ERROR ANALYSIS OF THE FINITE-STRIP METHOD FOR PARABOLIC EQUATIONS

Stanley S. Smith Myron B. Allen

1993
Journal Article WWRC-93-24

In

Numerical Methods for Partial
Differential Equations

Stanley S. Smith
Department of Mathematics
Black Hills State University
Spearfish, South Dakota

Myron B. Allen
Department of Mathematics
University of Wyoming
Laramie, Wyoming

Error Analysis of the Finite-Strip Method for Parabolic Equations

Stanley S. Smith
Department of Mathematics, Black Hills State University, Spearfish, South Dakota 57799

Myron B. Allen
Department of Mathematics, University of Wyoming, Laramie, Wyoming 82071

Received 30 October 1992; revised manuscript received 13 February 1993

The finite-strip method (FSM) is a hybrid technique which combines spectral and finite-element methods. Finite-element approximations are made for each mode of a finite Fourier series expansion. The Galerkin formulated method is set apart from other weighted-residual techniques by the selection of two types of basis functions, a piecewise linear interpolating function and a trigonometric function. The efficiency of the FSM is due in part to the orthogonality of the complex exponential basis: The linear system which results from the weak formulation is decoupled into several smaller systems, each of which may be solved independently. An error analysis for the FSM applied to time-dependent, parabolic partial differential equations indicates the numerical solution error is $O(h^2 + M^{-r})$. M represents the Fourier truncation mode number and h represents the finite-element grid mesh. The exponent $r \geq 2$ increases with the exact solution smoothness in the respective dimension. This error estimate is verified computationally. Extending the result to the finite-layer method, where a two-dimensional trigonometric basis is used, the numerical solution error is $O(h^2 + M^{-r} + N^{-q})$. The N and q represent the truncation mode number and degree of exact solution smoothness in the additional dimension.

I. INTRODUCTION

The finite-strip method (FSM) is a hybrid of the finite-element and spectral methods. Its typical applications are in the numerical solution of partial differential equations in two spatial variables, especially in problems that are geometrically regular in one coordinate direction. Owing to its unusual efficiency, the technique is a familiar one in structural mechanics [1]. It is also useful in models of stratified groundwater flow [2, 3]. A three-dimensional extension of the method, the finite-layer method, has utility in groundwater-flow models [4, 5] as well as in other applications. This paper presents an error analysis for the FSM applied to time-dependent, parabolic partial differential equations. We also indicate how to extend the analysis to the finite-layer method.

The FSM generates an approximate solution that, at each time level, belongs to a peculiar finite-element trial space. This space consists of functions that are piecewise polynomial in the z direction and are truncated Fourier series in the x direction. The space has a

Numerical Methods for Partial Differential Equations, 9, 667–690 (1993)

CCC 0749-159X/93/060667-24

tensor-product basis, each element of which is a product of two types of one-dimensional basis functions. The first type is associated with traditional finite-element techniques. We partition the z dimension of the spatial domain by a grid and define piecewise polynomials, such as standard piecewise linear basis functions $l_j(z)$, over the grid. The basis functions used for the x dimension are the trigonometric functions associated with spectral methods [6]. If $\omega_m(x)$ represents a typical element of the trigonometric basis, indexed by the Fourier mode number m, then a typical basis function of the trial space for the FSM has the form $\omega_m(x)l_j(z)$. Section III discusses this basis in more depth.

We discretize a given initial-boundary-value problem in space by using a Galerkin formulation [7] in which basis functions $\omega_m(x)l_j(z)$ serve as weight functions in the weighted-residual equations. We discretize in time using finite differences. Section IV outlines this formulation in more detail.

In problems having sufficient geometric regularity, the FSM has several computational advantages over traditional finite-element and spectral methods. Chief among these is the fact that it yields a sparse linear system to solve for each Fourier mode of the approximate solution. As discussed briefly in Sec. IV, the matrix equations for different modes are independent and therefore are amenable to parallel processing. Several other papers [2, 4, 5] discuss such computational matters in detail. This paper focuses on the analysis of the FSM.

The key question in the error analysis is the following: How does the error in the FSM solution decay as we refine the mesh size h of the finite-element grid in z or increase the number $2M + 1$ of Fourier modes used in x? Our development shows that, when the trial function is piecewise linear in z, the FSM error is $O(h^2 + M^{-r})$. Here, the exponent $r \geq 2$ increases with the smoothness of the exact solution in the x direction.

Our paper is organized as follows. Section II describes the physical problem of interest and the mathematical assumptions and notation. Section III discusses the FSM trial space, and Sec. IV describes the FSM formulation. Section V estimates the approximation error associated with interpolation and projection maps into the trial space. Using these estimates, Sec. VI derives an L^2 estimate of the difference between the approximate FSM solution and the exact solution. This error estimate is then verified computationally in Sec. VII. In Sec. VIII we sketch the extension of the analysis to the finite-layer method.

II. PHYSICAL PROBLEM AND NOTATION

Our analysis involves a two-dimensional generalization of the heat equation. Consider a rectangular spatial domain $\Omega := (-\pi, \pi) \times (0, 1)$ with homogeneous Dirichlet boundary conditions and coefficients that vary with z:

$$\begin{aligned} S(z)\partial_t u - K_x(z)\partial_x^2 u - \partial_z[K_z(z)\partial_z u] &= f(x,z,t), \qquad (x,y) \in \Omega,\ t \in (0,T], \\ u(x,z,t) &= 0, \qquad (x,z) \in \partial\Omega,\ t \in [0,T], \\ u(x,z,0) &= u^0(x,z), \qquad (x,z) \in \Omega. \end{aligned} \tag{1}$$

Here, $\partial_x u := \partial u/\partial x$, $\partial_x^2 u := \partial^2 u/\partial x^2$, and so forth. We adopt the following notation to describe the spatial domain: $X := (-\pi, \pi)$; $Z := (0, 1)$; $\Omega := X \times Z$. Also, $\partial\Omega$ denotes the boundary of Ω.

The problem (1) occurs in several applications. In two-dimensional saturated groundwater flow, the coefficient $S(z)$ represents specific storage. The coefficients $K_x(z)$ and $K_z(z)$ in this context denote hydraulic conductivities in the x and z directions, respectively. Huyakorn and Pinder [8], for example, discuss this application in detail. All three

coefficients may vary with the vertical coordinate z, as occurs in horizontally uniform sedimentary beds. The function $f(x,z,t)$ accounts for sources and $u(x,z,t)$ represents the unknown hydraulic head. The boundary-value problem (1) also has applications to conductive heat flow. For a two-dimensional, layered composite slab, $S(z) = 1.0$; $K_x(z)$ and $K_z(z)$ stand for thermal diffusivities, and $u(x,z,t)$ represents temperature. In realistic problems, it is generally necessary to rescale the domain $\Omega = (-\pi, \pi) \times (0, 1)$ to physical dimensions. Linear scalings may change the multiplicative constants in our error estimates but do not affect their asymptotic orders.

We assume that K_x and K_z, and S are piecewise constant with respect to z. We also assume that they are positive, bounded away from zero, and bounded above:

$$0 < c \leq K_x(z) \leq K, \tag{2}$$

$$0 < c \leq K_z(z) \leq K, \tag{3}$$

$$0 < s \leq S(z) \leq S^*. \tag{4}$$

We assume that the forcing function f and the initial condition u^0 are smooth enough to guarantee that the solution $u(x,z,t)$ exists, is unique, and depends continuously on these data.

We use a variety of normed function spaces in our analysis. Denote by $L^2(\Omega)$ the space of square-integrable, complex-valued functions defined on Ω. The quantity

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega |v|^2 \, dx \, dz \tag{5}$$

defines the standard norm on this space. Here, $|v(x,z)|^2 := v(x,z)\overline{v(x,z)}$, the overbar indicating complex conjugation. We use analogous notation for the one-dimensional domains X and Z. For example, the space of square-integrable functions on X is $L^2(X)$, and the corresponding norm is

$$\|v\|^2_{L^2(X)} = \int_X |v|^2 \, dx. \tag{6}$$

Given $v \in L^2(\Omega)$, $v(x,\cdot)$ represents a family of functions in $L^2(Z)$ (that is, functions of z), where x is a parameter. Similarly, $v(\cdot,z)$ represents a family of functions in $L^2(X)$ indexed by the parameter z. Thus $\|v(x,\cdot)\|_{L^2(Z)}$ represents a function in $L^2(X)$. We sometimes abbreviate this function by writing $\|v\|_{L^2(Z)}$. Likewise, when $v \in L^2(\Omega)$, $\|v\|_{L^2(X)}$ serves as shorthand for the function $\|v(\cdot,z)\|_{L^2(X)}$.

We denote by $\langle \cdot, \cdot \rangle$ the inner product associated with $L^2(\Omega)$. In working with this inner product we occasionally employ Fubini's theorem (see Royden [9]) and interchange the order of integration. Thus, if $v_1, v_2 \in L^2(\Omega)$, then

$$\langle v_1(x,z), v_2(x,z) \rangle = \int_X \int_Z v_1(x,z)\overline{v_2(x,z)} \, dz \, dx = \int_Z \int_X v_1(x,z)\overline{v_2(x,z)} \, dx \, dz. \tag{7}$$

We define Sobolev spaces associated with X and Z and then use these definitions to define function spaces over the two-dimensional domain Ω. The Sobolev spaces $H^2(Z)$, $H_0^2(Z)$, and $H_p^r(X)$ are defined in the usual way:

$$H^2(Z) := \{v \in L^2(X) : \partial_z^a v \in L^2(Z), \text{ for } 0 \leq a \leq 2\}, \tag{8}$$

$$H_0^2(Z) := \{v \in H^2(Z) : v(0) = v(1) = 0\}, \tag{9}$$

$$H_p^r(X) := \{v \in L^2(X) : \partial_x^a v \in L^2(Z) \text{ and is periodic for } 0 \leq a \leq r\}. \tag{10}$$

Following Canuto, Maday, and Quarteroni [6], we define the nonisotropic Hilbert space $H_p^{r,2}(\Omega)$ as the space containing all functions $v \in L^2(\Omega)$ such that

$$\int_X \sum_{\alpha=0}^{2} \|\partial_z^\alpha v\|_{L^2(Z)}^2 \, dx < \infty \tag{11}$$

and

$$\int_Z \sum_{\alpha=0}^{r} \|\partial_x^\alpha v\|_{L^2(X)}^2 \, dz < \infty. \tag{12}$$

We assume that $r \geq 1$, and we denote by $\mathcal{H}$ the space containing functions $v \in H_p^{(r+1),2}(\Omega)$ such that $\partial_z^2 \partial_x v \in L^2(\Omega)$ and $v(x,z) = 0$ when $(x,z) \in \partial\Omega$.

III. FINITE-STRIP TRIAL SPACE

What distinguishes the FSM from other weighted-residual techniques is its trial space. This space $\tilde{\mathcal{H}}$ is a finite-dimensional subspace of $\mathcal{H}$ whose standard basis contains products $\omega_m(x) l_j(z)$ of functions defined on X and Z. For the functions $l_j(z)$, we use basis functions for piecewise linear interpolation over a grid defined on Z. Trigonometric functions, defined below, serve as the basis functions $\omega_m(x)$ defined on X. We now describe this trial space in detail.

The piecewise linear basis $\{l_j(z)\}_{j=1}^{J-1}$ requires that Z be partitioned by a grid. Figure 1 depicts the nodal lines associated with the grid $0 = z_0 < z_1 < \cdots < z_J = 1$. We demand that the grid contain all loci of the jump discontinuities in the coefficients K_x, K_z, and S. The mesh size of this grid is

$$h = \max_{j=1,\ldots,J} |z_j - z_{j-1}|. \tag{13}$$

A typical piecewise linear basis function, shown in Fig. 2, has local support and satisfies the conditions

$$l_j(z_i) = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases} \qquad i = 0,1,\ldots,J, \qquad j = 1,2,\ldots,J-1. \tag{14}$$

These functions span a $(J-1)$-dimensional subspace $\mathcal{V}$ of $L^2(Z)$, namely,

$$\mathcal{V} = \left\{ v \in L^2(Z) \colon v(z) = \sum_{j=1}^{J-1} v_j l_j(z) \right\}. \tag{15}$$

Thus $\mathcal{V}$ contains all functions that are piecewise linear with respect to the given grid and that vanish at the endpoints $z_0 = 0$ and $z_1 = 1$.

The basis for approximation along the horizontal direction consists of trigonometric functions associated with truncated Fourier series on X. Figure 3 depicts one such function. Although Fourier sine-cosine series are typically used in FSM computations, for succinctness we use the complex exponential form. Letting $i^2 = -1$, we have

$$v(x) = \sum_{m=-\infty}^{\infty} \hat{v}_m \omega_m(x) \tag{16}$$

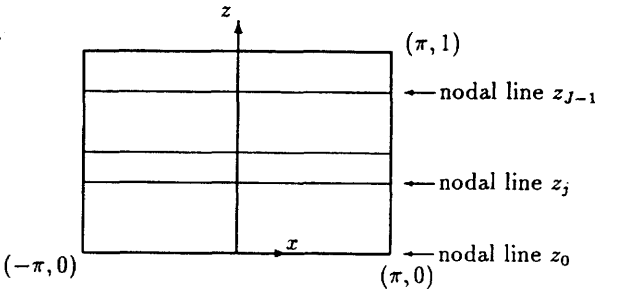

FIG. 1. Rectangular domain, Ω.

Here, $\omega_m(x) := \exp(imx)$ and $\hat{v}_m$ denotes the Fourier coefficient,

$$\hat{v}_m := \frac{1}{2\pi} \int_X v(x) \overline{\omega_m(x)} \, dx. \tag{17}$$

We denote by $\mathcal{U}$ the following $(2M+1)$-dimensional subspace of $L^2(X)$:

$$\mathcal{U} := \{ v \in L^2(X) \colon \hat{v}_m = 0 \text{ for } |m| > M \}. \tag{18}$$

Thus $\mathcal{U}$ contains all Fourier series on X that are truncated at mode number M.

Functions in the trial space $\tilde{\mathcal{H}}$ are bilinear combinations of basis functions associated with $\mathcal{U}$ and $\mathcal{V}$, that is,

$$\tilde{\mathcal{H}} = \left\{ v \in \mathcal{H} \colon v = \sum_{j=1}^{J-1} \sum_{|m| \leq M} v_{m,j} \omega_m(x) l_j(z) \right\} = \mathcal{U} \otimes \mathcal{V}. \tag{19}$$

Functions in $\tilde{\mathcal{H}}$ are thus piecewise linear in z and vary as truncated Fourier series in x. The dimension of $\tilde{\mathcal{H}}$ is therefore $(J-1) \cdot (2M+1)$.

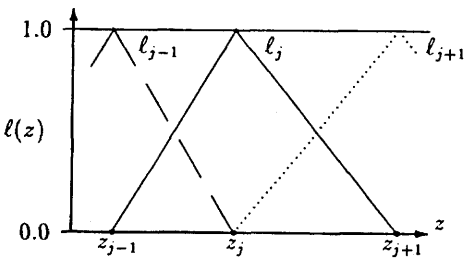

FIG. 2. Linear basis $l_j(z)$.

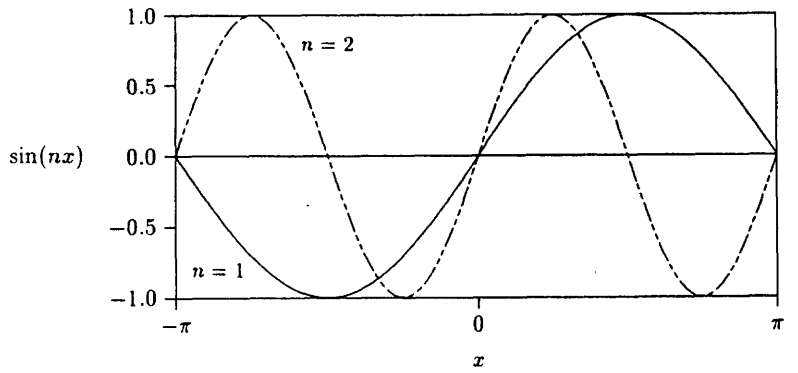

FIG. 3. Global basis functions, $\{\sin nx\}_{n=1}^{2}$.

IV. FORMULATION OF THE FSM

The FSM arises from the following weak form of the exact problem (1): Find a one-parameter family $u(\cdot,\cdot,t)$ in $\mathcal{H}$ such that, for all test functions $w \in \mathcal{H}$ and all times $t \in (0,T]$,

$$\langle S\partial_t u, w\rangle + \langle K_x\partial_x u, \partial_x w\rangle + \langle K_z\partial_z u, \partial_z w\rangle = \langle f, w\rangle. \tag{20}$$

To discretize this problem in space, we restrict $u(\cdot,\cdot,t)$ and w to a finite-dimensional subspace of $\mathcal{H}$: Find a one-parameter family of functions $\tilde{u}(\cdot,\cdot,t)$ in $\tilde{\mathcal{H}}$ such that, for all $w \in \tilde{\mathcal{H}}$ and all $t \in (0,T]$,

$$\langle S\partial_t \tilde{u}, w\rangle + \langle K_x\partial_x \tilde{u}, \partial_x w\rangle + \langle K_z\partial_z \tilde{u}, \partial_z w\rangle = \langle f, w\rangle. \tag{21}$$

This condition yields a set of $(J-1)\cdot(2M+1)$ ordinary differential equations in time.

Instead of solving these ordinary differential equations exactly, we use a temporally discrete approximation. We replace the function $\tilde{u}(x,z,t)$ by a sequence of functions $\tilde{u}^k(x,z) \approx \tilde{u}(x,z,k\tau)$ in $\tilde{\mathcal{H}}$. Here, τ represents the time step. Similarly, $u^k(x,z)$ signifies the exact solution value $u(x,z,k\tau)$. To solve for $\tilde{u}^k(x,z)$, we introduce the backward difference scheme

$$\left\langle S\frac{\tilde{u}^k - \tilde{u}^{k-1}}{\tau}, w\right\rangle + \langle K_x\partial_x\tilde{u}^k, \partial_x w\rangle + \langle K_z\partial_z\tilde{u}^k, \partial_z w\rangle = \langle f^k, w\rangle. \tag{22}$$

Since $\tilde{u}^k$ has the form

$$\tilde{u}^k(x,z,t) = \sum_{j=1}^{J-1}\sum_{|m|\le M}\Phi_{m,j}^k\omega_m(x)l_j(z), \tag{23}$$

our objective is to determine the coefficients $\Phi_{m,j}^k$ at each time level k. To start the calculations, we must choose an appropriate initial function $\tilde{u}^0(x,z)$. In practice, we project the exact initial condition $u^0(x,z)$ into the trial space $\tilde{\mathcal{H}}$ using projection operators defined

in the next section. This procedure amounts to interpolating the exact initial data in z and truncating its Fourier series in x.

We determine the unknown coefficients $\Phi_{m,j}^k$ at time level k by solving linear systems obtained using the basis functions $l_j(z)\omega_m(x)$ as weight functions w. If we order the weighted-residual equations lexicographically according to the index pairs (m,j), then the choice of the linear basis functions $l_j(z)$ for the vertical dimension implies that the linear system is tridiagonal. Our assumptions that K_x and K_z are strictly positive and bounded guarantee that the system is symmetric and positive definite and hence nonsingular at each time level. The system therefore generates a unique sequence $\tilde{u}^k$ in $\tilde{\mathcal{H}}$.

One benefit of the FSM is its efficiency in parallel-computing environments. This benefit owes its existence to the orthogonality of the trigonometric basis $\{\omega_m(x)\}_{|m|\le M}$:

$$\frac{1}{2\pi}\int_X \omega_n\overline{\omega_m}\,dx = \begin{cases}0 & \text{for } m \ne n\\ 1 & \text{for } m = n.\end{cases} \tag{24}$$

We also have

$$\int_X \partial_x\omega_n\overline{\partial_x\omega_m}\,dx = \int_X mn\omega_n\overline{\omega_m}\,dx = \begin{cases}2\pi m^2 & \text{for } m = n\\ 0 & \text{for } m \ne n.\end{cases} \tag{25}$$

Thus the tridiagonal system to be solved at each time level decouples into $(2M+1)$ independent matrix equations of size $J-1$, one system for each Fourier mode. This decoupling allows one to solve for distinct Fourier modes in parallel, as demonstrated computationally in [2, 4, 5].

Some further remarks about practical implementation are in order before we discuss the analysis. The theory presented here applies to a linear problem in which the spatial domain has a rectangular geometry and the spatial part of the differential operator is self-adjoint. Other geometries and boundary conditions may be accommodated if the appropriate eigenfunctions are used. Accommodating non-self-adjoint spatial operators is not such a straightforward matter. In such problems, the Fourier modes typically do not decouple, as discussed above, and much of the method's natural parallelism is lost. The use of superposition in the formulation of the FSM formally precludes nonlinear problems. However, as in ordinary finite-element and spectral methods, one can often approximate a nonlinear problem by an iterative sequence of linear ones, as in Newton's method. In these cases the use of the FSM at each iteration may be feasible. We do not explore these extensions of the method here.

V. APPROXIMATION ERROR ESTIMATES

In this section, we review error estimates for interpolation and projection into the trial space $\tilde{\mathcal{H}}$. We use these estimates in the error analysis presented later.

Define the interpolation map $I: L^2(Z) \to \mathcal{V}$ as follows:

$$(Iv)(z) := \sum_{j=1}^{J-1} v(z_j)l_j(z). \tag{26}$$

For functions $v \in \mathcal{H}$, we extend this map in the straightforward way:

$$(Iv)(x,z) := \sum_{j=1}^{J-1} v(x,z_j)l_j(z). \tag{27}$$

We denote by $\mathcal{P}$: $L^2(X) \to \mathcal{U}$ the projection that truncates Fourier series to $2M + 1$ terms. Provided that $M \geq 1$, we have

$$(\mathcal{P}v)(x) := \sum_{|m|\leq M} \hat{v}_m \omega_m(x). \tag{28}$$

Again, extension to functions of two variables is straightforward: For $v \in \mathcal{H}$,

$$(\mathcal{P}v)(x,z) := \sum_{|m|\leq M} \hat{v}_m(z)\omega_m(x), \tag{29}$$

where $\hat{v}_m(z) := (2\pi)^{-1} \int_X v(x,z)\omega_m(x)\,dx$.

Composition of these maps yields the *approximation map* $I\mathcal{P}$: $\mathcal{H} \to \tilde{\mathcal{H}}$. For $v \in \mathcal{H}$,

$$(I\mathcal{P}v)(x,z) = \sum_{j=1}^{J-1} \sum_{|m|\leq M} \hat{v}_m(z_j) l_j(z) \omega_m(x). \tag{30}$$

In estimating the FSM error $\|u^k - \tilde{u}^k\|_{L^2(\Omega)}$ in the next section, we need an estimate of $\|v - I\mathcal{P}v\|_{L^2(\Omega)}$, which we call the *approximation error*. To develop this estimate, we first discuss the errors associated with I and $\mathcal{P}$. Strang and Fix [10] show that the interpolation error for $v \in H^2(Z)$ obeys the bound

$$\|v - Iv\|_{L^2(Z)} \leq \pi^{-2}h^2 \|\partial_z^2 v\|_{L^2(Z)}. \tag{31}$$

Analogous estimates exist for the projection error associated with $\mathcal{P}$. If $v \in H_p^{r,2}(\Omega)$, where $r \geq 1$ is an integer, then

$$\|v - \mathcal{P}v\|_{L^2(\Omega)} \leq \frac{\sqrt{2\pi}}{M^r} \|\partial_x^r v\|_{L^2(\Omega)}. \tag{32}$$

Canuto *et al.* [11] outline a proof of this estimate, which we detail in Lemma 10 of the Appendix.

We now prove two lemmas giving an estimate of $\|v - I\mathcal{P}v\|_{L^2(\Omega)}$. In the proofs, we indicate parenthetically the steps where we use the Parseval equality, the Bessel inequality [12], and Fubini's theorem [9]. The first lemma estimates the interpolation error when we apply I to the truncated Fourier series $\mathcal{P}v$.

Lemma 1. *If $v \in \mathcal{H}$, then*

$$\|\mathcal{P}v - I\mathcal{P}v\|_{L^2(\Omega)} \leq \sqrt{\frac{2}{\pi^3}}\, h^2 \|\partial_z^2 v\|_{L^2(\Omega)}. \tag{33}$$

Proof. Using the definition of $\|\cdot\|_{L^2(\Omega)}$, we have

$$\begin{aligned}
\|\mathcal{P}v - I\mathcal{P}v\|_{L^2(\Omega)}^2 &= \int_X \int_Z |\mathcal{P}v(x,z) - I\mathcal{P}v(x,z)|^2\,dz\,dx \\
&= \int_X \|\mathcal{P}v(x,\cdot) - I\mathcal{P}v(x,\cdot)\|_{L^2(Z)}^2\,dx \\
\text{[Eq. (31)]} \quad &\leq \int_X (\pi^{-2}h^2)^2 \|\partial_z^2 \mathcal{P}v(x,\cdot)\|_{L^2(Z)}^2\,dx \\
&= (\pi^{-2}h^2)^2 \int_X \int_Z \left| \sum_{|m|\leq M} \partial_z^2 \hat{v}_m(z)\omega_m(x) \right|^2 dz\,dx \\
\text{(Fubini's theorem)} \quad &= (\pi^{-2}h^2)^2 \int_Z \int_X \left| \sum_{|m|\leq M} \partial_z^2 \hat{v}_m(z)\omega_m(x) \right|^2 dx\,dz \\
\text{(orthogonality)} \quad &= \frac{2h^4}{\pi^3} \int_Z \sum_{|m|\leq M} |\partial_z^2 \hat{v}_m(z)|^2\,dz \\
\text{(Bessel inequality)} \quad &\leq \frac{2h^4}{\pi^3} \int_Z \|\partial_z^2 v(\cdot,z)\|_{L^2(X)}^2\,dz \\
&= \frac{2h^4}{\pi^3} \|\partial_z^2 v\|_{L^2(\Omega)}^2.
\end{aligned}$$

■

When we combine Eq. (32) and Lemma 1 using the triangle inequality, we get an estimate of the approximation error:

Lemma 2. *If $v \in \mathcal{H}$ then*

$$\|v - I\mathcal{P}v\|_{L^2(\Omega)} \leq \frac{\sqrt{2\pi}}{M^r} \|\partial_x^r v\|_{L^2(\Omega)} + \sqrt{\frac{2}{\pi^3}}\, h^2 \|\partial_z^2 v\|_{L^2(\Omega)}. \tag{34}$$

Proof. The triangle inequality gives

$$\|v - I\mathcal{P}v\|_{L^2(\Omega)} \leq \|v - \mathcal{P}v\|_{L^2(\Omega)} + \|\mathcal{P}v - I\mathcal{P}v\|_{L^2(\Omega)}.$$

The desired result follows from the estimates (32) and (33). ■

(Canuto, Maday, and Quarteroni [1] obtain a comparable estimate.)

VI. ERROR ANALYSIS OF THE FSM

We now estimate the difference between the exact solution $u^k(x,z)$ of problem (1) and the approximate solution $\tilde{u}^k(x,z)$ generated by the FSM. We begin by defining three error components,

$$e^k := u^k - \tilde{u}^k, \tag{35}$$

$$\eta^k := u^k - I\mathcal{P}u^k, \tag{36}$$

$$\xi^k := I\mathcal{P}u^k - \tilde{u}^k. \tag{37}$$

The objective is to estimate $\|e^k\|_{L^2(\Omega)}$. Since $e^k = \eta^k + \xi^k$, the triangle inequality yields

$$\|e^k\|_{L^2(\Omega)} \le \|\eta^k\|_{L^2(\Omega)} + \|\xi^k\|_{L^2(\Omega)}. \tag{38}$$

Lemma 2 provides an estimate for $\|\eta^k\|_{L^2(\Omega)}$, so an estimate for $\|\xi^k\|_{L^2(\Omega)}$ will suffice to bound $\|e^k\|_{L^2(\Omega)}$.

Our development proceeds by the following plan: We first derive an equation using ξ^k as the test function in the fully discretized weak formulation, Eq. (22). We then obtain estimates for individual terms in this equation. Finally, we apply a discrete form of Gronwall's lemma to yield the desired estimate for $\|\xi^k\|_{L^2(\Omega)}$.

We start by restricting the weight function w to $\tilde{\mathcal{H}}$ in Eq. (20) and subtract Eq. (22) from it. We also add the quantity

$$\left\langle S\frac{u^k - u^{k-1}}{\tau}, w\right\rangle$$

to both sides of the resulting sum. It follows that, for all test functions $w \in \tilde{\mathcal{H}}$ and all time levels $k \in (0, T/\tau]$,

$$\left\langle S\frac{e^k - e^{k-1}}{\tau}, w\right\rangle + \langle K_x\partial_x e^k, \partial_x w\rangle + \langle K_z\partial_z e^k, \partial_z w\rangle = \left\langle S\left(\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right), w\right\rangle. \tag{39}$$

Since $e^k = \eta^k + \xi^k$, we may rearrange Eq. (39) to get

$$\begin{aligned}\left\langle S\frac{\xi^k - \xi^{k-1}}{\tau}, w\right\rangle + \langle K_x\partial_x\xi^k, \partial_x w\rangle + \langle K_z\partial_z\xi^k, \partial_z w\rangle &= \left\langle S\left(\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right), w\right\rangle \\ &\quad - \left\langle S\frac{\eta^k - \eta^{k-1}}{\tau}, w\right\rangle - \langle K_x\partial_x\eta^k, \partial_x w\rangle - \langle K_z\partial_z\eta^k, \partial_z w\rangle. \end{aligned}\tag{40}$$

Setting $w = \xi^k$ and multiplying through by τ yields

$$\begin{aligned}&\langle S\xi^k, \xi^k\rangle - \langle S\xi^{k-1}, \xi^k\rangle + \tau\langle K_x\partial_x\xi^k, \partial_x\xi^k\rangle + \tau\langle K_z\partial_z\xi^k, \partial_z\xi^k\rangle \\ &\quad = \tau\left\langle S\left(\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right), \xi^k\right\rangle - \langle S(\eta^k - \eta^{k-1}), \xi^k\rangle \\ &\qquad - \tau\langle K_x\partial_x\eta^k, \partial_x\xi^k\rangle - \tau\langle K_z\partial_z\eta^k, \partial_z\xi^k\rangle \\ &\quad \le \tau\left\langle S\left(\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right), \xi^k\right\rangle - \langle S(\eta^k - \eta^{k-1}), \xi^k\rangle \\ &\qquad + \tau|\langle K_x\partial_x\eta^k, \partial_x\xi^k\rangle| + \tau|\langle K_z\partial_z\eta^k, \partial_z\xi^k\rangle|. \end{aligned}\tag{41}$$

We now analyze individual terms in Eq. (41), beginning with $\langle S\xi^{k-1}, \xi^k\rangle$. The inequality $2\langle a, b\rangle \le \langle a, a\rangle + \langle b, b\rangle$ and the assumption that $0 < S$ imply that

$$|\langle S\xi^{k-1}, \xi^k\rangle| \le \tfrac{1}{2}\langle S\xi^{k-1}, \xi^{k-1}\rangle + \tfrac{1}{2}\langle S\xi^k, \xi^k\rangle. \tag{42}$$

Next we obtain an estimate for $\tau\langle K_x\partial_x\eta^k, \partial_x\xi^k\rangle$. Using the inequality $2\langle a, b\rangle \le \langle a, a\rangle + \langle b, b\rangle$, the definition of η, and the assumption that $0 < K_x \le K$, we find that

$$\begin{aligned}\langle K_x\partial_x\eta^k, \partial_x\xi^k\rangle &\le \tfrac{1}{2}\langle K_x[\partial_x u^k - I\mathcal{P}(\partial_x u^k)], \partial_x u^k - I\mathcal{P}(\partial_x u^k)\rangle \\ &\quad + \tfrac{1}{2}\langle K_x\partial_x\xi^k, \partial_x\xi^k\rangle \\ &\le \frac{K}{2}\|\partial_x u^k - I\mathcal{P}(\partial_x u^k)\|^2_{L^2(\Omega)} + \tfrac{1}{2}\langle K_x\partial_x\xi^k, \partial_x\xi^k\rangle. \end{aligned}\tag{43}$$

Applying Lemma 2 then yields

$$\begin{aligned}\langle K_x\partial_x\eta^k, \partial_x\xi^k\rangle &\le \frac{K}{2}\left(\frac{\sqrt{2\pi}}{M^r}\|\partial_x^{r+1}u^k\|_{L^2(\Omega)} + h^2\sqrt{\frac{2}{\pi^3}}\|\partial_z^2\partial_x u^k\|_{L^2(\Omega)}\right)^2 \\ &\quad + \tfrac{1}{2}\langle K_x\partial_x\xi^k, \partial_x\xi^k\rangle. \end{aligned}\tag{44}$$

Since $t \in (0, T]$, this last inequality allows us to deduce that

$$\langle K_x\partial_x\eta^k, \partial_x\xi^k\rangle \le \tfrac{1}{2}(M^{-r}\Gamma_1 + h^2\Gamma_2)^2 + \tfrac{1}{2}\langle K_x\partial_x\xi^k, \partial_x\xi^k\rangle, \tag{45}$$

where

$$\Gamma_1 := \sup_{t\in(0,T]}\sqrt{2\pi K}\,\|\partial_x^{r+1}u^k\|_{L^2(\Omega)}, \tag{46}$$

$$\Gamma_2 := \sup_{t\in(0,T]}\sqrt{\frac{2K}{\pi^3}}\,\|\partial_z^2\partial_x u^k\|_{L^2(\Omega)}. \tag{47}$$

Although the term $\langle K_z\partial_z\eta^k, \partial_z\xi^k\rangle$ may be analyzed similarly, we use a different approach to show that it vanishes. For any node z_j on the z axis,

$$\begin{aligned}\eta^k(x, z_j) &= \sum_{m=-\infty}^{\infty}\hat{u}_m(z_j)\omega_m(x) - \sum_{|m|\le M}\hat{u}_m(z_j)\omega_m(x) \\ &= \sum_{|m|>M}\hat{u}_m(z_j)\omega_m(x). \end{aligned}\tag{48}$$

Using the expansion (23) of $\bar{u}^k \in \tilde{\mathcal{H}}$, we write the quantity ξ^k as follows:

$$\begin{aligned}\xi^k(x, z) &= I\mathcal{P}u^k(x, z) - \bar{u}^k(x, z) \\ &= -\sum_{j=1}^{J-1}\sum_{|m|\le M}\left[\Phi^k_{m,j} - \hat{u}^k_m(z_j)\right]\omega_m(x)l_j(z). \end{aligned}$$

Differentiation with respect to z yields

$$\partial_z\xi^k(x, z) = -\sum_{|m|\le M}C^k_{m,j}\omega_m(x),$$

for any $z \in (z_{j-1}, z_j)$. Here,

$$C^k_{m,j} := \frac{[\Phi^k_{m,j} - \bar{u}^k_m(z_j)] - [\Phi^k_{m,j-1} - \bar{u}^k_m(z_{j-1})]}{z_j - z_{j-1}}.$$

Because the value of $K_z(z)$ is a constant $K_{z,j}$ for $z \in (z_{j-1}, z_j)$, the integral over Z in $\langle K_z\partial_z\eta^k, \partial_z\xi^k\rangle$ decomposes into a sum over the intervals formed by the finite-element

grid. Using Eq. (48), we get

$$\begin{aligned}\langle K_z\partial_z\eta^k, \partial_z\xi^k\rangle &= -\int_X \sum_{j=1}^{J} K_{2,j} \int_{z_{j-1}}^{z_j} \sum_{|m|\le M} C_{m,j}^k \omega_m(x)\overline{\partial_z\eta^k}\, dz\, dx \\ &= -\int_X \sum_{j=1}^{J} K_{2,j} \sum_{|m|\le M} C_{m,j}^k \omega_m(x) \int_{z_{j-1}}^{z_j} \overline{\partial_z\eta^k}\, dz\, dx \\ &= -\sum_{j=1}^{J} K_{2,j} \int_X \left(\sum_{|m|\le M} C_{m,j}^k \omega_m(x)\right)\left(\sum_{|m|>M} d_{m,j}^k \omega_m(x)\right) dx,\end{aligned}$$

where

$$d_{m,j}^k := \overline{\hat{u}_m^k(z_j) - \hat{u}_m^k(z_{j-1})}.$$

The orthogonality property (25) then implies that

$$\langle K_z\partial_z\eta^k, \partial_z\xi^k\rangle = 0, \tag{49}$$

as claimed.

In addition, since K_x and K_z are positive, the third and fourth terms on the left-hand side of Eq. (41) are non-negative:

$$0 \le \tau\langle K_x\partial_x\xi^k, \partial_x\xi^k\rangle \tag{50}$$

and

$$0 \le \tau\langle K_z\partial_z\xi^k, \partial_z\xi^k\rangle. \tag{51}$$

Incorporating the estimates (42)–(51) into Eq. (41) yields the inequality

$$\begin{aligned}\tfrac{1}{2}\langle S\xi^k, \xi^k\rangle - \tfrac{1}{2}\langle S\xi^{k-1}, \xi^{k-1}\rangle \le{}& \tau\left|\left\langle S\left(\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right), \xi^k\right\rangle\right| \\ &+ |\langle S(\eta^k - \eta^{k-1}), \xi^k\rangle| + \frac{\tau}{2}(M^{-r}\Gamma_1 + h^2\Gamma_2)^2.\end{aligned} \tag{52}$$

We now estimate the first two terms on the right-hand side of Eq. (52). Lemmas 3–5 concern the first term on the right-hand side, which involves the truncation error associated with the time-stepping scheme.

Lemma 3. *Let $u^k \in \mathcal{H}$ for $0 \le k \le T/\tau$. Then for all $(x,z) \in \Omega$,*

$$\frac{u^k(x,z) - u^{k-1}(x,z)}{\tau} - \partial_t u^k(x,z) = \frac{-1}{\tau}\int_{(k-1)\tau}^{k\tau} [t - (k-1)\tau]\partial_t^2 u(x,z,t)\, dt. \tag{53}$$

Proof. The fundamental theorem of calculus and integration by parts yield

$$\begin{aligned}u^k(x,z) - u^{k-1}(x,z) &= \int_{(k-1)\tau}^{k\tau} \partial_t u(x,z,t)\, dt \\ &= [t - (k-1)\tau]\partial_t u(x,z,t)\,\big|_{t=(k-1)\tau}^{t=k\tau} \\ &\quad - \int_{(k-1)\tau}^{k\tau} [t - (k-1)\tau]\partial_t^2 u(x,z,t)\, dt \\ &= [k\tau - (k-1)\tau]\partial_t u^k(x,z) \\ &\quad - \int_{(k-1)\tau}^{k\tau} [t - (k-1)\tau]\partial_t^2 u(x,z,t)\, dt.\end{aligned}$$

The desired result follows upon rearrangement. ∎

Lemma 4. *Let $u^k \in \mathcal{H}$ for $0 \le k \le T/\tau$. Then*

$$\left\|\partial_t u^k - \frac{u^k - u^{k-1}}{\tau}\right\|_{L^2(\Omega)}^2 \le \frac{\tau}{3}\int_{(k-1)\tau}^{k\tau} \|\partial_t^2 u\|_{L^2(\Omega)}^2\, dt. \tag{54}$$

Proof. Lemma 3 and the Cauchy-Schwarz inequality imply that

$$\begin{aligned}\left\|\partial_t u^k - \frac{u^k - u^{k-1}}{\tau}\right\|_{L^2(\Omega)}^2 &\le \frac{1}{\tau^2}\left\|\int_{(k-1)\tau}^{k\tau} [t - (k-1)\tau]\partial_t^2 u(x,z,t)\, dt\right\|_{L^2(\Omega)}^2 \\ &\le \frac{1}{\tau^2}\left\|\sqrt{\int_{(k-1)\tau}^{k\tau} [t - (k-1)\tau]^2\, dt}\sqrt{\int_{(k-1)\tau}^{k\tau} (\partial_t^2 u)^2\, dt}\right\|_{L^2(\Omega)}^2 \\ &= \frac{1}{\tau^2}\int_\Omega \left[\int_{(k-1)\tau}^{k\tau} [t - (k-1)\tau]^2\, dt \right. \\ &\quad \left. \times \int_{(k-1)\tau}^{k\tau} (\partial_t^2 u)^2\, dt\right] dx\, dz \\ &= \frac{1}{\tau^2}\int_\Omega \frac{\tau^3}{3}\int_{(k-1)\tau}^{k\tau} (\partial_t^2 u)^2\, dt\, dx\, dz \\ &= \frac{\tau}{3}\int_{(k-1)\tau}^{k\tau} \|\partial_t^2 u\|_{L^2(\Omega)}^2\, dt.\end{aligned}$$

The last step follows from Fubini's theorem. ∎

Lemma 5. *Let $u^k \in \mathcal{H}$ for $0 \le k \le T/\tau$. Then*

$$\begin{aligned}\tau\left|\left\langle S\left(\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right), \xi^k\right\rangle\right| \le{}& \frac{S^*\tau^2}{6}\int_{(k-1)\tau}^{k\tau} \|\partial_t^2 u\|_{L^2(\Omega)}^2\, dt \\ &+ \frac{S^*\tau}{2}\langle \xi^k, \xi^k\rangle.\end{aligned} \tag{55}$$

Proof. The assumption that $0 < S(z) \le S^*$ and the inequality $2\langle a,b\rangle \le \langle a,a\rangle + \langle b,b\rangle$ imply that

$$\tau\left|\left\langle S\left[\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right], \xi^k\right\rangle\right| \le S^*\tau\left|\left\langle\left[\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right], \xi^k\right\rangle\right| \le \frac{S^*\tau}{2}\left\|\frac{u^k - u^{k-1}}{\tau} - \partial_t u^k\right\|^2_{L^2(\Omega)} + \frac{S^*\tau}{2}\langle\xi^k,\xi^k\rangle.$$

The desired result follows from lemma 4. ∎

We now analyze the second term on the right-hand side of Eq. (52).

Lemma 6. *If η^k and ξ^k are as defined in Eqs. (36) and (37), then*

$$|\langle S(\eta^k - \eta^{k-1}), \xi^k\rangle| \le \frac{S^*}{2}\int_{(k-1)\tau}^{k\tau} \|\partial_t\eta\|^2_{L^2(\Omega)}\,dt + \frac{S^*\tau}{2}\langle\xi^k,\xi^k\rangle. \tag{56}$$

Proof. The Cauchy-Schwarz inequality, the assumption that $S(z) \le S^*$, and the inequality $2\langle a,b\rangle \le \langle a,a\rangle + \langle b,b\rangle$ yield

$$\begin{aligned}
|\langle S(\eta^k - \eta^{k-1}), \xi^k\rangle| &= \left|\left\langle S\int_{(k-1)\tau}^{k\tau} \partial_t\eta(x,z,t)\,dt, \xi^k\right\rangle\right| \\
&\le \left\langle S\sqrt{\int_{(k-1)\tau}^{k\tau} dt}\sqrt{\int_{(k-1)\tau}^{k\tau}(\partial_t\eta)^2\,dt}, |\xi^k|\right\rangle \\
&= \left\langle S\sqrt{\int_{(k-1)\tau}^{k\tau}(\partial_t\eta)^2\,dt}, \sqrt{\tau}\,|\xi^k|\right\rangle \\
&\le \frac{S^*}{2}\int_\Omega\int_{(k-1)\tau}^{k\tau}(\partial_t\eta)^2\,dt\,dx\,dz + \frac{S^*\tau}{2}\langle\xi^k,\xi^k\rangle \\
&= \frac{S^*}{2}\int_{(k-1)\tau}^{k\tau}\|\partial_t\eta\|^2_{L^2(\Omega)}\,dt + \frac{S^*\tau}{2}\langle\xi^k,\xi^k\rangle.
\end{aligned}$$

(Fubini's theorem) ∎

Application of lemma 5 and lemma 6 to Eq. (52) now produces the inequality

$$\begin{aligned}
\tfrac{1}{2}\langle S\xi^k,\xi^k\rangle - \tfrac{1}{2}\langle S\xi^{k-1},\xi^{k-1}\rangle \le{}& S^*\tau\langle\xi^k,\xi^k\rangle + \frac{S^*\tau^2}{6}\int_{(k-1)\tau}^{k\tau}\|\partial_t^2 u\|^2_{L^2(\Omega)}\,dt \\
&+ \frac{S^*}{2}\int_{(k-1)\tau}^{k\tau}\|\partial_t\eta\|^2_{L^2(\Omega)}\,dt \\
&+ \frac{\tau}{2}(M^{-r}\Gamma_1 + h^2\Gamma_2)^2.
\end{aligned} \tag{57}$$

We now make three observations to prepare for the application of the discrete Gronwall lemma. First, if p is any positive integer such that $p\tau \le T$ and if we sum Eq. (57) from

$k = 1$ through $k = p$, then we obtain the inequality

$$\begin{aligned}
\tfrac{1}{2}(\langle S\xi^p,\xi^p\rangle - \langle S\xi^0,\xi^0\rangle) \le{}& S^*\tau\sum_{k=1}^{p}\langle\xi^k,\xi^k\rangle + \frac{S^*\tau^2}{6}\int_0^{p\tau}\|\partial_t^2 u\|^2_{L^2(\Omega)}\,dt \\
&+ \frac{S^*}{2}\int_0^{p\tau}\|\partial_t\eta\|^2_{L^2(\Omega)}\,dt \\
&+ \sum_{k=1}^{p}\frac{\tau}{2}(M^{-r}\Gamma_1 + h^2\Gamma_2)^2.
\end{aligned} \tag{58}$$

Let us use the numerical initial condition $\hat{u}^0 = I\mathcal{P}u^0$, so that $\xi^0 = 0$ and thus $\langle\xi^0,\xi^0\rangle = 0$ and $\langle S\xi^0,\xi^0\rangle = 0$. In this case, we can multiply Eq. (58) by 2 and extend the integrations to the full time interval $(0,T]$ to get

$$\langle S\xi^p,\xi^p\rangle \le 2S^*\tau\sum_{k=0}^{p}\langle\xi^k,\xi^k\rangle + \beta', \tag{59}$$

where

$$\beta' := \frac{S^*\tau^2}{3}\int_0^T\|\partial_t^2 u\|^2_{L^2(\Omega)}\,dt + S^*\int_0^T\|\partial_t\eta\|^2_{L^2(\Omega)}\,dt + T(M^{-r}\Gamma_1 + h^2\Gamma_2)^2. \tag{60}$$

Second, the Fourier series for $\partial_t u$ may be written in terms of the Fourier coefficients of u. In particular, if $u(\cdot,\cdot,t) \in \mathcal{H}$ for $t \in (0,T]$, then

$$u(x,z,t) = \sum_{m=-\infty}^{\infty}\hat{u}_m(z,t)\omega_m(x),$$

and the series converges uniformly. Therefore

$$\partial_t u(x,z,t) = \sum_{m=-\infty}^{\infty}\partial_t\hat{u}_m(z,t)\omega_m(x).$$

Thus we can estimate the term in (60) involving $\partial_t\eta$ using lemma 2:

$$\|\partial_t\eta\|_{L^2(\Omega)} \le \frac{\sqrt{2\pi}}{M^r}\|\partial_t\partial_x^r u\|_{L^2(\Omega)} + \sqrt{\frac{2}{\pi^3}}\,h^2\|\partial_t\partial_z^2 u\|_{L^2(\Omega)}. \tag{61}$$

Third, utilizing the assumption that $0 < s \le S$, we can move the last term of the sum in Eq. (59) to the left-hand side, getting

$$(s - 2S^*\tau)\langle\xi^p,\xi^p\rangle \le 2S^*\tau\sum_{k=0}^{p-1}\langle\xi^k,\xi^k\rangle + \beta'. \tag{62}$$

Let us choose the time step τ small enough so that $s - 2S^*\tau > 0$. Defining

$$\lambda := \frac{2S^*\tau}{s - 2S^*\tau} \tag{63}$$

and

$$\beta := \frac{\beta'}{s - 2S^*\tau}, \tag{64}$$

we obtain

$$\langle \xi^p, \xi^p \rangle \le \lambda\tau \sum_{k=0}^{p-1} \langle \xi^k, \xi^k \rangle + \beta . \tag{65}$$

We now use a discrete form of Gronwall's lemma (reviewed in the Appendix) to establish the estimate on $\|\xi^k\|_{L^2(\Omega)}$. If p is any integer such that $p\tau \le T$, then

$$\|\xi^k\|^2_{L^2(\Omega)} \le \beta e^{\lambda T}, \quad \text{for } k = 0, 1, \ldots, P . \tag{66}$$

Finally, the main error estimate for the FSM results when we use the estimate (66) in the triangle inequality (38):

Theorem 1. *(FSM Error). Let $u(\cdot,\cdot,t) \in \mathcal{H}$ satisfy the initial-boundary-value problem (1) for $t \in (0, T)$. Let $\{\tilde{u}^k\}$ be a sequence of functions in $\tilde{\mathcal{H}}$ determined using the FSM, Eq. (22). If p is any integer such that $p\tau \le T$, then, for time levels $k = 0, 1, \ldots, p$,*

$$\|\tilde{u}^k - u^k\|_{L^2(\Omega)} \le \frac{\sqrt{2\pi}}{M^r} \|\partial_x^r u^k\|_{L^2(\Omega)} + \sqrt{\frac{2}{\pi^3}}\, h^2 \|\partial_z^2 u^k\|_{L^2(\Omega)} + \sqrt{\beta e^{\lambda T}} . \tag{67}$$

Here,

$$\lambda := \frac{2S^*\tau}{s - 2S^*\tau},$$

$$\beta := \frac{\beta'}{s - 2S^*\tau},$$

$$\beta' := S^* \int_0^T \|\partial_t \eta\|^2_{L^2(\Omega)}\, dt + \frac{S^*\tau^2}{3} \int_0^T \|\partial_t^2 u\|^2_{L^2(\Omega)}\, dt + T(M^{-r}\Gamma_1 + h^2\Gamma_2)^2$$

$$\le S^* \int_0^T \left[\frac{\sqrt{2\pi}}{M^r} \|\partial_t \partial_x^r u\|^2_{L^2(\Omega)} + \sqrt{\frac{2}{\pi^3}}\, h^2 \|\partial_t \partial_z^2 u\|^2_{L^2(\Omega)} \right]^2 dt$$

$$+ \frac{S^*\tau^2}{3} \int_0^T \|\partial_t^2 u\|^2_{L^2(\Omega)}\, dt$$

$$+ T\left(M^{-r} \sup_{t\in(0,T]} \sqrt{2K\pi}\, \|\partial_x^{r+1} u\|_{L^2(\Omega)} + h^2 \sup_{t\in(0,T]} \sqrt{\frac{2K}{\pi^3}}\, \|\partial_z^2 \partial_x u\|_{L^2(\Omega)} \right)^2 .$$

This theorem asserts that the L^2 error in the backward-Euler FSM applied to the problem (1) is $O(M^{-r} + h^2 + \tau)$. Here, r is the degree of smoothness of the exact solution in the x direction. The order of the estimate, M^{-r} in the Fourier direction and h^2 in the finite-element direction, remains unchanged if we scale the spatial domain to a more general rectangle $\Omega = (a, b) \times (c, d)$. In particular, the FSM converges in the sense that $\|\tilde{u}^k - u^k\|_{L^2(\Omega)} \to 0$ as $\max\{h, M^{-1}, \tau\} \to 0$.

VII. COMPUTATIONAL RESULTS

We test theorem 1 computationally with a dimensionless quenching problem from the classical theory of heat transfer. We solve the following model problem on $\Omega = (0, 1) \times (0, 1)$ with the FSM:

$$\begin{aligned} \partial_t u &= K(\partial_x^2 u + \partial_z^2 u), \qquad (x, z) \in \Omega,\ t \in (0, T], \\ u(x, z, 0) &= u^0 = 1, \qquad (x, z) \in \Omega , \\ u(x, z, t) &= 0, \qquad (x, z) \in \partial\Omega,\ t > 0 . \end{aligned} \tag{68}$$

We use a uniform finite-element grid on Z, the mesh size of which varies among different tests, as discussed below. Figure 4 depicts the decomposition of the domain Ω into strips.

The exact solution to the problem (68) has a double Fourier series,

$$u(x, z, k\tau) = \sum_{m=1,3,5,\ldots}^{\infty} \frac{4e^{-Kk\tau(\pi m)^2}}{m\pi} \sin(m\pi x) \sum_{n=1,3,5,\ldots}^{\infty} \frac{4e^{-Kk\tau(\pi n)^2}}{n\pi} \sin(n\pi z) . \tag{69}$$

The symmetry of the problem implies that only odd-numbered Fourier modes have nonzero amplitudes. This solution is continuously differentiable to all orders in both x and z for $t > 0$ ([13], Chap. 4).

Using the exact solution, we compute the error term by term as follows:

$$\|u^k - \tilde{u}^k\|^2_{L^2(\Omega)} = \|u^k\|^2_{L^2(\Omega)} - 2\langle u^k, \tilde{u}^k \rangle + \|\tilde{u}^k\|^2_{L^2(\Omega)} . \tag{70}$$

Orthogonality implies that the first term on the right-hand side of this expansion collapses to the infinite sum

$$\|u^k\|^2_{L^2(\Omega)} = \sum_{n=1,3,5,\ldots}^{\infty} \left[\frac{e^{-Kk\tau(\pi n)^2}}{n} \right]^2 \approx \sum_{n=1,3,5,\ldots}^{N} \left[\frac{e^{-Kk\tau(\pi n)^2}}{n} \right]^2 . \tag{71}$$

Here, N is a positive integer at which we truncate the series in the computations. To determine an appropriate value of N, we observe that $e^{-Kk\tau(\pi n)^2}$ decays quickly with n. We pick N such that $e^{-Kk\tau(\pi N)^2} \le 10^{-10}$, or

$$N \ge \sqrt{\frac{10 \ln 10}{\pi^2 Kk\tau}} , \tag{72}$$

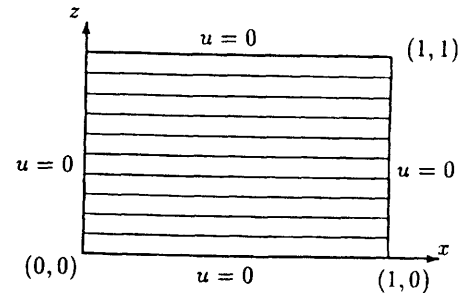

FIG. 4. Partition of domain into strips.

for $1 \le k \le T/\tau$. We also use the same value of N for the truncated series that arises from the second term on the right-hand side of (70):

$$\langle u^k, \tilde{u}^k \rangle \approx \frac{8}{\pi^2} \sum_{m=1,3,5,\ldots}^{M} \left\{ \sum_{j=1}^{J-1} \frac{e^{-Kk\tau(\pi m)^2}}{m} \Phi_{m,j}^k \left[\sum_{n=1,3,5,\ldots}^{N} \frac{e^{-Kk\tau(\pi n)^2}}{n} \left(\frac{1}{n\pi}\right)^2 \frac{c_j}{h} \right] \right\}, \quad (73)$$

where $c_j := 2\sin(n\pi z_j) - \sin(n\pi z_{j-1}) - \sin(n\pi z_{j+1})$. We use all the terms of $\tilde{u}^k$ to calculate its norm. Owing to orthogonality, mixed products of modes do not survive integration, and we obtain

$$\|\tilde{u}_\blacklozenge^k\|_{L^2(\Omega)}^2 = \frac{1}{2} \sum_{m=1,3,5,\ldots}^{M} \sum_{j=1}^{J-1} h \left(\frac{\Phi_{m,j-1}\Phi_{m,j}}{6} + \frac{2\Phi_{m,j}^2}{3} + \frac{\Phi_{m,j+1}\Phi_{m,j}}{6} \right). \quad (74)$$

Since $u^k \in H_p^{r,2}(\Omega)$ for all $r \ge 1$, theorem 1 indicates that $\|u^k - \tilde{u}^k\|_{L^2(\Omega)} = O(h^2 + M^{-r} + \tau)$ for all $r \ge 1$. The idea behind the following tests is to generate numerical solutions using an extremely small time step τ and to plot $\ln\|u^k - \tilde{u}^k\|_{L^2(\Omega)}$ vs $\ln h$ and $\ln M^{-1}$. The slopes of the resulting plots should confirm theorem 1.

The first computational test considers the effect of varying the finite-element mesh size h. The parameters for this test are summarized in Table I. We use $K = 0.02$ and a final time $T = 0.5$. To make the time-stepping error negligible, we choose $\tau = 0.0005$. To render the $O(M^{-r})$ error terms negligible, we choose $M = 65$ for the total number of Fourier modes. However, only the 32 odd-numbered modes contribute to the expansion of $\tilde{u}^k$. With this fixed value of M, we vary h from $\frac{1}{2}$ to $\frac{1}{28}$. Figure 5 depicts the results. The graph indicates that, as h shrinks, the FSM error is indeed $O(h^2)$.

Next we examine the effect of varying the total number M of Fourier modes. In this test problem, the Fourier coefficients decay rapidly as t increases. While this phenomenon is beneficial in computational practice, in numerical testing it requires us to look at early solutions to distinguish the FSM error from errors associated with finite machine precision. Table II summarizes the parameters of this test. We present results for $t = 0.03$, 0.1, and 0.3. To render the $O(h^2)$ portion of the error negligible, we fix $h = 0.002$.

The efficiency of the FSM becomes apparent in computations of this magnitude. At each time level, the problem decouples into 32 separate tridiagonal problems, each of which determines 499 values $\Phi_{m,j}^k$, $j = 1, 2, \ldots, 499$, for a distinct mode number m. Also calculated for each mode, using results of the lower-numbered modes, is the error $\|u^k - \tilde{u}^k\|_{L^2(\Omega)}$. To exploit the increasing smoothness of the solution in time, we increase the size of the time step τ as the calculations progress. Specifically, τ ranges from 0.0001 initially to a maximum value of 0.0025, which is still small enough to keep the time-stepping error negligible. Figure 6 shows a convergence plot of the errors computed for the three output times. The plot indicates convergence beyond all orders in r, until the machine's precision limits have been reached. This result is consistent with the fact that the exact solution in this test problem is smooth in x, belonging to $H_p^{r,2}(\Omega)$ for all $r \ge 1$.

TABLE I. Parameter summary for test 1 (varying h).

Diffusivity:	$K = 0.02$
Output time:	$T = 0.5$
Time step:	$\tau = 0.0005$

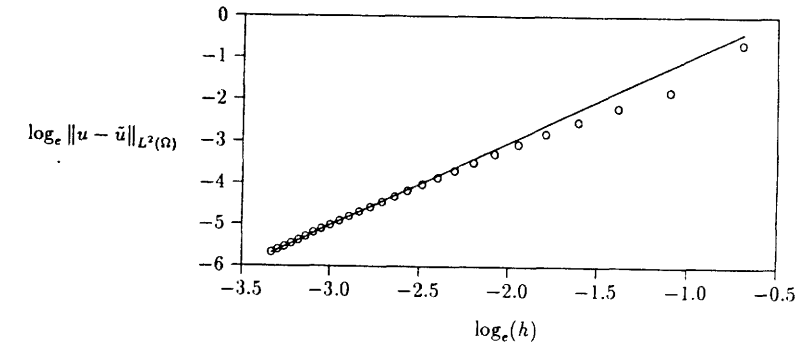

FIG. 5. Convergence plot for changing mesh size h.

These computational tests verify that it is possible in practice to obtain $O(M^{-r} + h^2)$ errors using the FSM, in accordance with theorem 1.

VIII. EXTENSION TO THE FINITE-LAYER METHOD

It is possible to extend the error estimate of theorem 1 to problems on three-dimensional domains $\Omega = X \times Y \times Z$ in a straightforward way. We now sketch this extension. By analogy with the FSM, we consider problems that are geometrically regular and periodic in x and y. Consider the following initial-boundary-value problem:

$$\begin{aligned} S\partial_t u - K_x\partial_x^2 u - K_y\partial_y^2 u - \partial_z(K_z\partial_z u) &= f \quad \text{on } \Omega \times (0,T], \\ u(x,y,z,t) &= 0, \qquad (x,y,z) \in \partial\Omega,\ t \in [0,T], \\ u(x,y,z,0) &= u^0(x,y,z), \qquad (x,y,z) \in \Omega. \end{aligned} \quad (75)$$

Here, the coefficients S, K_x, K_y, and K_z vary as functions of z and obey bounds similar to those given in the inequalities (2), (3), and (4).

Discretization in the finite-layer method is analogous to that used in the FSM. To discretize the problem in the z direction, we again use the piecewise linear basis

TABLE II. Parameter summary for test 2 (varying M). Diffusivity: $K = 0.02$.

Time data:		
number of steps	time step τ	total time t
100	0.000 10	0.01
80	0.000 25	0.03[a]
80	0.000 25	0.05
100	0.000 50	0.10[a]
80	0.002 50	0.30[a]

[a]Results included in Fig. 6.

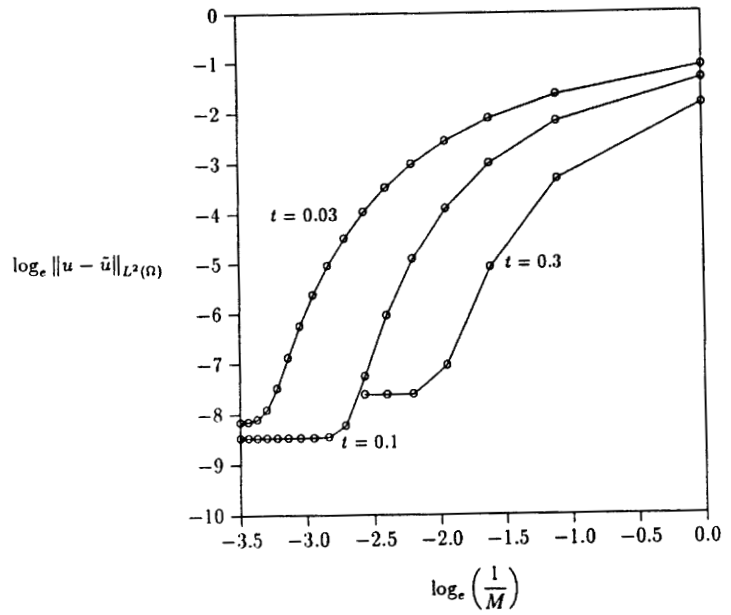

FIG. 6. Convergence plot for changing number M of Fourier modes.

functions $\{l_j(z)\}_{j=1}^{J-1}$. For the x and y directions, we use truncated Fourier series. The exponential basis functions in this case have the form

$$\omega_{m,n}(x,y) := \omega_m(x)\omega_n(y) = e^{i(mx+ny)}. \tag{76}$$

By orthogonality, we have

$$\frac{1}{4\pi^2}\int_{X\times Y} \omega_{nm}\overline{\omega_{m',n'}}\,dx = \begin{cases} 0 & \text{for } m \neq m' \text{ or } n \neq n' \\ 1 & \text{for } m = m' \text{ and } n = n'. \end{cases} \tag{77}$$

We again use backward differences to approximate time derivatives.

The appropriate nonisotropic Hilbert space $H_p^{r,q,2}(\Omega)$ in this setting contains all $v \in L^2(\Omega)$ such that

$$\int_{X\times Y}\sum_{a=0}^{2}\|\partial_z^a v\|_{L^2(Z)}^2\,dx\,dy < \infty, \tag{78}$$

$$\int_{Z\times Y}\sum_{a=0}^{r}\|\partial_x^a v\|_{L^2(X)}^2\,dz\,dy < \infty, \tag{79}$$

and

$$\int_{Z\times X}\sum_{a=0}^{q}\left\|\partial_y^a v\right\|_{L^2(Y)}^2 dz\,dx < \infty. \tag{80}$$

By analogy with the FSM, the space $\mathcal{H}$ contains all functions $v \in H_p^{r,q,2}(\Omega)$ for which $\partial_z^2\partial_x v, \partial_z^2\partial_y v, \partial_x^r\partial_y^q v, \partial_x^{r+1}\partial_y^q v, \partial_x^r\partial_y^{q+1} v \in L^2(\Omega)$ and v vanishes on $\partial\Omega$. The trial space $\tilde{\mathcal{H}}$ is the span of the tensor-product basis functions $l_j(z)\omega_{m,n}(x,y)$, where $j = 1,2,\ldots,J-1$, $|m| \le M$, and $|n| \le N$. The interpolation operator I is analogous to that used in the analysis of the FSM. The projection $\mathcal{P}$ in this context truncates double Fourier series:

$$(\mathcal{P}v)(x,y) := \sum_{|m|\le M}\sum_{|n|\le N}\hat{v}_{m,n}\omega_{m,n}(x,y), \tag{81}$$

where

$$\hat{v}_{m,n} := \frac{1}{4\pi^2}\int_{X\times Y} v(x,y)\omega_{m,n}(x,y)\,dx\,dy. \tag{82}$$

When we extend $\mathcal{P}$ to functions $v \in \mathcal{H}$, we have a projection error estimate comparable to Eq. (32).

Lemma 7. *If $v \in \mathcal{H}$, then*

$$\|v - \mathcal{P}v\|_{L^2(\Omega)} \le \frac{2\pi}{M^r N^q}\left\|\partial_x^r\partial_y^q v\right\|_{L^2(\Omega)}. \tag{83}$$

Proof.

$$\begin{aligned}
\|v - \mathcal{P}v\|_{L^2(\Omega)}^2 &= \int_Z\int_Y\int_X |v(x,y,z) - \mathcal{P}v(x,y,z)|^2\,dx\,dy\,dz \\
&= \int_Z \|v(\cdot,z) - \mathcal{P}v(\cdot,z)\|_{L^2(X\times Y)}^2\,dz \\
\text{(Parseval equality)}\quad &= 4\pi^2\int_Z\sum_{|m|>M}\sum_{|n|>N}|\hat{v}_{m,n}(z)|^2\,dz \\
&= 4\pi^2\int_Z\sum_{|m|>M}\sum_{|n|>N}(m^2)^{-r+r}(n^2)^{-q+q}|\hat{v}_{m,n}(z)|^2\,dz \\
&\le \frac{4\pi^2}{M^{2r}N^{2q}}\int_Z\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}|m^r n^q\hat{v}_{m,n}(z)|^2\,dz \\
&= \frac{4\pi^2}{M^{2r}N^{2q}}\int_Z\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\left|\left(\widehat{\partial_x^r\partial_y^q v}\right)_{m,n}(z)\right|^2 dz \\
\text{(Parseval equality)}\quad &= \frac{4\pi^2}{M^{2r}N^{2q}}\int_Z\left\|\partial_x^r\partial_y^q v\right\|_{L^2(X\times Y)}^2 dz \\
&= \frac{4\pi^2}{M^{2r}N^{2q}}\left\|\partial_x^r\partial_y^q v\right\|_{L^2(\Omega)}^2.
\end{aligned}$$

■

We now state the approximation error estimate corresponding to lemma 1 and lemma 2. The proofs of the next two lemmas are identical to those of the earlier lemmas, except for the following changes: Integrations over X become integrations over $X \times Y$; the basis function $\omega_m(x)$ is replaced by $\omega_{m,n}(x,y)$; and the sums over m are replaced by double sums over m and n.

Lemma 8. *If $v \in \mathcal{H}$, then*

$$\|\mathcal{P}v - I\mathcal{P}v\|_{L^2(\Omega)} \le \frac{2h^2}{\pi}\|\partial_z^2 v\|_{L^2(\Omega)}. \tag{84}$$

Lemma 9. *If $v \in \mathcal{H}$, then*

$$\|v - I\mathcal{P}v\|_{L^2(\Omega)} \le \frac{2\pi}{M^r N^q}\left\|\partial_x^r \partial_y^q v\right\|_{L^2(\Omega)} + \frac{2h^2}{\pi}\|\partial_z^2 v\|_{L^2(\Omega)}. \tag{85}$$

We obtain an error estimate for the finite-layer method by a sequence of arguments analogous to those leading to Theorem 1, incorporating the following changes:

1. Replace integration over X by integration over $X \times Y$.
2. Replace sums over m by double sums over m and n and use the respective truncation limits M and N where appropriate.
3. Manipulate the term $\langle K_y \partial_y \tilde{u}^k, \partial_y w\rangle$ in the error equation in a manner identical to that used for the term $\langle K_x \partial_x \tilde{u}^k, \partial_x w\rangle$ in the FSM analysis.

The following theorem results.

Theorem 2. *(Finite-Layer Error). Let $u^k \in \mathcal{H}$ denote the solution to the problem (75) at $t = k\tau$, and let $\tilde{u}^k \in \tilde{\mathcal{H}}$ be the corresponding solution to the finite-layer method. If p is any integer such that $p\tau \le T$, then*

$$\|\tilde{u}^k - u^k\|_{L^2(\Omega)} \le \frac{2\pi}{M^r N^q}\left\|\partial_x^r \partial_y^q u\right\|_{L^2(\Omega)} + \frac{2h^2}{\pi}\|\partial_z^2 u\|_{L^2(\Omega)} + \sqrt{\beta e^{\lambda T}}, \tag{86}$$

where

$$\lambda = \frac{2S^*\tau}{s - 2S^*\tau},$$

$$\beta \le \frac{S^*}{s - 2S^*\tau}\int_0^T \left[\frac{2\pi}{M^r N^q}\left\|\partial_t \partial_x^r \partial_y^q u\right\|_{L^2(\Omega)}^2 + \frac{2h^2}{\pi}\|\partial_t \partial_z^2 u\|_{L^2(\Omega)}^2\right]^2 dt$$
$$+ \frac{S^*\tau^2}{3(s - 2S^*\tau)}\int_0^T \|\partial_t^2 u\|_{L^2(\Omega)}^2\, dt + \frac{TK}{s - 2S^*\tau}$$
$$\times \left[\left(\sup_{t\in(0,T]} \frac{2\pi}{M^r N^q}\left\|\partial_x^{r+1}\partial_y^q u\right\|_{L^2(\Omega)} + \sup_{t\in(0,T]}\frac{2h^2}{\pi}\|\partial_z^2 \partial_x u\|_{L^2(\Omega)}\right)^2\right.$$
$$\left. + \left(\sup_{t\in(0,T]} \frac{2\pi}{M^r N^q}\left\|\partial_x^r \partial_y^{q+1} u\right\|_{L^2(\Omega)} + \sup_{t\in(0,T]}\frac{2h^2}{\pi}\|\partial_z^2 \partial_y u\|_{L^2(\Omega)}\right)^2\right].$$

Thus the error is $O(M^{-r} + N^{-q} + h^2 + \tau)$, in close analogy with the error estimate of theorem 1.

The Wyoming Water Research Center supported this work through a grant-in-aid. We also received support through NSF Grant No. RII-8610680 and ONR Grant No. 0014-88-K-0370. The authors thank Professor Jay Puckett and Professor Thomas Edgar of the Department of Civil Engineering, University of Wyoming, for their practical insights into the FSM.

APPENDIX

Projection Error

Lemma 10. *Let r and M be positive integers, and let $v \in H_p^{r,2}(\Omega)$. Then*

$$\|v - \mathcal{P}v\|_{L^2(\Omega)} \le \frac{\sqrt{2\pi}}{M^r}\|\partial_x^r v\|_{L^2(\Omega)}. \tag{87}$$

Proof. By definition,

$$\|v - \mathcal{P}v\|_{L^2(\Omega)}^2 = \int_Z \int_X |v(x,z) - (\mathcal{P}v)(x,z)|^2\, dx\, dz$$
$$= \int_Z \|v(\cdot,z) - (\mathcal{P}v)(\cdot,z)\|_{L^2(X)}^2\, dz$$
(Parseval equality) $$= 2\pi \int_Z \sum_{m>M} |\hat{v}_m(z)|^2\, dz$$
$$= 2\pi \int_Z \sum_{m>M} (m^2)^{-r+r}|\hat{v}_m(z)|^2\, dz$$
$$\le \frac{2\pi}{M^{2r}}\int_Z \sum_{m=-\infty}^{\infty} |m^r \hat{v}_m(z)|^2\, dz$$
$$= \frac{2\pi}{M^{2r}}\int_Z \sum_{m=-\infty}^{\infty} \left|\left(\widehat{\partial_x^r v}\right)_m(z)\right|^2 dz$$
(Parseval equality) $$= \frac{2\pi}{M^{2r}}\int_Z \|\partial_x^r v\|_{L^2(X)}^2\, dz$$
$$= \frac{2\pi}{M^{2r}}\|\partial_x^r v\|_{L^2(\Omega)}^2. \quad \blacksquare$$

Discrete Form of Gronwall's Lemma

Lemma 11. *Suppose that the real sequence $\{V_k\}_{k=0}^P$ satisfies the inequality*

$$|V_k| \le \beta + \lambda\tau\sum_{j=0}^{k-1}|V_j| \quad \text{for } k = 0, 1, \ldots, P,$$

where λ, β, and τ are non-negative real numbers. Then

$$|V_k| \le \beta e^{\lambda P\tau} \quad \text{for } k = 0, 1, \ldots, P.$$

Proof. Define the sequence $\{Z_k\}_{k=0}^P$ by

$$Z_k = \beta + \lambda\tau\sum_{j=0}^{k}|V_j|.$$

The definition of Z_k and the inequality of the hypothesis imply that

$$Z_0 = \beta + \lambda\tau|V_0| \le \beta + \lambda\tau\beta$$

and

$$Z_j - Z_{j-1} = \lambda\tau|V_j| \le \lambda\tau Z_{j-1} \quad \text{for } j = 1, 2, \ldots, P;$$

that is,

$$Z_0 \le (1 + \lambda\tau)\beta,$$
$$Z_j \le (1 + \lambda\tau)Z_{j-1},$$

for $j = 1, 2, \ldots, P$. Apply the above result $k - 1$ times. Since $(1 + \lambda\tau)^k \le e^{\lambda P\tau}$ for any integer k, $0 \le k \le P$, we have

$$Z_{k-1} \le (1 + \lambda\tau)Z_{k-2} \le \cdots \le (1 + \lambda\tau)^{k-1}Z_0 \le (1 + \lambda\tau)^k\beta \le \beta e^{\lambda P\tau},$$

or

$$\beta + \lambda\tau \sum_{j=0}^{k-1} |V_j| \le \beta e^{\lambda P\tau}, \quad k = 1, 2, \ldots, P.$$

The inequality in the hypothesis implies the desired result. ■

References

1. Yau-kai Cheung, *Finite Strip Method in Structural Analysis*, Pergamon, Oxford, England, 1976.
2. Jay A. Puckett and R. J. Schmidt, "Finite strip method for groundwater modeling in a parallel computing environment," *Eng. Comput.* 7, 167 (1990).
3. James E. Slattery, "The finite strip method in groundwater hydrology," M.S. thesis, Colorado State University, 1986.
4. S. S. Smith, M. B. Allen, J. A. Puckett, and T. Edgar, "Three-dimensional model of multi-well field using finite layer methods," in *Proceedings of Eleventh Annual American Geophysical Union Hydrology Days*, Hydrology Days Publications, Fort Collins, 1991, p. 23.
5. S. S. Smith, M. B. Allen, J. A. Puckett, and T. Edgar, "The finite-layer method for groundwater flow models," *Water Res. Res.* **28**, 1715 (1992).
6. Claudio Canuto, Y. Maday, and Alfio Quarteroni, "Analysis of the combined finite element and Fourier interpolation," *Numer. Math.* **39**, 205 (1982).
7. Karel Rektorys, *Variational Methods in Mathematics, Science and Engineering*, 2nd ed., Reidel, Boston, 1980.
8. Peter S. Huyakorn and George F. Pinder, *Computational Methods in Subsurface Flow*, Academic, New York, 1983.
9. H. L. Royden, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.
10. Gilbert Strang and George J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
11. Claudio Canuto, M. Yousuff Hussaini, Alfio Quarteroni, and Thomas A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.
12. Erwin Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, New York, 1989.
13. G. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, 1976.