Optimal filtering for the backward heat equation [1]

Thomas I. Seidman
Department of Mathematics and Statistics
University of Maryland Baltimore County
Baltimore, MD 21228, USA
e-mail: ⟨seidman@math.umbc.edu⟩

ABSTRACT: For the backwards heat equation, stabilized by an *a priori* initial bound, an estimator is determined for intermediate values which is optimal with respect to the bound and the observation accuracy. It is shown how this may be implemented computationally with error estimates for the computed approximation which can be made arbitrarily close to the uncertainty level induced by the ill-posedness of the underlying problem. Thus, the feasibility of this for practical computation, inevitably severely limited by that inherent uncertainty, is as good as possible.

KEY WORDS: *partial differential equation, ill-posed, backwards heat equation, filtering, error estimates.*

AMS SUBJECT CLASSIFICATIONS: 65M30, 35K05, 49E20.

[1] SIAM J. Num. Anal. **33**, pp. 162–170 (1996)

1. INTRODUCTION

One of the classic ill-posed problems — cf., e.g., [6] and numerous further references there — is the *backwards heat equation* (BHE): if we consider the heat equation

$$u_t = \nabla \cdot a\nabla u - qu \quad \text{on} \quad (0,T) \times \Omega \quad \text{with} \quad u\Big|_{\partial\Omega} = 0, \tag{1.1}$$

then rough initial data smooths out as t increases, making the forward solution map $\mathbf{S}(\tau) : u\Big|_{t=0} \mapsto u\Big|_{t=\tau}$ $(\tau > 0)$ compact so the inverse backward map $\mathbf{S}(-\tau) : u\Big|_{t=\tau} \mapsto u\Big|_{t=0}$ (although uniquely determined where it is defined at all) cannot possibly be continuous.[2] On the other hand, it has long been known [8], [4] that this may stabilized by the presence of a suitable *a priori* bound

$$\|u_0(\cdot)\| \leq M_0 \tag{1.2}$$

on the unknown 'initial data' $u_0(\cdot) = u\Big|_{t=0}$. This can be expressed quantitatively [6]: if u_1 and u_2 are two solutions of (1.1), each constrained as in (1.2), for which one has

$$\|u_1(T,\cdot) - u_2(T,\cdot)\| \leq 2\bar{\varepsilon}, \tag{1.3}$$

then one has, for arbitrary $t \in (0,T]$,

$$\|u_1(t,\cdot) - u_2(t,\cdot)\| \leq 2\,[M_0]^{1-t/T}\,[\bar{\varepsilon}]^{t/T}\,. \tag{1.4}$$

We note that the right hand side of (1.4) goes to 0 as the measurement accuracy is improved ($\bar{\varepsilon} \to 0$) so this does mean that one can achieve arbitrary accuracy in determining the unknown $u\Big|_t$ if one will measure with adequate accuracy. Of course, the ill-posedness of the underlying problem is reflected in the rapidity with which this will become infeasible for small t. It is not too difficult to see that this is sharp in that, given $0 < t < T$, there exist sequences $\bar{\varepsilon} \to 0$ and u_1, u_2 satisfying (1.2) and (1.3) for which (1.4) becomes an equality.

In this paper our problem is to determine a good approximation to $u\Big|_t$ (for some given $t > 0$) as an element of $\mathcal{X} = L^2(\Omega)$ for some unknown solution

[2] This is true for a wide variety of topologies but we will here be concerned exclusively with the context of $\mathcal{X} = L^2(\Omega)$.

of (1.1). Available as data is an 'observation' $\bar{v} \approx u\big|_T$ (for some $T > t$) with an estimate

$$\|u(T,\cdot) - \bar{v}\| \leq \bar{\varepsilon} \tag{1.5}$$

for the $\mathcal{X}$-norm of the observation error. We assume known that u is a solution of (1.1) on a time interval including $(0, T)$ with the known bound (1.2). The data for the problem are then $M_0 > 0$, $\bar{\varepsilon} > 0$, and $\bar{v} \in \mathcal{X}$ — and, of course, t, T, and the operator

$$\begin{gathered} \mathbf{A} : u \longmapsto -\nabla \cdot a\nabla u + qu \\ \mathcal{X} \supset \mathcal{D}(\mathbf{A}) = H_0^1(\Omega) \cap H^2(\Omega) \to \mathcal{X}. \end{gathered} \tag{1.6}$$

We assume, of course, that these data are consistent — that there are, indeed, some solutions $u(\cdot)$ satisfying (1.2) and (1.1) for which (1.5) holds.

Our objective is to present an implementable algorithm to determine a computational approximation $\hat{v}(t)$ to the unknown $u(t)$ with an error estimate

$$\|u(t) - \hat{v}(t)\| \leq \hat{\varepsilon}(t) \tag{1.7}$$

reflecting the effects of *both* sources of error/uncertainty: the inherent uncertainty induced by the measurement uncertainty — which is just

$$\varepsilon(t) := [M_0]^{1-t/T}\,[\bar{\varepsilon}]^{t/T} \tag{1.8}$$

since (1.4) is sharp — and the computational error which we must estimate. Our goal is to show that the appropriate use of relatively standard computational techniques enables us to come arbitrarily close to (1.8) in (1.7).

As an intermediate step, we construct a continuous approximation $\mathbf{R} = \mathbf{R}(t)$ to the discontinuous operator $\mathbf{S}(-[T-t])$ in a way which takes advantage of the *a priori* bound (1.2) to achieve the minimal loss of resolution so $v(t) := \mathbf{R}(t)\bar{v}$ is an *optimal* approximation to the unknown $u(t) \in \mathcal{X}$ in terms of the available data:

$$\|u(t,\cdot) - v(t,\cdot)\| \leq \varepsilon(t) \quad \forall u \text{ consistent with the data.} \tag{1.9}$$

From this one actually recovers, somewhat more constructively, the known result (1.4). Our approach, presented in Sections 2 and 3, is closely related to the construction discussed in [10] and, although we will make our presentation here independent of [10], [9], we note that it is an application of a general construction proposed and analyzed in [9] for the quasi-inversion of a C_0 semigroup of compact normal operators on a Hilbert space.

We then continue by describing the computational implementation,[3] leading to the realizable approximation $\hat{v}$ with the estimate (1.7). The analysis is made possible partly by the observation that the general estimates leading to (1.8) also apply, with re-interpretation, to a comparison problem associated with the implementation, and partly to availability in the requisite form of certain auxiliary estimates. This discussion is presented in Sections 4 and 5.

2. FORMULATION

We follow Fourier in treating the heat equation by series expansion. Consider (1.6) as an operator on the Hilbert space $\mathcal{X} = L^2(\Omega)$, assuming $0 < \underline{\alpha} \leq a(\cdot) \leq \overline{\alpha}$ and $q(\cdot) \in L^\infty(\Omega)$; without loss of generality, we also assume $q \geq 0$. We note that $\mathbf{A}$ is then densely defined, self adjoint, and positive definite with compact resolvent so one has an orthonormal basis of eigenfunctions $\{w_k(\cdot)\}$:

$$\mathbf{A}w_k = -\lambda_k w_k \qquad (\langle w_j, w_k\rangle = \delta_{jk}) \tag{2.1}$$

with

$$\lambda_k \in \mathbb{R}_+ \qquad \lambda_k \to +\infty. \tag{2.2}$$

One then has the usual 'forward representation' for solutions of (1.1):

$$u(t,\cdot) = \sum_k c_k e^{-\lambda_k t} w_k(\cdot). \tag{2.3}$$

One has, in particular,

$$u(0,\cdot) = \sum_k c_k w_k(\cdot). \tag{2.4}$$

At $t = T$ we have the expansions

$$u(T,\cdot) = \sum_k \overset{\circ}{\gamma}_k\, w_k(\cdot) \qquad \text{with } \overset{\circ}{\gamma}_k := e^{-\lambda_k T} c_k \tag{2.5}$$

and

$$\sum_k \bar{\gamma}_k w_k(\cdot) = \bar{v} \qquad \text{with } \bar{\gamma}_k := \langle \bar{v}(\cdot), w_k(\cdot)\rangle\ . \tag{2.6}$$

[3] This implementation and the results of some computational experience will be discussed in further detail in [3].

Since $\bar{v}$ is known, the coefficients $\{\bar{\gamma}_k\}$ in (2.6) are also known. Note that, using the orthonormality of $\{w_k\}$, the *a priori* bound (1.2) gives

$$\left[\sum_k |c_k|^2\right]^{1/2} = \|u(0,\cdot)\| \le M_0 \tag{2.7}$$

whereas (1.5) just gives

$$\sum_k \left|\overset{\circ}{\gamma}_k - \bar{\gamma}_k\right|^2 \le \varepsilon_T^2 \,. \tag{2.8}$$

The coefficients $\{c_k\}$ or $\{\overset{\circ}{\gamma}_k\}$ are, of course, otherwise unknown.

From (2.3), we have the 'exact backward representation'

$$u(t) = \sum_k \overset{\circ}{\rho}_k(t)\,\overset{\circ}{\gamma}_k\, w_k = [\mathbf{S}(T-t)]^{-1}u(T) \tag{2.9}$$

with

$$\overset{\circ}{\rho}_k(t) := e^{+\lambda_k(T-t)} = e^{+\lambda_k T(1-\tau)} \qquad \text{where } \tau := t/T. \tag{2.10}$$

If we were to attempt to apply this similarly to (2.6), one would obtain

$$\begin{aligned} &\overset{\circ}{v}(t) := \textstyle\sum_k \overset{\circ}{\rho}_k(t)\bar{\gamma}_k w_k \\ &\qquad\qquad \Rightarrow \quad \|\overset{\circ}{v}(t) - u(t)\|^2 = \textstyle\sum_k \left[\overset{\circ}{\rho}_k(t)\right]^2 \left|\overset{\circ}{\gamma}_k - \bar{\gamma}_k\right|^2 \end{aligned} \tag{2.11}$$

using the orthonormality of $\{w_k\}$. Since (2.2) gives $\overset{\circ}{\rho}_k \to \infty$ as $k \to \infty$ and one only has (2.8) to work with, we see that this $\overset{\circ}{v}(t) = [\mathbf{S}(T-t)]^{-1}\bar{v}$ is a completely useless approximation: its unbounded amplification of the errors of (1.5) is likely to be disastrous. This is the essence of the ill-posedness of the problem.

We must modify the ill-posed procedure leading to (2.11) and construct a bounded quasi-reversal (compare [5]) of $\mathbf{S}(T-t)$ — i.e., $\mathbf{R}(t) \approx [\mathbf{S}(T-t)]^{-1}$. We take this in the form

$$v(t,\cdot) = \mathbf{R}(t)\bar{v} := \sum_k \rho_k(t)\bar{\gamma}_k w_k(\cdot) \tag{2.12}$$

with a suitable choice of factors $\{\rho_k(t)\}$ such that

$$|\rho_k(t)| \le \beta \qquad \text{uniformly in } k \tag{2.13}$$

for some $\beta = \beta(t) \geq 0$. Thus, the 'filtered recovery' operator given by

$$\mathbf{R}(t) : w_k \longmapsto \rho_k(t) w_k \tag{2.14}$$

will be continuous (with $\|\mathbf{R}(t)\| \leq \beta$) and one has a stable error amplification in obtaining $v(t) := \mathbf{R}(t)\bar{v}$ in (2.12).

Here, $\beta(t)$ and $\rho_k(t)$ (suitably related to $\overset{\circ}{\rho}_k\ (t)$, of course) are to be determined so as to minimize $\varepsilon(t)$ in (1.9); this is one rationale for the term 'method of optimal filtering'. We also note that one may think of viewing $\mathbf{R}(t)$ as $\mathbf{S}(-[T-t])\mathbf{F}(t)$ where

$$\mathbf{F}(t) : w_k \mapsto \varphi_k w_k \qquad \text{with } \ \varphi_k := \rho_k(t)/\overset{\circ}{\rho}_k\ (t) \tag{2.15}$$

defines a 'filter' with respect to the spectral decomposition (2.1) with 'form factor' $[k \mapsto \varphi_k]$ — we will have $\rho \leq \overset{\circ}{\rho}$ so $0 \leq \varphi_k \leq 1$. The point of our construction will be to use knowledge of the 'noise–to–signal ratio' $\nu := \bar{\varepsilon}/M_0$, which we view as part of the available data, to make the filtering $\bar{v} \mapsto \mathbf{F}(t)\bar{v}$ produce a minimal loss of resolution at t — i.e., minimal $\varepsilon(t)$ in (1.9) — in this quasi-inversion of $\mathbf{S}$ while damping the observational noise (1.5) in the data to stabilize (2.14).

3. OPTIMAL FILTERING

We will use the forms of (2.12), (2.14) in an estimate of the error (1.9). An inner optimization, subject to (2.13), will then determine each $\rho_k = \rho_k(t)$ and give an error estimate in terms of β as a parameter; a subsequent outer optimization then determines $\beta = \beta(t)$. Each $\rho_k(t)$ is to depend only on T, on $\nu := \bar{\varepsilon}/M_0$, and on the eigenvalue λ_k so the mapping: $\bar{v} \mapsto v(t,\cdot)$ will be linear; each coefficient in the expansion of $v(t)$ will depend on the particular observation $\bar{v}$ through the corresponding coefficient $\bar{\gamma}_k$ appearing in (2.6) but the operator $\mathbf{R}(t)$ is independent of this.

Setting $v(t) := \mathbf{R}(t)\bar{v}$, as above, we decompose $\|u(t) - v(t)\|$ as

$$\begin{aligned} \|u(t) - v(t)\| &= \|[u(t) - \mathbf{R}(t)u(T)] + \mathbf{R}(t)[u(T) - \bar{v}]\| \\ &\leq \|u(t) - \mathbf{R}(t)u(T)\| + \beta(t)\bar{\varepsilon}. \end{aligned} \tag{3.1}$$

We also have, using (2.3), (2.14), etc.,

$$\begin{aligned} \|u(t) - \mathbf{R}(t)u(T)\|^2 &= \textstyle\sum_k \left|\rho_k(t) - \overset{\circ}{\rho}_k\ (t)\right|^2 \left|\overset{\circ}{\gamma}_k\right|^2 \\ &= \textstyle\sum_k \left\{\left|\rho_k(t) - e^{\lambda_k(T-t)}\right| e^{-\lambda_k T}\right\}^2 |c_k|^2 \end{aligned}$$

so, from (2.7), one has

$$\|u(t) - \mathbf{R}(t)u(T)\| \le M_0 \sup_k \left\{ \left| \rho_k(t) - e^{\lambda_k(T-t)} \right| e^{-\lambda_k T} \right\}. \tag{3.2}$$

Assuming, for the inner optimization, that $\beta = \beta(t)$ is already specified as a parameter, we define $\rho = \rho_k(t)$, separately for (each $t \in (0,T)$ and) each $k = 1, 2, \ldots$, by considering the problem (with $\lambda = \lambda_k$)

$$\begin{array}{ll} \text{minimize}_\rho & \left\{ \left| \rho - e^{\lambda(T-t)} \right| e^{-\lambda T} \right\} \\ \text{subject to:} & |\rho| \le \beta. \end{array} \tag{3.3}$$

This minimization is trivial: elementary calculus gives the choice

$$\rho_k(t) := \begin{cases} \overset{\circ}{\rho}_k(t) = e^{\lambda_k(T-t)} & \text{when } \overset{\circ}{\rho}_k(t) \le \beta \\ \beta & \text{when } \overset{\circ}{\rho}_k(t) \ge \beta \end{cases} \qquad k = 1, 2, \ldots, \tag{3.4}$$

i.e., $\rho_k = \min\left\{ e^{\lambda_k(T-t)}, \beta \right\}$. Thus, provided we will have $0 < \beta$,

$$\begin{array}{rl} \varphi_k = 1 & \text{when } \lambda_k \le \frac{\log \beta}{T-t} \\ 0 < \varphi_k < 1 & \text{else: } \rho_k = \beta < \overset{\circ}{\rho}_k . \end{array}$$

Using (3.4) in (3.2), we have

$$\begin{array}{rcl} \|u(t) - \mathbf{R}(t)u(T)\| & \le & M_0 \sup_k \left\{ \left[e^{-\lambda_k t} - \beta e^{-\lambda_k T} \right] : e^{\lambda_k(T-t)} > \beta \right\} \\ & \le & M_0 \sup_\lambda \left\{ [e^{-\lambda t} - \beta e^{-\lambda T}] : \lambda \in \mathbb{R} \right\} \\ & = & M_0 (1-\tau)[\tau/\beta]^{\tau/(1-\tau)} \qquad (\tau := t/T), \end{array} \tag{3.5}$$

by elementary calculus. Using this in (3.1) gives

$$\|u(t) - v(t)\| \le M_0 \left\{ (1-\tau)[\tau/\beta]^{\tau/(1-\tau)} + \nu\beta \right\} \tag{3.6}$$

with $\nu := \bar{\varepsilon}/M_0$. At this point we must do the outer optimization: choose $\beta = \beta(t)$ so as to minimize the right-hand side of (3.6), noting that all other ingredients there are known. Thus, dividing by M_0, we consider the problem

$$\text{minimize}_\beta \left\{ \mu\beta^{-s} + \nu\beta : \beta > 0 \right\} \quad \left(s := \frac{\tau}{1-\tau}, \; \mu := (1-\tau)\tau^s \right). \tag{3.7}$$

By elementary calculus once again, we obtain the optimal choice

$$\beta = \beta(t) := (1/\nu)^{1-\tau}\,\tau \qquad (\tau := t/T), \tag{3.8}$$

which completes the construction.

With the aid of the estimates above, we have:

THEOREM 1: *Let $u(\cdot)$ be any solution of (1.1) satisfying (1.5) and (1.2). Then one has*

$$\|u(t) - v(t)\| \le \varepsilon(t) := [M_0]^{1-t/T}\,[\bar\varepsilon]^{t/T} \quad (0 \le t \le T) \tag{3.9}$$

with the approximation $v(t) = \mathbf{R}(t)\bar v$ to $u(t)$ defined as

$$\begin{array}{l} v(t,\cdot) := \displaystyle\sum_{k=1}^{\infty} \rho_k(t)\bar\gamma_k w_k(\cdot) \qquad \textit{where} \\ \left\{\begin{array}{rcl} \rho_k(t) &:=& \min\left\{e^{\lambda_k(T-t)}, \frac{t}{T}\left(\frac{M_0}{\bar\varepsilon}\right)^{1-t/T}\right\} \\ \bar\gamma_k &:=& \langle \bar v, w_k\rangle \end{array}\right. \qquad \textit{for } k = 1,2,\ldots \end{array} \tag{3.10}$$

Proof: Use (3.4), (3.8) in (2.14), (2.6) and then (3.8) in (3.6). ■

We note that at T we have $\rho_k(T) = 1$ for each k so that (2.6), (3.10) give $v(T) = \bar v$ consistent with (3.9) giving $\varepsilon(T) = \bar\varepsilon$; similarly, at 0 we see that $\varepsilon(0) = M_0$ with $v(0) = 0$. The estimate (3.9) stabilizes for $0 < t < T$ in that the error bound goes to 0 as $\bar\varepsilon \to 0$, but is not directly[4] useful at $t = 0$. We observe also that (3.9) subsumes the previously known estimate (1.4) by taking $\bar v = [u_1(T) + u_2(T)]/2$.

4. COMPUTATIONAL IMPLEMENTATION

We now wish to consider the computational implementation of (3.10) in somewhat greater detail: effectively, we must obtain approximations to the eigenpairs $\{\lambda_k, w_k(\cdot)\}$ with a truncation to $\{k \le K\}$ for some finite K. To obtain explicit direct estimates of the approximation errors and then

[4] One can, in fact, approximate u_0 itself by a slight re-interpretation: Consider a sequence of increasingly accurate measurements $\bar v_n$ with $\bar\varepsilon = \bar\varepsilon_n \to 0$ and let v_n^* be obtained as $v(t_n)$ in (3.10), using $\bar v_n, \bar\varepsilon_n$ and times $t_n \to 0$ chosen so $t_n \log\bar\varepsilon_n \to -\infty$. Then $v_n^* \to u_0$ since $u(t_n) \to u_0$ by definition and (3.9) gives $\|u(t_n) - v_n^*\| \to 0$. This gives no convergence rate, which would only be available with stronger *a priori* information about u_0.

to estimate the propagated effects of these on the computation of $v(t,\cdot)$ is prohibitively complicated for analysis and gives an unrealistically poor estimate.

It is much better to proceed from a different viewpoint, noting that a quite standard procedure for approximately computing $\{\lambda_k, w_k(\cdot)\}$ would consist of selecting a finite element basis to define an approximation $\tilde{\mathbf{A}}$ to the operator $\mathbf{A}$ of (1.6), realized in terms of the selected basis by a matrix A, and then using some algorithm to compute the eigenpairs for A approximately for use in (3.10). We artificially introduce, as a comparison problem, the Galerkin approximation to (1.1) using $\tilde{\mathbf{A}}$ for which we can take advantage of available error estimates; we will also use backward error estimates [12] for the spectral computation. Some preliminary comments are in order before we see how to use these estimates.

Letting $\tilde{\mathcal{X}}$ be the finite element subspace of $\mathcal{X} = L^2(\Omega)$ and $\tilde{\mathbf{P}}$ be the orthonormal projection on $\mathcal{X}$ to $\tilde{\mathcal{X}}$, one sets $\tilde{\mathbf{A}} := \tilde{\mathbf{P}}\mathbf{A}\tilde{\mathbf{P}}$ which may be viewed alternatively as an operator on $\mathcal{X}$ or on $\tilde{\mathcal{X}}$, as convenient. The product $\tilde{\mathbf{P}}\mathbf{A}\tilde{\mathbf{P}}$ makes sense directly if one has $\tilde{\mathcal{X}} = \mathcal{R}(\tilde{\mathbf{P}}) \subset \mathcal{D}(\mathbf{A})$ and this can be extended[5] to considering $\tilde{\mathcal{X}} \subset H_0^1(\Omega)$. The Galerkin approximation is now obtained as an ordinary differential equation on $\tilde{\mathcal{X}}$ (i.e., on $\mathbb{R}^K$ by parametrization):

$$\tilde{u}^{\cdot} = \tilde{\mathbf{A}}\tilde{u}, \qquad \tilde{u}(0) = \tilde{u}_0 := \tilde{\mathbf{P}}u(0) \tag{4.3}$$

so $\tilde{u}(t) = \tilde{\mathbf{S}}(t)u(0)$ where $\tilde{\mathbf{S}}(t) = e^{t\tilde{\mathbf{A}}}$ is the semigroup generated by $\tilde{\mathbf{A}}$. For our purposes we will need an estimate of the form

$$\|\tilde{u}(t) - u(t)\| \leq \delta_f(t)\|u(0)\| \tag{4.4}$$

where $\delta_f = \delta_f(t;\tilde{\mathcal{X}})$ is expected to go to 0 as $\tilde{\mathcal{X}} \to \mathcal{X}$, i.e., as one refines the discretization. This is slightly non-standard in its use of L^2-norms throughout

[5] $\mathbf{A}$ extends as a continuous (coercive, self adjoint) operator: $H_0^1(\Omega) \to H^{-1}(\Omega)$ and for $\tilde{\mathcal{X}} \subset H_0^1(\Omega)$ we may consider

$$\begin{aligned} \tilde{\mathbf{P}}: &\quad \mathcal{X} \to \tilde{\mathcal{X}} \hookrightarrow H_0^1 \\ \tilde{\mathbf{P}} = \tilde{\mathbf{P}}^*: &\quad H^{-1} \to \tilde{\mathcal{X}}^* = \tilde{\mathcal{X}} \hookrightarrow \mathcal{X} \end{aligned} \tag{4.1}$$

and, again, the product $\tilde{\mathbf{P}}\mathbf{A}\tilde{\mathbf{P}}$ makes sense and represents a continuous self adjoint positive operator on $\mathcal{X}$ or on $\tilde{\mathcal{X}}$. The Galerkin approximation for (1.1) is then usually formulated as

$$\langle \xi, \dot{U}\rangle = \langle \xi, \mathbf{A}U\rangle \quad \forall \xi \in \tilde{\mathcal{X}} \quad \text{with } U(t) \in \tilde{\mathcal{X}} \tag{4.2}$$

taking $U(0) = \tilde{\mathbf{P}}u(0)$ and (4.1) shows that this is just (4.3).

but we note, e.g., that [1] (see also [11]) gives an estimate of the requisite form (4.4) with

$$\delta_f(t : \tilde{\mathcal{X}}) = C_r h^r t^{-r/2} \tag{4.5}$$

for certain finite element schemes where $h \approx$ [mesh scale] → 0 in refinement; this is closely related to approximation results for $H^r(\Omega)$, noting the equivalence $\|\mathbf{A}^{r/2}\xi\| \approx \|\xi\|_{H^r(\Omega)}$ and the inequality $\|\mathbf{A}^{r/2}\mathbf{S}(t)\| \leq Ct^{-r/2}$ since $\mathbf{S}(\cdot)$ is an analytic semigroup.

The approximation (4.3) above is usually constructed in terms of a specific finite element basis $\tilde{\mathcal{E}} = \{\tilde{e}_1, \ldots, \tilde{e}_K\}$ for $\tilde{\mathcal{X}}$ so $\tilde{\mathbf{A}}$ is represented by a (sparse) symmetric positive $K \times K$ matrix A. To approximate the (first K) eigenpairs of $\mathbf{A}$, one computes those of this matrix A and then maps the eigenvectors from $\mathbb{R}^K$ to $\tilde{\mathcal{X}}$ using the given basis $\tilde{\mathcal{E}}$. In computing the eigenpairs for A one must cope with round-off and propagated error effects, especially for large K, so this computation is not exact. A typical form for the error analysis here (compare [12], which also suggests that one may reasonably hope[6] $\|B\| = \mathcal{O}(K)\delta_p$ with δ_p giving the numerical precision) is particularly useful for our purposes: the computed approximations to the eigenpairs of A are exactly those of a perturbed matrix $[A + B]$ with the matrix B 'small'. This perturbation may, of course, be taken as including also the inaccuracies in computational specification of the entries $\langle \tilde{e}_j, \mathbf{A}\tilde{e}_k\rangle$ of A due to numerical quadrature, etc. The perturbation B of A corresponds to a perturbation $\mathbf{B}$ of $\tilde{\mathbf{A}}$ — in fact, one has $\mathbf{B} = \tilde{\mathbf{E}}^{-1}B\tilde{\mathbf{E}}$ where $\tilde{\mathbf{E}} : \tilde{\mathcal{X}} \to \mathbb{R}^K$ is the coordinate map for the basis $\tilde{\mathcal{E}}$. Typically $\tilde{\mathcal{E}}$ will not be orthonormal but is 'almost' so (due to controlled overlap of supports of the basis elements) so that $\tilde{\mathbf{E}}$ has a condition number of (uniformly) moderate size. Thus, any available estimate for B may reasonably be translated into an estimate

$$\|\mathbf{B}\| \leq \delta_s \tag{4.6}$$

in terms of the $\tilde{\mathcal{X}}$-norm (i.e., the $L^2(\Omega)$-norm) so the eigenpairs $\{(\hat{\lambda}_k, \hat{w}_k) : k = 1, \ldots, K\}$ which we will actually use are exactly those of $\hat{\mathbf{A}} := [\tilde{\mathbf{A}} + \mathbf{B}]$ with $\mathbf{B}$ satisfying (4.6).

For simplicity, we ignore, as an additional source of error, the implemented computation of $\hat{\gamma}_k := \langle \bar{v}, \hat{w}_k\rangle$ and assume either that this is exact or that the quadrature method used to obtain $\{\hat{\gamma}_k\}$ is embedded in the specifica-

[6] E.g., using LAPACK [2].

tion of the finite element method, so that its treatment is already subsumed by the error analysis provided above.

It is clear that the numbers δ_f and δ_s (respectively estimating the accuracy of the finite element and spectral computations) can, in principle, be made arbitrarily small: δ_f is made small by, e.g., refining the mesh (h small in (4.5)) at the expense of having large K and δ_s is then made small by using higher precision, etc., for the spectral computation. While this always works in principle, we do remark that the form of the (sharp) estimate (1.8) makes it clear that this will become infeasible quite rapidly for small $t > 0$. This is the inevitable price associated with ill-posedness.

5. COMPOSITE ERROR ESTIMATES

Given $0 < t < T$, our *computed* approximation to $u(t)$, replacing that of (3.10), will actually be

$$\begin{aligned} &\hat{v}(t,\cdot) := \sum_{k=1}^{K} \hat{\rho}_k(t)\hat{\gamma}_k\hat{w}_k(\cdot) \qquad \text{where} \\ (5.1)\qquad &\left\{ \begin{array}{rcl} \hat{\rho}_k(t) & := & \min\left\{ e^{\hat{\lambda}_k(T-t)}, \frac{t}{T}\left(\frac{M_0}{\hat{\varepsilon}}\right)^{1-t/T} \right\} \\ \hat{\gamma}_k & := & \langle \bar{v}, \hat{w}_k \rangle \end{array} \right. \quad \text{for } k = 1,\dots,K \end{aligned}$$

with $\hat{\varepsilon}$ yet to be specified. We now wish to estimate $\|u(t) - \hat{v}(t)\|$ for $\hat{v}$ given by (5.1).

We begin with the important observation that the *form* of (5.1) is exactly that of the previous approximation formula (3.10) — the only differences are the replacements of the operator $\mathbf{A}$ on $\mathcal{X}$ by the new operator $\hat{\mathbf{A}}$ on $\tilde{\mathcal{X}}$, of $\bar{\varepsilon}$ by $\hat{\varepsilon}$, and of $\bar{v}$ by $\tilde{\mathbf{P}}\bar{v} = \hat{v}(T)$ (noting that only $\tilde{\mathbf{P}}\bar{v}$ is needed to obtain $\{\hat{\gamma}_k\}$ in (5.1)). Thus the entire analysis of Section 3 again applies here to the comparison problem: If $\hat{u}(\cdot)$ is any solution of

$$(5.2)\qquad \ddot{\hat{u}} = \hat{\mathbf{A}}\hat{u} = \tilde{\mathbf{A}}\hat{u} + \mathbf{B}\hat{u} \qquad \text{with } \|u(0)\| \le M_0,$$

and if we have at T the estimate

$$(5.3)\qquad \|\hat{u}(T) - \tilde{\mathbf{P}}\bar{v}\| \le \hat{\varepsilon},$$

then the computed approximation $\hat{v}(\cdot)$, given by (5.1), satisfies the error estimate

$$(5.4)\qquad \|\hat{u}(t) - \hat{v}(t)\| \le [M_0]^{1-t/T}\,[\hat{\varepsilon}]^{t/T}$$

for $0 \leq t \leq T$.

In terms of the unknown initial data u_0 for (1.1), we will take

$$\tilde{u}_0 := \tilde{\mathbf{P}} u_0, \qquad \tilde{u}(t) := \tilde{\mathbf{S}}(t)\tilde{u}_0 \qquad \hat{u}(t) := \hat{\mathbf{S}}(t)\tilde{u}_0\ . \tag{5.5}$$

Note that (1.2) gives $\|\tilde{u}_0\| \leq M_0$ so $\hat{u}(\cdot)$ is a solution of (5.2). Further, since $\mathbf{A}$, hence also $\tilde{\mathbf{A}}$, is positive, one has $\|\tilde{\mathbf{S}}(t)\| \leq 1$ so $\|\tilde{u}(t)\| \leq M_0$. The 'variation of parameters' formula for (5.2) in terms of $\tilde{\mathbf{S}}(\cdot)$ gives

$$\begin{aligned}\hat{u}(t) &= \tilde{\mathbf{S}}(t)\hat{u}(0) + \textstyle\int_0^t \tilde{\mathbf{S}}(t-s)\mathbf{B}\hat{u}(s)\,ds \\ &= \tilde{u}(t) + \textstyle\int_0^t \tilde{\mathbf{S}}(t-s)\mathbf{B}\tilde{u}(s)\,ds + \int_0^t \tilde{\mathbf{S}}(t-s)\mathbf{B}[\hat{u}(s) - \tilde{u}(s)]\,ds\end{aligned}$$

whence, using (4.6),

$$\begin{aligned}&\|\hat{u}(t) - \tilde{u}(t)\| \\ &\qquad \leq \textstyle\int_0^t \|\tilde{\mathbf{S}}(t-s)\|\|\mathbf{B}\|\|\tilde{u}(s)\|\,ds + \int_0^t \|\tilde{\mathbf{S}}(t-s)\|\|\mathbf{B}\|\|\hat{u}(s) - \tilde{u}(s)\|\,ds \\ &\qquad \leq M_0\delta_s t + \delta_s \textstyle\int_0^t \|\hat{u}(s) - \tilde{u}(s)\|\,ds.\end{aligned}$$

Using an argument much as for the Gronwall inequality, one sees that this integral inequality bounds $\|\hat{u}(t) - \tilde{u}(t)\|$ by the solution for the equality $[\eta(t) = M_0\delta_s t + \delta_s \int_0^t \eta]$, i.e.,

$$\|\hat{u}(t) - \tilde{u}(t)\| \leq M_0 \left[e^{\delta_s t} - 1\right]. \tag{5.6}$$

Combining (5.6) with (4.4) then gives

$$\|\hat{u}(t) - u(t)\| \leq \delta_*(t) := M_0 \left(\left[e^{\delta_s t} - 1\right] + \delta_f(t)\right). \tag{5.7}$$

Note that for each $t \in (0, T]$ one has $\delta_*(t) \to 0$ as $\delta_s \to 0$ and $\delta_f(t) \to 0$.

Now consider (5.3). From (5.7) at T one has

$$\begin{aligned}\|\hat{u}(T) - \hat{v}(T)\| &= \|\tilde{\mathbf{P}}[\hat{u}(T) - \bar{v}]\| \\ &\leq \|\hat{u}(T) - u(T)\| + \|u(T) - \bar{v}\| \leq \delta_*(T) + \bar{\varepsilon}\end{aligned}$$

and we obtain (5.3) on taking

$$\hat{\varepsilon} := \bar{\varepsilon} + M_0 \left(\left[e^{\delta_s t} - 1\right] + \delta_f(t)\right) \tag{5.8}$$

for use both in (5.1) and in (5.4).

THEOREM 2: *Let $u(\cdot)$ be any solution of (1.1) satisfying (1.5) and (1.2). Let the approximation $\hat{v}(t)$ to $u(t)$ be defined by (5.1) with $\hat{\varepsilon}$ given by (5.8) and with $\{(\hat{\lambda}_k, \hat{w}_k\}$ the eigenpairs of $\hat{\mathbf{A}} := [\tilde{\mathbf{A}} + \mathbf{B}]$ where $\tilde{\mathbf{A}}$ is such as to give[7] (4.4), i.e.,*

$$\left\| e^{t\tilde{\mathbf{A}}} - e^{t\mathbf{A}} \right\| \leq \delta_f(t) \tag{5.9}$$

and the perturbation $\mathbf{B}$ satisfies (4.6). Then one has

$$\|u(t) - \hat{v}(t)\| \leq \delta_*(t) + [M_0]^{1-t/T} \left[\bar{\varepsilon} + \delta_*(T)\right]^{t/T} \tag{5.10}$$

with δ_* as in (5.7).

PROOF: Combine (5.4) and (5.8) with (5.7). ■

ACKNOWLEDGMENTS: Some support for this research has been provided by the NSF under ECS-8814788 and by the AFOSR under AFOSR-91-0008. I would like to thank R. Ewing for a number of especially stimulating discussions about this problem and to thank J. Demmel and J. Dongarra for [2]. Thanks also are due to J. Cannon, L. Eldén, T. Janik, and W. Rundell.

References

[1] J.H. Bramble, A.H. Schatz, V. Thomée, and L.B. Wahlbin, *Some convergence estimates for Galerkin type approximations for parabolic equations,* SIAM J Numer Anal. **14** (1977), pp. 218–241.

[2] J. Demmel, J. Dongarra, personal communications (also referring to Ch. 4 of the *LAPACK User's Guide*, SIAM, 1992).

[3] T. Janik and T.I. Seidman, *Computational aspects of the method of 'Optimal Filtering'*, in preparation.

[4] F. John, *Continuous dependence on data for solutions with a prescribed bound*, Comm. Pure Appl. Math. **13** (1960), pp. 551–585.

[7] We remark here that the argument is also valid in contexts in which $\tilde{\mathbf{A}}$, in (5.9), does not arise from Galerkin approximation — e.g., one could at this point be considering perturbation of the coefficients of (1.1) or even (compare [7]) perturbation of the region Ω.

[5] R. Lattes and J.-L. Lions, *Méthode de quasi-réversibilité et applications,* Dunod, Paris, 1967.

[6] L.E. Payne, *Improperly posed Problems in Partial Differential Equations,* SIAM, Philadelphia, 1975.

[7] L.E. Payne, *On stabilizing ill-posed problems against errors in geometry and modeling,* in *Inverse and Ill-Posed Problems,* (Heinz W. Engl and C.W. Groetsch, eds.), Academic Press, Orlando, 1987.

[8] C. Pucci, *Sui problemi Cauchy non "ben posti"*, Atti Acad. Naz. Lincei (Rend. Cl. Sc. fis. mat. et natur.) **18** (1955), pp. 473–477.

[9] T.I. Seidman, *'Optimal Filtering' for some ill–posed problems,* in *Wave Propagation and Inversion* (W. Fitzgibbon and M. Wheeler, eds.), SIAM (1992), pp. 108–123.

[10] T.I. Seidman and L. Eldén, *An 'Optimal Filtering' method for the sideways heat equation*, Inverse Problems **6** (1990), pp. 681–696.

[11] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Mathematics **#1054**, Springer, New York, 1984.

[12] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.