z}_x, \tilde{z}_x) + \lambda(\tilde{z}, \tilde{z}) > 0$, unless $\tilde{z}$ and hence f vanish. □

We will also frequently use the inequality

$$\int_0^H [(\sigma z_y, z) + \lambda(z, z)]\mathrm{d}y \geq 0. \tag{5.8}$$

We first show the existence and uniqueness of the Galerkin solution as well as a stability inequality, analogous to the two energy inequalities given above for the exact solution.

Theorem 5.1. *The system (5.7) has a unique solution, and the following stability inequalities hold:*

$$\int_0^H [\|\tilde{u}_x\|^2 + \|\tilde{u}\|^2 + \|T_{\lambda,h}^{1/2} P_h x \tilde{u}_y\|^2]\mathrm{d}y \leq C \int_0^H [\|f\|^2 + \|T_{\lambda,h}^{1/2} P_h f\|^2]\mathrm{d}y. \tag{5.9}$$

Proof. In (5.7), first set $\tilde{w} = \tilde{u}$ to obtain the identity

$$\int_0^H [(\sigma \tilde{u}_y, \tilde{u}) + (\tilde{u}_x, \tilde{u}_x) + \lambda(\tilde{u}, \tilde{u})]\mathrm{d}y = \int_0^H (f, \tilde{u})\mathrm{d}y.$$

Using (5.8) and the Poincaré inequality, we immediately get

$$\int_\Omega [u_x^2 + u^2]\mathrm{d}x\mathrm{d}y \leq C \int_\Omega f^2 \mathrm{d}x\mathrm{d}y. \tag{5.10}$$

In (5.7), next set $\tilde{w} = T_{\lambda,h}\sigma\tilde{u}_y$, and note that $(\tilde{u}_x, (T_{\lambda,h}\sigma\tilde{u}_y)_x) + \lambda(\tilde{u}, T_{\lambda,h}\sigma\tilde{u}_y) = (\tilde{u}, \sigma\tilde{u}_x)$. We then have

$$\int_0^H [(\sigma\tilde{u}_y, T_{\lambda,h}\sigma\tilde{u}_y) + (\tilde{u}, \sigma\tilde{u}_y)]\mathrm{d}y = \int_0^H (f, T_{\lambda,h}\sigma\tilde{u}_y)\mathrm{d}y.$$

Using the fact that

$$(\sigma\tilde{u}_y, T_{\lambda,h}\sigma\tilde{u}_y) = (P_h\sigma\tilde{u}_y, P_h T_{\lambda,h}\sigma\tilde{u}_y) = (P_h\sigma\tilde{u}_y, T_{\lambda,h}P_h\sigma\tilde{u}_y) = \|T_{\lambda,h}^{1/2}P_h\sigma\tilde{u}_y\|^2,$$

which follows from Lemmas 5.1 and 5.2, and also using (5.8), we get

$$\int_0^H [\|T_{\lambda,h}^{1/2} P_h x \tilde{u}_y\|^2]\mathrm{d}y \leq C \int_0^H [\|T_{\lambda,h}^{1/2} P_h f\|^2 + \|\tilde{u}\|^2]\mathrm{d}y. \tag{5.11}$$

Combining (5.10, 5.11), we get (5.9). From either of these inequalities we obtain the uniqueness, and hence the existence, of a solution of (5.7). □

We now consider error estimates for the finite element approximation. We first make a naive attempt to get an error estimate. Notice that the true solution satisfies a relation analogous to (5.7)

$$\int_0^H [(\sigma u_y, \tilde{w}) + (u_x, \tilde{w}_x) + \lambda(u, \tilde{w})]\mathrm{d}y = \int_0^H (f, \tilde{w})\mathrm{d}y, \quad \forall \tilde{w} \in \mathcal{S}. \tag{5.12}$$

Let $\tilde{e} = u - \tilde{u}$ be the error in the finite element solution. Subtracting (5.12) from (5.7), we get

$$\int_0^H [(\sigma\tilde{e}_y, \tilde{w}) + (\tilde{e}_x, \tilde{w}_x) + \lambda(\tilde{e}, \tilde{w})]\mathrm{d}y = 0, \quad \forall \tilde{w} \in \mathcal{S}. \tag{5.13}$$

Let $\hat{u}$ be an approximation to u in $\mathcal{S}$, such as an interpolant of u or a projection of u onto $\mathcal{S}$ using some inner product. Let $\hat{e} = \tilde{u} - \hat{u}$, so

$$\tilde{e} = (u - \hat{u}) - \hat{e}. \tag{5.14}$$

Inserting this into this (5.12) we then get

$$\int_0^H [(\sigma \hat{e}_y, \tilde{w}) + (\hat{e}_x, \tilde{w}_x) + \lambda(\hat{e}, \tilde{w})]\mathrm{d}y = \int_0^H [(\sigma(u_y - \hat{u}_y), \tilde{w}) + ((u_x - \hat{u}_x), \tilde{w}_x) + \lambda(u - \hat{u}, \tilde{w})]\mathrm{d}y. \tag{5.15}$$

Theorem 5.2. *One has*

$$\|u - \tilde{u}\|_{1,0,\Omega} \leq C \|u - \hat{u}\|_{1,\Omega}.$$

Proof. Setting $\tilde{w} = \hat{e}$ in (5.15) and using (5.8) we obtain

$$\int_\Omega \hat{e}_x^2 \mathrm{d}x\mathrm{d}y \leq \int_\Omega [\sigma \hat{e}(u_y - \hat{u}_y) + \hat{e}_x(u_x - \hat{u}_x) + \lambda \hat{e}(u - \hat{u})]\mathrm{d}x\mathrm{d}y.$$

Hence, using also the Poincaré inequality, $\|\hat{e}\|_{1,0,\Omega}^2 \leq C\|\hat{e}\|_{1,0,\Omega}\|u - \hat{u}\|_{1,\Omega}$, so $\|\hat{e}\|_{1,0,\Omega} \leq C\|u - \hat{u}\|_{1,\Omega}$. Using (5.14) and the triangle inequality, we get the result. □

Theorem 5.2 gives a bound for one portion of the error. It is also of interest to bound the other portion, $\{\int \|\sigma \hat{e}_y(\cdot, y)\|_{-1}^2 \mathrm{d}y\}^{1/2}$. For this we require several lemmas.

Lemma 5.3. *$T_{\lambda,h}^{1/2} P_h$ is a bounded operator on $\mathcal{S}$ with bound independent of h. More precisely:*

$$\int_0^H \|T_{\lambda,h}^{1/2}(f)\|^2 \mathrm{d}y \leq \frac{4L^2}{\pi^2} \int_0^H \|f\|^2 \mathrm{d}y \qquad \forall f \in \mathcal{S} \quad \forall h > 0.$$

Furthermore, $T_{\lambda,h} = (T_{\lambda,h}^{1/2} P_h)^2$.

Proof. Suppose $f \in \mathcal{S}$, then $\|T_{\lambda,h}^{1/2}(f)\|^2 = (T_{\lambda,h} f, f) \leq \|T_{\lambda,h}(f)\| \, \|f\|$. We next use that $\|T_{\lambda,h}(f)\|_0 \leq 2L/\pi |T_{\lambda,h}(f)|_1$ by Poincaré's inequality and then that $|T_{\lambda,h}(f)|_1^2 \leq ((T_{\lambda,h} f)_x, (T_{\lambda,h} f)_x) + \lambda \|T_{\lambda,h}(f)\|_0^2 = (T_{\lambda,h} f, f) \leq \|T_{\lambda,h}(f)\|_0 \, \|f\|_0$. This implies the asserted bound. Since $P_h f = f$ for $f \in \mathcal{S}$, $T_{\lambda,h}^{1/2} P_h$ satisfies the same bound. To connect to P_h, observe that, for $f \in \mathcal{S}^\star$, $T_{\lambda,h}(f) = P_h T_{\lambda,h}(f) = T_{\lambda,h} P_h(f) = T_{\lambda,h}^{1/2}(T_{\lambda,h}^{1/2} P_h(f)) = T_{\lambda,h}^{1/2} P_h(T_{\lambda,h}^{1/2} P_h(f))$. □

Lemma 5.4. *$T_\lambda^{1/2}$ is an isomorphism from $H^{-1}(0, L)$ onto $L^2(0, L)$ with the following equivalence:*

$$\|T_\lambda^{1/2} f\| \approx \|f\|_{-1}.$$

Proof. Suppose $f \in H^{-1}$, then $\|T_\lambda^{1/2}(f)\|^2 = (T_\lambda f, f) \leq \|T_\lambda f\|_1 \|f\|_{-1}$ and, at the same time, $\|T_\lambda^{1/2}(f)\|^2 = (T_\lambda f, f) = ((T_\lambda f)_x, (T_\lambda f)_x) + \lambda(T_\lambda f, T_\lambda f) \geq ((T_\lambda f)_x, (T_\lambda f)_x) = \|T_\lambda f\|_1^2$ so that $\|T_\lambda^{1/2} f\| \leq \|f\|_{-1}$. Conversely,

$$\|f\|_{-1} = \sup_{v \in H_0^1} \frac{(f, v)}{\|v\|_1} = \sup_{v \in H_0^1} \frac{((T_\lambda f)_x, v_x) + \lambda(T_\lambda f, v)}{\|v\|_1} \leq (1 + C\lambda)\|T_\lambda f\|_1,$$

which, together with the above identities, yields the inequality sought. □

We now consider the other error bound. Setting $\tilde{w} = T_{\lambda,h}(\sigma\hat{e}_y)$ in (5.15) and using the identity $(\hat{e}_x, [T_{\lambda,h}(\sigma\hat{e}_y)]_x) + \lambda(\hat{e}, T_{\lambda,h}(\sigma\hat{e}_y)) = (\sigma\hat{e}_y, \hat{e})$, we obtain

$$\int_0^H [(\sigma\hat{e}_y, T_{\lambda,h}(\sigma\hat{e}_y)) + (\sigma\hat{e}_y, \hat{e})]\mathrm{d}y = \int_0^H [(\sigma(u_y - \hat{u}_y), T_{\lambda,h}(\sigma\hat{e}_y)) + (u_x - \hat{u}_x, [T_{\lambda,h}(\sigma\hat{e}_y)]_x) + \lambda(u - \hat{u}, T_{\lambda,h}(\sigma\hat{e}_y))]\mathrm{d}y.$$

Using (5.8),

$$\begin{aligned}\int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma\hat{e}_y)\|^2 \mathrm{d}y \;\le\; & \int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma(u_y - \hat{u}_y))\| \; \|T_{\lambda,h}^{1/2} P_h(\sigma\hat{e}_y)\| \mathrm{d}y + \int_0^H (u_x - \hat{u}_x, [T_{\lambda,h}(\sigma\hat{e}_y)]_x) \\ & + \lambda \int_0^H \|T_{\lambda,h}^{1/2} P_h(u_y - \hat{u}_y))\| \; \|T_{\lambda,h}^{1/2} P_h(\sigma\hat{e}_y)\| \mathrm{d}y + C\|\hat{e}\|_{0,\Omega}^2.\end{aligned}$$

The first and third integrals on the right can be estimated using Schwarz's inequality and the arithmetic-geometric mean inequality, to obtain

$$\int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma\hat{e}_y)\|^2 \mathrm{d}y \le C \int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma(u_y - \hat{u}_y))\|^2 \mathrm{d}y + C \int_0^H (u_x - \hat{u}_x, [T_{\lambda,h}(\sigma\hat{e}_y)]_x) \mathrm{d}y + C\|u - \hat{u}\|_{0,\Omega}^2. \tag{5.16}$$

The difficulty now comes in estimating the second term. Using an inverse inequality we have

$$\begin{aligned}\int_0^H (u_x - \hat{u}_x, [T_{\lambda,h}(\sigma\hat{e}_y)]_x) \mathrm{d}y \;&\le\; \left\{\int_0^H \|u_x - \hat{u}_x\|^2 \mathrm{d}y\right\}^{1/2} \left\{\int_0^H \|[T_{\lambda,h}(\sigma\hat{e}_y)]_x\|^2 \mathrm{d}y\right\}^{1/2} \\ &\le\; Ch^{-1} \left\{\int_0^H \|u_x - \hat{u}_x\|^2 \mathrm{d}y\right\}^{1/2} \left\{\int_0^H \|T_{\lambda,h}(\sigma\hat{e}_y)\|^2 \mathrm{d}y\right\}^{1/2} \\ &\le\; Ch^{-1} \left\{\int_0^H \|u_x - \hat{u}_x\|^2 \mathrm{d}y\right\}^{1/2} \left\{\int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma\hat{e}_y)\|^2 \mathrm{d}y\right\}^{1/2}.\end{aligned}$$

In the last step we have used Lemma 5.3 with bound independent of h. Inserting this into (5.16), we obtain

$$\int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma\hat{e}_y)\|^2 \mathrm{d}y \le C \int_0^H \|T_{\lambda,h}^{1/2} P_h(\sigma(u_y - \hat{u}_y))\|^2 \mathrm{d}y + Ch^{-2} \int_0^H \|u_x - \hat{u}_x\|^2 \mathrm{d}y. \tag{5.17}$$

From (5.14),

$$\int_0^H \|T_\lambda^{1/2}(\sigma\tilde{e}_y)\|^2 \mathrm{d}y \le 2 \int_0^H \|T_\lambda^{1/2}(\sigma\hat{e}_y)\|^2 \mathrm{d}y + 2 \int_0^H \|T_\lambda^{1/2}(\sigma(u_y - \hat{u}_y))\|^2 \mathrm{d}y. \tag{5.18}$$

To bound the first term on the right, we need to relate $T_\lambda^{1/2}$ to $T_{\lambda,h}^{1/2}$:

$$\|T_\lambda^{1/2} g\|^2 = \|T_{\lambda,h}^{1/2} P_h g\|^2 + ((T_\lambda - T_{\lambda,h})g, g), \tag{5.19}$$

as $\|T_\lambda^{1/2}g\|^2 = (T_\lambda g, g)$ and $\|T_{\lambda,h}^{1/2}P_h g\|^2 = (T_{\lambda,h}g, g)$ hold. Since $T_{\lambda,h}g$ is the Galerkin approximation to the function $z = T_\lambda g$ which satisfies $-z_{xx} + \lambda z = g$, $z(0) = z(1) = 0$,

$$\|(T_\lambda - T_{\lambda,h})g\| \leq Ch\|T_\lambda g\|_{H^{1,0}} = Ch\left\{((T_\lambda g)_x, (T_\lambda g)_x) + \lambda(T_\lambda g, T_\lambda g)\right\}^{1/2} = Ch(T_\lambda g, g)^{1/2} = Ch\|T_\lambda^{1/2}g\|. \tag{5.20}$$

Hence we obtain, with the use of the Cauchy-Schwarz inequality,

$$\|T_\lambda^{1/2}g\|^2 \leq \|T_{\lambda,h}^{1/2}P_h g\|^2 + Ch^2\|T_\lambda^{1/2}g\|^2$$

so that for sufficiently small h,

$$\|T_\lambda^{1/2}g\|^2 \leq C\|T_{\lambda,h}^{1/2}P_h g\|^2.$$

Setting $g = \sigma\tilde{e}_y(\cdot, y)$ and integrating over y, we obtain

$$\int_0^H \|T_\lambda^{1/2}\sigma\tilde{e}_y\|^2 \mathrm{d}y \leq C\int_0^H \|T_{\lambda,h}^{1/2}P_h\sigma\tilde{e}_y\|^2 \mathrm{d}y.$$

Using (5.17) to bound the integral on the right, we get

$$\int_0^H \|T_\lambda^{1/2}\sigma\tilde{e}_y\|^2 \mathrm{d}y \leq \int_0^H \|T_{\lambda,h}^{1/2}P_h(\sigma(u_y - \hat{u}_y))\|^2 \mathrm{d}y + Ch^{-2}\int_0^H \|u_x - \hat{u}_x\|^2 \mathrm{d}y.$$

From Lemma 5.3,

$$\int_0^H \|T_\lambda^{1/2}\sigma\tilde{e}_y\|^2 \mathrm{d}y \leq \|(\sigma(u_y - \hat{u}_y))\|_{0,\Omega}^2 + Ch^{-2}\|u_x - \hat{u}_x\|_{0,\Omega}^2.$$

Finally, using Lemma 5.4 we obtain

Theorem 5.3. *One has*

$$\int_0^H \|\sigma\tilde{e}_y(\cdot, y)\|_{-1}^2 \mathrm{d}y \leq \|\sigma(u_y - \hat{u}_y)\|_{0,\Omega}^2 + Ch^{-2}\|u_x - \hat{u}_x\|_{0,\Omega}^2. \tag{5.21}$$

Following this, as a consequence of interpolation estimates, is

Corollary 5.1. *Under the assumptions made above, the following error estimate holds*

$$\|u - \tilde{u}\|_{1,0,\Omega} \leq C(h^p + k^q)\|u\|_{\max\{p,q\}+1,\Omega} \tag{5.22}$$

$$\int_0^H \|\sigma\tilde{e}_y(\cdot, y)\|_{-1}^2 \mathrm{d}y \leq C(h^{p-1} + (k/h)k^{q-1})\|u\|_{\max\{p,q\}+1,\Omega}. \tag{5.23}$$

Proof. The interpolation estimates used are the usual ones, see *e.g.* [9]. □

TABLE 1. Errors and rates of convergence for Aziz-Liu example.

N	$\|(u-\tilde{u})_x\|_{L^2(\Omega)}$	Rate	$\|u-\tilde{u}\|_{L^2(\Omega)}$	Rate
4	0.576(-1)	—	0.192(-2)	—
8	0.144(-1)	2.00	0.245(-3)	2.97
16	0.361(-2)	2.00	0.356(-4)	2.78
32	0.902(-3)	2.00	0.830(-5)	2.10

6. Numerical experiments

In this section we describe some numerical experiments using the numerical method described in Section 5.2. Our accuracy measurements will be compared with the theoretical convergence estimate (5.7). The existence-uniqueness theory of Section 4 applies to all the examples in this section. We also note that assumption (5.1) holds for all the examples in this section except for Example 4.

To form the mesh for this implementation we subdivide Ω into an $N \times N$ grid of rectangles so that $h \approx 1/N$. We take the approximation space $\mathcal{S}$ to be as described in Section 5.2 with $p = q = 2$. Thus, $\mathcal{S}$ is a subspace of $\{u \in H^1(\Omega) \cap C(\Omega) : u = 0 \text{ on } \Gamma_-^H \cup \Gamma_+^0 \cup \Gamma_0\}$. On each mesh rectangle a function in $\mathcal{S}$ has the form

$$a_{00} + a_{10}x + a_{01}y + a_{11}xy + a_{20}x^2 + a_{02}y^2 + a_{21}x^2y + a_{12}xy^2.$$

Using the interpolation results from Section 3.1 of [9], we conclude that

$$\|(u-\tilde{u})_x\|_{L^2(\Omega)} \le Ch^2\|u\|_{H^3(\Omega)}. \tag{6.1}$$

We use 8 nodes on the boundary of each rectangle positioned at the vertices and the midpoints to define the approximation functions uniquely. The matrix and right hand side entries are evaluated with a 3 point Gaussian quadrature.

Aziz-Liu Example. We apply our approximation scheme to Example 1 from [1]. Here $\Omega = (-1,1) \times (0,1)$, $\sigma(x,y) = x, \lambda = 0$, and f is defined so the true solution is

$$u(x,y) = \begin{cases} (x^2-1)y^2[(y-1)^2 - 4x^2] & \text{if } 0 \le x \le 1 \\ (x^2-1)(y-1)^2[y^2 - 4x^2] & \text{if } -1 \le x < 0. \end{cases}$$

Note that $u = 0$ on $\Gamma_+^0 = [0,1] \times \{0\}$ and $\Gamma_-^1 = [-1,0] \times \{1\}$. Also, u is smooth on the line $x = 0$ where the equation changes type from forward to backward. Table 1 has the results for several runs at different values of N.

The header in the last column, *Rate* is the exponent p for the rate of convergence for observed data,

$$\text{error} = Ch^p.$$

It is calculated by the formula

$$p = \frac{\log(e(N_1)) - \log(e(N_2))}{\log(N_2) - \log(N_1)}$$

with the number of subintervals N_1 and N_2 and errors are $e(N_1)$ and $e(N_2)$. We also observed in this example that the convergence rate in $L^2(\Omega)$ was tending to $O(h^2)$ which was also observed in [1].

Example 2. Here we take $\sigma(x,y) = 1 - 2y$ as was discussed in Section 2 (see Fig. 3). The domain in this example is $\Omega = (0,1) \times (0,1)$. The boundary conditions (1.2–1.4) imply that $u = 0$ on the entire boundary.

TABLE 2. Errors and rates of convergence for Example 2.

N	$\|(u-\tilde{u})_x\|_{L^2(\Omega)}$	Rate	$\|u-\tilde{u}\|_{L^2(\Omega)}$	Rate
4	0.814(-1)	—	0.148(-1)	—
8	0.164(-1)	2.31	0.331(-2)	2.16
16	0.382(-2)	2.10	0.811(-3)	2.03
32	0.943(-3)	2.02	0.202(-3)	2.01

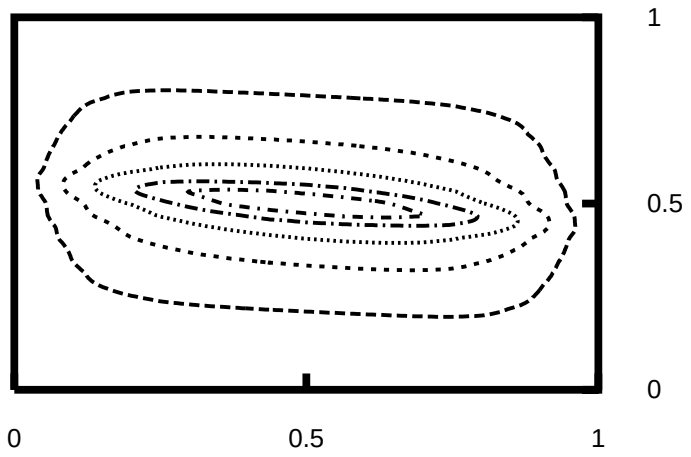

FIGURE 8. Contour lines of the approximation to the solution of the problem with $\sigma(x,y) = 10(3-x-5y)$ and $f(x,y) = 10$ in Example 2. The solution is negative in this case and the value of the function decreases from the outer to inner contours.

We take $f(x,y) = -\pi^2 \sin \pi x$. The true solution can be found by making a change of variables on y and then using separation of variables for the resulting heat equation. With this approach we find

$$u(x,y) = \begin{cases} ((1-2y)^{\pi^2/2} - 1)\sin \pi x & 0 \le x < 1/2 \\ ((2y-1)^{\pi^2/2} - 1)\sin \pi x & 1/2 \le x \le 1. \end{cases}$$

Note that since $\pi^2/2 \approx 4.97$, $u \in C^4(\bar{\Omega})$. Table 2 has the convergence results for this problem.

Again we observe the expected rate of convergence, $O(h^2)$. For this example, the $L^2(\Omega)$ norm was $O(h^2)$.

We also computed an approximation in the related case when the zero line of σ is tilted (see Sect. 2 and Fig. 4). In this computation we multiplied the σ-term by 10 to exaggerate its effects and set $f(x,y) = 10$. Figure 8 has the contour plot of the approximation on a 16×16 grid.

In the next two examples we examine the solution behavior in cases where the true solution is not known.

Example 3. Here we take $\sigma(x,y) = x$ as in the Aziz-Liu example, set $f(x,y) = 1$ and have $\Omega = [-1,1] \times [0,1]$. We required $u = 0$ on $\Gamma_+^0 = [0,1] \times \{0\}$, $\Gamma_-^1 = [-1,0] \times \{1\}$, and, as usual on Γ_0. A 16×16 grid is used. To the right of the y-axis is the forward region with initial data zero and to the left is the backward region with data also zero on the line $y = 1$. It is seen that the differential equation and the boundary conditions are not consistent at the points $(0,0)$ and $(0,1)$. One therefore expects that the solution will have singularities at these points. Some solution contours are shown in Figure 9.

Example 4. Finally we compute the approximation in case when $\sigma(x,y) = 16[(x-1/2)^2 + (y-1/2)^2 - 1/16]$ and $f(x,y) = 1$ which was described in Section 2, see Figure 5. In this case $\Omega = [0,1] \times [0,1]$ and we took $u = 0$

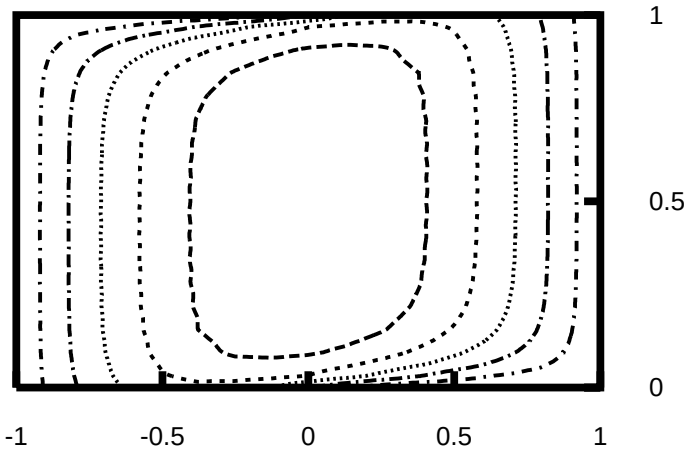

FIGURE 9. Contour lines of the approximation to the solution of the problem with $\sigma(x, y) = x$ and $f(x, y) = 1$ in Example 3. The solution is non-negative and the value of the function increases from the outer to inner contours.

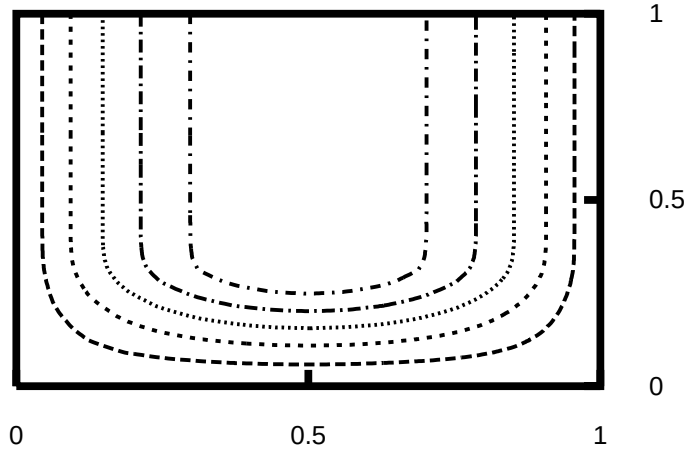

FIGURE 10. Contour lines of the approximation to the solution of the problem with $\sigma(x, y) = 16[(x-1/2)^2 + (y-1/2)^2 - 1/16]$ and $f(x, y) = 1$ in Example 4. The solution is negative again with the inner contour being the minimum.

on $\Gamma_+^0 = [0, 1] \times \{0\}$, and on Γ_0. Here there is a backward region in the shape of a disc that is imbedded in the square. The contours of this are displayed in Figure 10 which was computed on a 16×16 grid.

7. Some regularity conjectures

The regularity behavior of solutions to (1.1) seems to be rather intricate. In a series of papers (see [36, 37]), Pagani investigated these issues and established as an example the following theorem for the case: $\sigma(x, y) = x$.

Theorem 7.1 (Pagani)**.** *Let $\sigma(x, y) = x$. The solution to the system (1.1–1.4) satisfies the following a priori estimate:*

$$\int_0^H \int_0^L \left[|\sigma| u_y^2 + u_{xx}^2\right] \, \mathrm{d}x\mathrm{d}y \leq \int_0^H \int_0^L f^2 \, \mathrm{d}x\mathrm{d}y. \tag{7.1}$$

Gor'kov made various observations about the regularity near the intersection of the zero set of σ and $y = H$ in [17]. Again, for $\sigma(x, y) = x$, he states for the case of (1.1) being posed on the half-plane $y < H$, that if $u(x, H) = \phi(x) = |x|^\delta$ for $x < 0$ – with $0 < \delta < 1/2$, then the form of the solution is:

$$u(x, y) = \overline{\mu}(x, y)(|x|^\delta + |H - y|^{\delta/3}),$$

with $\overline{\mu}$ bounded below and above by positive constants (*cf.* p. 906 of [17]). Here, $f = 0$. Hence, even if we give proper terminal data for $u \in H^{s-1/2}$ — or $|x|^{1/2}u$ in H^s — for $s < \delta + 1$, the regularity pick-up is merely that $u \in H^r$ for $r < \delta/3 + 1$.

The structure of the solution for general σ has not been investigated to our knowledge. To understand some of the possibilities we consider two simple examples. Let $y^* \in (0, T)$ and $\alpha > 0$. The first example is $\sigma(x, y) = \alpha(y - y^*)$. The boundary conditions (1.2–1.4) in this case are $u(0, y) = u(L, y) = 0$. In this case, the variables separate. Expanding u and f in a Fourier sine series, we write

$$u(x, y) = \sum_1^\infty u_n(y) \sin n\pi L^{-1}x, \;\; f(x, y) = \sum_1^\infty f_n(y) \sin n\pi L^{-1}x.$$

One obtains for u_n the ordinary differential equation

$$(y - y^*)u_n'(y) + \beta_n u_n(y) = \alpha^{-1} f_n(y), \beta_n = \alpha^{-1}(k^2 + n^2\pi^2 L^{-2}). \tag{7.2}$$

This equation may be written $\big((y - y^*)^{\beta_n} u_n\big)' = \alpha^{-1}(y - y^*)^{\beta_n - 1} f_n$. Integrating, we have

$$u_n(y) = \alpha^{-1}(y - y^*)^{-\beta_n} \int_{y^*}^y (s - y^*)^{\beta_n - 1} f_n(s)\mathrm{d}s. \tag{7.3}$$

If f_n is smooth, we may expand f_n in a Taylor series around y^* and conclude that u_n is smooth. Since $(y-y^*)^{-\beta_n}$ is the solution of the homogeneous equation associated with (7.2), it follows that (7.3) gives the unique bounded solution of (7.2). Note that we have not imposed any boundary conditions on u_n; the requirement that u_n is bounded plays the role of a boundary condition in fixing the solution. The solution of (7.2) can then be obtained by assembling the sine series with these formulas for u_n. We conclude that, in this case, the regularity of the solution is dictated by the regularity of the function f.

For our second example we take $\sigma(x, y) = -\alpha(y - y^*)$. The boundary conditions (1.2–1.4) in this case are $u(0, y) = u(L, y) = 0$, $u(x, 0) = u(x, H) = 0$. We again expanding u and f in a Fourier sine series. We obtain for u_n the differential equation

$$-(y - y^*)u_n'(y) + \beta_n u_n(y) = \alpha^{-1} f_n(y), \beta_n = \alpha^{-1}(k^2 + n^2\pi^2 L^{-2}). \tag{7.4}$$

This equation may be written $-\big((y^* - y)^{-\beta_n} u_n\big)' = \alpha^{-1}(y^* - y)^{-\beta_n - 1} f_n$. We now integrate the equation and use the boundary conditions to obtain

$$u_n(y) = \alpha^{-1}(y - y^*)^{\beta_n} \int_{y^*}^y (s - y^*)^{-\beta_n - 1} f_n(s)\mathrm{d}s \qquad \text{for } \; y^* < s < 1,$$
$$u_n(y) = \alpha^{-1}(y^* - y)^{\beta_n} \int_{y^*}^y (y^* - s)^{-\beta_n - 1} f_n(s)\mathrm{d}s \qquad \text{for } \; 0 < s < y^*.$$

Thus, a different formula is obtained for u_n in each of the two regions $(0, y^*)$ and (y^*, H). The function u_n satisfies $u_n(0) = u_n(H) = 0$. To understand the behavior of u_n near $y = y^*$ we expand f_n in a Taylor series

around y^* to obtain

$$u_n(y) = \alpha^{-1}(y-y^*)^{\beta_n}\int_{y^*}^{y}(s-y^*)^{-\beta_n-1}[f_n(y^*)+(s-y^*)f_n'(y^*)+\cdots]\mathrm{d}s \quad \text{for } y^* < s < 1,$$

$$u_n(y) = \alpha^{-1}(y^*-y)^{\beta_n}\int_{y^*}^{y}(y^*-s)^{-\beta_n-1}[f_n(y^*)+(s-y^*)f_n'(y^*)+\cdots]\mathrm{d}s \quad \text{for } 0 < s < y^*.$$

We find from these expressions that $\lim_{y\to y^*\pm 0} u_n(u) = \alpha^{-1}f_n(y^*)$. We also find that u_n contains expressions like $(y-y^*)^{\beta_n}$, which are singular if $\beta_n \neq$ integer. (If $\beta_n =$ integer, u_n contains logarithmic terms.) The solution of (7.2) can again be obtained by assembling the sine series with these formulas for u_n. We conclude, in this case, that u is continuous across the line $y = y^*$ but that the regularity of u on $y = y^*$ is limited, regardless of the regularity of f.

Inspired by these examples, we formulate some conjectures for the solution u of (1.1). (1) In the open set where $\sigma \neq 0$, the regularity of the solution at a point (x^*, y^*) is dictated by the regularity of f in a neighborhood of this point. (2) At points (x^*, y^*) where $\sigma(x^*, y^*) = 0$, $\sigma_y(x^*, y^*) > 0$, the regularity of u is dictated by the regularity of f in a neighborhood of (x^*, y^*). (3) At a point (x^*, y^*) where $\sigma(x^*, y^*) = 0$ and $\sigma_y(x^*, y^*) < 0$, u has a singular behavior even if f is smooth in a neighborhood of (x^*, y^*). (4) At points on the boundary where $\sigma = 0$, or at points in the interior where $\sigma = \sigma_y = 0$, the solution u may have a particular singular behavior.

The third author, Søren Jensen, passed away during the course of this work. He was primarily responsible for putting this collaboration together and is greatly missed by all of us. We acknowledge his significant contributions and participation in all aspects of this project except the final editing. Don French's research was supported in part by the Taft foundation at the University of Cincinnati through their Grants-in-aid.

References

[1] A.K. Aziz and J.-L. Liu, A Galerkin method for the forward-backward heat equation. *Math. Comp.* **56** (1991) 35–44.
[2] A.K. Aziz and J.-L. Liu, A weighted least squares method for the backward-forward heat equation. *SIAM J. Numer. Anal.* **28** (1991) 156–167.
[3] M.S. Baouendi and P. Grisvard, Sur une équation d'évolution changeant de type. *J. Funct. Anal.* **2** (1968) 352–367.
[4] R. Beals, On an equation of mixed type from electron scattering. *J. Math. Anal. Appl.* **58** (1977) 32–45.
[5] H.A. Bethe, M.E. Rose and L.P. Smith, The multiple scattering of electrons. *Proc. Amer. Philos. Soc.* **78** (1938) 573–585.
[6] W. Bothe, Die Streuabsorption der Elektronenstrahlen (electron scattering). *Zeit. Physik* **54** (1929) 161–178.
[7] T. Cebeci, H.B. Keller and P.G. Williams, Separating boundary-layer flow calculations. *J. Comput. Physics* (1979) 363–378.
[8] T. Cebeci and K. Stewartson, On the calculation of separation of bubbles. *J. Fluid Mech.* **133** (1983) 287–296.
[9] P.G. Ciarlet, *The Finite Element Methods for Elliptic Problems*. North Holland (1980).
[10] C.N. Dawson, C.J. Van Duijn and R.E. Grundy, Large time asymptotics in contaminant transport in porous media. *SIAM J. Appl. Math.* **56** (1996) 965–993.
[11] J.A. Franklin and E.R. Rodemich, Numerical analysis of an elliptic-parabolic partial differential equation. *SIAM J. Numer. Anal.* **5** (1968) 680–716.
[12] M. Freidlin and H. Weinberger, On a backward-forward parabolic equation and its regularization. *J. Differential Equation* **105** (1993) 264–295.
[13] D.A. French, Continuous Galerkin Finite Element Methods for a Forward-Backward Heat Equation. *Numer. Methods Partial Differential Equations* **15** (1999) 257–265.
[14] D.A. French, Discontinuous Galerkin Finite Element Methods for a Forward-Backward Heat Equation. *Appl. Numer. Math.* **27** (1998) 1–8.
[15] M. Gevrey, Sur certaines équations aux dérivées partielles du type parabolique. *Comptes Rendues* **154** (1912) 1785–1788.
[16] M. Gevrey, Sur les équations aux dérivées partielles du type parabolique, IV. *J. Math. Pures Appl.* **10** (1914) 105–137.
[17] Ju.P. Gor'kov, A formula for the solution of a boundary-value problem for the stationary equation of Brownian motion. *Dokl. Akad. Nauk SSSR* **223** (1975), also *Soviet Math. Dokl.* **16** (1975) 904–908.
[18] P.S. Hagan and J.R. Ockendon, Half-range analysis of a counter-current separator. *J. Math. Anal. Appl.* **160** (1991) 358–378.
[19] E. Hopf, *Mathematical problems of radiative equilibrium*, Cambridge tracts in mathematics and mathematical physics, No. 31. Cambridge University Press (1934).

[20] J.B. Keller and H.F. Weinberger, Boundary and initial boundary-value problems for separable backward-forward parabolic problems. *J. Math. Phys. B* **38** (1997) 4343–4353.
[21] S. Kepinski, Über die Differentialgleichung $\frac{\partial^2 z}{\partial x^2} + \frac{m+1}{x}\frac{\partial z}{\partial x} - \frac{n}{x}\frac{\partial z}{\partial t} = 0$. *Math. Annal.* **61** (1905) 397–405.
[22] J.J. Kohn and L. Nirenberg, Degenerate elliptic-parabolic equations of second order. *Comm. Pure Appl. Math.* **20** (1967) 797–872.
[23] T. LaRosa, *The Propagation of an Electron Beam Through the Solar Corona*. Ph.D. thesis, Physics and Astronomy, University of Maryland, College Park (1986).
[24] T. LaRosa, The spatial structure of a nonthermal electron beam: conditions for stabilization *via* strong turbulence. *Astrophys. J.* **335** (1988) 425–440.
[25] J.-L. Lions, *Équations différentielles opérationelles et problèmes aux limites*. Springer, Berlin (1961).
[26] J.-L. Liu, Weak residual error estimates for symmetric positive systems. *Numer. Funct. Anal. Optim.* **14** (1993) 607–619.
[27] J.-L. Liu, A finite difference method for symmetric positive differential equations. *Math. Comp.* **62** (1994) 105–118.
[28] H. Lu, Galerkin and weighted Galerkin methods for the forward-backward heat equation. *Numer. Math.* **75** (1997) 339–56.
[29] H. Lu and Z.-H. Wen, Solution of a forward-backward heat equation, preprint, Department of Mathematics, University of Nijmegen, NL (1994).
[30] D.B. Melrose, *Plasma astrophysics: non-thermal processes in diffuse magnetized plasmas*. Vol. 1: The emission, absorption, and transfer of waves in plasmas. Vol. 2: Astrophysical application. Gordon and Breach, New York (1978).
[31] A.F. Messiter and R.L. Enlow, A model for laminar boundary-layer flow near a separation point. *SIAM J. Appl. Math.* **25** (1973) 655–670.
[32] W. Myller-Lebedeff, Die Theorie der Integralgleichungen in Anwendung auf einige Reihenentwicklungen. *Math. Annal.* **66** (1910) 388–416.
[33] K. Nickel, Die Prandtlschen Grenzschichtdifferentialgleichungen als asymptotischer Grenzfall der Navier-Stokesschen und der Eulerschen differentialgleichungen. *Arch. Rational Mech. Anal.* **13** (1963) 1–14.
[34] O.A. Oleĭnik, Mathematical problems of boundary layer theory. *Uspekhi Mat. Nauk* **23** (1968) 3–65.
[35] O.A. Oleĭnik and E.V. Radkevič, *Second order equations with nonnegative characteristic form*. Plenum Press, New York-London (1973).
[36] C.M. Pagani, On an initial-boundary value problem for the equation $w_t = w_{xx} - xw_y$. *Ann. Scuola Norm. Sup. Pisa, Ser. IV* **2** (1975) 219–263.
[37] C.M. Pagani, On forward-backward parabolic equations in bounded domains. *Boll. Un. Mat. Ita. B* (5) **13** (1976) 336–354.
[38] W.R.C. Phillips and J.T. Ratnanather, The outer region of a turbulent boundary layer. *Phys. Fluids A, Fluid dynamics* **2** (1990) 427.
[39] J.T. Ratnanather and P.G Daniels, Solution of the thermal boundary layer equations in regions of flow reversal. *SIAM J. Appl. Math.* **55** (1995) 192–204.
[40] L.B. Schiff and J.L. Steger, Numerical simulation of steady supersonic viscous flow. *AIAA J.* **18** (1980) 1421–1430.
[41] F.T. Smith, Steady and unsteady boundary-layer separation. *Annu. Rev. Fluid Mech* **18** (1986) 197–220.
[42] F.T. Smith and P.W. Duck, Separation of jets or thermal boundary layers from a wall. *Quart. J. Mech. Appl. Math.* **30** (1977) 143–156.
[43] K. Stewartson, Multistructural boundary layers on flat plates and related bodies. *Adv. in Appl. Mech.* **14** (1974) 145–239.
[44] K. Stewartson, D'Alembert's paradox. *SIAM Rev.* **23** (1981) 308–343.
[45] K. Stewartson and P.G. Williams, Self-induced separation. *Proc. Royal Soc. Ser. A* **312** (1969) 181–206.
[46] V. Vanaja, Numerical solution of a simple Fokker-Planck equation. *Appl. Numer. Math.* **9** (1992) 533–540.
[47] V. Vanaja and R.B. Kellogg, Iterative methods for a forward-backward heat equation. *SIAM J. Numer. Anal.* **27** (1990) 622–635.
[48] G. Milton Wing, *An introduction to transport theory*. John Wiley and Sons Inc., New York-London (1962).