W. E. Bosarge, Jr.
C. L. Smith

Numerical Properties of a Multivariate Ritz-Trefftz Method

Abstract: In this paper the numerical properties of the Ritz-Trefftz algorithm are discussed in the context of the numerical approximation to the linear parabolic regulator problem using multivariate splines. The algorithm is first derived in the problem context and the resulting linear algebraic system is discussed. Such properties as definiteness and band structure are treated. The algorithm is applied to a number of sample control problems, and it is shown that the method yields efficient and highly accurate continuous approximations to the solutions of the selected sample problems. Computer implementation of the general algorithm is also discussed.

Introduction

In [1] the authors developed the Ritz-Trefftz algorithm for the numerical treatment of the linear parabolic regulator problem. We describe the problem again in the present context of numerical properties of the algorithm.

Minimize (over u and f) the quadratic cost functional

$$J[u, f] = \tfrac{1}{2}\int_0^T \int_0^1 \{\langle v(x, t), Q(x, t)v(x, t)\rangle + \langle u(x, t), R(x, t)u(x, t)\rangle\}dxdt + \tfrac{1}{2}\int_0^T \langle f(t), S(t)f(t)\rangle dt^{\dagger} \tag{1}$$

subject to linear partial differential equation constraint

$$\frac{\partial v(x, t)}{\partial t} = A(x, t)\frac{\partial^2 v(x, t)}{\partial x^2} + B(x, t)u(x, t) \tag{2}$$

and associated boundary and initial conditions given by

$$v(x, 0) = v_0(x), \qquad 0 \le x \le 1 \tag{3}$$

and

$$\alpha v(0, t) + \frac{\partial v(0, t)}{\partial x} = cf(t), \qquad 0 \le t \le T \tag{4}$$

$$\beta v(1, t) + \frac{\partial v(1, t)}{\partial x} = 0, \qquad 0 \le t \le T. \tag{5}$$

We assume that v is an n-dimensional state vector and that u is an r-dimensional boundary control vector. Further, we assume that $u \varepsilon A_u$, $v \varepsilon A_v$, and $f \varepsilon A_f$, where A_u, A_v, and A_f are suitable Sobolev spaces. Here A is an $n \times n$ matrix function, Q and S are $n \times n$ positive definite matrix functions (we note here that either Q or S could be identically zero, the development for these cases is straightfoward and not presented in this paper), B an $n \times r$ matrix function, and R an $r \times r$ positive definite matrix function. We require that α, β and c be scalars, that A, B, Q, R be (elementwise) in $C^{\gamma}[0, 1] \times C^{\gamma-1}[0, T]$ and that $v_0(x)$ be in $C^{\gamma}[0, 1]$ (here C^{γ} is the space of γ-order continuously differentiable functions on $[0, 1]$). We require that γ be greater than one. The terminal time, T, is assumed fixed and $0 \le r \le n$.

The basic notion of the algorithm is the replacement of the infinite dimensional constraints (2) to (5) by the relaxed (finite dimensional) constraints given by

$$\int_0^T \int_0^1 w_i(x, t)\left(-\frac{\partial v}{\partial t} + A\frac{\partial^2 v}{\partial x^2} + Bu\right)_j dxdt = 0 \tag{6}$$

$$\int_0^1 w_i(x, t)[v(x, 0) - v_0(x)]_j dx = 0 \tag{7}$$

$$\int_0^T w_i(0, t)\left[-cf + \alpha v(0, t) + \frac{\partial v(0, t)}{\partial x}\right]_j dt = 0 \tag{8}$$

$$\int_0^T w_i(1, t)\left[\beta v(1, t) + \frac{\partial v(1, t)}{\partial x}\right]_j dt = 0 \tag{9}$$

for $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n$. The functions $w_i(x, t)$ $(i = 1, 2, \cdots, m)$ form a basis for any conven-

†Angular brackets indicate a vector interproduct notation.

ient finite dimensional space S_m over the rectangle $[0, 1] \times [0, T]$. Although conditions (6) to (9) are in the Galerkin tradition, we do not require that the elements v, u, and f lie in S_m as in the standard Galerkin approach. The concept is closely related to that of Trefftz, who first used a similar idea to produce approximate solutions to partial differential equations.

In [1] the authors presented error bounds for the cost functional as well as for the state and control. In addition they pointed out in detail the theoretical advantages of the Ritz-Trefftz method over the standard Ritz procedure.

In the present paper the numerical properties of the algorithm are our chief interest. In the next section we present a short derivation of the algorithm in the problem context. Subsequently we establish the definiteness of the resulting matrix and discuss some resulting numerical implications of this property. We spend some time discussing the band structure of the matrix and the computational complexity of the elements of the matrix. Then we present some numerical examples to illustrate the computational efficiency and accuracy of the algorithm. We consider a simple problem for which we have an analytic solution for accuracy comparisons and two slightly more difficult problems from mathematical physics, i.e., a laser cooling problem and the optimal heating of a slab (see Sage [2]). The last section contains some remarks concerning the implications of the results presented in this paper.

Ritz-Trefftz algorithm for parabolic regulator

The problem (1) to (5) is known to be equivalent to computing

$$\max_{\lambda \epsilon A_\lambda} \min_{\substack{u \epsilon A_u \\ f \epsilon A_f \\ v \epsilon A_v}} L[u, f, v; \lambda], \tag{10}$$

where $L[u, f, v; \lambda]$ is defined by

$$\begin{aligned} L[u, f, v; \lambda] = J[u, f] &+ \int_0^T \int_0^1 \left\langle \lambda_1(x, t), -\frac{\partial v}{\partial t} \right. \\ &\left. + A \frac{\partial^2 v}{\partial x^2} + Bu \right\rangle dx dt \\ &+ \int_0^1 \langle \lambda_2(x), v(x, 0) - v_0(x) \rangle dx \\ &+ \int_0^T \left\langle \lambda_3(t), -cf + \alpha v(0, t) \right. \\ &\left. + \frac{\partial v(0, t)}{\partial x} \right\rangle dt \\ &+ \int_0^T \left\langle \lambda_4(t), \beta v(1, t) + \frac{\partial v(1, t)}{\partial x} \right\rangle dt, \end{aligned} \tag{11}$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ and A_λ represents suitably choosen dual spaces. Replacement of (2) to (5) with (6) to (9) is then equivalent to computing

$$\max_{\lambda \epsilon S_m} \min_{\substack{u \epsilon A_u \\ f \epsilon A_f \\ v \epsilon A_v}} L[u, f, v; \lambda]. \tag{12}$$

It is easily shown, by standard variational arguments, that this problem is equivalent to computing

$$\max_{\lambda \epsilon S_m^0} L[u_\lambda, f_\lambda, v_\lambda; \lambda], \tag{13}$$

where

$$u_\lambda(x, t) = -R^{-1} B^t \lambda(x, t) \tag{14}$$

$$f_\lambda(t) = cS^{-1} A^t(0, t) \lambda(0, t) \tag{15}$$

$$v_\lambda(x, t) = -Q^{-1} \left\{ \frac{\partial \lambda}{\partial t} + \frac{\partial^2 (A^t \lambda)}{\partial x^2} \right\} \tag{16}$$

$$\lambda_2(x) = -\lambda(x, 0) \tag{13}$$

$$\lambda_3(t) = +A^t(0, t) \lambda(0, t) \tag{18}$$

$$\lambda_4(t) = -A^t(1, t) \lambda(1, t). \tag{19}$$

Here we have taken $\lambda(x, t) = \lambda_1(x, t)$,

$$S_m^0 = \{\lambda \epsilon S_m : \lambda(x, T) = 0, \alpha(A^t \lambda)(0, t) + \frac{\partial (A^t \lambda)(0, t)}{\partial x} = 0,$$

and

$$\beta(A^t \lambda)(1, t) + \frac{\partial (A^t \lambda)(0, t)}{\partial x} = 0\}. \tag{20}$$

Problem formulation (13) serves as the basis for the computational procedure. Let $w_i(x, t)$ $(i = 1, 2, \cdots, m)$ form a basis for S_m^0. Then if $\lambda_i \epsilon S_m^0$ (notice that λ is an n-vector), we conclude that

$$\lambda(x, t) = \sum_{j=1}^{m} \eta_j w_j(x, t), \ (x, t) \epsilon [0, 1] \times [0, T], \tag{21}$$

where $\eta_j (j = 1, 2, \cdots, m)$ is some set of n-vector coefficients.

We now derive an expression for L in terms of the η_j. We first note that

$$\begin{aligned} L[u, f, v; \lambda] = J[u, f] &- \int_0^1 \langle \lambda, v \rangle |_0^T dx \\ &+ \int_0^T \left\{ \left\langle A^t \lambda, \frac{\partial v}{\partial x} \right\rangle - \left\langle v, \frac{\partial A^t \lambda}{\partial x} \right\rangle \right\} \Big|_0^1 dt \\ &+ \int_0^T \int_0^1 \left\{ \left\langle v, \frac{\partial \lambda}{\partial t} + \frac{\partial^2 A^t \lambda}{\partial x^2} \right\rangle \right. \\ &\left. + \langle \lambda, Bu \rangle \right\} dx dt \end{aligned}$$

$$+\int_0^T \left\{\left\langle \lambda_3(t)\,, -cf+\alpha v(0\,,t)+\frac{\partial v(0\,,t)}{\partial x}\right\rangle dt + \left\langle \lambda_4\,, \beta v(1\,,t)+\frac{\partial v(1\,,t)}{\partial x}\right\rangle\right\} dt + \int_0^1 \langle \lambda_2\,, v(x\,,0)-v_0(x)\rangle dx\,. \tag{22}$$

From conditions (14) to (19) we deduce the result

$$L[u_\lambda\,, f_\lambda\,, v_\lambda\,; \lambda] = -J[u\,, f] + \int_0^1 \langle \lambda(x\,,0)\,, v_0(x)\rangle dx\,. \tag{23}$$

From (21) we now obtain

$$\int_0^1\int_0^T \langle u\,, Ru\rangle dxdt = \sum_{i=1}^m \sum_{j=1}^m \eta_i^t \left[\int_0^1\int_0^T E_{ij}(x\,,t)\,dxdt\right]\eta_j \tag{24}$$

where E_{ij} is an $n\times n$ symmetric matrix defined by $E_{ij}(x\,,t) = w_i(x\,,t)w_j(x\,,t)B(x\,,t)R^{-1}(x\,,t)B^t(x\,,t)$. Similarly, we find that

$$\int_0^1\int_0^T \langle v\,, Qv\rangle dxdt = \sum_{i=1}^m \sum_{j=1}^m \eta_i^t \left[\int_0^1\int_0^T F_{ij}(x\,,t)\,dxdt\right]\eta_j \tag{25}$$

and

$$\int_0^T \langle f\,, Sf\rangle dt = \sum_{i=1}^m \sum_{j=1}^m \eta_i^t \left[\int_0^T G_{ij}(t)\,dt\right]\eta_j\,, \tag{26}$$

where

$$F_{ij}(x\,,t) = \dot w_i(x\,,t)\dot w_j(x\,,t)Q^{-1}(x\,,t) + \dot w_i(x\,,t)Q^{-1}\frac{\partial^2 w_j A^t(x\,,t)}{\partial x^2} + \dot w_j(x\,,t)Q^{-1}\frac{\partial^2 w_i A^t(x\,,t)}{\partial x^2} + \frac{\partial^2 w_i A^t}{\partial x^2}Q^{-1}\frac{\partial^2 w_j A^t(x\,,t)}{\partial x^2} \tag{27}$$

and

$$G_{ij}(x\,,t) = c^2 w_i(0\,,t)w_j(0\,,t)A(0\,,t)S^{-1}A^t(0\,,t)\,. \tag{28}$$

Finally, we have that

$$\int_0^1 \langle \lambda(x\,,0)\,, v_0(x)\rangle dx = \sum_{i=1}^m \left[\int_0^1 w_i(x\,,0)v_0(x)\,dx\right]\eta_i\,. \tag{29}$$

We now perform the maximization indicated in (13). A necessary condition for the maximization of $L[u_\lambda\,, f_\lambda\,, v_\lambda\,; \lambda] \equiv \tilde L[\eta_1\,, \eta_2\,, \cdots, \eta_m]$ over η_i is that

$$\frac{\partial \tilde L}{\partial \eta_i} = -\frac{1}{2}\sum_{j=1}^m \left[\int_0^T\int_0^1 \{E_{ij}+E_{ij}^t+F_{ij}+F_{ij}^t\}dxdt + \int_0^T \{G_{ij}+G_{ij}^t\}dt\right]\eta_j + \int_0^1 w_i(x\,,0)v_0(x)\,dx = 0 \tag{30}$$

for $i = 1\,, 2\,, \cdots, m$. Define M_{ij} by

$$M_{ij} = \frac{1}{2}\int_0^T\int_0^1 \{E_{ij}+E_{ij}^t+F_{ij}+F_{ij}^t\}dxdt + \frac{1}{2}\int_0^T \{G_{ij}+G_{ij}^t\}dt\,. \tag{31}$$

Then, in terms of the η_i, Eq. (30) is equivalent to the linear system

$$My = b\,, \tag{32}$$

where the ($mn\times mn$ symmetric) matrix M is defined by

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1m} \\ M_{21} & M_{22} & \cdots & M_{2m} \\ \vdots & \vdots & & \\ M_{m1} & M_{m2} & \cdots & M_{mm} \end{bmatrix}, \quad M_{ij} = M_{ji}^t \tag{33}$$

and y and b are given by

$$y = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_m \end{bmatrix}, \qquad b = \begin{bmatrix} \int_0^1 w_1(x\,,0)v_0(x)\,dx \\ \vdots \\ \int_0^1 w_m(x\,,0)v_0(x)\,dx \end{bmatrix}. \tag{34}$$

Numerical properties of algebraic system

In this section we establish the positive definiteness of M, and show that the matrix is sparse and has band structure, providing we select the patch basis for the space S_m^0. We first prove a theorem in which the definiteness of M is established.

• *Theorem*

Assume that M is given by Eqs. (31) and (33) for the linear parabolic regulator problem. Then the matrix M is positive definite.

Proof: Suppose $y\varepsilon R^{mn}$, $y\neq 0$. Let λ correspond to y according to Eqs. (32) to (34) and (21) and let u, f, and v correspond to λ according to Eqs. (14) to (19). Then we have by (24), (25), and (26)

$$y^tMy = \tfrac{1}{2}y^t\left\{\int_0^T\int_0^1 [E+E^t+F+F^t]dxdt + \int_0^T [G(1\,,t)+G^t(1\,,t)-G(0\,,t)\right.$$

$$
\begin{aligned}
&- G^t(0, t)]dt\Big\}y \\
&= \int_0^T \int_0^1 \langle u, Ru\rangle dxdt + \int_0^T \int_0^1 \langle v, Qv\rangle dxdt \\
&\quad + \int_0^T \langle f, Sf\rangle dt \\
&= 2J[u, f]. \qquad (35)
\end{aligned}
$$

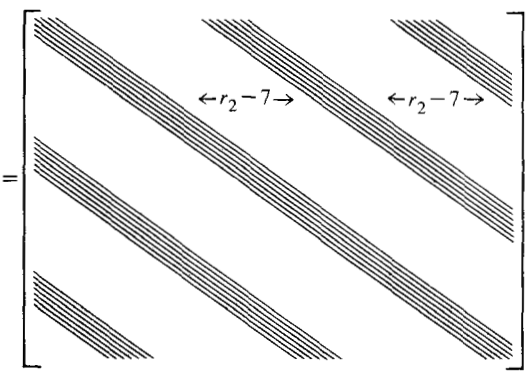
The proof is now complete since J is a positive definite form in the variables u, f, and v.

We remark here that the standard Ritz method does not utilize relationships (24) to (26). The functions u, f, v, and λ are each restricted to S_m, and it can be shown that the resulting matrix is indefinite.

For numerical work the definiteness of M is obviously extremely significant, especially in large problems. For example, suppose S_m is a space of multivariate cubic spline on some appropriate mesh. Let $\{w_j(x, t)\}_{j=1}^m$ be the patch basis for S_m. Then we have

$$
\int_0^T \int_0^1 (w_i w_j)(x, t)dxdt = 0, \qquad (36)
$$

where each pair $(i, j)\varepsilon\{(i, j): d(i, j) \geq 4\}$. Here $d(i, j) = \max\{|i_1 - j_1|, |i_2 - j_2|, \cdots, |i_n - j_n|\}$ and $i = (i_1, i_2, \cdots, i_n), j = (j_1, j_2, \cdots, j_n)$ are n-tuples which represent a coordinate index set for the basis functions. For example, if we consider a vector problem of a single spatial variable with r_1 mesh points in x, r_2 mesh points in time, and index the components of M along the spatial coordinate, M assumes the form

$M =$ 

, (37)

where the lines represent the nonzero matrix subdiagonal. Each element on the diagonals is an $(n \times n)$ block matrix. Since M is positive definite, we can now apply a Cholesky decomposition to (32). It is well known that the Cholesky algorithm is extremely stable, and we retain the band structure in the decomposition, i.e., zeros outside the band remain zeros throughout the decomposition. Usually however, zeros within the band are filled in. It can be shown that the number of multiply-adds required for the Cholesky decomposition of a matrix with band width $2b + 1$ is approximately $b^2 + b$ times the order of the matrix. Thus, in our case, taking $b = (2r_2 + 3)n$, we obtain approximately $[(2r_2 + 3)^2 n^2 + (2r_2 + 3)n][nr_2(r_1 + 1)] = O(4r_2^3 n^3 r_1)$ multiply-adds instead of $O(\frac{1}{3}r_1^3 r_2^3 n^3)$ as is the case with Gaussian elimination. We remark that, in the standard Ritz procedure, Gaussian elimination is normally used since the resulting matrix is indefinite.

We next consider the difficulties involved in computing the matrix. We discuss the computational requirements of generating the M matrix for several choices of uniform grids.

In Fig. 1 we present data generated from a computer code designed to count the number of integrations required to accurately generate the M matrix. We count only integrations which are possibly nonzero [i.e, those along the diagonals in (37)]. Also only half the matrix need be generated since M is symmetric. Since splines are not homogeneous polynomials over the entire region of integration, all regions of integration are assumed to be the subrectangles of the mesh so that a single element of M may involve several independent integrations. We count in thousands of integrations per state variable.

Figure 1 is approximately expressible by no. ops. = $[r_1(r_2 - 1)/16]$ thousands. If we use a Cartesian four-point Gauss integration scheme, we have sixteen functional evaluations per integration so that the number of functional evaluations required to generate M is approximately $O(r_1 r_2)$ in thousands per state variable.

In the next section we apply the Ritz-Trefftz algorithm to three sample problems.

Some numerical examples

• *Sample problem*

We begin this section by considering a simple parabolic system of one scalar variable with additive control only. Consider the system

$$
\frac{\partial v(x, t)}{\partial t} = \frac{\partial^2 v(x, t)}{\partial x^2} + u(x, t), \quad (x, t)\varepsilon[0, 1] \times [0, 1] \qquad (38)
$$

with the initial condition

$$
v(x, 0) = 1, \qquad x\varepsilon[0, 1] \qquad (39)
$$

and boundary conditions

$$
\frac{\partial v(0, t)}{\partial x} = \frac{\partial v(1, t)}{\partial x} = 0, \qquad t\varepsilon[0, 1]. \qquad (40)
$$

The cost functional is given by

$$
J[u, v] = \tfrac{1}{2}\int_0^1 \int_0^1 \{v^2(x, t) + u^2(x, t)\}dxdt. \qquad (41)
$$

We take the set of admissible controls, A_u, and the set of

admissible states, A_v, to be the set of Lebesque square integrable functions over the rectangle $[0, 1] \times [0, 1]$.

We can easily solve this problem analytically. From the necessary conditions (14) to (19) and the calculus of variations, we deduce that in order that u^* and v^* be an optimal solution to (38) to (41), there must exist a multiplier $\lambda^* \varepsilon L_2$ corresponding to (u^*, v^*) such that

$$\frac{\partial \lambda^*(x, t)}{\partial t} + \frac{\partial^2 \lambda^*(x, t)}{\partial x^2} + v^*(x, t) = 0,$$

$$u^*(x, t) + \lambda^*(x, t) = 0, \qquad (x, t) \varepsilon [0, 1] \times [0, 1],$$

$$\frac{\partial \lambda^*(0, t)}{\partial x} = \frac{\partial \lambda^*(1, t)}{\partial x} = 0, \qquad t \varepsilon [0, 1], \quad \text{and}$$

$$\lambda^*(x, 1) = 0, \qquad x \varepsilon [0, 1]. \tag{42}$$

Using the equations (38) to (42) and exploiting the symmetry in x, we obtain the following solution for u^* and v^*:

$$u^*(x, t) = \sinh(t) - \cosh(t) \sinh(1)/\cosh(1)$$

$$v^*(x, t) = \cosh(t) - \sinh(t) \sinh(1)/\cosh(1). \tag{43}$$

We list in Table 1 the results to five decimal places obtained from the Ritz-Trefftz procedure over a 4×4 grid using a bicubic spline basis with boundary conditions enforced. All calculations were done in eleven plus decimal digits accuracy and the system $My = b$ was solved using a Cholesky decomposition without iterative improvement.

The L_2 norms $\|u^* - \bar{u}\|_2$ and $\|v^* - \bar{v}\|_2$ (where $\bar{u}$ and $\bar{v}$ denote the computed control and state, respectively) were also calculated using a Cartesian four-point Gauss integration scheme over each subrectangle. The results are

$$\|u^* - \bar{u}\|_2 = 7.9066 \times 10^{-6}$$

$$\|v^* - \bar{v}\|_2 = 7.4660 \times 10^{-5} \tag{44}$$

$$J[\bar{u}, \bar{v}] = 0.76159.$$

We notice that the error bound on $\bar{u}$ is an order of magnitude better than the bound on $\bar{v}$. This is to be expected and is consistent with other spline applications in numerical analysis, since our problem has no spatial variations and therefore the spatial derivatives do not affect the accuracy of the solution. The computed state, however, is time varying and is calculated from the time derivative of the Lagrange multiplier λ.

From the considerations of the previous paragraph (problem symmetry, smoothness, etc.) and the results obtained by Bosarge and Johnson [3], we should expect the error bound on $\bar{u}$ to be $O(h^4)$ and the bound on $\bar{v}$ to be $O(h^3)$. Using these error bounds we calculate the empirical order constants to be 6.4×10^{-4} for $\bar{u}$ and 2×10^{-4} for $\bar{v}$. These estimates are in good agreement with order

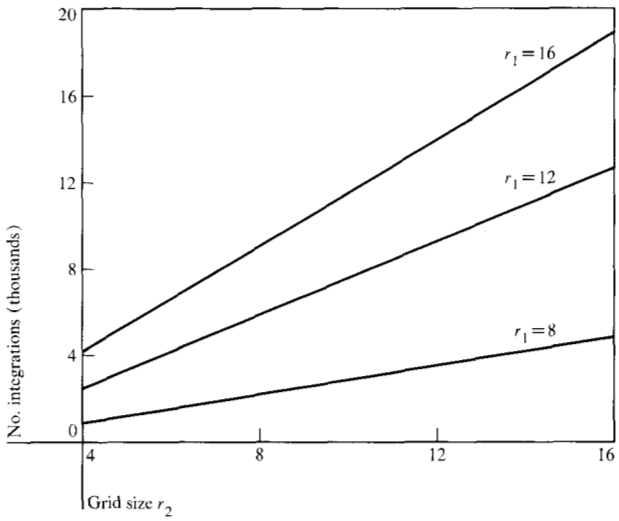


Figure 1 The relationship between the number of integration per problem and the grid size.

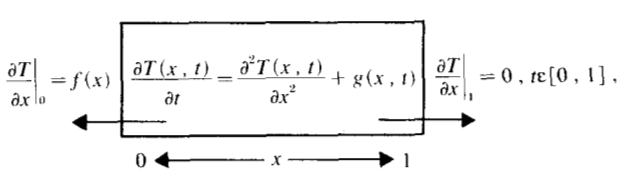
Table 1 The analytical and computational solution to the sample problem.

t	u^*	$\bar{u}$	v^*	$\bar{v}$
0	−0.76159	−0.76159	1.0	0.99974
0.1	−0.66524	−0.66524	0.92872	0.92883
0.2	−0.57554	−0.57554	0.86673	0.86669
0.3	−0.49160	−0.49161	0.81342	0.81331
0.4	−0.41259	−0.41259	0.76824	0.76837
0.5	−0.53770	−0.33769	0.73076	0.73079
0.6	−0.26619	−0.26619	0.70060	0.70050
0.7	−0.19735	−0.19735	0.67744	0.67744
0.8	−0.13048	−0.13048	0.66105	0.66110
0.9	−0.64913	−0.64912	0.65130	0.65130
1.0	0	0	0.64805	0.64804

constants computed from spline applications in other areas of numerical analysis. The constants can be estimated from approximation theory using the Peano Kernel theorem. For the natural cubic spline the constants can be shown to be somewhere near 0.001 times the norm of the fourth derivative of the function being interpolated.

• *Laser heating problem*

Here we consider the optimal cooling of a high-power (one megawatt) cw laser. After normalization and assuming that all heat transfer is by diffusion, we obtain the system illustrated in the figure. The arrows indicate the heat diffusion direction at the ends of the crystal:



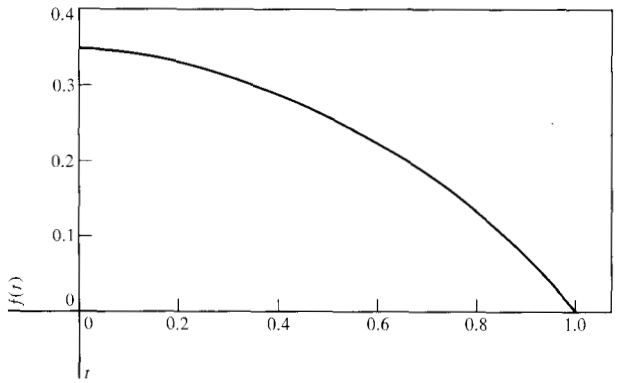


Figure 2 Temperature distribution for the laser heating problem.

Figure 3 Optimal boundary control for the laser heating problem.

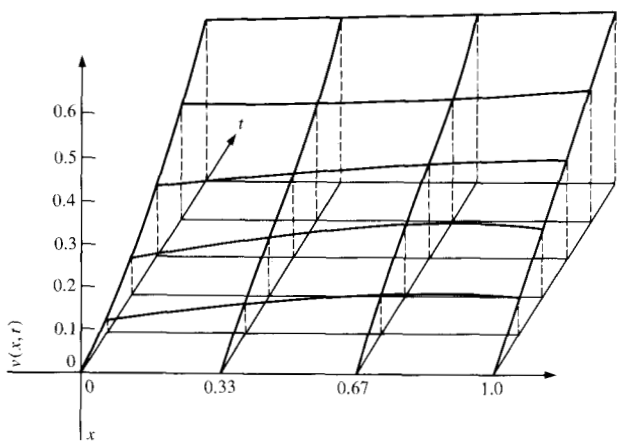


Table 2 Temperature distribution and optimal boundary control for the laser heating problem.

t	$f(t)$	$v(0,t)$	$v(\frac{1}{3},t)$	$v(\frac{2}{3},t)$	$v(1,t)$
0	0.35000	−0.02855	0.00940	0.00410	0.00212
0.2	0.33397	0.02905	0.12106	0.16793	0.18063
0.4	0.29482	0.17062	0.25275	0.30417	0.32135
0.6	0.22809	0.33440	0.40213	0.44622	0.46108
0.8	0.13118	9.52500	0.56814	0.59972	0.61085
1.0	0.0	0.75855	0.75855	0.77164	0.77741

where $T(x,0) = T_a$, $T(x,t)$ is the temperature at the point (x,t), T_a is the ambient air temperature, and $f(t)$ is the control. We take as the performance index the functional

$$J[f,T] = \tfrac{1}{2}\int_0^1\int_0^1 [T(x,t) - T_a]^2\,dxdt + \tfrac{1}{2}\int_0^1 f^2(t)dt\,, \tag{45}$$

and seek a solution $T\varepsilon L_2$ and $f\varepsilon L_2$.

If we take $g(x,t) = 1$ and consider the transformation $v(x,t) = T(x,t) - T_a$, we obtain the system

$$\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v(x,t)}{\partial x^2} + 1\,, \qquad (x,t)\varepsilon[0,1]\times[0,1] \tag{46}$$

with boundary conditions

$$\frac{\partial v(0,t)}{\partial x} = f(t)\,, \qquad \frac{\partial v(1,t)}{\partial x} = 0\,, \tag{47}$$

initial condition

$$v(x,0) = 0\,, \tag{48}$$

and cost functional

$$J[f,v] = \tfrac{1}{2}\int_0^1\int_0^1 v^2(x,t)\,dxdt + \tfrac{1}{2}\int_0^1 f^2(t)dt\,. \tag{49}$$

The problem stated in equations (46) to (49) is a standard parabolic regulator problem with boundary control. From the necessary conditions (14) to (19) and the calculus of variations there must exist Lagrange multipliers $\lambda_1{}^*\varepsilon L_2$ and $\lambda_2{}^*\varepsilon L_2$ such that

$$\frac{\partial \lambda_1{}^*(x,t)}{\partial t} + \frac{\partial^2 \lambda_1{}^*(x,t)}{\partial x^2} + v^*(x,t) = 0\,, \qquad (x,t)\varepsilon[0,1]\times[0,1] \tag{50}$$

$$\lambda_1{}^*(x,1) = 0\,, \qquad x\varepsilon[0,1]\,,$$

$$\lambda_1{}^*(0,t) + \lambda_2{}^*(0,t) = f(t)\,, \qquad t\varepsilon[0,1]\,, \tag{51}$$

and

$$\frac{\partial \lambda_1{}^*(0,t)}{\partial x} = \frac{\partial \lambda_1{}^*(1,t)}{\partial x} = 0\,, \qquad t\varepsilon[0,1]\,. \tag{52}$$

Applying the Ritz-Trefftz procedure we obtain Table 2. The values of Table 2 are plotted in Figs. 2 and 3.

We remark here that the Ritz-Trefftz algorithm represents a novel approach to solving a problem where boundary control is present. We note that neither the a priori error bounds presented in [1] nor the order constants of the previous section are affected by the presence of a boundary control.

• *Optimal heating of a slab*

Consider the optimal heating of a slab as it passes through a furnace. We consider heat transport by diffusion only and assume that our control is additive. The heat transfer from the ends of the rod is zero. The system is described by the set of equations.

$$\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v(x,t)}{\partial x^2} + u(x,t)\,, \qquad (x,t)\varepsilon[0,1]\times[0,1]\,, \tag{53}$$

with initial condition

$$\lim_{t \to 0^+} v(x, t) = 4x^3 - 6x^2 + 1, \qquad x \varepsilon [0, 1], \qquad (54)$$

and boundary conditions

$$\frac{\partial v(0, t)}{\partial x} = \frac{\partial v(1, t)}{\partial x} = 0, \qquad t \varepsilon [0, 1]. \qquad (55)$$

We assume that the cost functional is given by Eq. (41).

Again we seek a solution $u \varepsilon L_2$ and $v \varepsilon L_2$. The necessary conditions are given in equations (42).

We apply the Ritz-Trefftz procedure to this problem and present the numerical results in Tables 3 and 4.

A plot of Table 3, the control functional, is presented in Figure 4. We note finally that the M matrix generated for the first example serves just as well in this example since only the initial state $v_0(x)$ was changed for the two problems.

Conclusions

The numerical properties of the Ritz-Trefftz algorithm have been presented in the context of the numerical approximation to the linear parabolic regulator problem using multivariate splines. The algorithm was first derived in the general problem context and the definiteness, sparseness, and band structure of the resulting linear algebraic system were established.

We conclude that the method represents an efficient and highly accurate numerical approach to solving parabolic regulator problems. In fact, for equal computing time, the method yields superior accuracy compared with finite difference approximations to the integro-partial differential Ricatti equation, and the standard Ritz approach. That is to say, given a fixed number of operations, the predictable error bounds for the Ritz-Trefftz method are superior to known error bounds for the other methods assuming the same operation count.

Further, computer implementation of the method for large parabolic systems yields definite, sparse, band matrices, which allows large parabolic control systems to be treated numerically.

The method can be used in various control applications requiring real time numerical treatment of parabolic problems. Certain applications involving the use of this Galerkin-spline direct numerical treatment are discussed in the forthcoming paper.

References

1. W. E. Bosarge, Jr., O. G. Johnson, and C. L. Smith, "A Direct Method Approximation to the Linear Parabolic Regulator Problem Over Multivariate Spline Bases," IBM Publication No. 320.2401, December 17, 1970. (To appear, *SIAM Numerical Analysis*)
2. A. P. Sage, and S. P. Chadhuri, "Gradient and Quasilinearization Computational Techniques for Distributed Parameter Systems", *International Journal au Control*, **6**, No. 1, 81–98 (1967).

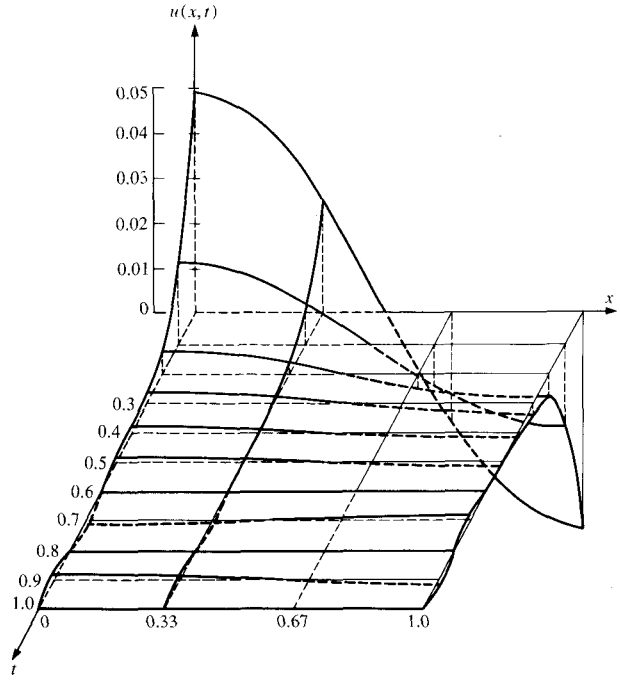


Figure 4 Plot of optimal control for the heating of a slab.

Table 3 Optimal control for the heating of a slab.

t	$u(0, t)$	$u(\frac{1}{3}, t)$	$u(\frac{2}{3}, t)$	$u(1, t)$
0	4.9742×10^{-2}	2.4818×10^{-2}	-2.4818×10^{-2}	-4.9715×10^{-2}
0.1	1.8866×10^{-2}	9.4306×10^{-3}	-9.4306×10^{-3}	-1.8856×10^{-2}
0.2	5.6532×10^{-3}	2.8349×10^{-3}	-2.8349×10^{-3}	-5.6501×10^{-3}
0.3	2.3553×10^{-3}	1.1788×10^{-3}	-1.1788×10^{-3}	-2.3541×10^{-3}
0.4	1.6239×10^{-3}	8.0768×10^{-4}	-8.0768×10^{-4}	-1.6230×10^{-3}
0.5	7.4666×10^{-4}	3.7062×10^{-4}	-3.7062×10^{-4}	-7.4625×10^{-4}
0.6	0	0	0	0
0.7	-2.8444×10^{-4}	-1.3957×10^{-4}	1.3957×10^{-4}	2.8428×10^{-4}
0.8	-8.6526×10^{-5}	-4.2191×10^{-5}	4.2191×10^{-5}	8.6478×10^{-5}
0.9	1.6516×10^{-4}	8.0628×10^{-4}	-8.0628×10^{-4}	-1.6507×10^{-4}
1.0	0	0	0	0

Table 4 Optimal temperature distribution for the heating of a slab.

t	$v(0, t)$	$v(\frac{1}{3}, t)$	$v(\frac{2}{3}, t)$	$v(1, t)$
0	$+9.6354 \times 10^{-1}$	$+4.7529 \times 10^{-1}$	-4.7529×10^{-1}	-9.5503×10^{-1}
0.1	$+4.1145 \times 10^{-1}$	$+2.0518 \times 10^{-1}$	-2.0518×10^{-1}	-4.0827×10^{-1}
0.2	$+1.3015 \times 10^{-1}$	$+6.5998 \times 10^{-2}$	-6.5998×10^{-2}	-1.2924×10^{-1}
0.3	$+3.4594 \times 10^{-2}$	$+1.7508 \times 10^{-2}$	-1.7058×10^{-2}	-3.4218×10^{-2}
0.4	$+2.6158 \times 10^{-2}$	$+1.2629 \times 10^{-2}$	-1.2629×10^{-2}	-2.5872×10^{-2}
0.5	$+1.6895 \times 10^{-2}$	$+8.1191 \times 10^{-3}$	-8.1191×10^{-3}	-1.6758×10^{-2}
0.6	$+5.6308 \times 10^{-3}$	$+2.9073 \times 10^{-3}$	-2.9073×10^{-3}	-5.6321×10^{-3}
0.7	-3.3853×10^{-3}	-1.3974×10^{-3}	$+1.3974 \times 10^{-3}$	$+3.3279 \times 10^{-3}$
0.8	-3.9892×10^{-3}	-1.9274×10^{-3}	$+1.9274 \times 10^{-3}$	$+3.9723 \times 10^{-3}$
0.9	$+6.9338 \times 10^{-4}$	$+1.4446 \times 10^{-4}$	-1.4446×10^{-4}	-6.5908×10^{-4}
1.0	$+5.3047 \times 10^{-3}$	$+2.5974 \times 10^{-3}$	-2.5974×10^{-3}	-5.3017×10^{-3}

3. W. E. Bosarge, Jr. and O. G. Johnson, "Error Bounds of High Order Accuracy for the State Regulator Problem via Piecewise Polynomial Approximations," *SIAM Journal of Control*, **9**, No. 1, (1971).

Bibliography

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L., *The Theory of Splines and Their Applications*, Academic Press, New York, 1967.

Bosarge, W. E., Jr. and Johnson, O. G., "Direct Method Approximation to the State Regulator Problem Using a Ritz-Trefftz Suboptimal Control," *IEEE Transactions on Automatic Control* **AC-15**, 627 (1970).

Bosarge, W. E., Jr. and Johnson, O. G., "Numerical Properties of the Ritz-Trefftz Algorithm for Optimal Control, "IBM DPD Technical Report 320.2376, February 20, 1970, *Comm. ACM* **14**, No. 6, (1971).

Bosarge, W. E., Jr., Johnson, O. G., McKnight, R. S., and Timlake, W. P., "The Ritz-Galerkin Procedure for Nonlinear Control Problems." (To appear, *SIAM Numerical Analysis*).

Lions, J. L., *Controls Optimal de Systems Gouvernes par des Equations aux derives Partielles*, Dunod, Paris, 1968.

Received November 15, 1971

W. E. Bosarge, Jr. is located at the IBM Scientific Center, 6900 Fannin Street, Houston, Texas 77025; C. L. Smith is at the International Mathematical and Statistical Libraries, Inc. Suite 510/6200 Hillcroft, Houston, Texas 77036.