1992

Preconditioning for Domain Decomposition Through Function Approximation

Mo Mu

John R. Rice
Purdue University, jrr@cs.purdue.edu

Report Number:
92-091

PRECONDITIONING FOR DOMAIN DECOMPOSITION THROUGH FUNCTION APPROXIMATION

Mo Mu
John R. Rice

CSD-TR-92-091
November 1992

PRECONDITIONING FOR DOMAIN DECOMPOSITION THROUGH FUNCTION APPROXIMATION

MO MU* AND JOHN R. RICE†

Abstract. A new approach was presented in [11] for constructing preconditioners through a function approximation for the domain decomposition-based preconditioned conjugate gradient method. This work extends the approach to more general cases where grids may be nonuniform; elliptic operators may have variable coefficients (but are separable and self-adjoint); and geometric domains may be nonrectangular. The theory of expressing the Schur complement as a function of a simple interface matrix is established. The approximation to this complicated function by a simple function is discussed and the corresponding error bound is given. Preconditioning a nonrectangular domain problem is done by first reducing it to a rectangular domain problem, and then applying the theory developed here for the rectangular domain case. Accurate error bounds are given by using the results in [6] for typical domains, such as $L-$, $T-$, and $C-$ shaped ones. Numerical results are also reported to illustrate the efficiency of this approach.

Key words. domain decomposition, preconditioners, preconditioned conjugate gradient methods, iterative methods, partial differential equations, parallel computation

AMS(MOS) subject classifications. 65N55, 65F10, 65Y05

1. Introduction. A new approach of constructing preconditioners for the domain decomposition-based preconditioned conjugate gradient *(PCG)* method through a function approximation is proposed in [11] by making the observation that the interface *capacitance matrix* S, or the *Schur complement*, can be viewed as a matrix function

$$S = f(T) \tag{1.1}$$

where $f(t)$ is a complicated function and T is a simple interface matrix. The approach is to find a simple approximation $r(t)$ to $f(t)$, such that, with $q(t) \equiv f(t)/r(t)$,

$$\begin{aligned} &\text{(a)} && R \equiv r(T) \text{ is easily invertible;} \\ &\text{(b)} && \frac{\max_i |q(t_i)|}{\min_i |q(t_i)|} \sim 1, \text{ or } \{q(t_i)\} \text{ are clustered,} \end{aligned} \tag{1.2}$$

where $\{t_i\} = \sigma(T)$ is the spectrum of T. The convergence rate of the *PCG* method is governed by the quantity

$$\kappa(R^{-1}S) \equiv \frac{\lambda_{\max}(R^{-1}S)}{\lambda_{\min}(R^{-1}S)} = \frac{\max_i |q(t_i)|}{\min_i |q(t_i)|} \tag{1.3}$$

* Computer Sciences Department, Purdue University, West Lafayette, IN 47907 U.S.A., mu@cs.purdue.edu, or na.mu@na-net.ornl.gov. Work supported in part by National Science Foundation grant CCR-8619817.

† Computer Sciences Department, Purdue University, West Lafayette, IN 47907 U.S.A., rice@cs.purdue.edu, or na.rice@na-net.ornl.gov. Work supported in part by the Air Force Office of Scientific Research grants, 88-0243, F49620-92-J-0069 and the Strategic Defense Initiative through Army Research Office contract DAAL03-86-K-0106.

where λ_{max} is the maximum eigenvalue and λ_{min} is the minimum eigenvalue, or by the spectrum distribution of the preconditioned matrix $R^{-1}S$ given by

$$\sigma(R^{-1}S) = \{q(t_i)\}. \tag{1.4}$$

It is easily seen that the conditions (1.2) imply that R is a good preconditioner for the *PCG* method.

We begin with a review of previous work and relate it to this approach. Relation (1.1) is essential to this approach. This relation is established in [2] for the standard model problem of a Poisson equation on a rectangle with Dirichlet condition discretized by the *5-point-star* stencil with a uniform square grid; where $f(t)$ is shown to be a rational function in terms of the Chebyshev polynomials and $T = 2I + K$ where K is the discrete one-dimensional Laplacian on the interface. An equivalent expression in terms of the eigendecomposition of K is obtained in [5]. An extension to the variable coefficient case in which the elliptic operator is separable and self-adjoint is implied in [1], where $f(t)$ is still a rational function but in terms of other orthogonal polynomials that play a role analogous to that of the Chebyshev polynomials in the constant coefficient case. These orthogonal polynomials are defined by the three term recurrence relation in terms of the discrete one-dimensional operator in the direction perpendicular to the interface. The roots of such a polynomial are the eigenvalues of the corresponding tridiagonal matrix from the theory of orthogonal polynomials. An equivalent expression is also used in [14] for the reciprocal $1/f(t)$ and is expanded into a sum of partial fractions to approximate the inverse of the Schur complement for a nonrectangular domain, which is referred as to the *rational approximation to the Schur complement of a nonrectangular domain* because of the rational function $f(t)$. The rational expressions in terms of these orthogonal polynomials developed in [1] and [2] are also used in [9], by being expanded into a sum of partial fractions, to devise a fast direct solver in a parallel setting for the original two-dimensional discrete operator. All the above assume a uniform square grid.

The rational expression theory for the Schur complement is extended in [11] to the nonuniform grid case in which the grid is nonuniform on the interface and uniform in the other direction and the elliptic operator has constant coefficients. In this case, (1.1) is modified to the form

$$S = \Theta^{1/2} f(\Theta^{-1/2} T \Theta^{-1/2}) \Theta^{1/2} \tag{1.5}$$

where Θ is a diagonal scaling operator corresponding to the spacings of the nonuniform grid on the interface, $f(t)$ is, within a constant factor, the same rational function as in the uniform case, and T corresponds to certain discrete one-dimensional operator on the interface.

For a nonrectangular domain Ω and mixed boundary value conditions, [14] shows a spectral equivalence in the sense that there exist constants c_1 and c_2, such that, for any vector $\mathbf{v}$ of a proper dimension,

$$c_1(S_0\mathbf{v}, \mathbf{v}) \leq (S\mathbf{v}, \mathbf{v}) \leq c_2(S_0\mathbf{v}, \mathbf{v}) \tag{1.6}$$

where S_0 is the Schur complement on the same interface Γ but corresponds to the Dirichlet condition for the rectangular region embedded in the original domain Ω by shifts of Γ up to $\partial\Omega$. Here $(\cdot, \cdot)$ is the inner product and ∂ is the boundary symbol. Similar results for Dirichlet boundary conditions are also obtained in [3], [4], and [6]. The relation (1.6) implies that S_0 can be taken as a preconditioner for S. This reduces, in principle, a non-Dirichlet problem on a nonrectangular region to a Dirichlet problem on a rectangular region. The efficiency of using S_0 depends on the constants c_1 and c_2 being close to 1.

For a rectangular region and the case of a constant coefficient and separable elliptic operator and a uniform grid on the interface, the Schur complement can be efficiently inverted using (1.1) and the *FFT* applied to the eigendecomposition of T. This makes the *PCG* method a direct solver in this case. Other well known preconditioners can be related to (1.1) by being viewed as approximating $f(t)$ with square-root like functions because $f(t)$ behaves like $t^{1/2}$ near the smallest eigenvalues of T. The matrix T is usually called K in this literature and they define the $K^{1/2}$-family of preconditioners, for example, see [3], [5], [8], and [10]. However, these preconditioners depend either on using the *FFT* for T or on using two- dimensional subdomain solvers, which makes their extension to general cases inefficient and ineffective. The approach proposed in [11] provides a general framework to construct preconditions for S using (1.1) and a function approximation to $f(t)$. Various approximations yield different preconditioners. The $K^{1/2}$-family of preconditioners can, of course, fall into this category. But they are not generally efficient because of the appearance of a square-root in the corresponding approximations. One of the basic principles in our approach is to have a simple form for the $r(t)$ such that the generated matrix R is easily invertible in terms of T. In [11] we illustrate how to construct such simple functions, such as a product of two first-degree interpolating rational functions, or a linear interpolation. By utilizing the special properties of $f(t)$ we can satisfy the conditions (1.2). Examples are given in [11] and [12] showing that this approach is very simple, effective and efficient. Independently, a similar idea is used in [14] by constructing another $m-$term sum of partial fractions as an approximation to the $n-$term sum expression for $1/f(t)$. A theoretical analysis is given showing that under certain conditions on the eigenvalues of two one-dimensional discrete operators in the x and y directions, the approximation error is of the order of $O(1/n^{\tau})$, for any $\tau > 0$, if $m = O(log\ n)$. However, for a real application it is not clear when those conditions can be satisfied, what number needs to be used for m, and so on. No numerical experiments are reported on the actual performance of the approach in [14].

The purpose of this paper is to extend our approach to general cases. Section 2 is devoted to establishing the relation (1.5) for a very general case on a rectangular region where the elliptic operator is separable and self-adjoint with variable coefficients and the grid may be nonuniform in both directions. An expression for efficiently evaluating $f(t)$ is presented and we note in our approach that $f(t)$ only needs to be evaluated at a few interpolating points. The function approximation and preconditioner construction are discussed in Section 3 with numerical results showing the efficiency and effectiveness of the approach. Section 4 considers the extension to a nonrectangular domain. An accurate estimate for the convergence rate of the *PCG* method is given with the help of the results in [6] from the relationship between overlapping and nonoverlapping for domain decomposition and from the dependence of the convergence rate on

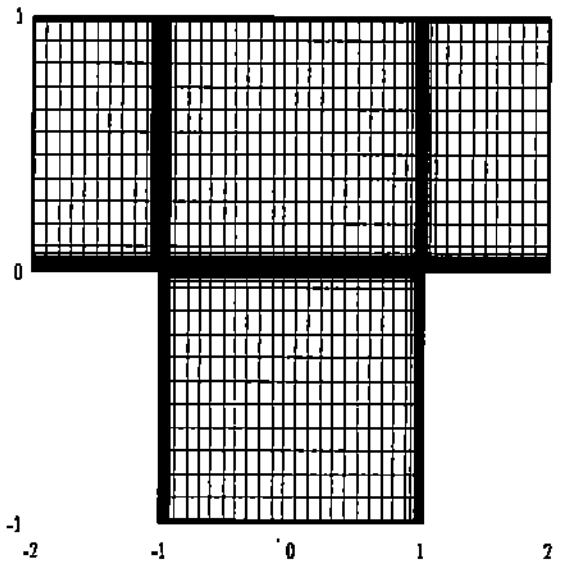

Figure 1.1. *T-shaped domain with an adaptive nonuniform grid suitable for singularities in the solution at the reentrant corners (-1, 0) and (1, 0).*

geometry for the Schwarz overlapping method. Finally, we give conclusions in Section 5.

A typical application of this work is for solving an elliptic boundary value problem on a concave domain, such as L-shaped or T-shaped, where certain geometry-related singularities are usually present in the solution. A common practice in discretization to efficiently handle this type of singularities is to use a nonuniform grid, for instance, generated from an adaptive procedure, so that the grid is much finer near reentrant corners where the singularities are located and coarser for other parts of the region where the solution is smooth. As an example, Fig.1.1 shows a T-shaped domain with two reentrant corners located at (-1, 0) and (1, 0). The solution of a boundary value problem with smooth data behaves like $\rho^{2/3}sin\frac{2}{3}\theta$ around each of these corners, where (ρ, θ) are local polar coordinates. Also seen from the figure is a nonuniform grid adapted to singularities at the two corners.

2. Expression of the Schur complement as a matrix function. We consider the elliptic Dirichlet boundary value problem on a rectangular domain Ω with a separable and self-adjoint operator of the form $L = L_x + L_y$,

$$
\begin{aligned}
L_x &= \frac{-\partial}{\partial x}\left(a(x)\frac{\partial}{\partial x}\right) + c(x); \\
L_y &= \frac{-\partial}{\partial y}\left(b(y)\frac{\partial}{\partial y}\right) + d(y).
\end{aligned} \tag{2.1}
$$

Assume that Ω is discretized by a nonuniform tensor product grid with $\{h_x^i\}_{i=1,\ldots,n_x+1}$ and $\{h_y^i\}_{i=1,\ldots,n_y+1}$ being the spacings for the x and y directions. Using the standard finite differences, L_x is discretized by a tridiagonal matrix A_{FD}^x:

$$\left[-\frac{2a\left(\frac{x_{i-1}+x_i}{2}\right)}{h_x^i\left(h_x^i+h_x^{i+1}\right)},\ \frac{2\left(h_x^i a\left(\frac{x_i+x_{i+1}}{2}\right)+h_x^{i+1}a\left(\frac{x_{i-1}+x_i}{2}\right)\right)}{h_x^i h_x^{i+1}\left(h_x^i+h_x^{i+1}\right)}+c(x_i),\ -\frac{2a\left(\frac{x_i+x_{i+1}}{2}\right)}{h_x^{i+1}\left(h_x^i+h_x^{i+1}\right)}\right]$$

and similarly L_y is discretized by A_{FD}^y. The discrete analog of L can be expressed

$$A_{FD} = A_{FD}^x \otimes I_{n_y} + I_{n_x} \otimes A_{FD}^y \tag{2.2}$$

where $\otimes$ denotes the Kronecker product and I_k is the identify matrix of order k. The matrix A_{FD} is nonsymmetric when the grid is nonuniform. To preserve the symmetric positive definite (SPD) property for obvious reasons, A_{FD} is usually scaled to become a SPD matrix A by

$$\begin{aligned} A &= (\Theta_x \otimes \Theta_y)A_{FD} \\ &= (\Theta_x \otimes \Theta_y)(A_{FD}^x \otimes I_{n_y}) + (\Theta_x \otimes \Theta_y)(I_{n_x} \otimes A_{FD}^y) \\ &= (\Theta_x A_{FD}^x) \otimes \Theta_y + \Theta_x \otimes (\Theta_y A_{FD}^y) \\ &= A_x \otimes \Theta_y + \Theta_x \otimes A_y \end{aligned} \tag{2.3}$$

where

$$\Theta_x = diag\left(\frac{h_x^i + h_x^{i+1}}{2}\right);$$

$$\Theta_y = diag\left(\frac{h_y^i + h_y^{i+1}}{2}\right),$$

and $A_x \equiv \Theta_x A_{FD}^x$, $A_y \equiv \Theta_y A_{FD}^y$ are tridiagonal SPD matrices. When $c(x) \equiv 0$ and $d(y) \equiv 0$, the 5-point-start finite difference stencil (2.3) is identical to the linear finite element stiffness matrix.

Suppose Ω is decomposed into two subdomains Ω_1 and Ω_2 by an interface Γ which is a horizontal grid line and there are m_1 and m_2 interior horizontal grid lines in Ω_1 and Ω_2, respectively. It is easy to see that $m_1 + m_2 + 1 = n_y$. Assume that the horizontal grid lines are ordered from the boundary towards Γ for each subdomain, then we can correspondingly write A_y and Θ_y as

$$A_y = \begin{bmatrix} A_y^1 & 0 & \beta_y^{10}\mathbf{e}_{m_1} \\ 0 & A_y^2 & \beta_y^{20}\mathbf{e}_{m_2} \\ \beta_y^{10}\mathbf{e}_{m_1}^T & \beta_y^{20}\mathbf{e}_{m_2}^T & \alpha_y^0 \end{bmatrix} \tag{2.4}$$

and

$$\Theta_y = \begin{pmatrix} \Theta_y^1 & 0 & 0 \\ 0 & \Theta_y^2 & 0 \\ 0 & 0 & \theta_y^0 \end{pmatrix} \tag{2.5}$$

where A_y^i and Θ_y^i are the corresponding tridiagonal and diagonal matrices for Ω_i with the proper ordering, and $\mathbf{e}_k$ is the unit vector of order k. The matrix A also has the block form

$$A = \begin{bmatrix} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ B_1^T & B_2^T & D \end{bmatrix} \tag{2.6}$$

where

$$\begin{aligned} A_i &= A_x \otimes \Theta_y^i + \Theta_x \otimes A_y^i, \quad i = 1,2, \\ D &= \theta_y^0 A_x + \alpha_y^0 \Theta_x, \\ B_i &= \beta_y^{i0} \Theta_x \otimes \mathbf{e}_{m_i}, \qquad i = 1,2. \end{aligned} \tag{2.7}$$

The interface matrix:

$$S \equiv D - \sum_{i=1}^{2} B_i^T A_i^{-1} B_i, \tag{2.8}$$

which is called the *Schur Complement of* $diag(A_i)$ *in* A and denoted by $(A/diag(A_i))$ [7], plays a key role in domain decomposition-based methods. Theorem 2.1 states that S can be expressed as a matrix function.

THEOREM 2.1. *Let*

$$\begin{aligned} T_x &\equiv \Theta_x^{-1/2} A_x \Theta_x^{-1/2}, \\ T_y^i(t) &\equiv t\Theta_y^i + A_y^i, \qquad i = 1,2, \end{aligned} \tag{2.9}$$

then the Schur complement can be expressed as

$$S = \Theta_x^{1/2} f(T_x) \Theta_x^{1/2}, \tag{2.10}$$

where

$$f(t) = (\theta_y^0 t + \alpha_y^0) - \sum_{i=1}^{2} \left(\frac{\beta_y^{i0}}{l^i_{m_i m_i}} \right)^2, \tag{2.11}$$

and $l^i_{m_i m_i}$ is the last diagonal element of the Cholesky factor of $T_y^i(t)$.

Proof. From (2.7) and (2.8) we have

$$S = (\theta_y^0 A_x + \alpha_y^0 \Theta_x) - \sum_{i=1}^{2} (\beta_y^{i0} \Theta_x \otimes \mathbf{e}_{m_i}^T) A_i^{-1} (\beta_y^{i0} \Theta_x \otimes \mathbf{e}_{m_i}). \tag{2.12}$$

To express S as a matrix function we want to change Θ_x to I_{n_x} in the right hand side of (2.12) in order to use the Kronecker product properties. Multiplying (2.12) by $\Theta_x^{-1/2}$ symmetrically, we have

$$\Theta_x^{-1/2} S \Theta_x^{-1/2} = (\theta_y^0 T_x + \alpha_y^0 I_{n_x}) - \sum_{i=1}^{2} (\beta_y^{i0})^2 (I_{n_x} \otimes \mathbf{e}_{m_i}^T) \tilde{A}_i^{-1} (I_{n_x} \otimes \mathbf{e}_{m_i}), \tag{2.13}$$

where

$$\begin{aligned} \tilde{A}_i &= (\Theta_x^{-1/2} \otimes I_{m_i}) A_i (\Theta_x^{-1/2} \otimes I_{m_i}) \\ &= (\Theta_x^{-1/2} \otimes I_{m_i})(A_x \otimes \Theta_y^i + \Theta_x \otimes A_y^i)(\Theta_x^{-1/2} \otimes I_{m_i}) \\ &= T_x \otimes \Theta_y^i + I_{n_x} \otimes A_y^i. \end{aligned} \tag{2.14}$$

Using the properties of the Kronecker product, we can express the right hand side of (2.13) as a function of T_x by formally substituting T_x by t, I_{n_x} by 1, and $\otimes$ by a normal product, which leads to

$$\Theta_x^{-1/2} S \Theta_x^{-1/2} = f(T_x) \tag{2.15}$$

with $f(t)$ defined by

$$f(t) = (\theta_y^0 t + \alpha_y^0) - \sum_{i=1}^{2} (\beta_y^{i0})^2 \mathbf{e}_{m_i}^T \left[T_y^i(t) \right]^{-1} \mathbf{e}_{m_i}. \tag{2.16}$$

It is easy to verify by forward and back substitutions that

$$\mathbf{e}_{m_i}^T \left[T_y^i(t) \right]^{-1} \mathbf{e}_{m_i} = (l^i_{m_i m_i})^{-2}, \quad i = 1, 2. \tag{2.17}$$

Combining (2.15) through (2.17) we thus complete the proof. □

We make several remarks about Theorem 2.1.

Remark 2.1. The function $f(t)$ only depends on the operators A_y, Θ_y, and the location of Γ, namely m_1 and m_2. It is independent of any information in the x-direction.

Remark 2.2. To evaluate $f(t)$ at a given point using (2.11) one only needs to factor two tridiagonal matrices. From the proof of Theorem 2.1, it is seen that actual forward and backward substitutions in (2.17) can also be avoided by the special ordering of horizontal grid lines for each subdomain. In addition, no storage is required in the Cholesky factorization since only the last diagonal element $l^i_{m_i m_i}$ is used in (2.11) for each i.

Remark 2.3. Instead of using the Cholesky factorization, one can use the eigen-decomposition for $\left[\Theta^i_y\right]^{-1/2} A^i_y \left[\Theta^i_y\right]^{-1/2}$ to get

$$T^i_y(t) = \left[\Theta^i_y\right]^{1/2} W_i(tI_{m_i} + \Lambda_i) W_i^T \left[\Theta^i_y\right]^{1/2} \tag{2.18}$$

where $\Lambda_i = diag(\lambda_i(j))$ and W_i are eigenvalues and eigenvectors for $\left[\Theta^i_y\right]^{-1/2} A^i_y \left[\Theta^i_y\right]^{-1/2}$. Let $\mathbf{z}_i = W_i^T \left[\Theta^i_y\right]^{-1/2} \mathbf{e}_{m_i}$, then we have, from (2.16),

$$f(t) = (\theta^0_y t + \alpha^0_y) - \sum_{i=1}^{2} (\beta^{i0}_y)^2 \sum_{j=1}^{m_i} \frac{(z_i(j))^2}{t + \lambda_i(j)} \tag{2.19}$$

where $z_i(j)$ is the $j-$th element of $\mathbf{z}_i$.

We can obtain another expression of $f(t)$ by introducing two sets of orthogonal polynomials $\{P^i_j(t)\}$ and $\{Q^i_j(t)\}$ defined in terms of the three-term recurrence relation, for $i = 1, 2$,

$$\begin{aligned}
P^i_{-1}(t) &= 0; \\
P^i_0(t) &= 1; \\
\beta^i_j P^i_j(t) &= (\theta^i_y(j)t + \alpha^i_j) P^i_{j-1}(t) - \beta^i_{j-1} P^i_{j-2}(t), \\
&\quad j = 1, 2, \ldots, m_i,
\end{aligned} \tag{2.20}$$

$$\begin{aligned}
Q^i_{-1}(t) &= 0; \\
Q^i_0(t) &= 1; \\
\beta^i_{m_i - j} Q^i_j(t) &= (\theta^i_y(m_i - j + 1)t + \alpha^i_{m_i - j + 1}) Q^i_{j-1}(t) - \beta^i_{m_i - j + 1} Q^i_{j-2}(t), \\
&\quad j = 1, \ldots, m_i,
\end{aligned}$$

where $A^i_y \equiv [\beta^i_{j-1},\ \alpha^i_j,\ \beta^i_j]$ and $\theta^i_y(j)$ is the $j-$th diagonal element of Θ^i_y. Then, similar to (2.12) in [1] (note Q^i_j here is R^i_j in [1]), we have

$$\{[T_y^i(t)]^{-1}\}_{qs} = \begin{cases} \dfrac{P_{s-1}^i(t)Q_{m_i-q}^i(t)}{\beta_{m_i}^i P_{m_i}^i(t)}, & q \geq s \\ \dfrac{P_{q-1}^i(t)Q_{m_i-s}^i(t)}{\beta_{m_i}^i P_{m_i}^i(t)}, & s \geq q. \end{cases} \tag{2.21}$$

Therefore, (2.16) can be written in terms of these polynomials as

$$f(t) = (\theta_y^0 t + \alpha_y^0) - \sum_{i=1}^{2} \left(\beta_y^{i0}\right)^2 \frac{P_{m_i-1}^i(t)}{\beta_{m_i}^i P_{m_i}^i(t)}. \tag{2.22}$$

By setting $\beta_{m_i}^i \equiv \beta_y^{i0}$ in (2.20), (2.22) becomes

$$f(t) = \frac{(\theta_y^0 t + \alpha_y^0)P_{m_1}^1(t)P_{m_2}^1(t) - \beta_y^{10}P_{m_1-1}^1(t) - \beta_y^{20}P_{m_2-1}^1(t)}{P_{m_1}^1(t)P_{m_2}^2(t)}. \tag{2.23}$$

Similar to (2.20), if we define another set of orthogonal polynomials $\{P_j(t)\}$ according to the global tridiagonal matrix for the operator $t\Theta_y + A_y$ with the natural bottom-to-top ordering for all interior horizontal grid lines in Ω, then by denoting $t\Theta_y + A_y \equiv [\beta_{j-1},\ \theta_j t + \alpha_j,\ \beta_j]$ with $\beta_{n_y} = 1$, it can be verified that (2.23) can be written as

$$f(t) = \frac{P_{n_y}(t)}{P_{m_1}^1(t)P_{m_2}^2(t)}. \tag{2.24}$$

Therefore, $\{P_j(t)\}$, $\{P_j^1(t)\}$ and $\{P_j^2(t)\}$ play a role analogous to the Chebyshev polynomials as in the constant coefficient and uniform grid case. The expressions (2.19) and (2.24) for $f(t)$, or similarly for $1/f(t)$, can be viewed as an extension of the work in [1] and [14] to the nonuniform grid case. They are all equivalently related to each other through the theory of orthogonal polynomials.

Remark 2.4. There are two reasons to favor the use of (2.11) to evaluate $f(t)$ in our approach. First, as shown in [11] and [12], $f(t)$ needs to be evaluated at only a few points for the interpolation. However, those approaches [14] using (2.19), where $f(t)$, or its reciprocal, is expanded into a sum of partial fractions, require the computation of all the eigenvalues and some of the eigenvectors of matrices that are related to A_y. Using (2.22) with the three-term recurrence requires about the same work as using (2.11). Expression (2.24) has a better mathematical form than (2.22), but it requires about as twice as much work as that of (2.22) because of computing $P_{n_y}(t)$. The second reason is the numerical stability problem, which is even more important. Numerical instability occurs especially when grids are very nonuniform. Numerical experiments show that the eigenvalue problem is often very ill-conditioned and that the three-term recurrence computation is also very unstable. This can be easily seen from (2.20) because usually $\theta_y^i(j)t + \alpha_j^i >> \beta_j^i$, namely, a wrong pivot is used in the computation. Fortunately, the matrices $T_y^i(t)$, $i = 1, 2$, are *SPD*, and therefore, the Cholesky factorization is numerically stable.

Remark 2.5. Finally, it is easy to see that all the theory developed in [1] can be similarly extended to the nonuniform grid case along the line of argument in Remark 2.3. Therefore, the marching algorithms in [1] and the parallel direct solvers in [9] can be correspondingly extended in a trivial way.

3. Function approximation and preconditioners. This section discusses finding a simple function $r(t)$ that approximates $f(t)$ in (2.11) such that conditions (1.2) are satisfied. Therefore, the matrix

$$M \equiv \Theta_x^{1/2} r(T_x) \Theta_x^{1/2} \tag{3.1}$$

is a good preconditioner for S in (2.10) when the *PCG* method is applied because

$$\begin{aligned} \kappa(M^{-1}S) &= \kappa(\Theta_x^{-1/2} q(T_x) \Theta_x^{1/2}) \\ &= \frac{\max\limits_{t_i \in \sigma(T_x)} |q(t_i)|}{\min\limits_{t_i \in \sigma(T_x)} |q(t_i)|}. \end{aligned} \tag{3.2}$$

As discussed in [11], a natural candidate for $r(t)$ is a rational function of low degree. If $r(t) = \Pi_k (t - a_k)/(t - b_k)$, then $M^{-1}S$ can be computed by a sequence of solves and multiplies with tridiagonal matrices since T_x is tridiagonal. We first describe a general approach to construct such a rational approximation by the *weighted rational Chebyshev approximation*

$$\min_{r(t) \in R_m^l} \max_{t \in [a,b]} \left| \frac{g(t) - r(t)}{w(t)} \right| \tag{3.3}$$

where $g(t)$ is a target function to be approximated, $w(t)$ is a weight function, and R_m^l is the approximation function space

$$\begin{aligned} R_m^l = \{ r(t) \;\; | \;\; & r(t) = \tfrac{p_l(t)}{q_m(t)}, \\ & p_l(t) \text{ and } q_m(t) \text{ are polynomials of degree } l \text{ and } m, \text{respectively} \}. \end{aligned}$$

THEOREM 3.1. *Assume that $r(t)$ is the optimal solution of (3.3) with $g(t) = f(t)$, $w(t) = f(t)$, $a = \lambda_{min}(T_x)$, $b = \lambda_{max}(T_x)$. Let*

$$\varepsilon = \max_{t \in [a,b]} \left| \frac{f(t) - r(t)}{f(t)} \right|, \tag{3.4}$$

and assume $\varepsilon < 1$, then we have

$$\kappa(M^{-1}S) \leq \frac{1 + \varepsilon}{1 - \varepsilon} \tag{3.5}$$

Proof. From (3.4), we have

$$|1 - \frac{r(t)}{f(t)}| \leq \varepsilon, \qquad \forall t \in [a, b]. \tag{3.6}$$

This can be written as

$$1 - \varepsilon \leq \frac{r(t)}{f(t)} \leq 1 + \varepsilon, \qquad \forall t \in [a, b]. \tag{3.7}$$

It follows that

$$\begin{aligned} \max_{t_i \in \sigma(T_x)} |q(t_i)| &\leq \frac{1}{1-\varepsilon}; \\ \min_{t_i \in \sigma(T_x)} |q(t_i)| &\geq \frac{1}{1+\varepsilon}. \end{aligned} \tag{3.8}$$

Notice that

$$\begin{aligned} \kappa(M^{-1}S) &\equiv \frac{\lambda_{\max}(M^{-1}S)}{\lambda_{\min}(M^{-1}S)} \\ &= \frac{\lambda_{\max}(\Theta_x^{-1/2} q(T_x) \Theta_x^{1/2})}{\lambda_{\min}(\Theta_x^{-1/2} q(T_x) \Theta_x^{1/2})} \\ &= \frac{\max\limits_{t_i \in \sigma(T_x)} |q(t_i)|}{\min\limits_{t_i \in \sigma(T_x)} |q(t_i)|}. \end{aligned} \tag{3.9}$$

Then (3.5) is obtained from (3.8) and (3.9). This completes the proof. □

There are many efficient algorithms devised for the weighted rational Chebyshev approximation, for example, see [13]. The error ε in (3.4) depends on R_m^l, i.e., the degrees l and m and, for fast convergence of the *PCG* method, one wishes ε to be as small as possible, which requires increasing l and m. On the other hand, using large l and m implies high expense in solving the preconditioning system and also, of less importance, in solving the weighted rational Chebyshev approximation problem. So there is a trade-off in choosing l and m properly.

Another approach as proposed in [11] is a more intuitive strategy by observing that $f(t)$ has a *two-part* property. That is, $f(t)$ looks mostly like a linear function, this part is called the *easy part*. At the left end region of $\sigma(T_x)$ the few smallest eigenvalues make $f(t)$ behave like $t^{1/2}$, this part is called the *hard part*. Furthermore, for a uniform y-direction grid, [11] shows that $f(t)$ does not depend very much on the location of Γ, i.e., on m_1 and m_2, the difference can only be seen in the hard part. If we use a simple rational function

$$z(t) \equiv \frac{at + b}{ct + d} \tag{3.10}$$

to construct $r(t)$ in two phases, then $z(t)$ can be easily determined by three interpolating points using divided differences. The corresponding preconditioning problem, after a scaling, reads

$$(Tx + e_1 I_{n_x})\mathbf{u} = (T_x + e_2 I_{n_x})\mathbf{v}, \tag{3.11}$$

where $e_1 = b/a$ and $e_2 = d/c$. The linear system (3.11) can be solved by

$$\mathbf{u} = \mathbf{v} + (e_2 - e_1)(T_x + e_1 I_{n_x})^{-1}\mathbf{v} \tag{3.12}$$

using a *SPD* tridiagonal solver. We construct $r(t)$ as follows: First use $r_1(t)$ of the form (3.10) to remove the hard part from $f(t)$ by

$$f_1(t) = f(t) \,/\, r_1(t) \tag{3.13}$$

where $r_1(t)$ approximates well the hard part of $f(t)$. It is natural to compute $r_1(t)$ by interpolating the first three smallest eigenvalues of T_x. Observe that for small eigenvalues we can use T_x^* instead of T_x to get fairly good estimates, where T_x^* is the analog of T_x when the x-direction grid is made uniform. For a constant coefficient operator L_x, there exists an analytic expression for the eigenvalues of T_x^* so that one can avoid completely computing eigenvalues for T_x. Now $f_1(t)$ is almost a linear function, so we can find a good approximation $r_2(t)$ to it of the form (3.10). We choose the first interpolating point the same as for $r_1(t)$, the other two are chosen as the two largest eigenvalues of T_x; only rough estimates of them are required. Thus, we define $r(t)$ as

$$r(t) \equiv r_1(t) r_2(t). \tag{3.14}$$

Examples in [11] show that this *two-phase* strategy is very effective for the case of constant coefficient operators and uniform y-direction grids. In general, the behavior of $f(t)$ depends on the y-direction information including the grid nonuniformity, the location of Γ, and the operator coefficients. A detailed experimental study about this dependence is found in [12]. In any case, a particular $f(t)$ has the so-called *two-part* property, therefore, the two-phase approximation strategy can also be applied. We give one example to illustrate the effectiveness of this approach and refer to [12] for more extensive experimental results. In this experiment, we solve a model problem of Poisson equation with Dirichlet condition on a unit square domain using a nonuniform grid as shown in Fig.3.1. The effects of variable coefficients of the operator are similar to those of the grid nonuniformity. The grid size is 61×33. The spacings in each direction are of an exponential distribution to account for an exponential type of singularity in the solution of (2.1). More specifically, for the x direction the distribution used is

$$h_x^i = min\{x_i^\alpha, 0.1\},$$

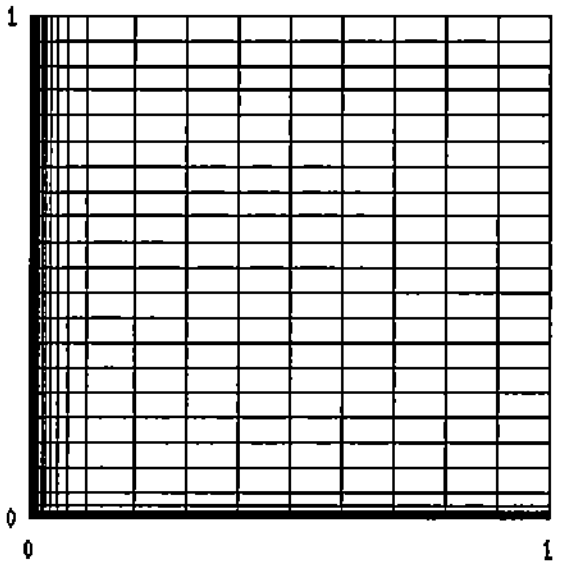

Figure 3.1. *The nonuniform grid used in the experiment is refined along the axes using the distributions of $y^{1.2}$ (vertical) and $x^{1.5}$ (horizontal).*

where $\alpha = 1.5$ and we start from $x = 1$ towards $x = 0$ until the specified number of grid points is reached, which is denoted as an *x^α-distribution.* Similarly for the y-direction, we take $\alpha = 1.2$ and $max_i\{h_y^i\} = 0.05$. The interface Γ is chosen such that $m1 = m2 = 15$, namely the work is equi-partitioned for two subdomains.

The two-phase approximation strategy is used to construct the rational approximation for the preconditioner in this example. The corresponding function curves involved in the rational approximation are shown in Fig.3.2 through Fig.3.4. They illustrate the two-part behavior of $f(t)$, its domain and range used in the approximation, and the behavior of $q(t)$ that determines the convergence behavior of the *PCG* method. The condition number of the preconditioned system is 1.106.

With this preconditioner, the *PCG* method converges in only 4 steps. The least squares error for the last two iterates is 2.9×10^{-6} (single precision 32 bits is used in the computation). In contrast, the ordinary *CG* method does not converge after 100 steps and its error is 4.2×10^{-3} at this point. It is also important to notice that the matrix T_x, or similarly T_y, is too ill-conditioned for the eigendecomposition approach so that no useful information is generated by standard *IMSL* eigenvalue subroutines. Therefore, any approach that requires eigendecomposition cannot be applied for this case.

4. Preconditioning for nonrectangular domains. To precondition the *PCG* method for a nonrectangular domain, it is natural to use an embedded rectangular domain to reduce the nonrectangular problem to a rectangular one because the remote parts of the domain have a less significant effect on the interface Schur complement. More specifically, let S be the Schur complement for the nonrectangular domain Ω, Ω_0 be the embedded rectangle by shifting Γ up to $\partial\Omega$ in both directions, and S_0 be the corresponding Schur complement for Ω_0. Further, let M be a preconditioner for Ω_0 (one of those discussed earlier). The combined effect of these two preconditioners is given by Theorem 4.1.

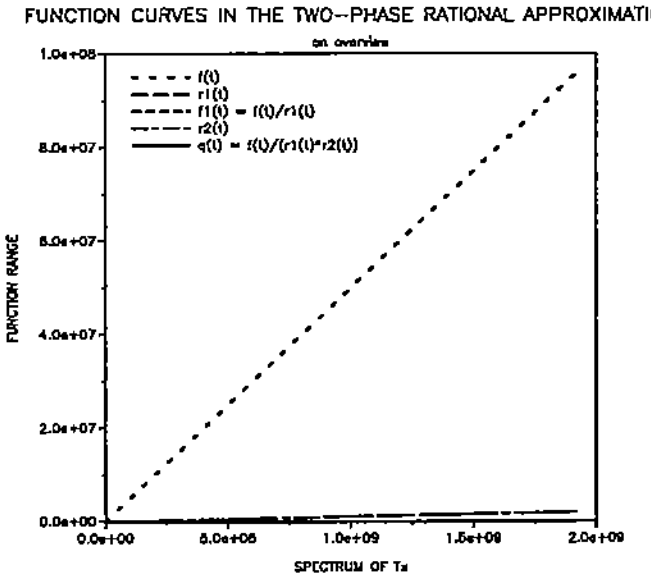

Figure 3.2. *The function f(t), its approximations and resulting q(t). On this scale one only sees the linear part of f(t); the curve q(t) is superimposed on the x-axis.*

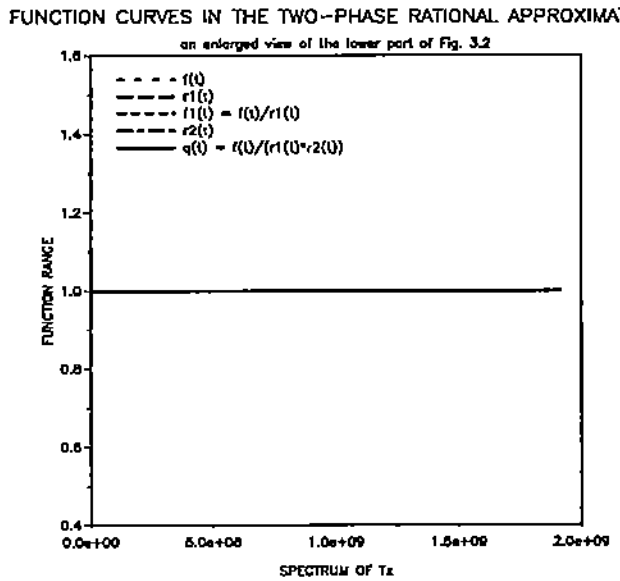

Figure 3.3. *An enlarged view of the lower part of Fig. 3.2 shows that q(t) is nearly 1.0 everywhere and the other functions rise along the y-axis and immediately go off the plot.*

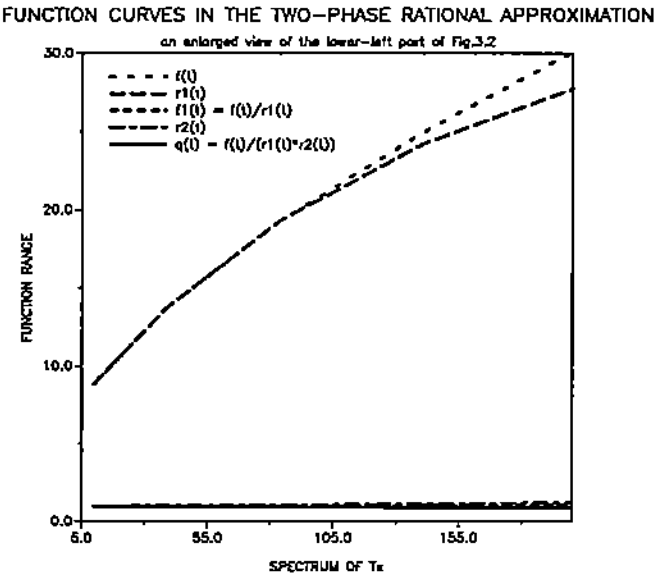

Figure 3.4. *An enlarged view of the lower left part of Fig. 3.2 showing the hard part of $f(t)$. The approximation $r_1(t)$ fits $f(t)$ well for the small eigenvalues; $f_1(t)$ is nearly linear and rises much more slowly than $f(t)$. The approximation $r_2(t)$ fits $f_1(t)$ well and the resulting $q(t)$ is very close to 1.0.*

THEOREM 4.1.

$$\kappa(M^{-1}S) \leq \kappa(S_0^{-1}S)\kappa(M^{-1}S_0). \tag{4.1}$$

Proof. From the definition, we have

$$\kappa(M^{-1}S) = \frac{\lambda_{\max}(M^{-1}S)}{\lambda_{\min}(M^{-1}S)}. \tag{4.2}$$

Notice that M, S_0 and S are all *SPD*. First, we have the estimate

$$\begin{aligned}
\lambda_{\max}(M^{-1}S) &= \lambda_{\max}(S_0^{1/2}M^{-1}SS_0^{-1/2}) \\
&\leq ||S_0^{1/2}M^{-1}SS_0^{-1/2}||_2 \\
&= ||(S_0^{1/2}M^{-1}S_0^{1/2})(S_0^{-1/2}SS_0^{-1/2})||_2 \\
&\leq ||S_0^{1/2}M^{-1}S_0^{1/2}||_2||S_0^{-1/2}SS_0^{-1/2}||_2 \\
&= \lambda_{\max}(S_0^{1/2}M^{-1}S_0^{1/2})\lambda_{\max}(S_0^{-1/2}SS_0^{-1/2}) \\
&= \lambda_{\max}(M^{-1}S_0)\lambda_{\max}(S_0^{-1}S).
\end{aligned} \tag{4.3}$$

Similarly, we have

$$
\begin{aligned}
1/\lambda_{\min}(M^{-1}S) &= \lambda_{\max}(S^{-1}M) \\
&\leq \lambda_{\max}(S^{-1}S_0)\lambda_{\max}(S_0^{-1}M) \\
&= \frac{1}{\lambda_{\min}(S_0^{-1}S)\lambda_{\min}(M^{-1}S_0)}.
\end{aligned} \tag{4.4}
$$

Therefore, we obtain

$$
\begin{aligned}
\kappa(M^{-1}S) &\leq \frac{\lambda_{\max}(S_0^{-1}S)\lambda_{\max}(M^{-1}S_0)}{\lambda_{\min}(S_0^{-1}S)\lambda_{\min}(M^{-1}S_0)} \\
&= \kappa(S_0^{-1}S)\kappa(M^{-1}S_0).
\end{aligned} \tag{4.5}
$$

The proof is complete. □

The bounds (1.6) imply that there is a generic constant upper bound for the factor $\kappa(S_0^{-1}S)$ in (4.1). More accurate bounds are derived in [6] for some particular domains. From those results, we have the following corollary.

COROLLARY 4.2. *Assume that the operator is a Laplacian. For all L-shaped and T-shaped domains, we have*

$$\kappa(M^{-1}S) \leq 2\kappa(M^{-1}S_0); \tag{4.6}$$

and for all C-shaped domains, we have

$$\kappa(M^{-1}S) \leq (2+\sqrt{2})\kappa(M^{-1}S_0). \tag{4.7}$$

For other operators, similar techniques in [6] can be applied to derive corresponding upper bounds.

5. Conclusions. This paper extends the approach of constructing preconditioners through a function approximation, presented in [11], to more general cases where grids can be nonuniform; operators can have variable coefficients but are separable and self-adjoint; and domains can be nonrectangular. Theoretical and experimental results show that this new approach is very simple, effective and efficient. The extended theory of expressing the Schur complement, or the original matrix, as a function of simple matrix can be applied for other purposes, such as fast direct solvers and in parallel computations.

Acknowledgment. We would like to thank Professor E. Gallopoulos for pointing us to the recent work [14] and his own work [9] that are related to our approach to some extent.

REFERENCES

[1] R.E. Bank (1977), *Marching algorithms for elliptic boundary value problems. II: the variable coefficient case*, SIAM J. Numer. Anal., 14, pp. 950–970.

[2] R.E. Bank and D.J. Rose (1977), *Marching algorithms for elliptic boundary value problems. I: the constant coefficient case*, SIAM J. Numer. Anal., 14, pp. 792–829.

[3] P.E. Bjorstad and O.B. Widlund (1984), *Solving elliptic problems on regions partitioned into substructures*, in Elliptic Problem Solvers II, (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, pp. 245–256.

[4] J.H. Bramble, J.E. Pasciak and A.H. Schatz (1986), *The construction of preconditioners for elliptic problems by substructuring, I*, Math. Comp., 46, pp. 361–369.

[5] T.F. Chan (1987), *Analysis of preconditioners for domain decomposition*, SIAM J. Numer. Anal., 24, pp. 382–390.

[6] T.F. Chan, T.Y. Hou and P.L. Lions (1991), *Geometry related convergence results for domain decomposition algorithms*, SIAM J. Numer. Anal., 28, pp. 378–391.

[7] R.W. Cottle (1974), *Manifestations of the Schur complement*, Linear Algebra and its Applications, 8, pp. 189–211.

[8] M. Dryja (1982), *A capacitance matrix method for Dirichlet problem on polygonal region*, Numer. Math., 39, pp. 51–64.

[9] E. Gallopoulos and J, Saad, (1989), *Some fast elliptic solvers on parallel architectures and their complexities*, International J. High Speed Computing, 1, pp. 113-141.

[10] G.H. Golub and D. Mayers (1983), *The use of pre-conditioning over irregular regions*, Lecture at Sixth Int. Conf. on Computing Methods in Applied Sciences and Engineering, Versailles, France.

[11] M. Mu (1992), *A new family of preconditioners for domain decomposition*, CSD-TR-92-064 and CER-92-27, Department of Computer Sciences, Purdue University, West Lafayette, IN47907, September, (submitted to SIAM J. Sci. Stat. Comput.).

[12] M. Mu and J.R. Rice (1993), *Constructing preconditioners with rational approximation*, Proceedings of the Sixth SIAM Conference on Parallel Processing and Scientific Computing, edited by R.F. Sincovec, SIAM, Philadelphia, PA, to appear.

[13] J. R. Rice (1992), *Numerical Methods, Software, and Analysis*, Chapter 11, Second Edition, Academic Press, Cambridge, MA.

[14] S. Sander (1992), *Domain decomposition and rational approximation problems*, Proceedings of the IMACS Symposium on Iterative Methods in Linear Algebra, April, 1991, Brussels, Belgium, P. De Groen and R. Beauwens ed., Elsevier Science Pub. Co., 1992.