Fast Fourier Transform Solvers and Preconditioners for Quadratic Spline Collocation

Christina C. Christara and Kit Sun Ng
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 3G4, Canada
{ccc,ngkit}@cs.utoronto.ca

submitted December 2000, revised October 2001, February 2002

Abstract

Quadratic Spline Collocation (QSC) methods of optimal order of convergence have been recently developed for the solution of elliptic Partial Differential Equations (PDEs). In this paper, linear solvers based on Fast Fourier Transforms (FFT) are developed for the solution of the QSC equations. The complexity of the FFT solvers is $O(N^2 \log N)$, where N is the gridsize in one dimension. These direct solvers can handle PDEs with coefficients in one variable or constant, and Dirichlet, Neumann and periodic boundary conditions, along at least one direction of a rectangular domain. General variable coefficient PDEs are handled by preconditioned iterative solvers. The preconditioner is the QSC matrix arising from a constant coefficient PDE. The convergence analysis of the preconditioner is presented. It is shown that, under certain conditions, the convergence rate is independent of the gridsize. The preconditioner is solved by FFT techniques, and integrated with one-step or acceleration methods, giving rise to asymptotically almost optimal linear solvers, with complexity $O(N^2 \log N)$. Numerical experiments verify the effectiveness of the solvers and preconditioners, even on problems more general than the analysis assumes. The development and analysis of FFT solvers and preconditioners is extended to QSC equations corresponding to systems of elliptic PDEs.

Keywords: spline collocation, elliptic boundary value problem, eigenvalue problem, fast Fourier transform, iterative solver, scaled Laplace preconditioner, system of PDEs.

AMS Subject Classification: 65N22,65N25,65N35,65T50,65F10,65F15.

1 Introduction

Optimal convergence order collocation methods based on smooth splines have been relatively recently developed [19], [4], for the solution of elliptic Boundary Value Problems (BVPs). These methods offer an alternative to Galerkin finite element methods as well as to Hermite spline collocation methods. It is known that collocation is a simple-to-implement and integration-free method. Its implementation is problem independent, requiring only one function evaluation per data point. Compared to collocation methods based on not-fully-smooth splines, e.g. Hermite cubic splines, smooth spline collocation methods give rise to smaller linear systems, since they use only one data point per subrectangle. Moreover, the linear systems arising from the so-called deferred-correction spline collocation methods are sparser than the respective ones from Hermite spline collocation and spline Galerkin methods.

In the numerical solution of (multi-dimensional) BVPs, the larger part of the overall computation time is spent in solving the resulting linear system of equations. Therefore, the development of a fast solver associated with a new discretisation method is essential for the success of the method. A variety of solvers for spline collocation equations has been studied [18], [5], including acceleration techniques with various

preconditioners and domain decomposition, but the analysis has been carried out for only a few solvers and for restricted classes of PDE operators. For example, in [8], multigrid methods for quadratic splines are developed and proven to have rate of convergence independent of the problem size for a certain model problem.

Hermite cubic spline collocation on the Gauss points gives fourth order of convergence approximations [11], [24]. Fast Fourier Transform (FFT) solvers for Hermite cubic spline collocation equations arising from Poisson's problem are developed in [3] and extended to PDEs with coefficients in one variable. These solvers are used as preconditioners for more general PDE problems with Dirichlet conditions in [1]. The convergence rate of the Richardson and Minimum Residual (MRES) preconditioned iterative methods is shown to be independent of the gridsize, by showing the spectral equivalence of the Hermite cubic spline operators corresponding to the Laplace and to a more general elliptic PDE operator with Dirichlet boundary conditions.

In its standard formulation, Quadratic Spline Collocation (QSC) on the midpoints of a uniform rectangular partition gives second (sub-optimal) order of convergence approximations [4]. In [4], optimal QSC methods are derived and analyzed; more specifically, the QSC collocation approximation is fourth order on the nodes and midpoints of a uniform rectangular partition, and third order globally. The derivation of optimal QSC methods is based on appropriate perturbations of either the operator, giving rise to the so-called *one-step* or *extrapolated* QSC method, or of the right side, giving rise to the so-called *two-step* or *deferred-correction* QSC method. In [4], the QSC linear system arising from the two-step QSC method applied to PDEs with constant coefficients and even derivative terms is written in a tensor product form and formulae for the eigenvalues and eigenvectors of the matrix are derived.

FFT solvers for the two-step QSC system arising from Helmholtz PDEs with constant coefficients are developed in [9]. The FFT is applied to both dimensions. Dirichlet, Neumann and periodic conditions are handled. The QSC linear system arising from general elliptic PDEs with variable coefficients is solved by preconditioned iterative methods, with the preconditioner being a diagonal scaling of the QSC matrix arising from a Helmholtz operator. The experiments show that the rate of convergence of the preconditioned iterative methods is independent of the grid size, however, no analysis is given.

In this paper, we consider the analysis of the convergence rate and present a number of additional results. We summarize the relevant properties of the QSC matrix in Section 2, and the formulation of the FFT solvers for QSC equations arising from Helmholtz PDEs with constant coefficients in Section 3. We extend the FFT solvers to PDEs with coefficients in one variable, and boundary conditions of any type along the direction of the other variable, by applying the FFT to one dimension and tridiagonal solves to the other. We also consider a 2×2 system of elliptic PDEs, and the optimal QSC discretisation method for this problem as developed in [22] and [7], and we formulate FFT solvers for the arising QSC matrix. These solvers can be easily extended to $n \times n$ systems of elliptic PDEs.

In Section 4, using a technique similar to the one used in [1] for Hermite cubic splines, we show that the QSC operators corresponding to the Laplace and to a general elliptic PDE operator without cross-derivative term and with Dirichlet boundary conditions are spectrally equivalent, and that the preconditioned Richardson and MRES iterative methods applied to the QSC matrix arising from a general elliptic PDE without cross-derivative term with preconditioner the QSC matrix corresponding to the Laplace operator have convergence rate independent of the gridsize. We extend the results to a general 2×2 system of elliptic PDEs.

The technique used in [1] to show the spectral equivalence of the Hermite cubic spline operators is based, among other, on a discrete inner product defined to match the two-point Gauss quadrature rule, on various relations regarding orthogonal spline collocation operators shown in other papers [12], [23], [10], [2], and on the eigenvalues and eigenfunctions for a model discrete eigenvalue problem, as presented in

[3].

For quadratic splines (and other smooth splines), these results are not given, and the proof techniques used for these results in the case of orthogonal collocation are not directly applicable to spline collocation. (Note that the results in [12], [23] hold for $\mathbf{C}^1$ piecewise polynomials of degree $r \geq 3$.) Reasons for this include the facts that spline collocation is not necessarily associated with orthogonal polynomials, and that it uses only one data point per subinterval. In general, there is less literature on the area of smooth spline collocation than on orthogonal $\mathbf{C}^1$ piecewise polynomial collocation, since optimal spline collocation methods have been relatively recently developed. Moreover, the fact that the basis functions used with quadratic (and other smooth) spline collocation are not nodal basis functions, i.e. the values of the coefficients do not represent function values at particular points of the grid, complicates matters even more.

In Section 4, we fill a part of this literature gap. We prove a number of new mathematical results regarding the QSC operator, similar to those proven for the Hermite cubic spline operator in [12], [23], [10], [2], and we develop the eigenvalues and eigenfunctions for a model QSC eigenvalue problem.

Finally, in Section 5, we present numerical results verifying the effectiveness of the solvers and of the preconditioners even when applied to PDEs more general than the ones assumed in the analysis.

2 Background

Consider a BVP described by the operator equation

$$\mathbf{L}u \equiv au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g \quad \text{for} \quad (x,y) \in \Omega \equiv (0,1) \times (0,1), \tag{1}$$

where a, b, c, d, e, f and g are given functions of x and y, and u is the unknown function of x and y, subject to some boundary conditions on the boundary $\partial\Omega$ of Ω. At each line of $\partial\Omega$ the boundary conditions may be any of the following types: *homogeneous Dirichlet*, *homogeneous Neumann*, or *periodic*. Note that some of the solvers that will be described are applicable to a wider range of boundary conditions (see Remark 1 of subsection 3.2 and Remark 3 of subsection 3.3). For brevity, in this section and subsection 3.1, we assume that the boundary conditions are homogeneous Neumann in the x direction and homogeneous Dirichlet in y, i.e.

$$u_x = 0 \text{ on } x = 0, x = 1 \text{ for } 0 \leq y \leq 1 \tag{2}$$

$$u = 0 \text{ on } y = 0, y = 1 \text{ for } 0 \leq x \leq 1 \tag{3}$$

Let $\Delta_x \equiv \{x_i = i/M, i = 0, \cdots, M\}$ and $\Delta_y \equiv \{y_j = j/N, j = 0, \cdots, N\}$ be uniform partitions of (0, 1) with step-sizes $h_x = \frac{1}{M}$ and $h_y = \frac{1}{N}$, respectively. We denote by S_{Δ_x} and S_{Δ_y} the quadratic spline spaces with respect to partitions Δ_x and Δ_y, respectively, constructed so that the splines satisfy the boundary conditions (2) and (3), respectively. The basis functions $\{\phi_i^x(x)\}_{i=1}^M$ and $\{\phi_j^y(y)\}_{j=1}^N$ for S_{Δ_x} and S_{Δ_y}, respectively, are generated through appropriate transformations of the model quadratic spline $\phi(x)$ defined by $\{\ \phi(x) \equiv x^2$ for $0 \leq x \leq 1$; $\phi(x) \equiv -3 + 6x - 2x^2$ for $1 \leq x \leq 2$; $\phi(x) \equiv 9 - 6x + x^2$ for $2 \leq x \leq 3$; $\phi(x) \equiv 0$ elsewhere$\}$, and appropriate adjustments to satisfy the boundary conditions. More specifically, let $\chi_i^x(x) \equiv \frac{1}{2}\phi(\frac{x}{h_x} - i + 2)$, for $i = 0, \cdots, M+1$, and $\chi_j^y(y) \equiv \frac{1}{2}\phi(\frac{y}{h_y} - j + 2)$ for $j = 0, \cdots, N+1$. Then $\phi_1^x = \chi_1^x + \chi_0^x$, $\phi_i^x = \chi_i^x$, $i = 2, \cdots, M-1$, $\phi_M^x = \chi_M^x + \chi_{M+1}^x$, $\phi_1^y = \chi_1^y - \chi_0^y$, $\phi_i^y = \chi_i^y$, $i = 2, \cdots, N-1$ and $\phi_N^y = \chi_N^y - \chi_{N+1}^y$. Let $\tau_i^x = (x_{i-1} + x_i)/2$, $i = 1, \cdots, M$ and $\tau_j^y = (y_{j-1} + y_j)/2$, $j = 1, \cdots, N$ be the midpoints of the partitions Δ_x and Δ_y, respectively.

Let $S_\Delta \equiv S_{\Delta_x} \otimes S_{\Delta_y}$ be the approximating space for the BVP (1)-(3). This space has dimension MN. Note that any $u_\Delta \in S_\Delta$ satisfies the boundary conditions by construction. The set of basis functions for S_Δ is chosen to be the tensor product $\{\phi_i^x(x)\phi_j^y(y)\}_{i=1,j=1}^{M,N}$ of quadratic B-splines in the x and y directions.

The two-step optimal quadratic spline collocation (QSC) method [4] determines an approximation $u_\Delta \in S_\Delta$ to u in two steps. In the first step, a bi-quadratic spline $U \in S_\Delta$ is computed so that it satisfies (1), on the set $\mathbf{T} \equiv \{(\tau_i^x, \tau_j^y), i = 1, \cdots, M, j = 1, \cdots, N\}$ of collocation points, i.e.

$$\mathbf{L}U = g \text{ on } \mathbf{T}. \tag{4}$$

The approximation U is of second order, i.e. non-optimal. In the second step, $u_\Delta \in S_\Delta$ is computed so that it satisfies a perturbed operator equation,

$$\mathbf{L}u_\Delta = g - \mathbf{P_L}U \text{ on } \mathbf{T}, \tag{5}$$

where $\mathbf{P_L}$ is a perturbation operator defined by stencils [4]. The resulting approximation u_Δ is of fourth order on the grid points and midpoints of the partition and third order globally, that is, it is of the same order as the bi-quadratic spline interpolant.

The linear equations resulting from (4), if ordered according to the natural ordering (without loss of generality, first bottom-up, then left-to-right), give rise to a $MN \times MN$ linear system $A\mathrm{x} = \mathrm{g}$, where A is a block-tridiagonal matrix with tridiagonal $N \times N$ blocks, the right-side vector g is a vector of values of $g(x, y)$ at the collocation points, and x is the vector of unknown coefficients (degrees of freedom) of the finite element representation of the bi-quadratic spline approximation U. It is instructive to note that equations (5) result in a linear system with the same matrix A and a perturbed right-side vector.

We next give the form of A in (4) and (5) for specific cases of operators $\mathbf{L}$. We will use the notation $trid\{\kappa_1, \kappa_2, \kappa_3\}$ to denote a tridiagonal matrix whose all sub-, main- and super-diagonal elements are equal to scalars κ_1, κ_2 and κ_3, respectively, except possibly a few elements which will be defined separately.

If the operator $\mathbf{L}$ is of Helmholtz type with constant coefficients, i.e. the PDE is

$$\mathbf{L}u \equiv au_{xx} + cu_{yy} + fu = g \text{ in } \Omega \tag{6}$$

where a, c and f are constants, then the QSC matrix takes the tensor product form

$$A \equiv \frac{1}{8}\left(\frac{a}{h_x^2} T_{-2}^{E,M} \otimes T_6^{D,N} + \frac{c}{h_y^2} T_6^{E,M} \otimes T_{-2}^{D,N} + \frac{f}{8} T_6^{E,M} \otimes T_6^{D,N}\right) \tag{7}$$

where $T_{-2}^{E,M}$ is a $M \times M$ tridiagonal matrix of the form $T_{-2}^{E,M} = trid\{1, -2, 1\}$, with $\left(T_{-2}^{E,M}\right)_{1,1} = -1$ and $\left(T_{-2}^{E,M}\right)_{M,M} = -1$, $T_6^{E,M} = T_{-2}^{E,M} + 8\mathbf{I}^M$, $\mathbf{I}^M$ is the $M \times M$ identity matrix, $T_{-2}^{D,N}$ is a $N \times N$ tridiagonal matrix of the form $T_{-2}^{D,N} = trid\{1, -2, 1\}$, with $\left(T_{-2}^{D,N}\right)_{1,1} = -3$ and $\left(T_{-2}^{D,N}\right)_{N,N} = -3$, $T_6^{D,N} = T_{-2}^{D,N} + 8\mathbf{I}^N$ and $\mathbf{I}^N$ is the $N \times N$ identity matrix. Note that the first superscript, D or E, of a matrix denotes the type of boundary conditions, Dirichlet or Neumann, respectively, inherited in the entries of the matrix.

Now consider a more general type of operator $\mathbf{L}$ than the one in (6). Let $\mathbf{L}$ have coefficients in one variable and no first order terms with respect to the other variable. Without loss of generality, assume that the PDE is

$$\mathbf{L}u \equiv au_{xx} + cu_{yy} + eu_y + fu = g \text{ in } \Omega \tag{8}$$

where a, c, e and f are functions of y. Then the QSC matrix takes the tensor product form

$$A \equiv \frac{1}{8}\left(\frac{1}{h_x^2} T_{-2}^{E,M} \otimes T_{6a}^{D,N} + \frac{1}{h_y^2} T_6^{E,M} \otimes T_{-2c}^{D,N} + \frac{1}{2h_y} T_6^{E,M} \otimes T_{0e}^{D,N} + \frac{1}{8} f T_6^{E,M} \otimes T_{6f}^{D,N}\right) \tag{9}$$

where $T_{6a}^{D,N} = D_a T_6^{D,N}$, $T_{-2c}^{D,N} = D_c T_{-2}^{D,N}$, $T_{0e}^{D,N} = D_e T_0^{D,N}$ and $T_{6f}^{D,N} = D_f T_6^{D,N}$, and where D_a, D_c, D_e and D_f are $N \times N$ diagonal matrices with the values of the functions a, c, e and f, respectively, on the points τ_j^y, $j = 1, \cdots, N$, and $T_0^{D,N}$ is a $N \times N$ tridiagonal matrix of the form $T_0^{D,N} = trid\{-1, 0, 1\}$, with $(T_0^{D,N})_{1,1} = 1$ and $(T_0^{D,N})_{N,N} = -1$.

3 FFT Solvers for Quadratic Spline Collocation Equations

3.1 Diagonalizations and Algorithms

In this section, we describe two algorithms for the direct solution of the QSC equations arising from (6) and (8). Each algorithm is based on a certain diagonalization (or block-diagonalization) of the matrix of QSC equations.

In [4] explicit formulae for the eigenvalues and eigenvectors of the matrices in (7) are derived. As noted in [9], both $T_{-2}^{E,M}$ and $T_6^{E,M}$ can be diagonalized by the inverse of the Discrete Cosine Transform II (DCT-II) [20] matrix C^M of size $M \times M$, and both $T_{-2}^{D,N}$ and $T_6^{D,N}$ can be diagonalized by the inverse of the Discrete Sine Transform II (DST-II) [20] matrix S^N of size $N \times N$, thus, the QSC matrix A of (7) can be diagonalized by the inverse of $\mathrm{C}^M \otimes \mathrm{S}^N$. That is,

$$A = ((\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1}) \Lambda (\mathrm{C}^M \otimes \mathrm{S}^N) \tag{10}$$

where Λ is a diagonal matrix with the eigenvalues of A. If $\Lambda_{-2}^{E,M}$ and $\Lambda_6^{E,M}$ are $M \times M$ diagonal matrices with the eigenvalues of $T_{-2}^{E,M}$ and $T_6^{E,M}$, respectively, on the diagonal, and $\Lambda_{-2}^{D,N}$ and $\Lambda_6^{D,N}$ are $N \times N$ diagonal matrices with the eigenvalues of $T_{-2}^{D,N}$ and $T_6^{D,N}$, respectively, on the diagonal, Λ takes the form

$$\Lambda \equiv \frac{1}{8}\left(\frac{a}{h_x^2}\Lambda_{-2}^{E,M} \otimes \Lambda_6^{D,N} + \frac{c}{h_y^2}\Lambda_6^{E,M} \otimes \Lambda_{-2}^{D,N} + \frac{f}{8}\Lambda_6^{E,M} \otimes \Lambda_6^{D,N}\right). \tag{11}$$

The diagonalization (10) of the QSC matrix A in (7) gives rise to an algorithm for computing the solution $\mathrm{x} = A^{-1}\mathrm{g}$, using the Fast Cosine Transform II (FCT-II), the Fast Sine Transform II (FST-II) and the respective inverse transforms, iFCT-II and iFST-II.

For describing this algorithm and the next one, we adopt a convenient notation from [20]. For any $MN \times 1$ vector g, let $\mathrm{g}_{N\times M}$ denote a $N \times M$ matrix with entries the components of g laid out in N rows and M columns, column-by-column. Also, for brevity and later convenience, we define the following modules:

Module $\mathrm{g}^{(2)} =$ **FCST(**$M, N,$ g**)**

Step 1: Perform FCT-II of size M to each of the N columns of $(\mathrm{g}_{N\times M})^T$ to obtain $\mathrm{g}_{M\times N}^{(1)} = \mathrm{C}^M(\mathrm{g}_{N\times M})^T$.

Step 2: Perform FST-II of size N to each of the M columns of $(\mathrm{g}_{M\times N}^{(1)})^T$ to obtain $\mathrm{g}_{N\times M}^{(2)} = \mathrm{S}^N(\mathrm{g}_{M\times N}^{(1)})^T$, or equivalently, $\mathrm{g}^{(2)} = (\mathrm{C}^M \otimes \mathrm{S}^N)\mathrm{g}$.

The above two steps require approximately $2.5MN\log_2 M$ and $2.5MN\log_2 N$ real single flops [20], respectively. Hence, FCST($M, N,$ g) requires approximately $2.5MN\log_2(MN)$ real single flops.

Module $\mathrm{x} =$ **iFCST(**$M, N, \mathrm{g}^{(3)}$**)**

Step 1: Perform iFCT-II of size M to each of the N columns of $(\mathrm{g}_{N\times M}^{(3)})^T$ to obtain $\mathrm{g}_{M\times N}^{(4)} = (\mathrm{C}^M)^{-1}(\mathrm{g}_{N\times M}^{(3)})^T$.

Step 2: Perform iFST-II of size N to each of the M columns of $(\mathrm{g}_{M\times N}^{(4)})^T$ to obtain $\mathrm{x}_{N\times M} = \mathrm{S}^N(\mathrm{g}_{M\times N}^{(4)})^T$, or equivalently, $\mathrm{x} = ((\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1})\mathrm{g}^{(3)}$.

The above two steps require approximately $2.5MN\log_2 M$ and $2.5MN\log_2 N$ real single flops, respectively. Hence, iFCST($M, N, \mathrm{g}^{(2)}$) requires approximately $2.5MN\log_2(MN)$ real single flops.

We now give the algorithm for computing $\mathrm{x} = A^{-1}\mathrm{g}$ based on the diagonalization (10) of the QSC matrix A in (7). Let Λ be a $MN \times MN$ diagonal matrix, with the eigenvalues of A on the diagonal.

Algorithm 2D-FFTQSC(M, N, g)

Step 1: Compute $\mathrm{g}^{(2)} = \mathrm{FCST}(M, N, \mathrm{g}) = (\mathrm{C}^M \otimes \mathrm{S}^N)\mathrm{g}$.

Step 2: Compute $\mathrm{g}^{(3)} = \Lambda^{-1}\mathrm{g}^{(2)}$.

Step 3: Compute $\mathrm{x} = \mathrm{iFCST}(M, N, \mathrm{g}^{(3)}) = ((\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1})\mathrm{g}^{(3)} = ((\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1})\Lambda^{-1}(\mathrm{C}^M \otimes \mathrm{S}^N)\mathrm{g} = A^{-1}\mathrm{g}$.

The 2D-FFTQSC algorithm requires approximately $5MN\log_2(MN)$ real single flops.

We now consider the QSC matrix A in (9). Since both $T_{-2}^{E,M}$ and $T_6^{E,M}$ can be diagonalized by the inverse of the DCT-II matrix C^M, the QSC matrix A of (9) can be block-diagonalized by the inverse of $\mathrm{C}^M \otimes \mathbf{I}^N$. That is,

$$A = ((\mathrm{C}^M)^{-1} \otimes \mathbf{I}^N)B(\mathrm{C}^M \otimes \mathbf{I}^N) \tag{12}$$

where

$$B \equiv \frac{1}{8}\left(\frac{1}{h_x^2}\Lambda_{-2}^{E,M} \otimes T_{6a}^{D,N} + \frac{1}{h_y^2}\Lambda_6^{E,M} \otimes T_{-2c}^{D,N} + \frac{1}{2h_y}\Lambda_6^{E,M} \otimes T_{0e}^{D,N} + \frac{1}{8}f\Lambda_6^{E,M} \otimes T_{6f}^{D,N}\right). \tag{13}$$

Note that the block-diagonal matrix B consists of M tridiagonal blocks of size $N \times N$. Also note that the form of A in (7) is a sub-case of the form of A in (9), therefore the block-diagonalization (12) holds for A in (7) too.

A second algorithm for computing $\mathrm{x} = A^{-1}\mathrm{g}$ based on the block-diagonalization (12) of the QSC matrix A in (9) is now presented and turns out to be asymptotically twice as fast as the 2D-FFTQSC algorithm.

Algorithm 1D-FFTQSC(M, N, g)

Step 1: Perform FCT-II of size M to each of the N columns of $(\mathrm{g}_{N\times M})^T$ to obtain $\mathrm{g}_{M\times N}^{(1)} = \mathrm{C}^M(\mathrm{g}_{N\times M})^T$, or equivalently, $\mathrm{g}^{(1)} = (\mathrm{C}^M \otimes \mathbf{I}^N)\mathrm{g}$.

Step 2: Solve the block-diagonal system $B\mathrm{g}^{(2)} = \mathrm{g}^{(1)}$, where B is given in (13).

Step 3: Perform iFCT-II of size M to each of the N columns of $(\mathrm{g}_{N\times M}^{(2)})^T$ to obtain $\mathrm{x}_{M\times N} = (\mathrm{C}^M)^{-1}(\mathrm{g}_{N\times M}^{(2)})^T$, or equivalently, $\mathrm{x} = ((\mathrm{C}^M)^{-1} \otimes \mathbf{I}^N)\mathrm{g}^{(2)} = ((\mathrm{C}^M)^{-1} \otimes \mathbf{I}^N)B^{-1}(\mathrm{C}^M \otimes \mathbf{I}^N)\mathrm{g} = A^{-1}\mathrm{g}$.

The block-diagonal matrix B consists of M tridiagonal blocks of size $N \times N$, therefore Step 2 consists of solving M tridiagonal systems of size $N \times N$, i.e. it requires approximately $8MN$ real single flops, of which $3MN$ are attributed to LU factorization and $5MN$ to back-and-forward (b/f) substitutions. Therefore, the 1D-FFTQSC algorithm requires approximately $5MN\log_2(M)$ (+ lower order terms of MN) real single flops.

If we assume that $M \approx N$, this is asymptotically twice as fast as the 2D-FFTQSC algorithm. However, for reasonable gridsizes, the advantage of the 1D over the 2D algorithm may not become visible, since the extra $\log_2(N)$ term of the 2D algorithm complexity is small. It is also worth noting that, if the boundary conditions are periodic in y, the block-diagonal matrix B consists of blocks that are no longer tridiagonal, but "almost" tridiagonal (tridiagonal with "corner" entries), in which case about twice as many flops are needed for the solution of the blocks.

It is further instructive to note that in both the 2D-FFTQSC and the 1D-FFTQSC algorithms, the intermediate data are accessed by rows and by columns, in an alternating way. For example, in Step 1

of 1D-FFTQSC, $\mathrm{g}_{N\times M}$ is accessed by rows (Fourier transforms are applied to each of its rows), while in Step 2, the intermediate result $\mathrm{g}^{(1)}_{M\times N}$ is accessed by columns (tridiagonal solves are applied to each of its columns).

Algorithms "2D-FFTQSC" and "1D-FFTQSC" are the QSC equivalent to Algorithms I and II, respectively, in [3]. Algorithms I and II in [3] are given in a generic form and the Fourier diagonalization and block-diagonalization are given for Hermite cubic spline collocation. It is worth noting that, for Hermite cubic spline collocation, four times as many Fourier transforms are needed as for QSC, and that the system of Hermite cubic spline collocation equations is four times as large as the QSC system. Moreover, the matrix B arising in Algorithm II from Hermite cubic spline collocation has 4 non-zero entries per row (it is penta-diagonal and almost block-diagonal) instead of 3 of the respective one from QSC.

For three-dimensional problems, FFT solvers that apply Fourier transforms in one, two or three dimensions can be used. In the most effective form, Fourier transforms are applied to two dimensions and tridiagonal solves in the third dimension. Thus the 2D-FFTQSC can be incorporated into a three-dimensional solver that applies the 2D-FFTQSC to block-diagonalize the linear system and uses tridiagonal solves for the third dimension, giving rise to a $O(N^3 \log_2 N)$ asymptotic complexity. In [9], a three-dimensional solver that applies Fourier transforms in all three dimensions is developed.

3.2 Other boundary conditions

The two algorithms given in the previous section were designed to handle boundary conditions that are homogeneous Neumann in the x direction and homogeneous Dirichlet in y. They can be easily adjusted to handle other boundary conditions that are Dirichlet, Neumann or periodic on any side of the rectangular domain. Further, algorithm 1D-FFTQSC can handle general conditions in one direction.

For periodic conditions in any of (or both) the two directions, the one-dimensional matrices arising can be diagonalized by the inverse of the Discrete Fourier Transform (DFT) matrix, and the respective computation is implemented by the Fast Fourier Transform (FFT) and its inverse (iFFT). Explicit formulae of the eigenvalues and eigenvectors for periodic conditions is given in [9]. When the 1D-FFTQSC algorithm is used for boundary conditions that are periodic in y and Dirichlet or Neumann in x, it is advisable to order the points, equations and unknowns first left-to-right, then bottom-up, so that the FFTs are applied to the y direction and the x direction is handled by the tridiagonal solves. In this way, the solution of the "almost" tridiagonal matrices is substituted by the solution of (purely) tridiagonal matrices.

We now consider the case of Dirichlet conditions on one side and Neumann on the opposite side of the same direction, give explicit formulae for the eigenvalues and eigenvectors of the arising matrix, and derive FFT solvers for it. The study of the FFT solution of the linear system arising in this case of boundary conditions was motivated by [15]. Without loss of generality, we assume that the Dirichlet boundary is ordered first and the Neumann one last (Dirichlet-Neumann boundary conditions).

The QSC matrix arising from the one-dimensional BVP

$$u_{xx} = g \quad \text{for} \quad x \in (0,1), \;\; u(0) = 0, \; u_x(1) = 0$$

takes the form $T_{-2}^{DN,M} = trid\{1,-2,1\}$, with $(T_{-2}^{DN,M})_{1,1} = -3$ and $(T_{-2}^{DN,M})_{M,M} = -1$. The eigenvalues of $T_{-2}^{DN,M}$ are

$$\lambda_i = -4\sin^2\frac{(2i-1)\pi}{4M}, i = 1,\cdots,M \tag{14}$$

and an orthonormal set of eigenvectors is

$$\{(\delta_i)_j = \sqrt{\frac{2}{M}}\sin\frac{(2i-1)(2j-1)\pi}{4M}, j = 1,\cdots,M\}_{i=1}^{M}.$$

We can show that $T_{-2}^{DN,M}$ has a Fourier diagonalization of the form

$$T_{-2}^{DN,M} = (\sqrt{2M}E^T(\mathrm{S}^{2M})^{-1}R^T)\Lambda_{-2}^{DN,M}(R\mathrm{S}^{2M}E\sqrt{\frac{2}{M}}) \tag{15}$$

where $\Lambda_{-2}^{DN,M}$ denotes the $M \times M$ diagonal matrix with the eigenvalues of $T_{-2}^{DN,M}$ on the diagonal (given by (14)), and E and R are $2M \times M$ extension and $M \times 2M$ restriction matrices, respectively, defined by

$$E \equiv \begin{bmatrix} \mathbf{I}^M \\ 0 \end{bmatrix} \text{ and } R \equiv \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdot & \cdot & 0 \\ & & & & \cdot & \cdot & \cdot & \cdot & & \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & 1 & 0 & 0 & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & 0 \end{bmatrix}.$$

Thus the computation associated with the Dirichlet-Neumann boundary conditions can be handled by the FST-II and iFST-II of double size, and the appropriate extension and restriction operations. This gives rise to an extra factor of 2 in the computational complexity of the algorithms. It should be noted that, if the 1D-FFTQSC algorithm is used for solving a problem with Dirichlet-Neumann boundary conditions in only one direction, this direction should be handled by the tridiagonal solves, with appropriate ordering of the points, equations and unknowns, thus avoiding the double size Fourier transforms. If the Neumann boundary is ordered first and the Dirichlet last, the eigenvectors are given by cosine formulae, and a diagonalization similar to (15) occurs, which uses the DCT-II matrix double size and its inverse, and slightly different extension and restriction operators.

Remark 1. Algorithm 1D-FFTQSC is applicable not only to a wider range of PDE operators, but also to a wider range of boundary conditions, than algorithm 2D-FFTQSC. More specifically, the block-diagonalization (12) and the applicability of algorithm 1D-FFTQSC are still valid, even if the boundary conditions in the y direction are not of the types listed in Section 2. Moreover, the roles of the x and y directions can be switched as follows. If a, c, d and f are functions of x, $b = 0$, $e = 0$, and the boundary conditions are of one of the types listed in Section 2 in the y direction and arbitrary in the x direction, with appropriate ordering of points, equations and unknowns, we can apply an algorithm similar to 1D-FFTQSC to solve the resulting linear system.

3.3 Extension to systems of PDEs

We consider the extension of the FFT solvers to QSC equations arising from a 2×2 system of linear second-order elliptic PDEs in two dimensions

$$\begin{bmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \text{ in } \Omega \tag{16}$$

where, for $i = 1, 2$ and $j = 1, 2$,

$$\mathbf{L}_{ij}u \equiv a_{ij}u_{xx} + b_{ij}u_{xy} + c_{ij}u_{yy} + d_{ij}u_x + e_{ij}u_y + f_{ij}u, \tag{17}$$

$a_{ij}, b_{ij}, c_{ij}, d_{ij}, e_{ij}, f_{ij}$ and g_i are given functions of x and y, and u and v are the unknown functions of x and y. We assume that both u and v are subject to same boundary conditions, any of those listed in

Section 2. For simplicity, in this subsection we will assume that the boundary conditions are homogeneous Neumann in the x direction and homogeneous Dirichlet in y for both u and v, i.e.

$$u_x = v_x = 0 \text{ on } x = 0, x = 1 \text{ for } 0 \le y \le 1 \tag{18}$$

$$u = v = 0 \text{ on } y = 0, y = 1 \text{ for } 0 \le x \le 1 \tag{19}$$

The optimal two-step QSC method applied to (16), (18)-(19) as described in [22], [7] gives rise to two linear systems. In the first step of the QSC method, a system $A\mathrm{x} = \mathrm{g}$ is to be solved, which, with appropriate ordering (block ordering), takes the 2×2 block form

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \end{bmatrix} = \begin{bmatrix} \mathrm{g}_1 \\ \mathrm{g}_2 \end{bmatrix} \tag{20}$$

Each submatrix A_{ij}, for $i = 1, 2$ and $j = 1, 2$, arises from the QSC discretization of the respective operator $\mathbf{L}_{ij}$ of (16). In the second step of the QSC method, a linear system with the same matrix and perturbed right-side vector arises.

If each of the operators $\mathbf{L}_{ij}$, for $i = 1, 2$ and $j = 1, 2$, is of Helmholtz type with constant coefficients, each of the blocks A_{ij} takes a tensor product form similar to that in (7) and the matrix A assumes the block-diagonalization

$$A = \begin{bmatrix} W & 0 \\ 0 & W \end{bmatrix} \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} W^{-1} & 0 \\ 0 & W^{-1} \end{bmatrix} \tag{21}$$

where $W = (\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1}$, and $\Lambda_{ij} = W^{-1} A_{ij} W$, for $i = 1, 2$ and $j = 1, 2$, are diagonal matrices, that take a tensor product form as in (11). The matrix

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \tag{22}$$

can be reordered to give a block-diagonal matrix with 2×2 blocks on the diagonal. It should be emphasized that (21) is not a point-diagonalization of the matrix in (20), as is (10) of the QSC matrix A in (7), in the scalar PDE case. A point-diagonalization of the matrix in (20) is possible [22], [7], but it leads to an FFT algorithm that requires more flops than the FFT algorithm 2D-FFTQSC2 arising from (21) and described further in this section.

If each of the operators $\mathbf{L}_{ij}$, for $i = 1, 2$ and $j = 1, 2$, has coefficients variable in y and no first order terms with respect to x, each of the blocks A_{ij} takes a tensor product form similar to that in (9) and the matrix A assumes the block-diagonalization

$$A = \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} Z^{-1} & 0 \\ 0 & Z^{-1} \end{bmatrix} \tag{23}$$

where $Z = (\mathrm{C}^M)^{-1} \otimes \mathbf{I}^N$, and $B_{ij} = Z^{-1} A_{ij} Z$, for $i = 1, 2$ and $j = 1, 2$, are block-diagonal matrices with tridiagonal blocks on the diagonal, that can take a tensor product form similar to that in (13). The matrix

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \tag{24}$$

can be reordered to give a block-diagonal matrix with septa-diagonal blocks on the diagonal and at most 6 non-zero entries per row.

The block-diagonalization (21) gives rise to an algorithm for computing the solution $\mathrm{x} = A^{-1}\mathrm{g}$, using the Fast Cosine Transform II (FCT-II), the Fast Sine Transform II (FST-II) and the respective inverse transforms, iFCT-II and iFST-II.

Algorithm 2D-FFTQSC2(M, N, g**)**

Step 1: Compute $\mathrm{g}_1^{(1)} = \mathrm{FCST}(M, N, \mathrm{g}_1) = (\mathrm{C}^M \otimes \mathrm{S}^N)\mathrm{g}_1$.
Compute $\mathrm{g}_2^{(1)} = \mathrm{FCST}(M, N, \mathrm{g}_2) = (\mathrm{C}^M \otimes \mathrm{S}^N)\mathrm{g}_2$.

Step 2: Let $\mathrm{g}^{(1)} = [(\mathrm{g}_1^{(1)})^T(\mathrm{g}_2^{(1)})^T]^T$. Solve the system $\Lambda \mathrm{g}^{(2)} = \mathrm{g}^{(1)}$, where Λ is given in (22).
Let $\mathrm{g}_1^{(2)} = [(\mathrm{g}^{(2)})_i, i = 1, \cdots, MN]^T$ and $\mathrm{g}_2^{(2)} = [(\mathrm{g}^{(2)})_i, i = MN+1, \cdots, 2MN]^T$.

Step 3: Compute $\mathrm{g}_1^{(3)} = \mathrm{iFCST}(M, N, \mathrm{g}_1^{(2)}) = ((\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1})\mathrm{g}_1^{(2)}$.
Compute $\mathrm{g}_2^{(3)} = \mathrm{iFCST}(M, N, \mathrm{g}_2^{(2)}) = ((\mathrm{C}^M)^{-1} \otimes (\mathrm{S}^N)^{-1})\mathrm{g}_2^{(2)}$.
Let $\mathrm{x} = [\mathrm{x}_1^T \mathrm{x}_2^T]^T$.

The dominant part of the computation in the above algorithm are two performances of FCST and two of iFCST. The solution of $\Lambda \mathrm{g}^{(2)} = \mathrm{g}^{(1)}$ in Step 2, is performed by reordering the rows and columns of Λ and the rows of $\mathrm{g}^{(1)}$ according to the alternating ordering [22], [7], so that Λ becomes a block-diagonal matrix with 2×2 blocks on the diagonal. Therefore, Step 2 requires $O(MN)$ flops, more precisely, approximately $8MN$ real single flops. Therefore, the 2D-FFTQSC2 algorithm requires approximately $10MN \log_2(MN)$ (+ lower order terms of MN) real single flops, that is, about twice as much as the 2D-FFTQSC algorithm for a single Helmholtz PDE problem (6), (2)-(3).

A second algorithm for computing $\mathrm{x} = A^{-1}\mathrm{g}$ based on the block-diagonalization (23) is now presented and turns out to be asymptotically twice as fast as the 2D-FFTQSC2 algorithm.

Algorithm 1D-FFTQSC2(M, N, g**)**

Step 1: Perform FCT-II of size M to each of the N columns of $((\mathrm{g}_1)_{N\times M})^T$ to obtain $(\mathrm{g}_1^{(1)})_{M\times N} = \mathrm{C}^M((\mathrm{g}_1)_{N\times M})^T$, or equivalently, $\mathrm{g}_1^{(1)} = (\mathrm{C}^M \otimes \mathbf{I}^N)\mathrm{g}_1$.
Perform FCT-II of size M to each of the N columns of $((\mathrm{g}_2)_{N\times M})^T$ to obtain $(\mathrm{g}_2^{(1)})_{M\times N} = \mathrm{C}^M((\mathrm{g}_2)_{N\times M})^T$, or equivalently, $\mathrm{g}_2^{(1)} = (\mathrm{C}^M \otimes \mathbf{I}^N)\mathrm{g}_2$.

Step 2: Let $\mathrm{g}^{(1)} = [(\mathrm{g}_1^{(1)})^T(\mathrm{g}_2^{(1)})^T]^T$. Solve the system $B\mathrm{g}^{(2)} = \mathrm{g}^{(1)}$, where B is given in (24).
Let $\mathrm{g}_1^{(2)} = [(\mathrm{g}^{(2)})_i, i = 1, \cdots, MN]^T$ and $\mathrm{g}_2^{(2)} = [(\mathrm{g}^{(2)})_i, i = MN+1, \cdots, 2MN]^T$.

Step 3: Perform iFCT-II of size M to each of the N columns of $((\mathrm{g}_1^{(2)})_{N\times M})^T$ to obtain $(\mathrm{x}_1)_{M\times N} = (\mathrm{C}^M)^{-1}((\mathrm{g}_1^{(2)})_{N\times M})^T$, or equivalently, $\mathrm{x}_1 = ((\mathrm{C}^M)^{-1} \otimes \mathbf{I}^N)\mathrm{g}_1^{(2)}$.
Perform iFCT-II of size M to each of the N columns of $((\mathrm{g}_2^{(2)})_{N\times M})^T$ to obtain $(\mathrm{x}_2)_{M\times N} = (\mathrm{C}^M)^{-1}((\mathrm{g}_2^{(2)})_{N\times M})^T$, or equivalently, $\mathrm{x}_2 = ((\mathrm{C}^M)^{-1} \otimes \mathbf{I}^N)\mathrm{g}_2^{(2)}$.
Let $\mathrm{x} = [\mathrm{x}_1^T \mathrm{x}_2^T]^T$.

The solution of $B\mathrm{g}^{(2)} = \mathrm{g}^{(1)}$ in Step 2, is performed by reordering the rows and columns of B and the rows of $\mathrm{g}^{(1)}$ according to the alternating ordering [22], [7], so that B becomes a block-diagonal matrix with septa-diagonal blocks on the diagonal. (In the implementation, we actually form B as a septa-diagonal matrix, and reorder the components of $\mathrm{g}^{(1)}$ and $\mathrm{g}^{(2)}$ appropriately.) Therefore, Step 2 requires $O(MN)$ flops, more precisely, approximately $44MN$ real single flops, of which $18MN$ are attributed to LU factorization and $26MN$ to b/f substitutions. Each of Steps 1 and 3 require $2 \times 2.5MN \log_2(M)$ real single flops. Therefore, the 1D-FFTQSC2 algorithm requires approximately $10MN \log_2(M)$ (+ lower order terms of MN) real single flops, that is, about twice as much as the 1D-FFTQSC algorithm for a single Helmholtz PDE problem (6), (2)-(3), which is (relative to the single PDE case) optimal. If we assume that $M \approx N$, 1D-FFTQSC2 is asymptotically twice as fast as the 2D-FFTQSC2 algorithm. However, we note again that, for reasonable gridsizes, the advantage of the 1D over the 2D algorithm may not become visible, since the extra $\log_2(N)$ term of the 2D algorithm complexity is small, and the factor for the lower order terms of MN of the 1D algorithm is relatively large.

Remark 2. Both the 1D-FFTQSC2 and the 2D-FFTQSC2 algorithms can be extended in a straightforward way to $n \times n$ systems of PDEs.
Remark 3. In describing the algorithms, we have assumed that both u and v satisfy the same boundary conditions all along $\partial\Omega$. While this assumption cannot be relaxed for algorithm 2D-FFTQSC2, algorithm 1D-FFTQSC2 is applicable in some other cases of boundary conditions. More specifically, if u and v satisfy the same boundary conditions of either Dirichlet, Neumann, periodic, or Dirichlet-Neumann type in the x-direction, and different in the y, then the diagonalization (23) is still valid, and the FFTs should be applied in the direction of the same boundary conditions, and the tridiagonal solves in the other. Moreover, the roles of x and y directions can be switched as explained in Remark 1 or subsection 3.2.
Remark 4. When converting the biharmonic equation subject to certain boundary conditions into a system of two second-order PDEs, we obtain a special case of (16), with $\mathbf{L}_{11} = \mathbf{L}_{22} = -\triangle$, $\mathbf{L}_{12} = 0$ and $\mathbf{L}_{21} = \mathbf{E}$, where $\triangle$ is the Laplacian and $\mathbf{E}$ is the identity operator, and where u and v satisfy Dirichlet or Neumann conditions. This system is decoupled and its solution can be obtained by solving two single PDEs. The single PDEs can be solved each by algorithm 1D-FFTQSC as long as the boundary conditions in one of the directions are among those listed in Section 2. (Algorithm 2D-FFTQSC is also applicable if the boundary conditions allow.) For the convergence analysis of QSC for systems of two PDEs see [7].

4 Preconditioners for Quadratic Spline Collocation Equations

In this section, we consider the solution of the QSC equations arising from general elliptic PDEs of the form (1) by preconditioned iterative methods. The analysis is carried out for the QSC equations arising from self-adjoint PDEs with homogeneous Dirichlet boundary conditions all along $\partial\Omega$. It is then extended to non-self-adjoint PDEs without cross-derivative term. The analysis assumes that the preconditioner is the QSC operator $\triangle_h$ arising from the Laplace operator and homogeneous Dirichlet conditions on $\partial\Omega$, therefore, S_{Δ_x}, the basis functions ϕ_i^x, and consequently S_Δ are adjusted appropriately.

4.1 Quadrature relations

Consider the BVP described by the operator equation (1), where $\mathbf{L}$ is a self-adjoint operator given by

$$\mathbf{L}u \equiv -(a(x,y)u_x)_x - (c(x,y)u_y)_y + f(x,y)u, \tag{25}$$

and homogeneous Dirichlet boundary conditions on $\partial\Omega$, i.e.

$$u(x,y) = 0 \ \text{ on } \partial\Omega. \tag{26}$$

In the following, let $(\cdot,\cdot)$ denote the standard inner product, that is, $(v,w) = \int_{\mathcal{D}} vw d\mathcal{D}$, where $\mathcal{D}$ may be an one- or two-dimensional domain, and let $||\cdot||_{L^2(\mathcal{D})}$ be the associated L^2 norm.
For any bounded functions $v(x,y)$ and $w(x,y)$, define the discrete pseudo-inner product $(v,w)_{xy}$ by two equivalent formulae

$$(v,w)_{xy} \equiv \sum_{j=1}^{N} h_y(v(\cdot,\tau_j^y), w(\cdot,\tau_j^y))_x \equiv \sum_{i=1}^{M} h_x(v(\tau_i^x,\cdot), w(\tau_i^x,\cdot))_y,$$

where $(v,w)_x$ and $(v,w)_y$ are defined by

$$(v,w)_x \equiv \sum_{i=1}^{M} h_x(vw)(\tau_i^x,\cdot) \text{ and } (v,w)_y \equiv \sum_{j=1}^{N} h_y(vw)(\cdot,\tau_j^y).$$

Note that, for $v, w \in S_\Delta$, $(v, w)_{xy}$ is an inner product, since any bi-quadratic spline can be uniquely determined by its values on the collocation points [4], and a bi-quadratic spline is the zero one, if and only if its values on all the collocation points are zero. Therefore, S_Δ is a Hilbert space and $(\cdot, \cdot)_{xy}$ the associated inner product.

Using the Peano representation for the midpoint quadrature rule error applied to a function $p \in \mathbf{C}^2[x_{i-1}, x_i]$, $i = 1, \cdots, M$, we get

$$h_x p(\tau_i^x) - \int_{x_{i-1}}^{x_i} p dx = -\int_{x_{i-1}}^{x_i} p_{xx} K_i(x) dx, \tag{27}$$

where $K_i(x)$ is the Peano kernel defined by $K_i(x) \equiv (x_{i-1} - x)^2/2$ for $x_{i-1} \leq x \leq \tau_i^x$ and $K_i(x) \equiv (x_i - x)^2/2$ for $\tau_i^x \leq x \leq x_i$. It is easy to show that $0 \leq K_i(x) \leq h_x^2 C_1$, for $i = 1, \cdots, M$, where $C_1 = 1/8$.

Let L_h and $\triangle_h$ be QSC operators from S_Δ into S_Δ corresponding to $\mathbf{L}$ in (25) and to $\triangle$ (Laplacian), respectively. That is, L_h and $\triangle_h$ are defined by

$$(L_h v)(\tau_i^x, \tau_j^y) = \mathbf{L} v(\tau_i^x, \tau_j^y) \text{ and } (\triangle_h v)(\tau_i^x, \tau_j^y) = \triangle v(\tau_i^x, \tau_j^y)$$

for $i = 1, \cdots, M$, $j = 1, \cdots, N$. Our goal is to show that L_h is spectrally equivalent to $-\triangle_h$, under the inner product $(\cdot, \cdot)_{xy}$. This is shown in Theorem 3, but to obtain this result we will need a number of other results which we show next.

Lemma 1 *Let $p \in S_{\Delta_x}$. Then $0 \leq ||p_x||^2_{L^2(0,1)} \leq (-p_{xx}, p)_x$.*

PROOF
Using integration by parts and applying (27) to $-p_{xx} p$ in each subinterval $[x_{i-1}, x_i]$, $i = 1, \cdots, M$, and summing up we have

$$0 \leq (p_x, p_x) = (-p_{xx}, p) = (-p_{xx}, p)_x - \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (p_{xx} p)_{xx} K_i(x) dx \leq (-p_{xx}, p)_x,$$

taking into account that, in each $[x_{i-1}, x_i]$, $(p_{xx} p)_{xx} = p_{xx}^2 \geq 0$ and $K_i(x) \geq 0$. QED.

Theorem 1 *Assume $a(x, y) \in \mathbf{C}^3(\overline{\Omega})$ with respect to x and $c(x, y) \in \mathbf{C}^3(\overline{\Omega})$ with respect to y, $f(x, y) \in \mathbf{C}$, and $0 < \alpha \leq a(x, y), c(x, y) \leq \gamma$, for all $(x, y) \in \overline{\Omega}$. Then, for all $v, w \in S_\Delta$,*

$$(L_h v, w)_{xy} = B_h^1(v, w) + B_h^2(v, w) + (fv, w)_{xy}, \tag{28}$$

where

$$B_h^1(v, w) = B_h^1(w, v) \tag{29}$$

$$\alpha(-\triangle_h v, v)_{xy} \leq B_h^1(v, v) \leq \gamma(-\triangle_h v, v)_{xy} \tag{30}$$

$$|B_h^2(v, w)| \leq C\delta(h_x, h_y)(-\triangle_h v, v)_{xy}^{1/2}(-\triangle_h w, w)_{xy}^{1/2}, \tag{31}$$

and where C is a positive constant independent of a, c, f, h_x and h_y, and

$$\delta(h_x, h_y) = \max(h_x \max(||a_x||_\infty, ||a_{xx}||_\infty) + h_x^2 ||a_{xxx}||_\infty, h_y \max(||c_y||_\infty, ||c_{yy}||_\infty) + h_y^2 ||c_{yyy}||_\infty).$$

PROOF

For any $j = 1, \cdots, N$, by applying (27) to $-(av_x)_x(\cdot, \tau_j^y)w(\cdot, \tau_j^y)$ in each subinterval $[x_{i-1}, x_i]$, $i = 1, \cdots, M$, summing up and using integration by parts and Leibnitz's differentiation formula, we get

$$\begin{aligned}(-(av_x)_x(\cdot, \tau_j^y), w(\cdot, \tau_j^y))_x &= (av_x(\cdot, \tau_j^y), w_x(\cdot, \tau_j^y)) + \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} ((av_x)_x w)_{xx}(x, \tau_j^y) K_i(x) dx \\ &= I_1(a, v, w, \tau_j^y) + I_2(a, v, w, \tau_j^y), \end{aligned} \tag{32}$$

where

$$\begin{aligned} I_1(a, v, w, \tau_j^y) &\equiv \int_0^1 (av_x w_x)(x, \tau_j^y) dx + \alpha_{022} \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (av_{xx} w_{xx})(x, \tau_j^y) K_i(x) dx, \\ I_2(a, v, w, \tau_j^y) &\equiv \sum_{i=1}^{M} \sum_{l=1}^{3} \sum_{m+n=4-l}^{1 \le m \le 2, 0 \le n \le 2} \alpha_{lmn} \int_{x_{i-1}}^{x_i} (a_x^{(l)} v_x^{(m)} w_x^{(n)})(x, \tau_j^y) K_i(x) dx \end{aligned}$$

where the constants α_{lmn} arise from Leibnitz's formula and are positive.

From the definition of I_1, the lower bound of a, the positiveness of α_{lmn} and of the integrals, and the fact that $I_2(1, v, v, \tau_j^y) = 0$, we have

$$\begin{aligned} I_1(a, v, v, \tau_j^y) &\ge \alpha \int_0^1 v_x^2 dx + \alpha \alpha_{022} \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} v_{xx}^2 K_i(x) dx \\ &= \alpha I_1(1, v, v, \tau_j^y) + \alpha I_2(1, v, v, \tau_j^y) = \alpha(-v_{xx}(\cdot, \tau_j^y), v(\cdot, \tau_j^y))_x. \end{aligned}$$

In a similar way, we can show an upper bound for $I_1(a, v, v, \tau_j^y)$. Thus

$$\alpha(-v_{xx}(\cdot, \tau_j^y), v(\cdot, \tau_j^y))_x \le I_1(a, v, v, \tau_j^y) \le \gamma(-v_{xx}(\cdot, \tau_j^y), v(\cdot, \tau_j^y))_x. \tag{33}$$

From the definition of I_2 and the triangle and Cauchy-Schwarz inequalities, we have

$$\begin{aligned} |I_2(a, v, w, \tau_j^y)| &\le h_x^2 C_2 \sum_{i=1}^{M} \sum_{l=1}^{3} ||a_x^{(l)}||_\infty \sum_{m+n=4-l}^{1 \le m \le 2, 0 \le n \le 2} ||v_x^{(m)}(\cdot, \tau_j^y)||_{L^2(x_{i-1}, x_i)} ||w_x^{(n)}(\cdot, \tau_j^y)||_{L^2(x_{i-1}, x_i)} \\ &\le h_x^2 C_2 \sum_{l=1}^{3} ||a_x^{(l)}||_\infty \sum_{m+n=4-l}^{1 \le m \le 2, 0 \le n \le 2} ||v_x^{(m)}(\cdot, \tau_j^y)||_{L^2(0,1)} ||w_x^{(n)}(\cdot, \tau_j^y)||_{L^2(0,1)} \end{aligned}$$

where C_2 is a positive constant arising from the constants α_{lmn} and C_1. Using the inverse inequality $||v_{xx}||_{L^2(0,1)} \le C_3 h_x^{-1} ||v_x||_{L^2(0,1)}$, for $v \in S_\Delta$, where C_3 a positive constant independent of h_x, and the Poincaré inequality $||v||_{L^2(0,1)} \le C_4 ||v_x||_{L^2(0,1)}$, $C_4 > 0$, we have

$$|I_2(a, v, w, \tau_j^y)| \le C_5 \delta_a(h_x) ||v_x(\cdot, \tau_j^y)||_{L^2(0,1)} ||w_x(\cdot, \tau_j^y)||_{L^2(0,1)}, \tag{34}$$

where

$$\delta_a(h_x) = h_x \max(||a_x||_\infty, ||a_{xx}||_\infty) + h_x^2 ||a_{xxx}||_\infty$$

and C_5 a positive constant arising from C_2, C_3 and C_4. By applying Lemma 1, (34) becomes

$$|I_2(a, v, w, \tau_j^y)| \le C_5 \delta_a(h_x)(-v_{xx}(\cdot, \tau_j^y), v(\cdot, \tau_j^y))_x^{1/2} (-w_{xx}(\cdot, \tau_j^y), w(\cdot, \tau_j^y))_x^{1/2}. \tag{35}$$

By (32) and the definition of $(\cdot, \cdot)_{xy}$, we have

$$(-(av_x)_x, w)_{xy} = \sum_{j=1}^{N} h_y(-(av_x)_x(\cdot, \tau_j^y), w(\cdot, \tau_j^y))_x = C_h^1(v, w) + C_h^2(v, w),$$

where $$C_h^1(v,w) \equiv \sum_{j=1}^{N} h_y I_1(a,v,w,\tau_j^y), \text{ and } C_h^2(v,w) \equiv \sum_{j=1}^{N} h_y I_2(a,v,w,\tau_j^y).$$

By (33) and the definition of I_1, we have

$$C_h^1(v,w) = C_h^1(w,v) \tag{36}$$

$$\alpha(-v_{xx},v)_{xy} \le C_h^1(v,v) \le \gamma(-v_{xx},v)_{xy}. \tag{37}$$

Furthermore by (35), Lemma 1 and the inequality $\sum s_i^{1/2} t_i^{1/2} \le (\sum s_i)^{1/2} (\sum t_i)^{1/2}$, for nonnegative scalars s_i and t_i, we have

$$\begin{aligned} |C_h^2(v,w)| &\le C_5 \delta_a(h_x) \sum_{j=1}^{N} h_y (-v_{xx}(\cdot,\tau_j^y), v(\cdot,\tau_j^y))_x^{1/2} (-w_{xx}(\cdot,\tau_j^y), w(\cdot,\tau_j^y))_x^{1/2} \\ &\le C_5 \delta_a(h_x) (\sum_{j=1}^{N} h_y (-v_{xx}(\cdot,\tau_j^y), v(\cdot,\tau_j^y))_x)^{1/2} (\sum_{j=1}^{N} h_y (-w_{xx}(\cdot,\tau_j^y), w(\cdot,\tau_j^y))_x)^{1/2} \\ &= C_5 \delta_a(h_x) (-v_{xx},v)_{xy}^{1/2} (-w_{xx},w)_{xy}^{1/2}. \end{aligned} \tag{38}$$

By symmetry, we can show that

$$D_h^1(v,w) = D_h^1(w,v) \tag{39}$$

$$\alpha(-v_{yy},v)_{xy} \le D_h^1(v,v) \le \gamma(-v_{yy},v)_{xy} \tag{40}$$

$$|D_h^2(v,w)| \le C_5 \delta_c(h_y) (-v_{yy},v)_{xy}^{1/2} (-w_{yy},w)_{xy}^{1/2}, \tag{41}$$

where

$$\begin{aligned} D_h^1(v,w) &\equiv \sum_{i=1}^{M} h_y J_1(c,v,w,\tau_i^x), \quad D_h^2(v,w) \equiv \sum_{i=1}^{M} h_y J_2(c,v,w,\tau_i^x), \\ J_1(c,v,w,\tau_i^x) &\equiv \int_0^1 (c v_y w_y)(\tau_i^x,y) dy + \alpha_{022} \sum_{j=1}^{N} \int_{y_{j-1}}^{y_j} (c v_{yy} w_{yy})(\tau_i^x,y) K_j(y) dy, \text{ and} \\ J_2(c,v,w,\tau_i^x) &\equiv \sum_{j=1}^{N} \sum_{l=1}^{3} \sum_{m+n=4-l}^{1\le m\le 2, 0\le n\le 2} \alpha_{lmn} \int_{y_{j-1}}^{y_j} (c_y^{(l)} v_y^{(m)} w_y^{(n)})(\tau_i^x,y) K_j(y) dy. \end{aligned}$$

Finally, let $B_h^1 \equiv C_h^1 + D_h^1$ and $B_h^2 \equiv C_h^2 + D_h^2$. Then $(L_h v, w)_{xy} = B_h^1(v,w) + B_h^2(v,w) + (f(x,y)v, w)_{xy}$. Conditions (29) and (30) follow easily from (36), (37), (39) and (40). Condition (31) follows from (38), (41) and the inequality used in the derivation of (38). QED.
Theorem 1 is the QSC counterpart of Theorem 3.1 in [1], which holds for Hermite cubic splines.

4.2 Eigenvalues and eigenfunctions of a QSC problem

In this section, we obtain the eigenvalues and an orthonormal set of eigenfunctions for a model one-dimensional QSC eigenproblem. The respective Hermite cubic spline results are found in [13], [14] and [3]. Consider the eigenvalue problem

$$-p_{xx}(\tau_i^x) = \lambda p(\tau_i^x),\ i = 1, \cdots, M,\ p \in S_{\Delta_x},\ p \ne 0 \tag{42}$$

$$p(0) = p(1) = 0. \tag{43}$$

Using the eigenvalue and eigenvector formulae for the QSC matrix as given in [4], it is easy to show that the eigenvalues for (42)-(43) are

$$\lambda_l = \frac{8\sin^2(l\pi h_x/2)}{h_x^2(-\sin^2(l\pi h_x/2)+2)},\ l = 1,\cdots,M$$

and the corresponding eigenfunctions are

$$p_l(x) = \frac{2}{-\sin^2(l\pi h_x/2)+2}\sum_{i=1}^{M}(q_l)_i\phi_i^x(x),\ l = 1,\cdots,M,$$

where the vectors q_l are defined by

$$(q_l)_i = \sqrt{2}\sin(l\pi h_x(2i-1)/2),\ i = 1,\cdots,M,\ l = 1,\cdots,M-1,$$

and

$$(q_M)_i = \sin(\pi(2i-1)/2),\ i = 1,\cdots,M.$$

It is also easy to show that the eigenvalues are distinct and positive.

Lemma 2 *Let $v, w \in S_{\Delta_x}$. Then $(v_{xx}, w)_x = (v, w_{xx})_x$.*

PROOF
As in the proof of Lemma 1

$$(v_{xx}, w)_x = (v_{xx}, w) - \sum_{i=1}^{M}\int_{x_{i-1}}^{x_i}(v_{xx}w)_{xx}K_i(x)dx,$$

and

$$(v, w_{xx})_x = (v, w_{xx}) - \sum_{i=1}^{M}\int_{x_{i-1}}^{x_i}(vw_{xx})_{xx}K_i(x)dx.$$

The result of the Lemma now follows by noting that $(v_{xx}w)_{xx} = v_{xx}w_{xx} = (vw_{xx})_{xx}$ for $v, w \in S_{\Delta_x}$, and that $(v_{xx}, w) = (v, w_{xx})$ due to the integration by parts rule. QED.
Applying Lemma 2 in the x and y dimensions, we get that the QSC operator $\triangle_h$ is self-adjoint.

Lemma 3 *The set $\{p_l\}_{l=1}^M$ is orthonormal with respect to the inner product $(\cdot,\cdot)_x$.*

PROOF
For $l \neq m$, we have

$$((p_l)_{xx}, p_m)_x = h_x\sum_{i=1}^{M}((p_l)_{xx}p_m)(\tau_i^x) = -\lambda_l h_x\sum_{i=1}^{M}(p_l p_m)(\tau_i^x) = -\lambda_l(p_l, p_m)_x$$

and

$$(p_l, (p_m)_{xx})_x = h_x\sum_{i=1}^{M}(p_l(p_m)_{xx})(\tau_i^x) = -\lambda_m h_x\sum_{i=1}^{M}(p_l p_m)(\tau_i^x) = -\lambda_m(p_l, p_m)_x.$$

By Lemma 2, $((p_l)_{xx}, p_m)_x = (p_l, (p_m)_{xx})_x$. But as the eigenvalues are distinct we must have $(p_l, p_m)_x = 0$, for $l \neq m$, which implies orthogonality. To show orthonormality, we have, for $l \neq M$, $q_l^T q_l = \sum_{i=1}^{M} 2\sin^2(l\pi h_x(2i-1)/2) = 2\cdot M/2 = M$, while $q_M^T q_M = \sum_{i=1}^{M}\sin^2(\pi(2i-1)/2) = M$. Let $\overline{p_l}$ be the vector of values of $p_l(x)$ on τ_i^x, $i = 1,\cdots,M$. Then, $\overline{p_l} = 2T_6^{E,M}q_l/(-\sin^2(l\pi h_x/2)+2) = q_l$. The orthonormality of p_l, $i = 1,\cdots,M$, follows from the fact that $(p_l, p_l)_x = h_x\overline{p_l}^T\overline{p_l}$. QED.

Any $v \in S_{\Delta_x}$ can be written as $v = \sum_{i=1}^{M}(v, p_i)_x p_i$. The following theorem gives bounds for the QSC operator $\triangle_h$ in terms of the identity operator.

Theorem 2 *For all* $v \in S_\Delta$,

$$\lambda_*(h_x,h_y)(v,v)_{xy} \le (-\triangle_h v, v)_{xy} \le \lambda^*(h_x,h_y)(v,v)_{xy}, \tag{44}$$

where

$$\lambda_*(h_x,h_y) = \lambda_{min}(h_x) + \lambda_{min}(h_y), \qquad \lambda^*(h_x,h_y) = \lambda_{max}(h_x) + \lambda_{max}(h_y),$$
$$\lambda_{min}(h) = \frac{8\sin^2(\pi h/2)}{h^2(-\sin^2(\pi h/2)+2)}, \quad \textit{and} \quad \lambda_{max}(h) = \frac{8}{h^2}, \ \textit{where } h = h_x \textit{ or } h = h_y.$$

PROOF
First, for $w \in S_{\Delta_x}$,

$$\begin{aligned}(-w_{xx},w)_x &= (\sum_{j=1}^M (w,p_j)_x(-p_j)_{xx}, \sum_{i=1}^M (w,p_i)_x p_i)_x = \sum_{i=1}^M\sum_{j=1}^M (w,p_j)_x(w,p_i)_x((-p_j)_{xx},p_i)_x \\ &= \sum_{i=1}^M\sum_{j=1}^M (w,p_j)_x(w,p_i)_x\lambda_j(p_j,p_i)_x = \sum_{i=1}^M (w,p_i)_x^2\lambda_i(p_i,p_i)_x = \sum_{i=1}^M \lambda_i(w,p_i)_x^2.\end{aligned}$$

Since $(w,w)_x = \sum_{i=1}^M (w,p_i)_x^2 \ge 0$, we have

$$\lambda_{min}(h_x)(w,w)_x \le (-w_{xx},w)_x \le \lambda_{max}(h_x)(w,w)_x.$$

Then, for $v \in S_\Delta$,

$$\begin{aligned}\lambda_{min}(h_x)(v,v)_{xy} &= h_y\sum_{j=1}^N \lambda_{min}(h_x)(v(\cdot,\tau_j^y), v(\cdot,\tau_j^y))_x \\ &\le h_y\sum_{j=1}^N (-v(\cdot,\tau_j^y)_{xx}, v(\cdot,\tau_j^y))_x = (-v_{xx},v)_{xy} \\ &\le h_y\sum_{j=1}^N \lambda_{max}(h_x)(v(\cdot,\tau_j^y), v(\cdot,\tau_j^y))_x = \lambda_{max}(h_x)(v,v)_{xy}.\end{aligned}$$

By symmetry, $\lambda_{min}(h_y)(v,v)_{xy} \le (-v_{yy},v)_{xy} \le \lambda_{max}(h_y)(v,v)_{xy}$. Hence,

$$\lambda_*(h_x,h_y)(v,v)_{xy} \le (-v_{xx}-v_{yy},v)_{xy} \le \lambda^*(h_x,h_y)(v,v)_{xy}.$$

QED.
From Theorem 2, it is clear that $-\triangle_h$ is positive definite. Also, $(-\triangle_h)^{-1}$ exists and is unique.

4.3 Spectral equivalence of QSC operators

The following theorem shows that the QSC operator L_h corresponding to $\mathbf{L}$ in (25) is spectrally equivalent to the QSC operator $-\triangle_h$ corresponding to the negative Laplacian. The Hermite cubic spline equivalent is Theorem 3.1 of [1]. For any two linear operators L_h^1 and L_h^2 from S_Δ into S_Δ, the notation $L_h^1 \le L_h^2$ means $(L_h^1 v, v)_{xy} \le (L_h^2 v, v)_{xy}, \forall v \in S_\Delta$.

Theorem 3 *Under the assumptions of Theorem 1,*

$$\left[\alpha + \frac{\eta_*}{\lambda_*(h_x,h_y)} - C\delta(h_x,h_y)\right](-\triangle_h) \le L_h \le \left[\gamma + \frac{\eta^*}{\lambda_*(h_x,h_y)} + C\delta(h_x,h_y)\right](-\triangle_h) \tag{45}$$

and

$$(L_h^* - L_h)(-\triangle_h)^{-1}(L_h - L_h^*) \leq 4C^2\delta^2(h_x, h_y)(-\triangle_h), \tag{46}$$

where

$$\eta_* = \min(0, f_*),\ \eta^* = \max(0, f^*),\ f_* = \min\{f(x,y)\},\ f^* = \max\{f(x,y)\},$$

C and $\delta(h_x, h_y)$ are defined in Theorem 1, and $\lambda_*(h_x, h_y)$ in Theorem 2.

PROOF
By using the left inequality in (44),

$$(fv, v)_{xy} \leq f^*(v, v)_{xy} \leq \eta^*(v, v)_{xy} \leq \eta^*(-\triangle_h v, v)_{xy}/\lambda_*(h_x, h_y),$$

and

$$(fv, v)_{xy} \geq f_*(v, v)_{xy} \geq \eta_*(v, v)_{xy} \geq \eta_*(-\triangle_h v, v)_{xy}/\lambda_*(h_x, h_y).$$

Relation (45) now follows from Theorem 1.

To show (46), consider $w \equiv (-\triangle_h)^{-1}(L_h - L_h^*)v$. From (28), (29) and (31) we have

$$(w, (L_h - L_h^*)v)_{xy} = B_h^2(v, w) - B_h^2(w, v) \leq 2C\delta(h_x, h_y)(-\triangle_h v, v)_{xy}^{1/2}(w, (L_h - L_h^*)v)_{xy}^{1/2},$$

and hence

$$(w, (L_h - L_h^*)v)_{xy} \leq 4C^2\delta^2(h_x, h_y)(-\triangle_h v, v)_{xy},$$

which implies (46). QED.

4.4 Preconditioned iterative methods for QSC

Let g_Δ be the quadratic spline interpolant of g at the midpoints. In this section, using the spectral equivalence of L_h and $-\triangle_h$, we formulate preconditioned iterative methods for solving $L_h u_\Delta = g_\Delta$, with preconditioner $-\triangle_h$, and convergence rate independent of h_x and h_y.

Let $||\cdot|| = \sqrt{(\cdot,\cdot)_{xy}}$ be the standard norm in S_Δ, $||L_h^1|| = \sup_{v\neq 0} ||L_h^1 v||/||v||$ be the induced norm of operator L_h^1 from S_Δ to S_Δ, $||\cdot||_{L_h^2} = \sqrt{(L_h^2\cdot,\cdot)}$ be the energy norm (or L_h^2-norm) associated with the self-adjoint and positive definite operator L_h^2 from S_Δ to S_Δ and E_h be the identity operator in S_Δ.

Let us assume that $f(x,y) > -2\pi^2\alpha$ for $(x,y) \in \overline{\Omega}$, and let also the assumptions of Theorem 1 be valid. Relations (45) and (46) of Theorem 3 show that the operators L_h and $-\triangle_h$ (A and B, respectively, in [1]) satisfy all the assumptions of Lemma 2.1 of [1], with

$$\gamma_1 = \alpha + \frac{\eta_*}{\lambda_*(h_x, h_y)} - C\delta(h_x, h_y),\ \gamma_2 = \gamma + \frac{\eta^*}{\lambda_*(h_x, h_y)} + C\delta(h_x, h_y),\ \gamma_3 = C\delta(h_x, h_y).$$

Note that, by using the fact that $\sin(x) = x + O(x^3)$, $\forall x$, we have $1/\lambda_*(h_x, h_y) = 1/(2\pi^2) + O(h_x^2 + h_y^2)$. Then,

$$\gamma_1 = \alpha + \frac{\eta_*}{2\pi^2} + O(h_x + h_y), \gamma_2 = \gamma + \frac{\eta^*}{2\pi^2} + O(h_x + h_y) \text{ and } \gamma_3 = O(h_x + h_y). \tag{47}$$

Furthermore, since $f > -2\pi^2\alpha$, $\gamma_1 > 0$ for sufficiently small h_x and h_y, and $\gamma_3 \geq 0$. Also note that the condition $f > -2\pi^2\alpha$ guarantees that the eigenvalues of the QSC matrix arising from a constant coefficients Helmholtz operator are of the same sign [4].

Applying Lemma 2.1 of [1] with D of the Lemma chosen to be $-\triangle_h$ gives

$$||E_h - \overline{\tau}\triangle_h^{-1/2} L_h \triangle_h^{-1/2}|| \leq \rho \tag{48}$$

where

$$\overline{\tau} = \frac{2}{\gamma_1+\gamma_2}(1-\rho\overline{\kappa}),\ \rho = \frac{(1+\overline{\kappa})\gamma_2-(1-\overline{\kappa})\gamma_1}{(1+\overline{\kappa})\gamma_2+(1-\overline{\kappa})\gamma_1},\ \overline{\kappa} = \frac{\gamma_3}{\sqrt{\gamma_1\gamma_2+\gamma_3^2}}. \tag{49}$$

Relation (48) shows that ρ is the bound for the norm of the iteration matrix of a one-step preconditioned iterative method applied to $L_h u_\Delta = g_\Delta$ with preconditioner $-\triangle_h$ and scaling factor $\overline{\tau}$. Thus ρ is the rate of convergence of the respective preconditioned Richardson and MRES methods as shown in Theorems 2.1 and 2.2 in [1]. More specifically,

$$||u_\Delta^{(k)} - u_\Delta||_{-\triangle_h} \leq \rho^k ||u_\Delta^{(0)} - u_\Delta||_{-\triangle_h} \tag{50}$$

where $u_\Delta^{(k)} = \sum_{i=1}^{M}\sum_{j=1}^{N} \mathrm{x}_{i,j}^{(k)} \phi_i^x \phi_j^y$ is the bi-quadratic spline approximation computed at the k-th iteration of Richardson's method applied to $L_h u_\Delta = g_\Delta$, with (symmetric) preconditioner $-\triangle_h$ and scaling factor $\overline{\tau}$, where $\overline{\tau}$ and ρ are given by (49). A similar relation can be shown for the MRES iterates and for the $(L_h^*(-\triangle_h)^{-1}L_h)$-norm of the error.

We next show that ρ is asymptotically independent of h_x and h_y. We also predict an approximation to ρ and the optimum scaling parameter $\overline{\tau}$ for Richardson's method according to Lemma 2.1 in [1]. From (47) and (49) we have

$$\overline{\tau} = \frac{2}{\gamma+\alpha+(\eta^*+\eta_*)/(2\pi^2)} + O(h_x+h_y) \text{ and } \rho = \frac{\gamma-\alpha+(\eta^*-\eta_*)/(2\pi^2)}{\gamma+\alpha+(\eta^*+\eta_*)/(2\pi^2)} + O(h_x+h_y). \tag{51}$$

It is interesting to note that the approximations for $\overline{\tau}$ and ρ obtained from (51) by disregarding the $O(h_x+h_y)$ terms are exactly the same as those in [1] for Hermite cubic spline collocation.
It is also worth noting that, if we make certain assumptions for the signs of f_* and f^*, tighter bounds for L_h than those in (45) can be obtained. For example, if $f_* \geq 0$,

$$\left[\alpha + \frac{f_*}{\lambda^*(h_x,h_y)} - C\delta(h_x,h_y)\right](-\triangle_h) \leq L_h \leq \left[\gamma + \frac{f^*}{\lambda_*(h_x,h_y)} + C\delta(h_x,h_y)\right](-\triangle_h). \tag{52}$$

However, (52) leads to a convergence rate improved by only $O(h_x^2+h_y^2)$ compared to ρ in (51), since $1/\lambda^*(h_x,h_y) = O(h_x^2+h_y^2)$.

4.5 H^1 norms of the QSC approximation error

Relation (50) gives the rate of convergence in the $(-\triangle_h)$-norm of the error in the bi-quadratic spline approximations $u_\Delta^{(k)} = \sum_{i=1}^{M}\sum_{j=1}^{N} \mathrm{x}_{i,j}^{(k)} \phi_i^x \phi_j^y$ to $u_\Delta = \sum_{i=1}^{M}\sum_{j=1}^{N} \mathrm{x}_{i,j} \phi_i^x \phi_j^y$ computed by the preconditioned Richardson method. We will now obtain a result in the H^1-norm of the error. For Hermite cubic splines, the equivalence of the $-\triangle_h$ and H^1-norms follows easily from earlier work on Hermite cubic spline collocation, mainly [12] and [23]. In the case of QSC, we need to show first several results.

Lemma 4 *Let $p \in S_{\Delta_x}$. Then $(-p_{xx},p)_x \leq \frac{3}{2}||p_x||^2_{L^2(0,1)}$.*

PROOF
Similarly as in the proof of Lemma 1, by applying (27) to $p_x p_x$ in each subinterval $[x_{i-1},x_i]$, $i=1,\cdots,M$, and summing up we have

$$0 \leq (p_x,p_x)_x = (p_x,p_x) - \sum_{i=1}^{M}\int_{x_{i-1}}^{x_i} (p_x^2)_{xx} K_i(x)dx.$$

Thus $(p_x, p_x) \geq \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (p_x^2)_{xx} K_i(x) dx$. Since p is a quadratic in $[x_{i-1}, x_i]$, $(p_x^2)_{xx} = 2p_{xx}^2 = 2(p_{xx}p)_{xx}$. Thus $(-p_{xx}, p)_x = (p_x, p_x) + \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (p_{xx}p)_{xx} K_i(x) dx = (p_x, p_x) + \frac{1}{2} \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (p_x^2)_{xx} K_i(x) dx \leq \frac{3}{2}(p_x, p_x)$. QED.

We establish bounds on the quadratic spline basis functions similar to those shown in Lemma 5.4 and Theorem 5.5 of [23], using the technique of that paper.

Lemma 5 *For $i = 1, \cdots, M$, $j = 1, \cdots, M$, $||\phi_i^x||_{L^\infty(x_{j-1}, x_j)} \leq \frac{3}{2} 3^{-|i-j|}$.*

PROOF
Using the form of the quadratic spline basis functions, we have $||\phi_i^x||_{L^\infty(x_{j-1}, x_j)} \leq \frac{1}{2} \max\{1, \frac{3}{2}, 1, 0\} = \frac{3}{4}$, which leads to the desired result. QED.

Corollary 1 *For $i = 1, \cdots, M$, $j = 1, \cdots, M$, $|(\phi_i^x, \phi_j^x)| \leq \frac{9}{2} h_x 2^{-|i-j|}$.*

PROOF
Without loss of generality, let $i \leq j$. Using Lemma 5, and a technique similar to that used for the proof of Theorem 5.5 in [23], we have
$|(\phi_i^x, \phi_j^x)| \leq \sum_{k=1}^{M} \int_{x_{k-1}}^{x_k} |\phi_i^x||\phi_j^x| dx \leq \frac{9}{4} h_x \sum_{k=1}^{M} 3^{-(|i-k|+|j-k|)} \leq \frac{9}{4} h_x 3^{-(j-i)}(2 + j - i)$.
Now using the inequality $2 + l \leq 2(\frac{3}{2})^l$, which holds for any integer $l \geq 0$, we get the desired result. QED.

The following lemma holds for quadratic splines and does not have a Hermite cubic spline equivalent, since quadratic splines are "non-nodal", that is, the degrees of freedom of a quadratic spline do not represent values of the spline at some particular points. The lemma is needed in the proof of Theorem 4.

Lemma 6 *Let x_i, $i = 1, \cdots, M$, be the coefficients (degrees of freedom) of the finite element representation of a quadratic spline $p \in S_{\Delta_x}$. Then*

$$\sum_{i=1}^{M} \mathrm{x}_i^2 \leq \frac{64}{5} \sum_{i=1}^{M} (p(\tau_i^x))^2$$

PROOF
Given that $p(\tau_1^x) = \frac{1}{8}(5\mathrm{x}_1 + \mathrm{x}_2)$, $p(\tau_i^x) = \frac{1}{8}(\mathrm{x}_{i-1} + 6\mathrm{x}_i + \mathrm{x}_{i+1})$, $i = 2, \cdots, M-1$, and $p(\tau_M^x) = \frac{1}{8}(\mathrm{x}_{M-1} + 5\mathrm{x}_M)$, the proof of this lemma, though tedious, uses only simple calculations. QED.

Theorem 4 *Let $p \in S_{\Delta_x}$. Then $\frac{1}{3}(p, p)_x \leq ||p||_{L^2(0,1)}^2 \leq \frac{864}{5}(p, p)_x$.*

PROOF
The left inequality is shown by considering an arbitrary subinterval (x_{j-1}, x_j), $j = 1, \cdots, M$, and showing that $\frac{1}{3} h_x (p(\tau_i^x))^2 \leq \int_{x_{j-1}}^{x_j} (p(x))^2 dx$ by doing simple calculations, since p is a quadratic. To show the right inequality, consider the finite element representation of $p = \sum_{i=1}^{M} \mathrm{x}_i \phi_i^x$. Then, using Corollary 1 and a technique similar to that used for the proof of Theorem 5.5 in [23], we have
$||p||_{L^2(0,1)}^2 = \sum_{i=1}^{M} \sum_{j=1}^{M} \mathrm{x}_i \mathrm{x}_j (\phi_i^x, \phi_j^y) \leq \frac{9}{2} h_x \sum_{i=1}^{M} \sum_{j=1}^{M} |\mathrm{x}_i||\mathrm{x}_j| 2^{-|i-j|} \leq \frac{9}{4} h_x \sum_{i=1}^{M} \sum_{j=1}^{M} (\mathrm{x}_i^2 + \mathrm{x}_j^2) 2^{-|i-j|} \leq \frac{9}{2} h_x \sum_{i=1}^{M} (\mathrm{x}_i^2 \sum_{j=1}^{M} 2^{-|i-j|}) \leq \frac{27}{2} h_x \sum_{i=1}^{M} \mathrm{x}_i^2$. Now using Lemma 6, we get the desired result. QED.

We now extend Lemma 4 to two dimensions. In order to do this, we follow the approach in [23] and define a semidiscrete norm in S_Δ by

$$|||w|||^2 \equiv \sum_{j=1}^{N} h_y \int_0^1 w_x^2(x, \tau_j^y) dx + \sum_{i=1}^{M} h_x \int_0^1 w_y^2(\tau_i^x, y) dy.$$

Note that $|||\cdot|||$ is a seminorm in the space of differentiable functions in Ω.

Corollary 2 *Let $w \in S_\Delta$. Then $\frac{1}{3}|||w||| \leq \|\nabla w\|_{L^2(\Omega)} \leq \frac{864}{5}|||w|||$.*

PROOF
The proof follows from Lemma 4 applied to both the x and y dimensions. QED.

The following theorem shows the equivalence of the $-\triangle_h$- and H^1-norms in the bi-quadratic spline space.

Theorem 5 *Let $w \in S_\Delta$. Then, for some positive constants C_6 and C_7 independent of Δ,*

$$C_6||w||_{H^1(\Omega)} \leq ||w||_{-\triangle_h} \leq C_7||w||_{H^1(\Omega)}.$$

PROOF
Consider the left inequality. By the definition of the $(\cdot,\cdot)_{xy}$, $(\cdot,\cdot)_x$ and $(\cdot,\cdot)_y$ inner products and by Lemma 1, we have $||w||^2_{-\triangle_h} = (-w_{xx}, w)_{xy} + (-w_{yy}, w)_{xy} = \sum_{j=1}^N h_y(-w_{xx}, w)_x(\cdot, \tau_j^y) + \sum_{i=1}^M h_x(-w_{yy}, w)_y(\tau_i^x, \cdot)$ $\geq \sum_{j=1}^N h_y(w_x, w_x)(\cdot, \tau_j^y) + \sum_{i=1}^M h_x(w_y, w_y)(\tau_i^x, \cdot) = |||w|||^2$. Employing Corollary 2, then the Poincaré inequality, we get $|||w||| \geq \frac{5}{864}\|\nabla w\|_{L^2(\Omega)} \geq C_6||w||_{H^1(\Omega)}$. The right inequality can be shown in a similar way, using Lemma 4 instead of Lemma 1. QED.

Having the equivalence of the $-\triangle_h$- and H^1-norms for bi-quadratic splines and using (50), we get

$$||u_\Delta^{(k)} - u_\Delta||_{H^1(\Omega)} \leq C\rho^k||u_\Delta^{(0)} - u_\Delta||_{H^1(\Omega)} \tag{53}$$

where $u_\Delta^{(k)} = \sum_{i=1}^M \sum_{j=1}^N \mathrm{x}_{i,j}^{(k)} \phi_i^x \phi_j^y$ is the bi-quadratic spline approximation computed at the k-th iteration of Richardson's method applied to $L_h u_\Delta = g_\Delta$, with preconditioner $-\triangle_h$ and scaling factor $\overline{\tau}$, where $\overline{\tau}$ and ρ are given by (51), and C is a constant independent of Δ, a, c and f.

We can obtain the relation (53) for the preconditioned MRES iterates too, by first establishing the spectral equivalence of the $-\triangle_h$ and $L_h^*(-\triangle_h)^{-1}L_h$ operators as in Lemma 3.1 of [1], then employing Theorem 2.2 in the same paper.

4.6 Non-self-adjoint operators

In this section, we extend the result of Theorem 3 to non-self-adjoint operators, without cross-derivative term. Consider the BVP described by the operator equation (1), where $\mathbf{L}$ is given by

$$\mathbf{L}u \equiv -(a(x,y)u_x)_x - (c(x,y)u_y)_y + d(x,y)u_x + e(x,y)u_y + f(x,y)u, \tag{54}$$

and homogeneous Dirichlet boundary conditions on $\partial\Omega$ (26). Let L_h be QSC operator from S_Δ into S_Δ corresponding to $\mathbf{L}$ in (54).

Theorem 6 *Assume $a(x,y)$ and $d(x,y) \in \mathbf{C}^3(\overline{\Omega})$ with respect to x and $c(x,y)$ and $e(x,y) \in \mathbf{C}^3(\overline{\Omega})$ with respect to y, $f(x,y) \in \mathbf{C}$, and $0 < \alpha \leq a(x,y), c(x,y) \leq \gamma$, for all $(x,y) \in \overline{\Omega}$. Then, for all $v, w \in S_\Delta$,*

$$\left(\alpha + \frac{\eta_*}{\lambda_*(h_x,h_y)} - C[\delta+\omega]\right)(-\triangle_h) \leq L_h \leq \left(\gamma + \frac{\eta^*}{\lambda_*(h_x,h_y)} + C[\delta+\omega]\right)(-\triangle_h) \tag{55}$$

and

$$(L_h^* - L_h)(-\triangle_h)^{-1}(L_h - L_h^*) \leq 4C^2[\sigma + \delta(h_x,h_y) + \omega(h_x,h_y)]^2(-\triangle_h), \tag{56}$$

where C is a positive constant independent of a, c, d, e, f, h_x and h_y, $\delta(h_x, h_y)$ is defined in Theorem 1, $\lambda_*(h_x, h_y)$ in Theorem 2, and

$$\begin{aligned}
\eta_* &= \min(0, \min \tilde{f}(x,y)), \;\; \eta^* = \max(0, \max \tilde{f}(x,y)), \\
\omega(h_x, h_y) &= \max(h_x \max(||d||_\infty, ||d_x||_\infty) + h_x^2 \max(||d_{xx}||_\infty, ||d_{xxx}||_\infty), \\
&\qquad h_y \max(||e||_\infty, ||e_y||_\infty) \; + h_y^2 \max(||e_{yy}||_\infty, ||e_{yyy}||_\infty)), \\
\sigma &= \max(||d||_\infty, ||d_x||_\infty, ||e||_\infty, ||e_y||_\infty),
\end{aligned}$$

where $\tilde{f}(x,y) = f(x,y) - [d_x(x,y) + e_y(x,y)]/2$.

PROOF
We can rewrite $\mathbf{L}u$ to get

$$\begin{aligned}
\mathbf{L}u &= -(a(x,y)u_x)_x - (c(x,y)u_y)_y \\
&\quad + \frac{1}{2}[d(x,y)u_x + (d(x,y)u)_x + e(x,y)u_y + (e(x,y)u)_y] + \tilde{f}(x,y)u.
\end{aligned}$$

For any $j = 1, \cdots, N$, by applying (27) to $((dv_x + (dv)_x)w)(x, \tau_j^y)$ in each subinterval $[x_{i-1}, x_i]$, $i = 1, \cdots, M$, summing up and using integration by parts and Leibnitz's differentiation formula, we get

$$\begin{aligned}
((dv_x + (dv)_x)(\cdot, \tau_j^y), w(\cdot, \tau_j^y))_x &= (dv_x(\cdot, \tau_j^y), w(\cdot, \tau_j^y)) - \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (dv_x w)_{xx}(x, \tau_j^y) K_i(x) dx \\
&+ ((dv)_x(\cdot, \tau_j^y), w(\cdot, \tau_j^y)) - \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} ((dv)_x w)_{xx}(x, \tau_j^y) K_i(x) dx \\
&= I_3(d, v, w, \tau_j^y) + I_4(d, v, w, \tau_j^y), \qquad (57)
\end{aligned}$$

where

$$\begin{aligned}
I_3(d, v, w, \tau_j^y) &\equiv \int_0^1 ((dv)_x w)(x, \tau_j^y) dx - \int_0^1 ((dw)_x v)(x, \tau_j^y) dx, \\
I_4(d, v, w, \tau_j^y) &\equiv \sum_{i=1}^{M} \sum_{l=1}^{3} \sum_{m+n=3-l}^{0 \le m,n \le 2} \alpha_{lmn} \int_{x_{i-1}}^{x_i} (d_x^{(l)} v_x^{(m)} w_x^{(n)})(x, \tau_j^y) K_i(x) dx.
\end{aligned}$$

By using (57) and a similar approach to that in the proof of Theorem 1, we can show that

$$(L_h v, w)_{xy} = \sum_{i=1}^{4} B_h^i(v, w) + (\tilde{f}(x,y)v, w),$$

where the bilinear forms B_h^1 and B_h^2 satisfy (29)-(31) and

$$\begin{aligned}
B_h^3(v, v) = 0, \; |B_h^3(v, w)| &\le C\sigma(-\triangle_h v, v)_{xy}^{1/2} (-\triangle_h w, w)_{xy}^{1/2}, \\
|B_h^4(v, w)| &\le C\omega(h_x, h_y)(-\triangle_h v, v)_{xy}^{1/2} (-\triangle_h w, w)_{xy}^{1/2}.
\end{aligned}$$

The proof of the inequalities (55) and (56) now follows in a similar way to that of the proof for the inequalities (45) and (46) in Theorem 3. QED.

Theorem 6 is the QSC counterpart of Theorem 4.1 in [1]. Note that, for QSC as for Hermite cubic spline collocation, we cannot predict an approximation to the optimal scaling factor $\overline{\tau}$ for Richardson's method, when the PDE operator is non-self-adjoint.

Theorem 6 allows relations (50) and (53) to hold even on QSC equations arising from (54), assuming that the condition $f(x,y) > -2\pi^2\alpha$ is replaced by $\tilde{f}(x,y) > -2\pi^2\alpha$, where $\tilde{f}$ is defined in Theorem 6.

4.7 Extension to systems of PDEs

In this section, we consider the solution of the QSC equations arising from general 2×2 systems of elliptic PDEs of the form (16) by preconditioned iterative methods.

Consider the space $S_\Delta \times S_\Delta$ of 2×1 vectors $[u, v]^T$ of bi-quadratic splines that satisfy homogeneous Dirichlet boundary conditions by construction. In this section, for convenience, we will denote $[u, v]^T$ by $[u, v]$. It is easy to show that, for $[u_1, v_1], [u_2, v_2] \in S_\Delta \times S_\Delta$, $([u_1, v_1], [u_2, v_2])_{xy} \equiv (u_1, u_2)_{xy} + (v_1, v_2)_{xy}$ defines an inner product, and that $S_\Delta \times S_\Delta$ is a Hilbert space, of dimension $2NM$.

The analysis is carried out for the QSC equations arising from (16), with $\mathbf{L}_{ij}$, $i = 1, 2$, $j = 1, 2$, given by

$$\mathbf{L}_{ij}u \equiv -(a_{ij}(x, y)u_x)_x - (c_{ij}(x, y)u_y)_y + f_{ij}(x, y)u. \tag{58}$$

We assume that both u and v satisfy homogeneous Dirichlet boundary conditions on $\partial\Omega$. The preconditioner is the QSC operator arising from

$$\hat{\triangle} \equiv \begin{bmatrix} \xi_1\triangle & \xi_2\triangle \\ \xi_2\triangle & \xi_1\triangle \end{bmatrix} \tag{59}$$

with $\xi_1 > \xi_2 \geq 0$. In Theorem 8 we give a formula to compute the "best" ξ_1 and ξ_2 a priori, using knowledge only from the coefficients of the PDE operators.

Let L_{hij} be the QSC operators corresponding to $\mathbf{L}_{ij}$, $i = 1, 2$, $j = 1, 2$, respectively. For $[u, v] \in S_\Delta \times S_\Delta$, define the operators $\hat{L}_h$ and $\hat{\triangle}_h$ to be the QSC operators corresponding to $\mathbf{L}$ and $\hat{\triangle}$, respectively. That is,

$$\hat{L}_h[u, v] \equiv [L_{h11}u + L_{h12}v, L_{h21}u + L_{h22}v] \text{ and } \hat{\triangle}_h[u, v] \equiv [\xi_1\triangle_h u + \xi_2\triangle_h v, \xi_2\triangle_h u + \xi_1\triangle_h v].$$

Thus, $(-\hat{\triangle}_h[u, v], [u, v])_{xy} = \xi_1(-\triangle_h u, u)_{xy} + \xi_2(-\triangle_h v, u)_{xy} + \xi_2(-\triangle_h u, v)_{xy} + \xi_1(-\triangle_h v, v)_{xy}$. It is easy to show that $\hat{\triangle}_h$ is self-adjoint and positive definite under $([\cdot, \cdot], [\cdot, \cdot])_{xy}$. More specifically, from Theorem 2 and the definition of $\hat{\triangle}_h$, we have

$$(\xi_1 - \xi_2)\lambda_*(h_x, h_y)([u, v], [u, v])_{xy} \leq (-\hat{\triangle}_h[u, v], [u, v])_{xy} \leq (\xi_1 + \xi_2)\lambda^*(h_x, h_y)([u, v], [u, v])_{xy}.$$

Note also that the adjoint $\hat{L}_h^*$ of $\hat{L}_h$ is given by $\hat{L}_h^*[u, v] \equiv [L_{h11}^* u + L_{h12}^* v, L_{h21}^* u + L_{h22}^* v]$ and the inverse $\hat{\triangle}_h^{-1}$ of $\hat{\triangle}_h$ is given by $\hat{\triangle}_h^{-1}[u, v] \equiv \frac{1}{\xi_1^2 - \xi_2^2}[\xi_1\triangle_h^{-1}u - \xi_2\triangle_h^{-1}v, -\xi_2\triangle_h^{-1}u + \xi_1\triangle_h^{-1}v]$.

We consider first the case $\xi_1 = 1$, $\xi_2 = 0$. In the following, we present a result similar to Theorem 3 for 2×2 systems of PDEs and the above preconditioner.

Theorem 7 *Assume that the operators $\mathbf{L}_{ij}$, for $i = 1, 2$ and $j = 1, 2$, satisfy assumptions similar to those of Theorem 1, with $0 < \alpha_{ij} \leq a_{ij}(x, y), c_{ij}(x, y) \leq \gamma_{ij}$, for all $(x, y) \in \overline{\Omega}$, and $f_{ij} > -2\pi^2\alpha_{ij}$, for $i = 1, 2$ and $j = 1, 2$, and L_{h12} and L_{h21} are self-adjoint under $(\cdot, \cdot)_{xy}$. Let $\mathrm{A}_{ij} = \alpha_{ij} + \frac{\eta_{ij*}}{\lambda_*(h_x,h_y)} - C\delta_{ij}(h_x, h_y)$, $\Gamma_{ij} = \gamma_{ij} + \frac{\eta_{ij}^*}{\lambda_*(h_x,h_y)} + C\delta_{ij}(h_x, h_y)$, where η_{ij*}, η_{ij}^* and δ_{ij} are defined similarly to η_*, η^* and δ in Theorem 3, respectively. Let $\theta_1 = \Gamma_{12}/\mathrm{A}_{11}$, $\theta_2 = \Gamma_{21}/\mathrm{A}_{22}$ and $\Theta = (\theta_1\Gamma_{11} + \theta_2\Gamma_{22})/2$. Then,*

$$\min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\}(-\hat{\triangle}_h) \leq \hat{L}_h \leq \max_{i=1,2}\{\Gamma_{ii} + \Theta\}(-\hat{\triangle}_h) \tag{60}$$

and

$$(\hat{L}_h^* - \hat{L}_h)(-\hat{\triangle}_h)^{-1}(\hat{L}_h - \hat{L}_h^*) \leq 4C^2 \max_{i=1,2}\{\delta_{ii}^2(h_x, h_y)\}(-\hat{\triangle}_h), \tag{61}$$

where C is a positive constant independent of $a_{ij}, c_{ij}, f_{ij}, h_x$ and h_y, and $\hat{\triangle}_h$ is constructed with $\xi_1 = 1$ and $\xi_2 = 0$.

PROOF
We note that, since $\mathbf{L}_{ij}$, for $i = 1, 2$ and $j = 1, 2$, satisfy assumptions similar to those of Theorem 1, Theorem 3 holds for each of the discrete operators L_{hij}, $i = 1, 2$ and $j = 1, 2$. Thus

$$\mathrm{A}_{ij}(-\triangle_h) \leq L_{hij} \leq \Gamma_{ij}(-\triangle_h), \tag{62}$$

and note that $\mathrm{A}_{ij} > 0$ for sufficiently small h_x and h_y. Thus, all four operators L_{hij}, $i = 1, 2$, $j = 1, 2$, are spectrally equivalent to each other and to the negative Laplacian. More specifically, $L_{h12} \leq \Gamma_{12}(-\triangle_h) \leq \Gamma_{12}/\mathrm{A}_{11}L_{h11} = \theta_1 L_{h11}$. Similarly, $L_{h21} \leq \Gamma_{21}/\mathrm{A}_{22}L_{h22} = \theta_2 L_{h22}$. From $L_{h12} \leq \theta_1 L_{h11}$, we get $(L_{h12}u, u)_{xy} \leq \theta_1(L_{h11}u, u)_{xy}$, which implies

$$(L_{h12}v, u)_{xy} \geq -\theta_1(L_{h11}u, u)_{xy} + (L_{h12}(u + v), u)_{xy}. \tag{63}$$

In a similar way, and using the self-adjointness of L_{h12}, we get

$$(L_{h12}v, u)_{xy} \geq -\theta_1(L_{h11}v, v)_{xy} + (L_{h12}(u + v), v)_{xy}. \tag{64}$$

From (63), (64), and (62) applied to L_{h11}, we get

$$\begin{aligned} 2(L_{h12}v, u)_{xy} &\geq -\theta_1\Gamma_{11}((-\triangle_h u, u)_{xy} + (-\triangle_h v, v)_{xy}) + \mathrm{A}_{12}(-\triangle_h(u + v), u + v)_{xy} \\ &\geq -\theta_1\Gamma_{11}((-\triangle_h u, u)_{xy} + (-\triangle_h v, v)_{xy}). \end{aligned} \tag{65}$$

From (65) and a similar relation for L_{h21} we have

$$\begin{aligned} 2((L_{h12}v, u)_{xy} + (L_{h21}u, v)_{xy}) &\geq -(\theta_1\Gamma_{11} + \theta_2\Gamma_{22})((-\triangle_h u, u)_{xy} + (-\triangle_h v, v)_{xy}) \\ &\quad +(\mathrm{A}_{12} + \mathrm{A}_{21})(-\triangle_h(u + v), u + v)_{xy} \qquad (66) \\ &\geq -(\theta_1\Gamma_{11} + \theta_2\Gamma_{22})((-\triangle_h u, u)_{xy} + (-\triangle_h v, v)_{xy}). \qquad (67) \end{aligned}$$

In a similar way, we can get

$$\begin{aligned} 2((L_{h12}v, u)_{xy} + (L_{h21}u, v)_{xy}) &\leq (\theta_1\Gamma_{11} + \theta_2\Gamma_{22})((-\triangle_h u, u)_{xy} + (-\triangle_h v, v)_{xy}) \\ &\quad -(\mathrm{A}_{12} + \mathrm{A}_{21})(-\triangle_h(u - v), u - v)_{xy} \qquad (68) \\ &\leq (\theta_1\Gamma_{11} + \theta_2\Gamma_{22})((-\triangle_h u, u)_{xy} + (-\triangle_h v, v)_{xy}). \qquad (69) \end{aligned}$$

From (67), (69), and (62) applied to L_{h11} and L_{h22}, we can obtain (60).

To show (61), consider $[u_1, v_1] \equiv (-\hat{\triangle}_h)^{-1}(\hat{L}_h - \hat{L}_h^*)[u, v]$. Using the self-adjointness of L_{h12} and L_{h21}, we have $([u_1, v_1], (\hat{L}_h - \hat{L}_h^*)[u, v])_{xy} = (u_1, (L_{h11} - L_{h11}^*)u)_{xy} + (v_1, (L_{h22} - L_{h22}^*)v)_{xy}$. Applying a technique similar to that used for the proof of (46) to each of the terms of the above sum we get (61). QED.

Remark 5. Theorem 7 shows that, if, for sufficiently small h_x and h_y, $\mathrm{A}_{ii} - \Theta > 0$, for $i = 1, 2$, then $\hat{L}_h$ is spectrally equivalent to $-\hat{\triangle}_h$. Thus, the preconditioner constructed from $-\hat{\triangle}_h$ gives rise to iterative methods with convergence rate independent of the problem size. Having sufficiently small h_x and h_y is needed in bounding not only A_{ii}, $i = 1, 2$, but also $\min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\}$ from below by a positive constant independent of h_x and h_y, and $\max_{i=1,2}\{\Gamma_{ii} + \Theta\}$ from above by a constant independent of h_x and h_y.

Remark 6. The conditions $\mathrm{A}_{ii} - \Theta > 0$, for $i = 1, 2$, force θ_1 and/or θ_2 to be strictly less than 1 (and possibly far less than 1).

Remark 7. The conditions $L_{h12} \leq \theta_1 L_{h11}$ and $L_{h21} \leq \theta_2 L_{h22}$, with $\theta_1 < 1$ and $\theta_2 < 1$ can be interpreted as strict diagonal dominance of $\mathbf{L}_{11}$ and $\mathbf{L}_{22}$ over $\mathbf{L}_{12}$ and $\mathbf{L}_{21}$, respectively.

Remark 8. If $\theta_1 > 1$, but $L_{h12} \geq \theta_1^* L_{h11}$ for $\theta_1^* > 1$, and $\theta_2 > 1$, but $L_{h21} \geq \theta_2^* L_{h22}$ for $\theta_2^* > 1$, with appropriate rearrangement of the PDE operators and/or the unknown functions u and v, we can obtain an equivalent 2×2 system of PDEs which satisfies the "diagonal dominance" condition.

Remark 9. The self-adjointness of L_{h12} and L_{h21} under $(\cdot,\cdot)_{xy}$ is a quite restrictive condition. It holds if $\mathbf{L}_{12}$ and $\mathbf{L}_{21}$ have constant coefficients, or in some other special cases. While we were not able to relax this condition for the analysis, our experiments show that the condition is only sufficient, not necessary.

We consider now the case of the preconditioner arising from (59) with $\xi_1 > \xi_2 > 0$. In the following, we present a result similar to Theorem 3 for 2×2 systems of PDEs and the above preconditioner.

Theorem 8 *Assume that the operators $\mathbf{L}_{ij}$, for $i = 1, 2$ and $j = 1, 2$, satisfy assumptions similar to those of Theorem 1, with $0 < \alpha_{ij} \leq a_{ij}(x,y), c_{ij}(x,y) \leq \gamma_{ij}$, for all $(x,y) \in \overline{\Omega}$, $f_{ij} > -2\pi^2\alpha_{ij}$, for $i = 1, 2$ and $j = 1, 2$, and L_{h12} and L_{h21} are self-adjoint under $(\cdot,\cdot)_{xy}$. Let $\mathrm{A}_{ij} = \alpha_{ij} + \frac{\eta_{ij*}}{\lambda_*(h_x,h_y)} - C\delta_{ij}(h_x,h_y)$, $\Gamma_{ij} = \gamma_{ij} + \frac{\eta_{ij}^*}{\lambda_*(h_x,h_y)} + C\delta_{ij}(h_x,h_y)$, where η_{ij*}, η_{ij}^* and δ_{ij} are defined similarly to η_*, η^* and δ in Theorem 3, respectively. Let $\theta_1 = \Gamma_{12}/\mathrm{A}_{11}$, $\theta_2 = \Gamma_{21}/\mathrm{A}_{22}$, $\theta_1' = \mathrm{A}_{12}/\Gamma_{11}$, $\theta_2' = \mathrm{A}_{21}/\Gamma_{22}$, $\Theta = (\theta_1\Gamma_{11} + \theta_2\Gamma_{22})/2$, $\mathrm{B} = (\mathrm{A}_{12} + \mathrm{A}_{21})/2$, $\Theta' = (\theta_1'\mathrm{A}_{11} + \theta_2'\mathrm{A}_{22})/2$, and $\mathrm{B}' = (\Gamma_{12} + \Gamma_{21})/2$. Then,*

$$\kappa(-\hat{\triangle}_h) \leq \hat{L}_h \leq (-\hat{\triangle}_h) \tag{70}$$

where $\kappa = \min\{\mathrm{B}/\mathrm{B}',\ \min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\}/\max_{i=1,2}\{\Gamma_{ii} - \Theta'\}\}$ and $\hat{\triangle}_h$ is constructed with

$$\xi_1 = \max_{i=1,2}\{\Gamma_{ii} + \mathrm{B}' - \Theta'\},\ \xi_2 = \mathrm{B}'. \tag{71}$$

Moreover,

$$(\hat{L}_h^* - \hat{L}_h)(-\hat{\triangle}_h)^{-1}(\hat{L}_h - \hat{L}_h^*) \leq 4C^2 \max_{i=1,2}\{\delta_{ii}^2(h_x,h_y)\}\frac{2\xi_1}{(\xi_1^2 - \xi_2^2)(\xi_1 - \xi_2)}(-\hat{\triangle}_h), \tag{72}$$

where C is a positive constant independent of $a_{ij}, c_{ij}, f_{ij}, h_x$ and h_y.

PROOF

As in the proof of Theorem 7 we can obtain $\theta_1' L_{h11} \leq L_{h12} \leq \theta_1 L_{h11}$ and $\theta_2' L_{h22} \leq L_{h21} \leq \theta_2 L_{h22}$. We re-write (66) as

$$\begin{aligned} 2((L_{h12}v,u)_{xy} + (L_{h21}u,v)_{xy}) &\geq -2\Theta((-\triangle_h u,u)_{xy} + (-\triangle_h v,v)_{xy}) + 2\mathrm{B}(-\triangle_h(u+v),u+v)_{xy} \\ &= 2(\mathrm{B}-\Theta)((-\triangle_h u,u)_{xy} + (-\triangle_h v,v)_{xy}) + 2\mathrm{B}((-\triangle_h u,v)_{xy} + (-\triangle_h v,u)_{xy}). \end{aligned} \tag{73}$$

Using the conditions $\theta_1' L_{h11} \leq L_{h12}$, $\theta_2' L_{h22} \leq L_{h21}$, and a technique similar to that used for the proof of (66) we get

$$\begin{aligned} 2((L_{h12}v,u)_{xy} + (L_{h21}u,v)_{xy}) \leq\ & -(\theta_1'\mathrm{A}_{11} + \theta_2'\mathrm{A}_{22})((-\triangle_h u,u)_{xy} + (-\triangle_h v,v)_{xy}) \\ & +(\Gamma_{12}+\Gamma_{21})(-\triangle_h(u+v),u+v)_{xy} \\ = 2(\mathrm{B}'-\Theta')((-\triangle_h u,u)_{xy} + (-\triangle_h v,v)_{xy}) &+ 2\mathrm{B}'((-\triangle_h u,v)_{xy} + (-\triangle_h v,u)_{xy}). \end{aligned} \tag{74}$$

From (73), (74), and (62) applied to L_{h11} and L_{h22}, we can obtain

$$\begin{aligned} &\min_{i=1,2}\{(\mathrm{A}_{ii} + \mathrm{B} - \Theta)\}((-\triangle_h u,u)_{xy} + (-\triangle_h v,v)_{xy}) + \mathrm{B}((-\triangle_h u,v)_{xy} + (-\triangle_h v,u)_{xy}) \\ \leq &(\hat{L}_h[u,v],[u,v])_{xy} \\ \leq &\max_{i=1,2}\{(\Gamma_{ii} + \mathrm{B}' - \Theta')\}((-\triangle_h u,u)_{xy} + (-\triangle_h v,v)_{xy}) + \mathrm{B}'((-\triangle_h u,v)_{xy} + (-\triangle_h v,u)_{xy}). \end{aligned} \tag{75}$$

For convenience, let $\kappa_1 = \mathrm{B}/\mathrm{B}'$, $\kappa_2 = \min_{i=1,2}\{(\mathrm{A}_{ii} - \Theta)\}/\max_{i=1,2}\{(\Gamma_{ii} - \Theta')\}$ and $\kappa_3 = \min_{i=1,2}\{(\mathrm{A}_{ii} + \mathrm{B} - \Theta)\}/\max_{i=1,2}\{(\Gamma_{ii} + \mathrm{B}' - \Theta')\}$. We now consider three cases.
Case 1: If $\kappa_3 = \kappa_1$, then $\kappa = \kappa_1 = \kappa_2 = \kappa_3$, and we get (70) directly.
Case 2: If $\kappa_3 > \kappa_1$, then $\kappa_2 > \kappa_3 > \kappa_1 = \kappa$, in which case we strengthen the left inequality (weaken the left side) of (75) by replacing $\min_{i=1,2}\{(\mathrm{A}_{ii} + \mathrm{B} - \Theta)\}$ by $\kappa \max_{i=1,2}\{(\Gamma_{ii} + \mathrm{B}' - \Theta')\}$, and then get (70).
Case 3: If $\kappa_3 < \kappa_1$, then $\kappa = \kappa_2 < \kappa_3 < \kappa_1$, in which case we strengthen the left inequality (weaken the left side) of (75) by replacing B by $\kappa_2 \mathrm{B}'$, which again leads to (70). Thus the proof of (70) is complete.
To show (72), consider $[u_1, v_1] \equiv (-\hat{\triangle}_h)^{-1}(\hat{L}_h - \hat{L}_h^*)[u, v]$. Then, we have

$$\begin{aligned}
& ([u_1, v_1], (\hat{L}_h - \hat{L}_h^*)[u, v])_{xy} = (u_1, (L_{h11} - L_{h11}^*)u)_{xy} + (v_1, (L_{h22} - L_{h11}^*)v)_{xy} \\
= \; & \tfrac{\xi_1}{\xi_1^2 - \xi_2^2}(-\triangle_h^{-1}(L_{h11} - L_{h11}^*)u, (L_{h11} - L_{h11}^*)u)_{xy} - \tfrac{\xi_2}{\xi_1^2 - \xi_2^2}(-\triangle_h^{-1}(L_{h22} - L_{h22}^*)v, (L_{h11} - L_{h11}^*)u)_{xy} \\
- \; & \tfrac{\xi_2}{\xi_1^2 - \xi_2^2}(-\triangle_h^{-1}(L_{h11} - L_{h11}^*)u, (L_{h22} - L_{h22}^*)v)_{xy} + \tfrac{\xi_1}{\xi_1^2 - \xi_2^2}(-\triangle_h^{-1}(L_{h22} - L_{h22}^*)v, (L_{h22} - L_{h22}^*)v)_{xy}.
\end{aligned}$$

Applying a technique similar to that used for the proof of (46) once to each of the first and fourth terms of the above sum, and twice to each of the second and third terms, we get

$$\begin{aligned}
& ([u_1, v_1], (\hat{L}_h - \hat{L}_h^*)[u, v])_{xy} \\
\leq \; & \frac{\xi_1}{\xi_1^2 - \xi_2^2} 4C^2 \delta_{11}^2 (-\triangle_h u, u)_{xy} + \frac{\xi_2}{\xi_1^2 - \xi_2^2} 2C\delta_{11}(-\triangle_h u, u)_{xy}^{1/2} 2C\delta_{22}(-\triangle_h v, v)_{xy}^{1/2} \\
+ \; & \frac{\xi_2}{\xi_1^2 - \xi_2^2} 2C\delta_{22}(-\triangle_h v, v)_{xy}^{1/2} 2C\delta_{11}(-\triangle_h u, u)_{xy}^{1/2} + \frac{\xi_1}{\xi_1^2 - \xi_2^2} 4C^2 \delta_{22}^2 (-\triangle_h v, v)_{xy} \\
\leq \; & \frac{4C^2 \max_{i=1,2}\{\delta_{ii}^2(h_x, h_y)\}}{\xi_1^2 - \xi_2^2}((-\hat{\triangle}_h[u, v], [u, v])_{xy} + \xi_2(-\triangle_h(u - v), u - v)_{xy})
\end{aligned}$$

which leads to (72), taking into account that $\frac{\xi_1 + \xi_2}{\xi_1 - \xi_2}(-\hat{\triangle}_h[u, v], [u, v])_{xy} \geq \xi_2(-\triangle_h(u - v), u - v)_{xy}$. QED.
Remark 10. Theorem 8 shows that, if, for sufficiently small h_x and h_y, $\mathrm{A}_{ii} - \Theta > 0$, for $i = 1, 2$, and $B > 0$, then $\hat{L}_h$ is spectrally equivalent to $-\hat{\triangle}_h$. Thus, the preconditioner constructed from $-\hat{\triangle}_h$ with the scalars ξ_1 and ξ_2 chosen as in (71) gives rise to iterative methods with convergence rate independent of the problem size. To obtain a computable approximation to ξ_1 and ξ_2, approximate A_{ij} and Γ_{ij} by ignoring the $C\delta_{ij}$ terms and approximating λ_* by $2\pi^2$. It is worth noting that what matters in constructing the preconditioner is the ratio ξ_2/ξ_1 and not the actual values of ξ_1 and ξ_2.
Remark 11. Theorem 8 gives tighter bounds for the spectral equivalence of $\hat{L}_h$ and $-\hat{\triangle}_h$ than Theorem 7, in Cases 1 and 3 of the proof, since $\kappa > \min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\}/\max_{i=1,2}\{\Gamma_{ii} + \Theta\} \equiv \kappa'$, and κ and κ' give the ratios of the bounding constants in (70) and (60), respectively. These ratios are also an indication of the convergence rate. The closer to 1 the κ or κ' are, the better. (In Case 2, we cannot tell by theory which bound is better.)
Remark 12. When $\xi_2 = 0$, the preconditioner is solved by two applications of the 1D-FFTQSC or 2D-FFTQSC algorithms and not by the 1D-FFTQSC2 or the 2D-FFTQSC2 algorithms. While the flops counts are of the same order, the preconditioner with $\xi_2 = 0$ requires less flops than the preconditioner with $\xi_2 > 0$.
Remark 13. The condition $\mathrm{A}_{ii} - \Theta > 0$ guarantees that $\Gamma_{ii} - \Theta' > 0$, which in turn guarantees that $\xi_1 > \xi_2$.
Remark 14. Using a technique similar to that of the proof of Theorem 8, we can show the relation $\kappa''(-\hat{\triangle}_h) \leq \hat{L}_h \leq (-\hat{\triangle}_h)$, with $\kappa'' = \min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\}/\max_{i=1,2}\{\Gamma_{ii} - 2\mathrm{B} + \Theta\}$, and $\hat{\triangle}_h$ constructed with $\xi_1 = \max_{i=1,2}\{\Gamma_{ii} - \mathrm{B} + \Theta\}$, $\xi_2 = \mathrm{B}$. However, Theorem 8 gives again (see Remark 11) tighter bounds in Cases 1 and 3, since $\kappa > \kappa''$. (Again, in Case 2, we cannot tell by theory which bound is better.)

5 Numerical results

We present results from numerical experiments to demonstrate the performance of the FFT solvers described. All experiments were run on a SUN Ultra-4 (CPU 400 MHz SUNW, UltraSPARC-II) in Fortran using double precision. The timings were obtained by the routine `etime()` (user CPU time). For the implementation of the FFT we used the package [21], while for the other transforms (FCT-II, FST-II, etc) our own code. The QSC method was implemented by us.

The first problem considered is a Helmholtz problem with constant coefficients, used to compare the performance of the 1D-FFTQSC and the 2D-FFTQSC algorithms as direct solvers.

Problem 1:

$$u_{xx} + 3u_{yy} - 2u = g \text{ in } \Omega \equiv (0, 2\pi) \times (0, \pi), \quad u = 0 \text{ on } y = 0, y = \pi, \quad u \text{ periodic in } x$$

The function g is chosen so that the solution u to the problem is $u(x, y) = \sin x \ \sin y$. Note that the 1D-FFTQSC algorithm is applicable as a direct solver to the QSC system arising from more general than constant coefficient Helmholtz problems. The computational complexity of 1D-FFTQSC remains the same whether it is applied to problems with operators of the type (6) or (8).

Table 1 shows results from the application of the 1D-FFTQSC and the 2D-FFTQSC algorithms to the QSC equations arising from the discretization of Problem 1. The error and convergence results are presented in order to verify the performance of the QSC method, as it is described in [4]. (Both solvers produced the same errors.) For each N shown in Table 1, the error "on grid points" corresponds to the maximum in absolute value error of the QSC approximation on the discretization grid, while the "global" error corresponds to the maximum in absolute value error of the QSC approximation on a uniform grid of 20×20 points. The "global" error is taken as an approximation to the uniform norm of the error. The results show that the QSC method is globally of third order, while on the grid points of fourth order, as expected.

The time shown corresponds to the application of the algorithms in the first step of the QSC method. The second step requires the same amount of time as the first one. The times to generate and update the right-side vector in the first and second steps, respectively, of the QSC method are not included. The timing results verify that the FFT solvers are asymptotically almost optimal. As a measure of the "drift" from optimality, we apply linear least squares fit of the form $time \approx \kappa_1 \times N^{\kappa_2}$ to the data, and obtain $time \approx 2.1 \times 10^{-7} N^{2.25}$ for 1D-FFTQSC, and $time \approx 2.1 \times 10^{-7} N^{2.30}$ for 2D-FFTQSC. When comparing the experimental timing results of 1D-FFTQSC and 2D-FFTQSC, the advantage of the former over the latter as obtained by the theoretically expected flops is verified. The difference in the performance of the two solvers is minimal for a small grid, while, as the gridsize increases, it becomes more apparent. However, the asymptotic factor of 2 obtained by the theory is not reached for the gridsizes considered in the experiments.

The second problem is a system of two constant coefficients Helmholtz PDEs.

Problem 2:

$$\begin{array}{l} 6u_{xx} + 3u_{yy} + u + 7v_{xx} + 4v_{yy} + v = g_1 \\ 3u_{xx} + 4u_{yy} + u + 2v_{xx} + 5v_{yy} + v = g_2 \end{array} \quad \text{in } \Omega \equiv (0,1) \times (0,1), \quad \begin{array}{l} u = 0 \\ v = 0 \end{array} \quad \text{on } \partial\Omega.$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x, y) = (x^2 - x)(y^2 - y)e^{x+y}$ and $v(x, y) = x^{9/2}(x - 1)^2 y^{9/2}(y - 1)^2$. Note again that the 1D-FFTQSC2 algorithm is applicable as a direct solver to the QSC system arising from more general than constant coefficient Helmholtz systems of PDEs.

Table 2 shows results from the application of the 1D-FFTQSC2 and the 2D-FFTQSC2 algorithms to the QSC equations arising from the discretization of Problem 2. The error on the gridpoints corresponds

Table 1: Errors, respective orders of convergence, and time in seconds, corresponding to Problem 1 discretized by the QSC method, for several gridsizes $N \times N$. The solution is obtained by 1D-FFTQSC and 2D-FFTQSC.

	on gridpoints		global		time	
N	error	order	error	order	1D-FFTQSC	2D-FFTQSC
32	1.2e-05		5.8e-05		0.00061	0.00077
64	7.7e-07	4.00	6.2e-06	3.23	0.00212	0.00225
128	4.8e-08	4.00	8.6e-07	2.85	0.01118	0.01534
256	3.0e-09	4.00	1.1e-07	2.95	0.05423	0.06356
512	1.9e-10	4.00	9.1e-09	3.61	0.29842	0.41723

Table 2: Errors, respective orders of convergence, and time in seconds, corresponding to Problem 2 discretized by the QSC method, for several gridsizes $N \times N$. The solution is obtained by 1D-FFTQSC2 and 2D-FFTQSC2.

	on gridpoints (u)		global (v)		time	
N	error	order	error	order	1D-FFTQSC2	2D-FFTQSC2
32	8.6e-08		1.1e-07		0.00270	0.00285
64	5.4e-09	4.00	1.0e-08	3.46	0.01172	0.01054
128	3.4e-10	4.00	1.2e-09	3.12	0.04412	0.05097
256	2.1e-11	4.00	5.3e-11	4.46	0.21107	0.25862
512	1.3e-12	4.00	8.8e-12	2.58	1.12897	1.31450

to u, while the "global" error to v. The advantage of the 1D over the 2D FFT solver is less apparent in the case of systems than in the case of single PDEs. We note that solving the septa-diagonal matrix B in Step 2 of the 1D-FFTQSC2 algorithm gives rise to a relatively large factor (approximately 44, counting real single flops) for the complexity of Step 2, which is not taken into account in the asymptotic performance. By linear least squares fit of the timing data, we obtain $time \approx 1.4 \times 10^{-6} N^{2.16}$ for 1D-FFTQSC2, and $time \approx 1.1 \times 10^{-6} N^{2.23}$ for 2D-FFTQSC2. The difference in the exponents 2.16 and 2.23 for the case of systems of PDEs is more significant than the difference in the exponents 2.25 and 2.30 for the case of single PDEs, but its effect will be felt for larger gridsizes, as the factors 1.4 and 1.1 suggest.

The relatively large factor in the solution of B may also be the reason for the large ratios between the timings of 1D-FFTQSC2 of Table 2 over the respective ones of 1D-FFTQSC of Table 1. The theory suggests an asymptotic factor of 2, while we experimentally get factors of 5.5 to 3.7, when N ranges between 32 and 512. The experiments suggest that the theoretical asymptotic factors are reached for gridsizes larger than the ones considered in the experiments.

We now present results from numerical experiments to demonstrate the performance of the preconditioners described. We consider general elliptic PDE problems, including some problems more general than the analysis assumes. The implementation of our solution methods can be extended in three aspects: (a) the integration of the preconditioners with acceleration methods, such as GMRES; (b) the application of an additional diagonal scaling preconditioner; and (c) the substitution of the Laplace preconditioner $-\triangle_h$ by the QSC operator H_h arising from the model Helmholtz operator (6) with $a = c = 1$ and

$f = -1$. (A similar substitution can be done for systems of PDEs.)

We briefly give the rationale for these extensions. The experimental study in [9], which considers several acceleration methods, as well as Richardson and MRES methods, shows that acceleration methods such as GMRES are in general faster solution methods than one-step methods, and that their convergence rate with the preconditioners considered is independent of the problem size, as is that of one-step methods. Although we do not provide an analysis of the convergence rate of GMRES in this paper, we present numerical results from the GMRES application, since the results of Theorem 2.2 of [1] are valid for the GMRES method as well (though GMRES may converge faster than ρ in (51) indicates), and since the overall times with GMRES are in general lower than those with either Richardson or MRES. In [9], a diagonal scaling preconditioner applied in a multiplicative way on top of the Laplace preconditioner is considered. Left, right and symmetric-left-right diagonal preconditionings are tested. These variations were motivated by the work in [16]. In [9], it is shown experimentally, that left diagonal preconditioning (DL-preconditioning) is in most cases more effective than no diagonal preconditioning or other forms of diagonal preconditioning applied on top of Laplace preconditioning. In this paper, the results of Table 3 are with the DL preconditioner, since, for Problem 3, this preconditioner requires a few less iterations than $-\triangle_h$. In [9], problems with boundary conditions other than Dirichlet are considered. When the boundary conditions are Neumann or periodic, the QSC matrix arising from the Laplace operator is not uniquely solvable, therefore, $-\triangle_h$ is substituted by H_h defined above. Through numerical experiments, we have found that for Dirichlet conditions the number of iterations with preconditioner H_h is the same as with $-\triangle_h$, and with Neumann or periodic conditions, H_h has convergence rate independent of the problem size. Therefore, Helmholtz preconditioning can be used irrespectively of the boundary conditions. In [9] results from three-dimensional problems are also presented.

In all experiments, the stopping criterion is the relative Euclidean norm of the residual and the tolerance is set to 10^{-8} for step 1 of the QSC method and to 10^{-6} for step 2. The "restart" of GMRES is set to 20. We use [17] for the implementation of the GMRES method.

We consider a test problem with variable coefficients.

Problem 3:

$$(x+y+1)u_{xx} + e^{x-y}u_{yy} + (x+1)u_x + (y-1)u_y - \zeta(xy+1)u = g \text{ in } \Omega \equiv (0,1)\times(0,1)$$
$$u = 0 \text{ on } \partial\Omega$$

The function g is chosen so that the solution u to the problem is $u(x,y) = x^{9/2}(x-1)^2 y^{9/2}(y-1)^2$. The parameter ζ controls the size of the u term. Several values of the parameter ζ were considered. Note that $\zeta = -10$, barely violates the condition $\tilde{f} > -2\pi^2\alpha$, while $\zeta = -15$ clearly violates it, and $\zeta = -50$ violates it further.

Table 3 shows results from the application of the preconditioned GMRES method applied to to the QSC equations arising from the discretization of Problem 3, for $\zeta = -15$ and $\zeta = -50$. Note that the number of iterations were the same for the cases $\zeta = 1$, $\zeta = -10$ and $\zeta = -15$. The 1D-FFTQSC and the 2D-FFTQSC algorithms were used for the solution of the preconditioner at each iteration. The convergence rate of the iterative method is independent of the problem size, for both $\zeta = -15$ and $\zeta = -50$, a fact that indicates the effectiveness of the preconditioner, even for problems that do not meet the (anyway sufficient but not necessary) conditions obtained from the theory. However, the absolute number of iterations is affected by ζ, as expected from theory. The timing results show the time spent per iteration, which includes the solution of the preconditioner by the FFT algorithms and any other computation required by the GMRES method, as well as the *total* time for the solution of the QSC system (both steps), which includes the time to generate the QSC matrix and right-side vector, some preprocessing necessary for the FFT, the solution of the first step of the QSC method, the update of the right-side in the second step, and the solution of the second step of the QSC method.

Table 3: Errors on the gridpoints, respective orders of convergence, number of iterations for convergence of the DL-preconditioned GMRES method and time in seconds, corresponding to Problem 3 discretized by the QSC method, for several gridsizes $N \times N$. The solution of the preconditioner is obtained by 1D-FFTQSC and 2D-FFTQSC.

	$\zeta = -15$								$\zeta = -50$		
	on gridpoints		no. of iter.		time				no. of iter.		time
N	error	order	step 1	step 2	per it.	total	per it.	total	step 1	step 2	total
					1D-FFTQSC		2D-FFTQSC				GE
32	3.1e-08		18	13	0.003	0.18	0.003	0.18	24	20	0.11
64	1.9e-09	4.00	18	13	0.013	0.73	0.013	0.73	26	20	0.95
128	1.2e-10	4.01	18	13	0.060	3.16	0.063	3.24	26	20	10.51
256	7.4e-12	4.01	18	13	0.281	13.90	0.297	14.31	26	20	162.28
512	4.6e-13	4.01	18	13	1.315	61.58	1.456	65.79	26	20	

By comparing the times of Table 1 with the per iteration times of Table 3, we note that the FFT solver is a small part (about 25%) of the GMRES iteration. From the data in Table 3, we also infer that the computation associated with the discretization of the problem (matrix and right-side vector generation) is a significant part (about 33%) of the overall computation. This is, of course, a "desirable" effect of a fast solution method. For comparison, we show the time taken by banded Gauss elimination.

Next, we consider a system of PDEs with cross-derivative terms. This system of PDEs arises in stress-analysis problems.

Problem 4:

$$\begin{array}{l} \nabla^2 u + \frac{1}{1-2\nu}(u_{xx} + v_{xy}) = g_1 \\ \nabla^2 v + \frac{1}{1-2\nu}(v_{yy} + u_{xy}) = g_2 \end{array} \quad \text{in } \Omega \equiv (0,1) \times (0,1), \quad \begin{array}{l} u = 0 \\ v = 0 \end{array} \quad \text{on } \partial\Omega.$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y) = (x^2 - x)(y^2 - y)e^{x+y}$ and $v(x,y) = x^{9/2}(x-1)^2 y^{9/2}(y-1)^2$. Although this problem has constant coefficients, it cannot be solved directly by the FFT solvers, because of the cross-derivative term. The parameter ν controls the size of this term. The physically acceptable values of ν are in (0, 0.5). The larger the ν, the more ill-conditioned the arising linear system.

Table 4 shows results from the application of the preconditioned GMRES method applied to to the QSC equations arising from the discretization of Problem 4, for $\nu = 0.25$ and $\nu = 0.35$. For this experiment, we chose $\xi_1 = 1$ and $\xi_2 = 0$.

For $\nu = 0.25$, that is, for relatively small cross-derivative term, the convergence rate of the iterative method is independent of the problem size, while for $\nu = 0.35$, it is slightly affected by the problem size. The number of iterations is dependent on the conditioning of the system, as expected. Even for ill-conditioned problems, though, the effectiveness of the preconditioned iterative solver over Gauss elimination is apparent. Note that the bandwidth of the QSC linear system arising from a 2×2 system of PDEs (with the alternating ordering [22]) is $2(N+2)+3$, so the memory and time requirements of Gauss elimination are 4 and 8 times as large, respectively, for a system of PDEs than for a single PDE.

The ratios of the respective per iteration timings of Tables 4 and 3 are around 2.3 to 2.5. In the iterative solution of the QSC equations by the GMRES method, a significant part of each iteration is the matrix-vector multiplication; another part is the solution of the preconditioner; and the rest are vector operations and the solution of a small least squares problem. The matrix-vector multiplication time for a system of

Table 4: Errors on the gridpoints, respective orders of convergence, number of iterations for convergence of the $-\hat{\triangle}_h$-preconditioned GMRES method with $(\xi_1, \xi_2) = (1, 0)$, and time in seconds, corresponding to Problem 4 discretized by the QSC method, for several gridsizes $N \times N$. The solution of the preconditioner is obtained by two applications of 1D-FFTQSC or of 2D-FFTQSC.

	$\nu = .25$								$\nu = .35$		
	on gridpoints (u)		no. of iter.		time				no. of iter.		time
N	error	order	step 1	step 2	per it.	total	per it.	total	step 1	step 2	total
					1D-FFTQSC2		2D-FFTQSC2				GE
32	1.8e-07		21	16	0.007	0.63	0.007	0.65	32	25	0.69
64	1.1e-08	3.97	22	17	0.033	2.55	0.033	2.62	34	28	5.83
128	7.8e-10	3.87	22	17	0.167	11.32	0.167	11.68	35	29	74.09
256	7.3e-11	3.43	22	17	0.710	46.67	0.718	48.59	36	29	
512	6.7e-12	3.44	22	17	3.310	205.12	3.482	218.39	36	30	

two PDEs over the respective time for a single PDE gives rise to a ratio of 4. The respective ratio for the solution of the preconditioner is 2, and for the vector operations 2 as well. Therefore, the experimentally obtained global ratios of about 2.3 to 2.5 are within the expected range.

Finally, we present some results to show the effect of ξ_1 and ξ_2 to the number of iterations required for convergence. We consider two systems of PDEs, namely Problems 5 and 6. In each problem, each of the four operators $\mathbf{L}_{ij}$, $i = 1, 2$, $j = 1, 2$, is of the form (58), with a_{ij}, c_{ij} and f_{ij} defined below.

Problem 5:

$$a_{11} = 5 + 2e^{xy},\ c_{11} = 6 + e^{-xy}, \quad a_{12} = 1 + 0.5(x^2 + y^2),\ c_{12} = 1 + 0.5(x^2 + 2y^2),$$
$$f_{11} = x + y, \qquad f_{12} = -1/(1 + x + y)$$
$$a_{21} = 0.3e^{x+y},\ c_{21} = 0.3e^{x-y}, \quad a_{22} = 7 + 2\sin x \sin y,\ c_{22} = 5 + \cos x \cos y,$$
$$f_{21} = -x^3y^3, \qquad f_{22} = 1/(1 + xy)$$

Problem 6:

$$a_{11} = 3 + e^{xy},\ c_{11} = 4 + e^{-xy}, \quad a_{12} = 3 + 0.5(x^2 + y^2),\ c_{12} = 3 + 0.5(x^2 + 2y^2),$$
$$f_{11} = x + y, \qquad f_{12} = -1/(1 + x + y)$$
$$a_{21} = 0.5e^{x+y},\ c_{21} = 0.5e^{x-y}, \quad a_{22} = 4 + \sin x \sin y,\ c_{22} = 3 + \cos x \cos y,$$
$$f_{21} = -x^3y^3, \qquad f_{22} = 1/(1 + xy)$$

For both problems, the domain is the unit square, the boundary conditions are homogeneous Dirichlet, and the functions g_1 and g_2 are chosen so that the solution to the problems is $u(x, y) = (x^2 - x)(y^2 - y)e^{x+y}$ and $v(x, y) = x^{9/2}(x - 1)^2y^{9/2}(y - 1)^2$.

Table 5 shows the number of iterations required for convergence of GMRES with the preconditioners arising from the indicated values of ξ_1 and ξ_2, for Problems 5 and 6. On Problem 5, the condition $\min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\} > 0$ is satisfied and the computed ratio ξ_1/ξ_2 according to Theorem 8 is 5.28. From the experiments, it is verified that the least number of iterations is obtained when ξ_1/ξ_2 is 5. (Values close to 5 also gave the same number of iterations.) For Problem 5, $\kappa = 0.15$ and $\kappa' = 0.10$ (see Remark 11), so it is not surprising that the preconditioner with $(\xi_1, \xi_2) = (1, 0)$ requires more iterations than the one with $(\xi_1, \xi_2) = (5, 1)$. We also calculated $\kappa'' = 0.11$ and $\xi_1/\xi_2 = 24.69$ according to Remark 14. Neither the bound nor the iterations are any better than Theorem 8. On Problem 6, the condition $\min_{i=1,2}\{\mathrm{A}_{ii} - \Theta\} > 0$ is not satisfied, therefore, formally speaking Theorems 7 and 8 are not applicable.

Table 5: Number of iterations for convergence of the $-\hat{\triangle}_h$-preconditioned GMRES method with (ξ_1, ξ_2) as shown, corresponding to Problems 5 and 6 discretized by the QSC method, for several gridsizes $N \times N$.

		Problem 5										Problem 6							
	(ξ_1, ξ_2)	(1,0)		(2,1)		(3,1)		(5,1)		(25,1)		(1,0)		(2,1)		(3,1)		(5,1)	
N	step	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2		
32		12	9	15	11	12	9	11	8	11	8	21	16	16	12	17	13	19	14
64		12	9	15	11	12	9	11	8	11	8	21	16	16	12	17	13	19	14
128		12	9	15	11	12	9	11	8	11	9	22	16	16	12	18	13	19	14
256		12	9	15	11	12	9	11	8	11	9	22	16	16	12	18	13	19	14

However, we calculated the ratio ξ_1/ξ_2 according to Theorem 8 and it was found to be 2.11. From the experiments, it is verified that the least number of iterations is obtained when ξ_1/ξ_2 is 2. The ratio ξ_1/ξ_2 according to Remark 14 is 5.9, which gives the same number of iterations as $(\xi_1, \xi_2) = (5, 1)$, clearly more than $(\xi_1, \xi_2) = (2, 1)$. In all cases, the convergence rate remains independent of the problem size, except for very minor deviations.

As a final remark, we note that the algorithms presented here have a high degree of parallelism (as far as the computation is considered). As noted in [6], though, on massively parallel distributed memory machines, they are not very scalable due to the high communication costs arising from the data transposition operations, which are used to implement the patterns of data access in the algorithms. Still, satisfactory parallel efficiencies can be obtained for large gridsizes and reasonable number of processors [6].

References

[1] B. Bialecki. Preconditioned Richardson and minimal residual iterative methods for piecewise Hermite bicubic orthogonal spline collocations equations. *SIAM J. Sci. Comput.*, 15(3):668–680, 1994.

[2] B. Bialecki and X. C. Cai. H^1-norm error bounds for piecewise Hermite bicubic orthogonal spline collocations schemes for elliptic boundary value problems. *SIAM J. Numer. Anal.*, 31(4):1128–1146, 1994.

[3] B. Bialecki, G. Fairweather, and K. Bennett. Fast direct solvers for piecewise Hermite bicubic orthogonal spline collocation equations. *SIAM J. Numer. Anal.*, 29(1):156–173, 1992.

[4] C. C. Christara. Quadratic spline collocation methods for elliptic partial differential equations. *BIT*, 34(1):33–61, 1994.

[5] C. C. Christara. Parallel solvers for spline collocation equations. *Advances in Engineering Software*, 27:71–89, 1996.

[6] C. C. Christara, X. Ding, and K. R. Jackson. An efficient transposition algorithm for distributed memory computers. In A. Pollard, D. J. K. Mewhort, and D. F. Weaver, editors, *High Performance Computing Systems and Applications*, pages 349–368. Kluwer Academic Publishers, 1999.

[7] C. C. Christara and K. S. Ng. Quadratic spline collocation methods for systems of PDEs. DCS Tech. Rep. 318, University of Toronto, Toronto, Ontario, Canada, 24 pgs., July 2001, `ftp://ftp.cs.utoronto.ca/na/Reports/ccc/sys.ps.Z`.

[8] C. C. Christara and B. Smith. Multigrid and multilevel methods for quadratic spline collocation. *BIT*, 37(4):781–803, 1997.

[9] A. Constas. Fast Fourier transform solvers for quadratic spline collocation. M.Sc. Thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 64 pgs., July 1996, `ftp://ftp.cs.utoronto.ca/na/Reports/ccc/constas-96-msc.ps.Z`.

[10] K. D. Cooper and P. M. Prenter. Alternating direction collocation for separable elliptic partial differential equations. *SIAM J. Numer. Anal.*, 28(3):711–727, 1991.

[11] C. de Boor and B. Swartz. Collocation at Gaussian points. *SIAM J. Numer. Anal.*, 10(4):582–606, 1973.

[12] J. Douglas and T. Dupont. Collocation methods for parabolic equations in a single-space variable. *Lecture Notes in Mathematics*, 385:1–147, 1974.

[13] W. R. Dyksen. *Tensor product generalized alternating direction implicit methods for solving separable linear elliptic partial differential equations*. PhD thesis, Department of Computer Science, Purdue University, IN, U.S.A., 1982.

[14] W. R. Dyksen. Tensor product generalized ADI methods for separable elliptic problems. *SIAM J. Numer. Anal.*, 24:59–76, 1987.

[15] Personal communication of G. Fairweather with the first author, April 2000.

[16] A. Greenbaum. Diagonal scalings of the Laplacian as preconditioners for other elliptic differential operators. *SIAM J. Matrix Anal. Appl.*, 13:826–846, 1992.

[17] W. Gropp and B. Smith. *Users Manual for KSP: Data-Structure-Neutral Codes Implementing Krylov Space Methods*. Argonne National Laboratory, August 1993.

[18] E. N. Houstis, J. R. Rice, C. C. Christara, and E. A. Vavalis. Performance of scientific software. *The IMA Volumes in Mathematics and its applications*, 14:123–156, 1988.

[19] E. N. Houstis, E. A. Vavalis, and J. R. Rice. Convergence of an $O(h^4)$ cubic spline collocation method for elliptic partial differential equations. *SIAM J. Numer. Anal.*, 25(1):54–74, 1988.

[20] C. Van Loan. *Computational Frameworks for the Fast Fourier Transform*. Society for Industrial and Applied Mathematics, 1992.

[21] B. Mossberg. *Documentation for the FFT subroutines available in four libraries*. NEC Systems Laboratory, Inc., August 1992.

[22] K. S. Ng. Quadratic spline collocation methods for systems of elliptic PDEs. M.Sc. Thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 67 pgs., April 2000, `ftp://ftp.cs.utoronto.ca/na/Reports/ccc/ngkit-00-msc.ps.Z`.

[23] P. Percell and M. F. Wheeler. A C^1 finite element collocation method for elliptic equations. *SIAM J. Numer. Anal.*, 17(5):605–622, 1980.

[24] P. M. Prenter and R. D. Russell. Orthogonal collocation for elliptic partial differential equations. *SIAM J. Numer. Anal.*, 13(6):923–939, 1976.