KYUNGPOOK Math. J. 50(2010), 419-431

Preconditioning Cubic Spline Collocation Methods for a Coupled Elliptic Equation

Byeong-Chun Shin
Department of Mathematics, Chonnam National University, Gwangju 500-757, Korea
e-mail : bcshin@jnu.ac.kr

Sang Dong Kim*
Department of Mathematics, Kyungpook National University, Daegu 702-701, Korea
e-mail : skim@knu.ac.kr

Abstract. A low-order finite element preconditioner is analyzed for a cubic spline collocation method which is used for discretizations of coupled elliptic problems derived from an optimal control problrm subject to an elliptic equation. Some numerical evidences are also provided.

1. Introduction

It is known that the problem to find the state variable u and the control function θ for a given target state $\hat{u}$ which minimizes the following functional

$$\mathcal{J}(u,\theta) = \frac{1}{2}\int_\Omega |u-\hat{u}|^2\, dx + \frac{\delta}{2}\int_\Omega |\theta|^2\, dx, \tag{1.1}$$

subject to

$$\begin{aligned} -\Delta u + u = \theta \quad &\text{in} \quad \Omega := (-1,1)^2, \\ u = 0 \quad &\text{on} \quad \partial\Omega, \end{aligned} \tag{1.2}$$

is called an optimal control problem (see [1], [2] and [6] for example). The constant δ $(0 < \delta \leq 1)$ in (1.1) is called a penalty parameter which will be used to change a relative importance of two terms in (1.1). Using the Lagrange multiplier rule, one may have the coupled optimality system of two elliptic type equations for state and

* Corresponding Author.

This Research was supported by Kyungpook National University Research Fund, 2009.
Received September 13, 2010; accepted September 17, 2010.
2000 Mathematics Subject Classification: 65M55, 65N30.
Key words and phrases: Coupled elliptic equations, Cubic spline collocation methods, Finite element preconditioner.

adjoint variables (see [6]):

$$\begin{cases} -\Delta u + u + \dfrac{1}{\delta} v = 0 & \text{in} \quad \Omega, \\ -\Delta v + v - u = -\hat{u} & \text{in} \quad \Omega, \\ u = v = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{1.3}$$

Here, the following optimality condition is used to get the coupled system of only state and adjoint variables:

$$-\frac{1}{\delta} v = \theta. \tag{1.4}$$

There are many ways to discretize (1.3). One of them is to use a C^1- interpolatory cubic spline for discretizing (1.3) (see [10], [12] and [13] for example). Such a discretization yields a linear system whose eigenvalue is increasing as the size of matrix becomes large. Hence, the goal of this paper is to provide a finite element preconditioner (see [8], [9], [11] and [12] for example) so that the resulting preconditioned systems $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$ (see (3.17)) with a finite element preconditioner $\mathcal{B}_h$ has eigenvalues which are independent of the size of the linear system. In this respect, we will show that the real parts of eigenvalues of $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$ are positive, uniformly bounded away from zero and the absolute values of eigenvalues are uniformly bounded whose bounds are only dependent on penalty parameter δ in (1.4).

This paper is organized as follows. In section 2, cubic spline collocation methods for coupled elliptic equations (1.2) is presented. In section 3, the finite element preconditioner is constructed using the symmetric part of (1.2). The validity of the finite element preconditioner is analyzed in terms of eigenvalues. Some numerical evidences for developed theory are provided in section 4. Finally, a conclusion is given in section 5.

2. Cubic spline collocation method

In this section we review the C^1 interpolatory cubic spline generated by Hermite cubic splines defined on the unit interval $I := [0, 1]$. Let N be a positive integer and $h = \frac{1}{N}$. With the knots $t_k = kh$, $k = 0, \cdots, N$, the Hermite cubic spline space $S_{h,3}$ on I is defined by the set of $C^1(I)$ functions whose restrictions on $I_k = (t_{k-1}, t_k)$ are cubic polynomials.

The local Legendre-Gauss([LG]) points $\{\xi_i\}_{i=0}^{2N+1}$ are given by

$$\xi_0 = 0, \quad \xi_{2i-1} = t_{i-1} + h\eta_1, \quad \xi_{2i} = t_{i-1} + h\eta_2, \quad \xi_{2N+1} = 1 \tag{2.1}$$

where $\eta_1 = \frac{1}{2}\Big(1 - \frac{1}{\sqrt{3}}\Big)$ and $\eta_2 = \frac{1}{2}\Big(1 + \frac{1}{\sqrt{3}}\Big)$.

Let $S_{h,3}^0$ be a subspace of $S_{h,3}$ whose functions vanish on the boundary. Let $\{\, \phi_i \,\}_{i=1}^{2N}$ be the C^1 Lagrange interpolatory basis for the space $S_{h,3}^0$ satisfying

$$\phi_i(\xi_k) = \delta_{i,k}, \quad 1 \le k, i \le 2N. \tag{2.2}$$

For the finite element preconditioner, let us denote by V_h^0 the space of continuous piecewise linear functions which break at the collocation points and vanish at end points. Denote by $\{\psi_i\}_{i=1}^{2N}$ the nodal basis for the space V_h^0.

For two dimensional function spaces, let $\mathbf{S}_h^0 := S_{h,3}^0 \otimes S_{h,3}^0$ and $\mathbf{V}_h^0 := V_h^0 \otimes V_h^0$ be tensor product function spaces of one-dimensional function spaces, respectively, and denote by $\boldsymbol{\mathcal{S}}_h^0 := \mathbf{S}_h^0 \times \mathbf{S}_h^0$ and $\boldsymbol{\mathcal{V}}_h^0 := \mathbf{V}_h^0 \times \mathbf{V}_h^0$. The collocation points are given by

$$\boldsymbol{\xi}_{i,j} = (\xi_i, \xi_j) \quad \text{for } \ i, j, = 1, \cdots, 2N.$$

Denote by the basis functions for $\mathbf{S}_h^0$ and $\mathbf{V}_h^0$

$$\Phi_{i,j}(x,y) := \phi_i(x)\phi_j(y) \quad \text{and} \quad \Psi_{i,j}(x,y) := \psi_i(x)\psi_j(y),$$

respectively.

We use the standard Sobolev spaces $H^1(\Omega)$ and $H_0^1(\Omega)$ with the usual Sobolev $H^1(\Omega)$-norm $\|\cdot\|_1$ and $H^1(\Omega)$-seminorm $|\cdot|_1$. Denote by $(\cdot,\cdot)$ and $\|\cdot\|$ be the usual L^2 inner product and L^2 norm, respectively. Define a discrete inner product $\langle\cdot,\cdot\rangle_N$ over the space $\mathbf{S}_h^0$ as

$$\langle u, v\rangle_N = \frac{h^2}{4}\sum_{i,j=1}^{2N} u(\xi_i,\xi_j)\, v(\xi_i,\xi_j).$$

For complex functions $u = p + iq$ and $v = r + is$, we use the same notation for the complex inner product and discrete inner product such as

$$(u,v) := (p+iq, r-is) \quad \text{and} \quad \langle u, v\rangle_N := \langle p+iq, r-is\rangle_N.$$

For matrix functions U and V, define

$$(U,V) = \sum_{k=1}^{4} (u_k, v_k), \quad \|U\|^2 := \sum_{k=1}^{4} \|u_k\|^2 \ \text{ where } U = \begin{bmatrix} u_1 & u_2 \\ u_3 & u_4 \end{bmatrix}, V = \begin{bmatrix} v_1 & v_2 \\ v_3 & v_4 \end{bmatrix}.$$

With a vector function $\mathbf{u} = [u, v]^T$, the optimality system given in (1.3) can be represented by

$$\mathcal{A}\,\mathbf{u} := -\mathsf{A}\,\Delta\mathbf{u} + (\mathsf{A} + \mathsf{C})\,\mathbf{u} = \mathbf{f} \quad \text{in} \quad \Omega, \tag{2.3}$$

where

$$\mathsf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\delta} \end{bmatrix}, \quad \mathsf{C} = \begin{bmatrix} 0 & \frac{1}{\delta} \\ -\frac{1}{\delta} & 0 \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} 0 \\ -\frac{1}{\delta}\hat{u} \end{bmatrix},$$

with the zero boundary condition $\mathbf{u} = \mathbf{0}$ on $\partial\Omega$. The differential operator $\mathcal{A}$ can be represented by

$$\mathcal{A}\,\mathbf{u} = \begin{bmatrix} \mathcal{A}^1\,\mathbf{u} \\ \mathcal{A}^2\,\mathbf{u} \end{bmatrix} := \begin{bmatrix} -\Delta u + u + \frac{1}{\delta}v \\ -\frac{1}{\delta}u - \frac{1}{\delta}\Delta v + \frac{1}{\delta}v \end{bmatrix}$$

Then, the C^1 cubic spline collocation method introduced in [12] is to find $\mathbf{u}_h = [u_1, u_2]^T \in \boldsymbol{\mathcal{S}}_h^0$ satisfying

$$\mathcal{A}\,\mathbf{u}_h(\xi_i, \xi_j) = \mathbf{f}(\xi_i, \xi_j), \quad 1 \le i, j \le 2N. \tag{2.4}$$

For one dimensional arrangement, we denote by, for $i, j = 1, 2, \cdots, 2N$,

$$\boldsymbol{\xi}_\mu := \boldsymbol{\xi}_{i,j}, \quad \Phi_\mu := \Phi_{i,j} \quad \text{and} \quad \Psi_\mu := \Psi_{i,j} \quad \text{with} \quad \mu = 2N(i-1) + j.$$

Denote by $N_h = (2N)^2$ the number of the interior collocation points. Let

$$\mathbf{R}_h(\mu, \nu) = (-\Delta\Phi_\nu + \Phi_\nu)(\boldsymbol{\xi}_\mu) \quad \text{for} \quad \mu, \nu = 1, 2, \cdots, N_h$$

and let I_{N_h} be the $N_h \times N_h$ identity matrix.

The algebraic linear system induced by the cubic spline collocation method (2.4) is given by

$$\mathcal{A}_h\,\mathbf{U} = \mathbf{F} \tag{2.5}$$

where

$$\mathcal{A}_h = \begin{bmatrix} \mathbf{R}_h & \frac{1}{\delta} I_{N_h} \\ -\frac{1}{\delta} I_{N_h} & \frac{1}{\delta}\mathbf{R}_h \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{u}_h(\boldsymbol{\xi}_\mu) \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \quad \text{and} \quad \mathbf{F} = \begin{bmatrix} \mathbf{f}(\boldsymbol{\xi}_\mu) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ F \end{bmatrix}$$

with

$$U_1 = \begin{bmatrix} u_1(\boldsymbol{\xi}_1) \\ \vdots \\ u_1(\boldsymbol{\xi}_{N_h}) \end{bmatrix}, \quad U_2 = \begin{bmatrix} u_2(\boldsymbol{\xi}_1) \\ \vdots \\ u_2(\boldsymbol{\xi}_{N_h}) \end{bmatrix}, \quad F = -\frac{1}{\delta} \begin{bmatrix} \hat{u}(\boldsymbol{\xi}_1) \\ \vdots \\ \hat{u}(\boldsymbol{\xi}_{N_h}) \end{bmatrix}.$$

3. Finite element preconditioner

Consider a simple decoupled elliptic operator $\mathcal{B}$ to investigate a finite element preconditioner:

$$\mathcal{B}\,\mathbf{u} := -\mathsf{A}\,\Delta\mathbf{u} + \mathsf{A}\,\mathbf{u} \quad \text{in} \quad \Omega \tag{3.1}$$

with the zero boundary condition. Let $\mathbf{P}_h$ be the finite element discretization in the space $\mathbf{V}_h^0$ given by

$$\mathbf{P}_h(\mu, \nu) = (\nabla\Psi_\mu, \nabla\Psi_\nu) + (\Psi_\mu, \Psi_\nu) \quad \text{for} \quad \mu, \nu = 1, 2, \cdots, N_h.$$

Then the finite element discretization for the problem (3.1) in the space $\boldsymbol{\mathcal{V}}_h^0$ is given by

$$\mathcal{B}_h := \begin{bmatrix} \mathbf{P}_h & 0 \\ 0 & \frac{1}{\delta}\mathbf{P}_h \end{bmatrix}. \tag{3.2}$$

In this section, we will show that the finite element discretization $\mathcal{B}_h$ is an optimal preconditioner for the matrix $\mathcal{A}_h$ induced by the cubic spline collocation method.

3.1. Interpolation operators

Let $J_h : V_h^0 \to S_{h,3}^0$ be the interpolation operator such that $(J_h\,\chi_h)(\xi_i) = \chi_h(\xi_i)$ for $i = 1, 2, \cdots, 2N$ and let $\mathbf{J}_h : \mathbf{V}_h^0 \to \mathbf{S}_h^0$ be the two dimensional interpolation operator such that $(\mathbf{J}_h\, u_h)(\boldsymbol{\xi}_\mu) = u_h(\boldsymbol{\xi}_\mu)$ for $\mu = 1, 2, \cdots, N_h$.

In [12], they showed that there exists a constant $C > 0$ such that

$$\frac{1}{C}\|\chi_h\| \le \|J_h\,\chi_h\| \le C\|\chi_h\|, \quad \frac{1}{C}|\chi_h|_1 \le |J_h\,\chi_h|_1 \le C|\chi_h|_1, \quad \forall\, \chi_h \in V_h^0, \tag{3.3}$$

and

$$\frac{1}{C}\|u_h\| \le \|\mathbf{J}_h\, u_h\| \le C\|u_h\|, \quad \frac{1}{C}|u_h|_1 \le |\mathbf{J}_h\, u_h|_1 \le C|u_h|_1, \quad \forall\, u_h \in \mathbf{V}_h^0. \tag{3.4}$$

In this paper we use the generic constant C at many places, which does not depend on mesh size h and N.

Let us define the vector interpolation operator $\boldsymbol{\mathcal{J}}_h : \boldsymbol{\mathcal{V}}_h^0 \to \boldsymbol{\mathcal{S}}_h^0$ such that, for $\mathbf{u}_h := [u_h, v_h]^T \in \boldsymbol{\mathcal{V}}_h^0$,

$$(\boldsymbol{\mathcal{J}}_h \mathbf{u}_h)(\boldsymbol{\xi}_\mu) := \left[(\mathbf{J}_h u_h)(\boldsymbol{\xi}_\mu), (\mathbf{J}_h v_h)(\boldsymbol{\xi}_\mu)\right]^T = \mathbf{u}_h(\boldsymbol{\xi}_\mu).$$

Using the equivalence in (3.4), we have the following theorem.

Theorem 3.1. *For all $\mathbf{u}_h \in \boldsymbol{\mathcal{V}}_h^0$, there exists a positive constant C independent of h such that*

$$\frac{1}{C}\|\mathbf{u}_h\| \le \|\boldsymbol{\mathcal{J}}_h \mathbf{u}_h\| \le C\|\mathbf{u}_h\| \quad \textit{and} \quad \frac{1}{C}\|\mathbf{u}_h\|_1 \le \|\boldsymbol{\mathcal{J}}_h \mathbf{u}_h\|_1 \le C\|\mathbf{u}_h\|_1.$$

3.2. Analysis on $\mathcal{P}_1$ finite element preconditioner

Let us define a bilinear form associating with the operator $\mathcal{A}$:

$$\begin{aligned} a_h(\mathbf{u}_h, \mathbf{v}_h) &:= \left\langle \mathcal{A}\,\mathbf{u}_h, \mathbf{v}_h \right\rangle_N = \left\langle -\mathsf{A}\Delta \mathbf{u}_h, \mathbf{v}_h \right\rangle_N + \left\langle (\mathsf{A}+\mathsf{C})\mathbf{u}_h, \mathbf{v}_h \right\rangle_N \\ &= \left\langle -\Delta u_1 + u_1 + \frac{1}{\delta}u_2, v_1 \right\rangle_N + \left\langle -\frac{1}{\delta}\Delta u_2 + \frac{1}{\delta}u_2 - \frac{1}{\delta}u_1, v_2 \right\rangle_N \end{aligned} \tag{3.5}$$

for $\mathbf{u}_h = [u_1, u_2]^T$, $\mathbf{v}_h = [v_1, v_2]^T \in \boldsymbol{\mathcal{S}}_h^0$. Then the variational problem associating with the collocation problem (2.4) is to find $\mathbf{u}_h = [u_1, u_2]^T \in \boldsymbol{\mathcal{S}}_h^0$ such that

$$a_h(\mathbf{u}_h, \mathbf{v}_h) = \langle \mathbf{f}, \mathbf{v}_h \rangle_N \quad \text{for all } \ \mathbf{v}_h \in \boldsymbol{\mathcal{S}}_h^0. \tag{3.6}$$

We also define a bilinear form associating with the operator $\mathcal{B}$:

$$\beta_h(\mathbf{u}_h, \mathbf{v}_h) = (\mathcal{B}\,\mathbf{u}_h, \mathbf{v}_h) = (\nabla u_1, \nabla v_1) + (u_1, v_1) + \frac{1}{\delta}\,(\nabla u_2, \nabla v_2) + \frac{1}{\delta}\,(u_2, v_2) \tag{3.7}$$

for $\mathbf{u}_h = [u_1, u_2]^T$, $\mathbf{v}_h = [v_1, v_2]^T \in \boldsymbol{\mathcal{V}}_h^0$. Note that the bilinear form $\beta_h(\cdot,\cdot)$ is symmetric but the bilinear form $a_h(\cdot,\cdot)$ is not symmetric.

Lemma 3.2. *It holds that*

$$\langle -A\Delta \mathbf{u}_h, \mathbf{v}_h\rangle_N = \langle \mathbf{u}_h, -A\Delta \mathbf{v}_h\rangle_N \quad \textit{for all} \quad \mathbf{u}_h, \mathbf{v}_h \in \boldsymbol{\mathcal{S}}_h^0. \tag{3.8}$$

Proof. It is enough to show that

$$\langle -\Delta u, v\rangle_N = \langle u, -\Delta v\rangle_N \quad \text{for all} \quad u, v \in \mathbf{S}_h^0.$$

Let us recall Lemma 3.1 in [3] or [12] such that, for $f, g \in S_{h,3}^0$,

$$\langle f, -g''\rangle_{N,1} := -\frac{h}{2}\sum_{i=1}^{2N} f(\xi_i)g''(\xi_i) = (f', g') + C\sum_{k=1}^{N} f_k^{(3)} g_k^{(3)} h^5 \tag{3.9}$$

where $f_k^{(3)}$ denotes the third derivative of f on I_k and C is an absolute positive constant. Hence, it follows that the following symmetry for one dimensional case:

$$\langle f, -g''\rangle_{N,1} = \langle g, -f''\rangle_{N,1}.$$

Using the above result and the definition of $\langle\cdot,\cdot\rangle_N$, one may easily show that

$$\begin{aligned}
\big\langle -\Delta u, v\big\rangle_N &= \sum_{i=1}^{2N}\frac{h}{2}\left(\sum_{j=1}^{2N}\frac{h}{2}\Big(-u_{xx}(\xi_i,\xi_j)v_(\xi_i,\xi_j) - u_{yy}(\xi_i,\xi_j)v(\xi_i,\xi_j)\Big)\right)\\
&= \sum_{j=1}^{2N}\frac{h}{2}\big\langle -u_{xx}(x,\xi_j), v(x,\xi_j)\big\rangle_{N,1} + \sum_{i=1}^{2N}\frac{h}{2}\big\langle -u_{yy}(\xi_i,y), v(\xi_i,y)\big\rangle_{N,1}\\
&= \sum_{j=1}^{2N}\frac{h}{2}\big\langle -u(x,\xi_j), v_{xx}(x,\xi_j)\big\rangle_{N,1} + \sum_{i=1}^{2N}\frac{h}{2}\big\langle -u(\xi_i,y), v_{yy}(\xi_i,y)\big\rangle_{N,1}\\
&= \langle u, -\Delta v\rangle_N.
\end{aligned}$$

This completes the proof. □

One may easily see from Lemma 3.2 that the symmetric part of $a_h(\mathbf{u}_h, \mathbf{v}_h)$ is given by

$$a_h^s(\mathbf{u}_h, \mathbf{v}_h) := \big\langle -\mathsf{A}\Delta \mathbf{u}_h, \mathbf{v}_h\big\rangle_N + \big\langle \mathsf{A}\mathbf{u}_h, \mathbf{v}_h\big\rangle_N.$$

Lemma 3.3. *For all $u_h \in \mathbf{S}_h^0$, there exists a positive constant C such that*

$$\frac{1}{C}\,\|u_h\|_1^2 \le \langle -\Delta u_h + u_h, u_h\rangle_N \le C\,\|u_h\|_1^2. \tag{3.10}$$

Proof. From Lemma 3.3 in [3], we have

$$\|\nabla u_h\|^2 \le \langle -\Delta u_h, u_h\rangle_N \le \frac{5}{3}\|\nabla u_h\|^2$$

and from Lemma 4.1 and 4.2 in [12] we have

$$\frac{1}{C}\|u_h\|^2 \le \langle u_h, u_h\rangle_N \le C\|\nabla u_h\|^2.$$

Combining above two inequalities completes the conclusion. □

The following norm equivalence guarantees the existence and uniqueness of the solution in $\boldsymbol{\mathcal{S}}_h^0$ for the variational problem (3.6).

Proposition 3.4. *For any real valued vector function* $\mathbf{u}_h = [u_1, u_2]^T \in \boldsymbol{\mathcal{S}}_h^0$*, there exists a positive constant* C*, independent of* h*, such that*

$$\frac{1}{C}\,\|\mathbf{u}_h\|_{1,\delta}^2 \le a_h(\mathbf{u}_h, \mathbf{u}_h) = a_h^s(\mathbf{u}_h, \mathbf{u}_h) \le C\,\|\mathbf{u}_h\|_{1,\delta}^2, \tag{3.11}$$

where the norm $\|\mathbf{u}_h\|_{1,\delta}^2$ *is defined by*

$$\|\mathbf{u}_h\|_{1,\delta}^2 := \|u_1\|_1^2 + \frac{1}{\delta}\|u_2\|_1^2.$$

Proof. For $\mathbf{u}_h = [u_1, u_2]^T \in \boldsymbol{\mathcal{S}}_h^0$, one may easily show that $\langle \mathsf{C}\mathbf{u}_h, \mathbf{u}_h\rangle_N = 0$. Hence we have

$$\begin{aligned} a_h(\mathbf{u}_h, \mathbf{u}_h) &= \big\langle -\mathsf{A}\Delta\mathbf{u}_h, \mathbf{u}_h\big\rangle_N + \big\langle \mathsf{A}\mathbf{u}_h, \mathbf{u}_h\big\rangle_N \\ &= \big\langle -\Delta u_1 + u_1, u_1\big\rangle_N + \big\langle -\frac{1}{\delta}\Delta u_2 + \frac{1}{\delta}u_2, u_2\big\rangle_N. \end{aligned}$$

Using Lemma 3.3 yields the conclusion. □

Lemma 3.5. *Let* $\mathbf{u}_h = [u_1, u_2]^T$ *be a complex vector valued function with* $u_1 = p_1 + iq_1$ *and* $u_2 = p_2 + iq_2$ *where* p_1*,* q_1*,* p_2*,* $q_2 \in \mathbf{S}_h^0$*. It follows that*

$$\mathrm{Re}\big(a_h(\mathbf{u}_h, \mathbf{u}_h)\big) = \langle -A\Delta\mathbf{u}_h, \mathbf{u}_h\rangle_N + \langle A\mathbf{u}_h, \mathbf{u}_h\rangle_N = a_h^s(\mathbf{u}_h, \mathbf{u}_h), \tag{3.12}$$

and there exists a positive constant C*, independent of* h*, such that*

$$\frac{1}{C}\,\|\mathbf{u}_h\|_{1,\delta}^2 \le \mathrm{Re}\big(a_h(\mathbf{u}_h, \mathbf{u}_h)\big) = a_h^s(\mathbf{u}_h, \mathbf{u}_h) \le C\,\|\mathbf{u}_h\|_{1,\delta}^2. \tag{3.13}$$

Proof. Since the symmetric part of $a_h(\mathbf{u}_h, \mathbf{v}_h)$ is $\big\langle -\mathsf{A}\Delta\mathbf{u}_h, \mathbf{v}_h\big\rangle_N + \big\langle \mathsf{A}\mathbf{u}_h, \mathbf{v}_h\big\rangle_N$, one may easily show that $\big\langle -\mathsf{A}\Delta\mathbf{u}_h, \mathbf{u}_h\big\rangle_N + \big\langle \mathsf{A}\mathbf{u}_h, \mathbf{u}_h\big\rangle_N$ is a positive real number. The following estimation

$$\langle \mathsf{C}\mathbf{u}_h, \mathbf{u}_h\rangle_N = \frac{1}{\delta}\Big(\langle u_2, u_1\rangle_N - \langle u_1, u_2\rangle_N\Big) = \frac{2i}{\delta}\Big(\langle p_1, q_2\rangle_N - \langle p_2, q_1\rangle_N\Big) \tag{3.14}$$

shows that $\langle \mathsf{C}\mathbf{u}_h, \mathbf{u}_h\rangle_N$ is a pure imaginary number. This completes (3.12). Now, combining (3.12) with Lemma 3.3, we have the conclusion (3.13). □

With a complex or real vector function $\mathbf{u}_h = [u_1, u_2]^T$ in the space $\boldsymbol{\mathcal{S}}_h^0$, denote by $\mathbf{U}$ the vector containing the nodal values of the functions u_1 and u_2, that is,

$$\mathbf{U} = \left(u_1(\boldsymbol{\xi}_1), \cdots, u_1(\boldsymbol{\xi}_{N_h}), u_2(\boldsymbol{\xi}_1), \cdots, u_2(\boldsymbol{\xi}_{N_h})\right)^T.$$

Then, one may easily check that, for $\mathbf{u}_h, \mathbf{v}_h \in \boldsymbol{\mathcal{S}}_h^0$,

$$a_h(\mathbf{u}_h, \mathbf{v}_h) = \mathbf{V}^H\left(\mathcal{W}_h \mathcal{A}_h\right)\mathbf{U}, \quad \text{with} \quad \mathcal{W}_h = \begin{bmatrix} W_h & 0 \\ 0 & W_h \end{bmatrix}. \tag{3.15}$$

where $\mathcal{A}_h$ is given in (2.5) and $W_h = \operatorname{diag}\left(\frac{h^2}{4}\right)$, and $\mathbf{V}^H$ denotes the conjugate transpose of $\mathbf{V}$.

Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest absolute eigenvalues of a square matrix A, respectively. The spectral radius and the spectrum of a square matrix A are denoted by $\rho(A)$ and $\sigma(A)$, respectively.

The generalized field of values of the matrix pair $\mathcal{W}_h\mathcal{A}_h$ and $\mathcal{B}_h$ is defined as

$$\mathcal{F}(\mathcal{W}_h\mathcal{A}_h, \mathcal{B}_h) := \left\{ \frac{\mathbf{V}^H \mathcal{W}_h \mathcal{A}_h \mathbf{V}}{\mathbf{V}^H \mathcal{B}_h \mathbf{V}} \mid \mathbf{V} \neq \mathbf{0},\ \mathbf{V} \in \mathbb{C}^{2N_h} \right\}.$$

Then it is well-known that

$$\sigma(\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h) \subset \mathcal{F}(\mathcal{W}_h\mathcal{A}_h, \mathcal{B}_h).$$

Theorem 3.6. *There exists a positive constants C, independent of h, such that, for arbitrary eigenvalue λ of the preconditioned matrix $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$,*

$$0 < \frac{1}{C} \leq \operatorname{Re}(\lambda) \leq C \quad \textit{and} \quad |\lambda| \leq C\left(1 + \frac{1}{\delta}\right). \tag{3.16}$$

Proof. Let $\mathbf{V}_\lambda = [V_1, V_2]^T \neq \mathbf{0} \in \mathbb{C}^{2N_h}$ with $V_i = (v_{i1}, v_{i2}, \cdots, v_{iN_h})^T$ for $i = 1, 2$ and let $\mathbf{v}_\lambda = [v_1, v_2]^T \in \boldsymbol{\mathcal{V}}_h^0$ with $v_i = \sum_{\mu=1}^{N_h} v_{i\mu}\Psi_\mu$ for $i = 1, 2$. From the definitions of the bilinear forms, we have

$$\mathbf{V}_\lambda^H\left(\mathcal{W}_h\mathcal{A}_h\right)\mathbf{V}_\lambda = a_h(\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda, \boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda) \quad \text{and} \quad \mathbf{V}_\lambda^H\mathcal{B}_h\mathbf{V}_\lambda = \beta_h(\mathbf{v}_\lambda, \mathbf{v}_\lambda)$$

so that

$$w_\lambda := \frac{\mathbf{V}_\lambda^H\left(\mathcal{W}_h\mathcal{A}_h\right)\mathbf{V}_\lambda}{\mathbf{V}_\lambda^H\mathcal{B}_h\mathbf{V}_\lambda} = \frac{a_h(\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda, \boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda)}{\beta_h(\mathbf{v}_\lambda, \mathbf{v}_\lambda)}.$$

Note from (3.7) that there exists a positive constant C such that

$$\frac{1}{C}\left\|\mathbf{v}_\lambda\right\|_{1,\delta}^2 \leq \beta_h(\mathbf{v}_\lambda, \mathbf{v}_\lambda) \leq C\left\|\mathbf{v}_\lambda\right\|_{1,\delta}^2.$$

By Theorem 3.1, Lemma 3.5 and (3.7), there exists a positive constant C, independent of h and δ, such that

$$\mathrm{Re}(w_\lambda) = \frac{\mathrm{Re}\big(a_h(\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda, \boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda)\big)}{\beta_h(\mathbf{v}_\lambda, \mathbf{v}_\lambda)} \le C\,\frac{\|\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda\|_{1,\delta}^2}{\|\mathbf{v}_\lambda\|_{1,\delta}^2} \le C$$

and

$$\mathrm{Re}(w_\lambda) \ge \frac{1}{C}\,\frac{\|\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda\|_{1,\delta}^2}{\|\mathbf{v}_\lambda\|_{1,\delta}^2} \ge \frac{1}{C}.$$

On the other hand, we obtain from (3.14) that

$$|\mathrm{Im}(w_\lambda)| = \frac{|\langle \mathsf{C}\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda, \boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda\rangle_N|}{\beta_h(\mathbf{v}_\lambda, \mathbf{v}_\lambda)} \le C\,\frac{1}{\delta}\,\frac{\|\boldsymbol{\mathcal{J}}_h\mathbf{v}_\lambda\|^2}{\|\mathbf{v}_\lambda\|_1^2} \le C\,\frac{1}{\delta}.$$

Thus we have

$$|w_\lambda| \le |\mathrm{Re}(w_\lambda)| + |\mathrm{Im}(w_\lambda)| \le C\left(1 + \frac{1}{\delta}\right).$$

From the fact that

$$\sigma(\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h) \subset \mathcal{F}(\mathcal{W}_h\mathcal{A}_h, \mathcal{B}_h),$$

we have the conclusion. □

Consider the following preconditioned system

$$\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h\ \mathbf{U} = \mathcal{B}_h^{-1}\mathcal{W}_h\mathbf{F}, \tag{3.17}$$

where the preconditioned matrix can be rewritten by

$$\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h = \begin{bmatrix} \mathbf{P}_h^{-1}W_h\mathbf{R}_h & 0 \\ 0 & \mathbf{P}_h^{-1}W_h\mathbf{R}_h \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{\delta}\mathbf{P}_h^{-1}W_h \\ -\mathbf{P}_h^{-1}W_h & 0 \end{bmatrix}.$$

From Theorem 3.6 one may find a constant $\theta > 0$ so that

$$\rho(I - \theta\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h) < 1.$$

Hence, the theorem guarantees the convergence of the following damped Jacobi iterative method for the preconditioned system:

$$\mathbf{U}^{k+1} = (I - \theta\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h)\mathbf{U}^k + \theta\mathcal{B}_h^{-1}\mathcal{W}_h\mathbf{F}.$$

One may also use the preconditioned Biconjugate Gradient Method (BiCG) or a generalized minimal residual method (GMRES). See [4], [5], [7], [14] and [15] for more details.

4. Numerical results

In this section, using the developed finite element preconditioners we will test an optimal control problem subject to elliptic differential equation (1.2) with a target function $\hat{u}(x, y) = \sin \pi x \sin \pi y$. Then, we have the following optimality system

$$(3.1) \qquad \begin{cases} -\Delta u + u + \dfrac{1}{\delta} v = 0 & \text{in} \quad \Omega, \\ -\dfrac{1}{\delta}\Delta v + \dfrac{1}{\delta} v - \dfrac{1}{\delta} u = -\dfrac{1}{\delta} \sin \pi x \sin \pi y & \text{in} \quad \Omega, \\ u = v = 0 & \text{on} \quad \partial\Omega. \end{cases}$$

Note that the optimality system (3.1) has the exact solution

$$(3.2) \qquad u = \frac{1}{1 + \delta(1 + 2\pi^2)^2} \sin \pi x \sin \pi y, \qquad v = \frac{-\delta(1 + 2\pi^2)}{1 + \delta(1 + 2\pi^2)^2} \sin \pi x \sin \pi y,$$

and the exact optimal control is given by

$$\theta = -\frac{1}{\delta} v = \frac{(1 + 2\pi^2)}{1 + \delta(1 + 2\pi^2)^2} \sin \pi x \sin \pi y.$$

In order to provide evidences of Theorem 3.6, we first consider the distributions of eigenvalues and condition numbers for the non-preconditioned matrix $\mathcal{W}_h\mathcal{A}_h$ and preconditioned matrix $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$. In FIG 1, one may see from the distributions of eigenvalues that real parts of eigenvalues of $\mathcal{W}_h\mathcal{A}_h$ spread abroad from 5.48e+002 to 1.80e+011 but those of $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$ have very small range from 1.00 to 6.59 where $h = 1/N = 1/16$ and $\delta = 10^{-10}$.

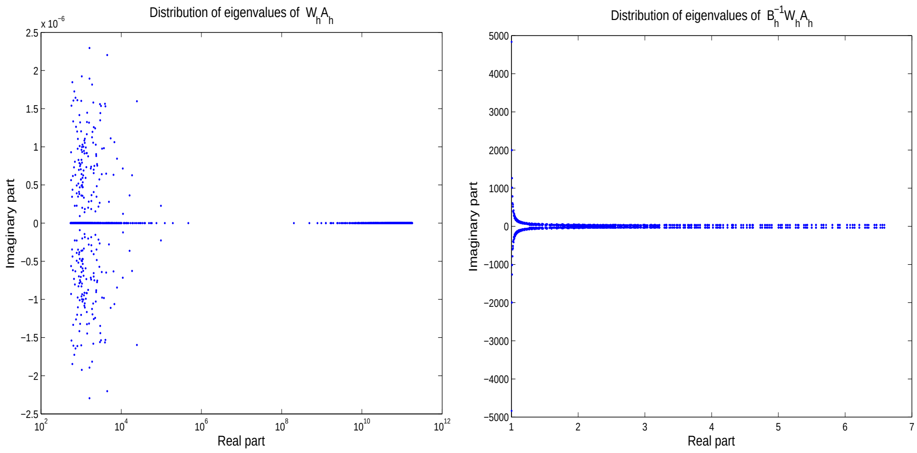

Figure 1: Distributions of eigenvalues for $\mathcal{W}_h\mathcal{A}_h$ and $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$

We also show that the range of real parts of eigenvalues for preconditioned matrix $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$ is independent of the penalty parameter δ in FIG 2. But, FIG 3 shows that the condition numbers of $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$ are slightly increased as δ decreases to 0.

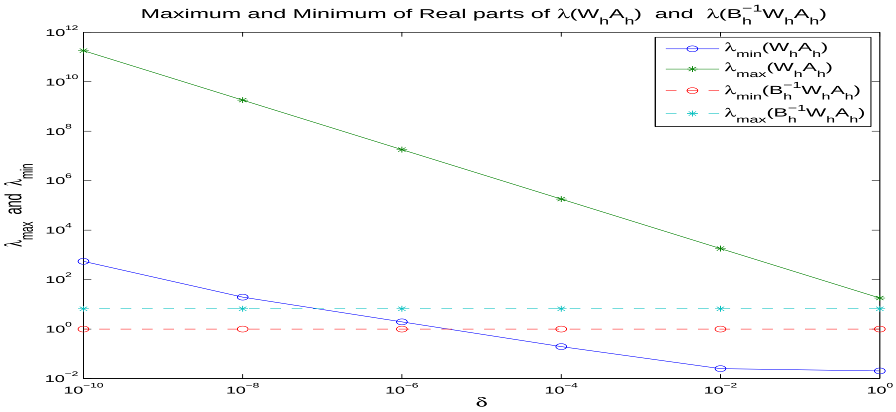

Figure 2: Maximum and minimum of real parts of eigenvalues for $\mathcal{W}_h\mathcal{A}_h$ and $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$

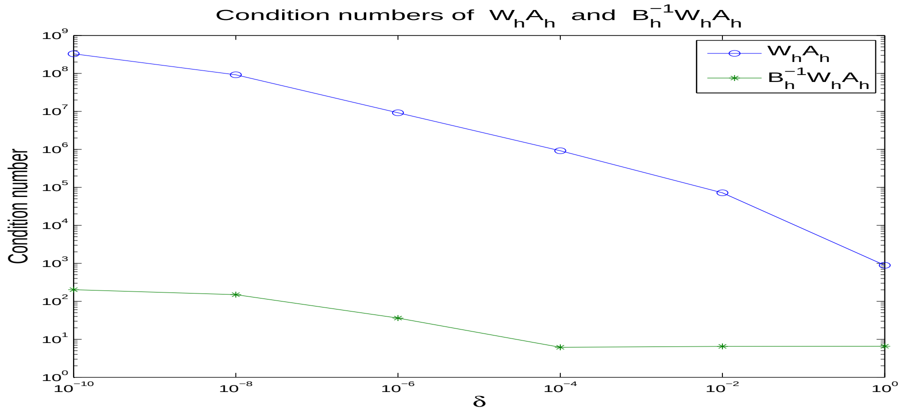

Figure 3: Condition numbers for $\mathcal{W}_h\mathcal{A}_h$ and $\mathcal{B}_h^{-1}\mathcal{W}_h\mathcal{A}_h$

Now we show the optimal controlability of the cubic spline collocation method. Let $\mathbf{u}_h = [u_h, v_h]^T$ be the approximate solution for the equation (3.1) by the cubic spline collocation method (2.5). To show the effects of the penalty parameter δ as $\delta \to 0$, we report the L^2-norm errors between the target state $\hat{u}$ and the controlled state u_h, the L^2-norm of optimal control $\theta_h = -\frac{1}{\delta}v_h$ and the value of the cost functional $\mathcal{J}(u_h, \theta_h)$ with a fixed $N = 20$ in Table 4.1. From the table, one may see that the smaller δ is, the better the controlability is, even though the condition numbers are increased as δ decreases.

δ	$\|u_h - \hat{u}\|$	$\|\theta_h\|$	$\mathcal{J}(u_h, \theta_h)$
10^{0}	4.9884e-001	2.4053e-002	1.2471e-001
10^{-3}	1.5038e-001	7.2509e+000	3.7594e-002
10^{-6}	2.1497e-004	1.0365e+001	5.3741e-005
10^{-9}	2.1510e-007	1.0370e+001	5.3764e-008
10^{-12}	2.1511e-010	1.0370e+001	5.3764e-011

Table. 4.1. *The numerical values of optimal control problem with $N = 20$.*

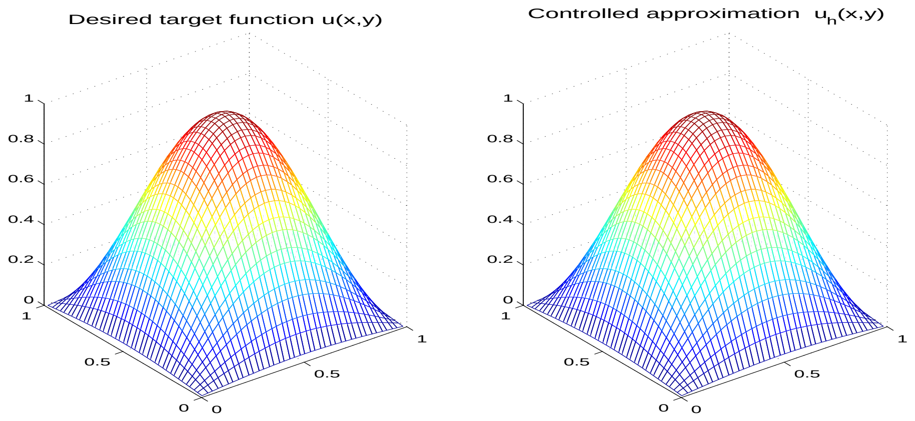

Figure 4: Desired target function $\hat{u}(x, y)$ and the controlled function $u_h(x, y)$

5. Concluding remarks

An optimal control problem subject to (1.2) yields coupled elliptic differential equations (1.3). Any kind of discretizations leads to a nonsymmetric linear systems. In this paper, the cubic spline collocation method is chosen because it is very accurate but the resulting linear systems have large condition numbers. This situation now becomes one of disadvantages if one aims at a fast and efficient numerical simulations for an optimal control problem subject to even a simple elliptic differential equation like (1.2). To overcome such a disadvantage, the lower-order finite element preconditioner is proposed so that the preconditioned linear system can be solved fast by iteration methods designed for nonsymmetric linear systems, like BiCG, GMRES, etc..

References

[1] P. Bochev and M. Gunzburger, *Least-squares finite element methods for optimality systems arising in optimization and control problems*, SIAM J. Numer. Anal., **43**(2006), 2517-2543.

[2] Y. Chen, N. Yi and W. Liu, *A Legendre-Galerkin spectral method for optimal control problems governed by elliptic equations*, SIAM J. Numer. Anal., **46**(2008), 2254-2275.

[3] J. Douglas and T. Dupont, Collocation methods for parabolic equations in a single space variable, Lecture Notee in Mathematics 385, Springer-Verlag Press, Cambridge, UK, 2002.

[4] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., **20**(1983), 345-357.

[5] A. Greenbaum, Iterative methods for solving linear systems, SIAM, Philadelphia, 1997.

[6] M. Gunzburger, Perspectives in Flow Control and Optimization, Adv. Des. Control 5, SIAM, Philadelphia, 2002.

[7] R. A. Horn and C. R. Johnson, Topics in Matrix Analysis. Cambridge University Press, Cambridge, 1994.

[8] S. D. Kim, *Piecewise bilinear preconditioning of high-order finite element methods*, ETNA, **26**(2007), 228-242.

[9] S. Kim and S. D. Kim, *Preconditioning on high-order element methods using Chebyshev-Gauss-Lobatto notes*, Applied Numerical Mathematics, **59**(2009) 316-333.

[10] S. D. Kim and S. Kim, *Exponential decay of C^1-cubic splines vanishiing at two symmetric points in each knot interval*, Numer. Math., **76**(1997), 470-488.

[11] S. D. Kim and S. V. Parter, *Preconditioning Chebyshev spectral collocation method for elliptic partial differential equations*, SIAM J. Numer. Anal., **33**(1996), 2375-2400.

[12] S. D. Kim and S. V. Parter, *Preconditioning cubic spline collocation discretizations of elliptic equations*, Numer. Math., **72**(1995), 39-72.

[13] S. D. Kim and B. C. Shin, *On the exponential decay of C^1 cubic Lagrange splines on non-uniform meshes and for non-uniform data points*, Houston J. of Math., **24**(1998), 173-183.

[14] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric lienar systems*, SIAM J. Sci. Comput., **7**(1986), 856-869.

[15] L. N. Trefethen and D. B. Bau, *Numerical linear algegra*, SIAM, Philadelphia, 1997.