A preconditioner with inexact element face solver for three dimensional $p$-version finite element methods

Weiming Cao$^a$, Benqi Guo$^b$,$^*$

$^a$Division of Mathematics and Statistics, University of Texas at San Antonio, TX 78249, USA
$^b$Department of Mathematics, University of Manitoba, Winnipeg, Man. Canada R3T 2N2

Received 6 March 2001; received in revised form 24 May 2001

Abstract

In this paper, we propose to modify a preconditioner developed in Pavarino and Widlund (Comput. Math. Appl. 33 (1997) 193) for three-dimensional $p$-version finite element or spectral element approximations by solving the sub-problems on element faces with an inexact element face solver. Such a modification reduces the cost to evaluate the action of the preconditioner from $O(p^4)$ to $O(p^3)$ per element face, where $p$ is the polynomial degree used for approximation. Furthermore, it is shown that the coefficient matrix of the inexact solver is spectrally equivalent to the original stiffness matrix on element faces. Therefore, such a change will not affect the polylogarithmic estimate given in Pavarino and Widlund for the condition number of the preconditioned system. Numerical results are also given to confirm the theoretical analysis. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 65F10; 65N30; 65N55

Keywords: $p$-version finite element methods; Spectral element methods; Precondition; Condition number; Iterative solver; Parallel solver

1. Introduction

It is well-known that for the numerical solution of partial differential equations, $p$-version finite element methods and spectral element methods can offer the advantage of higher order convergence [3,14,17]. For the solution of the linear algebraic system resulted from the $p$-version finite element...
or spectral element approximations, iterative methods are commonly used for their lower memory requirement. This is especially the case for three-dimensional problems, where the number of unknowns exceeding millions is not unusual [12]. A crucial issue for the efficiency of iterative methods is to use preconditioning techniques to reduce the condition numbers. Over the past two decades, there have been a large number of studies devoted to this topic. One of the most notable preconditioning methods is the so-called iterative substructuring method developed by Bramble et al. [2] and Dryja and their coworkers [6–8,18]. The basic idea of iterative substructuring method is to define the preconditioner as the block diagonal matrices, each diagonal block is associated with either a global problem on a coarse mesh or a subproblem on a subregion (e.g., the union of a number of elements). This method has been widely used in the $h$-version of finite element computation.

In [15,16], Pavarino and Widlund extended this idea to higher order $p$-version FEM and spectral element methods in three dimensions. They basically divided the entire approximation space into an exotic function space associated with the wirebasket (i.e., the union of element vertices and element edges), and a number of function spaces associated with the element faces. Then the preconditioner is defined by removing from the stiffness matrices all the coupling between basis functions associated with the wirebasket and element faces and between those associated with different element faces. Such a preconditioner leads to a condition number growing only polylogarithmically with $p$ for the preconditioned system, where $p$ is the polynomial degree used for the approximation in each spatial variable. Based on these works, Guo and Cao [9] designed successfully a comprehensive three-level preconditioner for the general $hp$ version finite element approximations.

In this paper, we consider to modify the preconditioner proposed by Pavarino and Widlund [16] for three-dimensional $p$-version finite element and spectral element approximations. The innovation is that when evaluating the action of the preconditioner, we do not solve the subproblems associated with element faces exactly, which costs at least $O(p^4)$ operations. Instead, we replace the stiffness matrix associated with element faces by the matrix defined with the $H^{1/2}_{\infty}$-norm, i.e., use an inexact solver for the subproblems on element faces. This technique was developed in a more general context in our previous work [5,10]. Here we incorporate it into the preconditioner developed in [16]. It is shown that the modified preconditioner using the inexact face solver is indeed spectrally equivalent to the original one, thus the overall performance of preconditioning will be kept as good as the original one as given in [16]. Besides, the coefficient matrix associated with the inexact face solver is well-structured. Its inversion admits a fast algorithm only requiring the total number of operations of order $O(p^3)$. Thus the work needed by the new preconditioner is an order of $p$ less than that by Pavarino and Widlund’s.

In this paper, we also present a refined proof for the spectral equivalence (with equivalence constant independent of $p$) between a quadratic form and the $H^{1/2}_{\infty}$-norm over certain finite dimensional polynomial subspaces. Such a conclusion was conjectured in [5], but the equivalence constant there was only shown to be bounded by a number of order $p$.

This paper is organized as follows. In Section 2 we introduce the model elliptic problem and its discretization by the $p$-version finite element method and spectral element method. In Section 3 we describe the preconditioner used for the iterative solution of the linear algebraic systems resulted from the FEM. We start from recalling the preconditioner proposed in [15,16]. Then we modify it to get a new preconditioner by using the inexact element interface solver based on the matrix associated with the $H^{1/2}_{\infty}$-norm on element faces. In Section 4, we present a theoretical analysis of
the condition number for the preconditioned system. Finally, an algorithm to implement the new preconditioner efficiently is described.

2. Model problem and $p$-version approximation

2.1. The model problem

Let $\Omega$ be a polyhedral domain in $\mathbb{R}^3$. $L^2(\Omega)$, $H^1(\Omega)$, and $H^1_0(\Omega)$ are the usual Sobolev spaces, and $(\cdot,\cdot)$ is the $L^2$-inner product. Given $f \in L^2(\Omega)$, consider the following Poisson equation with homogeneous Dirichlet condition:

$$ -\Delta u = f \quad \text{in } \Omega, $$

$$ u = 0 \quad \text{on } \partial \Omega. $$

Define a bilinear form $a(\cdot,\cdot): H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbb{R}$ as $a(u,v) = (\nabla u, \nabla v)$. Then the weak formulation of (2.1) is to find $u \in H^1_0(\Omega)$ such that

$$ a(u,v) = (f,v), \quad \forall v \in H^1_0(\Omega). $$

We consider the approximation of (2.2) by the $p$-version finite element method or spectral element method. First we partition $\Omega$ into a number of small elements, i.e., $\Omega = \bigcup K$. Each $K$ is required to be a shape regular hexahedral [4]. To build the approximation subspace, we introduce a special set of basis functions on the reference element $\hat{K} = [-1,1]^3$ as follows.

2.2. Shape functions

Let $p$ be a positive integer, and let $L_p(t)$ be the Legendre polynomial of degree $p$. Denote by $-1 = t_0 < t_1 < \cdots < t_p = 1$ the Gauss–Lobatto–Legendre (GLL) points, i.e., the zeros of $(1-t^2)L'_p(t)$.

Let $\hat{\ell}_j(t)$, $0 \leq j \leq p$, be the Lagrange interpolation polynomials such that $\hat{\ell}_j(t_j) = \delta_{jj'}$. They can be expressed as (see (2.3.24) of [3]):

$$ \hat{\ell}_j(t) = \frac{-1}{p(p+1)L'_p(t)} \frac{(1-t^2)L'_p(t)}{t-t_j}, \quad 0 \leq j \leq p. $$

There are four types of basis functions on the reference element $\hat{K}$. All of them are the tensor products of the above Lagrange interpolation polynomials in three spatial directions $\xi$, $\eta$ and $\zeta$.

1. Nodal modes $\hat{\gamma}_{j_1}(\xi)\hat{\gamma}_{j_2}(\eta)\hat{\gamma}_{j_3}(\zeta)$ with $j_1$, $j_2$ and $j_3$ being either 0 or $p$. Each of these functions takes value 1 at one of the vertex of $\hat{K}$ and vanishes at all the other GLL points on the edges of $\hat{K}$. We denote by $\Psi^{[N]}_p$ the set of all nodal modes;

2. Side modes associated with a side $\hat{\gamma}_m$ of $\hat{K}$. These functions vanish on all the edges of $\hat{K}$ except $\hat{\gamma}_m$. For instance, if $\hat{\gamma}_m = \{((\xi,-1,1), -1 < \xi < 1\}$, then the side modes on it are $\hat{\gamma}_{j_1}(\xi)\hat{\gamma}_0(\eta)\hat{\gamma}_0(\zeta)$ for $1 \leq j_1 \leq p - 1$. We use $\Psi^{[\hat{\gamma}_m]}_p$ to denote the set of all the side modes associated with $\hat{\gamma}_m$;
(3) Face modes associated with a face \( \tilde{f}_n \) of \( \tilde{K} \). These functions vanishes on all the faces of \( \tilde{K} \) except \( \tilde{f}_n \). For instance, if \( \tilde{f}_n = \{ (\xi, \eta; -1), -1 < \xi, \eta < 1 \} \), then the face modes on it are \( \hat{\psi}_{\tilde{f}_n}^{(j_1, j_2)}(\xi) \hat{\psi}_{\tilde{f}_n}^{(j_3)}(\eta) \), for \( 1 \leq j_1, j_2 \leq p - 1 \). We use \( \hat{\psi}_{\tilde{f}_n}^{(j_1)} \) to denote the set of all the face modes associated with \( \tilde{f}_n \).

(4) Internal modes \( \hat{\psi}_{\tilde{f}_n}^{(j_1)}(\xi) \hat{\psi}_{\tilde{f}_n}^{(j_2)}(\eta) \hat{\psi}_{\tilde{f}_n}^{(j_3)}(\zeta) \), \( 1 \leq j_1, j_2, j_3 \leq p - 1 \). These functions vanish on the boundary of \( \tilde{K} \). We denote by \( \hat{\psi}_{\tilde{f}_n}^{(j_1)} \) the set of all internal modes.

We define a polynomial space on \( \tilde{K} \) as follows:

\[
\hat{\psi}_{\tilde{f}_n} = \hat{\psi}_{\tilde{f}_n}^{(N)} \oplus \bigcup_{m=1}^{12} \hat{\psi}_{\tilde{f}_n}^{(y_{CR})} \oplus \bigcup_{n=1}^{6} \hat{\psi}_{\tilde{f}_n}^{(f_n)} \oplus \hat{\psi}_{\tilde{f}_n}^{(i)}. 
\]

2.3. Approximation subspace

Consider an element \( K \). Let \( F_K \) be the affine mapping from \( \tilde{K} \) onto \( K \), and let \( y_{m} \) and \( f_n \) be the images of \( y_{m} \) and \( f_n \) under \( F_K \), respectively. We define the function spaces associated with the nodes, sides, faces, and the interior of element \( K \) as follows:

\[
\hat{\psi}_p = \hat{\psi}_p^{(N)} \oplus \bigcup_{m=1}^{12} \hat{\psi}_p^{(y_{CR})} \oplus \bigcup_{n=1}^{6} \hat{\psi}_p^{(f_n)} \oplus \hat{\psi}_p^{(i)}. 
\]

Now an approximation subspace in the element \( K \) is defined as

\[
\hat{\psi}_p(K) = \hat{\psi}_p^{(N)}(K) \oplus \bigcup_{m=1}^{12} \hat{\psi}_p^{(y_{CR})}(K) \oplus \bigcup_{n=1}^{6} \hat{\psi}_p^{(f_n)}(K) \oplus \hat{\psi}_p^{(i)}(K). 
\]  

The approximation subspace for the \( p \)-version of FEM in \( \Omega \) is

\[
\hat{\psi}_p(\Omega) = \{ v | v|_K \in \hat{\psi}_p(K) \} \cap H^1_0(\Omega). 
\]

2.4. FE equation and its matrix form

The \( p \)-version finite element method for solving (2.2) is to find \( u \in \hat{\psi}_p(\Omega) \) such that

\[
a(u, v) = (f, v), \quad \forall v \in \hat{\psi}_p(\Omega). 
\]

Let us use \( \{ \phi_{u_\nu} \} \) to denote the set of all the basis functions associated with the wirebasket of the partition, i.e., the union of all element vertices and edges. Denote by \( \{ \phi_{f_i} \} \) all the basis functions associated with element faces, and by \( \{ \phi_{i} \} \) those associated with the element interiors. Correspondingly, we decompose the total degree of freedoms into \( \{ u_{w_\nu} \}, \{ u_{f_i} \} \) and \( \{ u_i \} \), and the
load vector into \( \{ F_{wj} \}, \{ F_{fk} \} \) and \( \{ F_{il} \} \), associated with the wirebasket, the faces, and the interiors, respectively, namely, \( \mathbf{u} = [u_w, u_f, u_i] \), and \( \mathbf{F} = [F_w, F_f, F_i] \). Then the linear system of algebraic equations for (2) can be expressed as

\[
\begin{bmatrix}
S_{ww} & S_{wf} & S_{wi} \\
S^T_{wf} & S_{ff} & S_{fi} \\
S^T_{wi} & S^T_{fi} & S_{ii}
\end{bmatrix}
\begin{bmatrix}
u_w \\
u_f \\
u_i
\end{bmatrix}
=
\begin{bmatrix}
F_w \\
F_f \\
F_i
\end{bmatrix}.
\]  

(7)

Each block in the coefficient matrix of (7) represents the coupling between basis functions associated with two geometric objects. Since the supports of the interior basis functions on different elements do not intersect, \( S_{ii} \) is indeed block diagonal with each diagonal block representing the coupling between the interior bases in one element. In practical computation, the linear algebraic system (7) is formed by assembling local stiffness matrices and local load vectors on all elements.

Now we attempt to reduce the size of (7) by eliminating all the unknowns in element interiors. Consider a change of basis functions as follows:

\[
(\psi_{wj}, \psi_{fk}, \psi_{li}) \rightarrow (\psi_{wj}, \psi_{fk}, \phi_{li}),
\]

where the new basis functions \( \psi_{wj} \) and \( \psi_{fk} \) are defined to be identical to \( \phi_{wj} \) and \( \phi_{fk} \) on element interfaces and orthogonal to all the interior basis functions, i.e.,

\[
\begin{aligned}
\psi_{wj} &= \phi_{wj} & \text{on } \partial K \text{ for all } K, \\
\langle \psi_{wj}, \phi_i \rangle &= 0 & \forall \phi_i \in \Psi_p(\Omega)
\end{aligned}
\]

and similarly for \( \psi_{fk} \). \( \{\psi_{wj}\} \) and \( \{\psi_{fk}\} \) are called discrete harmonic on each element.

Under the new basis functions, the linear system (7) will be changed into

\[
\begin{bmatrix}
\tilde{S}_{ww} & \tilde{S}_{wf} & 0 \\
\tilde{S}^T_{wf} & \tilde{S}_{ff} & 0 \\
0 & 0 & S_{ii}
\end{bmatrix}
\begin{bmatrix}
u_w \\
u_f \\
u_i
\end{bmatrix}
=
\begin{bmatrix}
\tilde{F}_w \\
\tilde{F}_f \\
\tilde{F}_i
\end{bmatrix}.
\]

(8)

where

\[
\begin{aligned}
\tilde{S}_{ww} &= S_{ww} - S_{wi}(S_{ii})^{-1}S^T_{wi}, \\
\tilde{S}_{wf} &= S_{wf} - S_{wi}(S_{ii})^{-1}S^T_{fi}, \\
\tilde{S}_{ff} &= S_{ff} - S_{fi}(S_{ii})^{-1}S^T_{fi}, \\
\tilde{u}_i &= u_i + (S_{ii})^{-1}S^T_{wi}u_w + (S_{ii})^{-1}S^T_{fi}u_f, \\
\tilde{F}_w &= F_w - S_{wi}(S_{ii})^{-1}F_i, \\
\tilde{F}_f &= F_f - S_{fi}(S_{ii})^{-1}F_i.
\end{aligned}
\]

It is clear from (8) that we can separate the solution for the unknowns in \( u_w \) and \( u_f \) from those in \( u_i \). More precisely, we can solve

\[
A
\begin{bmatrix}
u_w \\
u_f
\end{bmatrix}
=
\begin{bmatrix}
\tilde{S}_{ww} & \tilde{S}_{wf} \\
\tilde{S}^T_{wf} & \tilde{S}_{ff}
\end{bmatrix}
\begin{bmatrix}
u_w \\
u_f
\end{bmatrix}
=
\begin{bmatrix}
\tilde{F}_w \\
\tilde{F}_f
\end{bmatrix}.
\]

(9)
for \( \mathbf{u}_w \) and \( \mathbf{u}_f \) first. Then the interior unknowns in \( \mathbf{u}_i \) comes from the solution of

\[
S_{ii} \mathbf{u}_i = \mathbf{F}_i - S_{wi}^T \mathbf{u}_w - S_{fi}^T \mathbf{u}_f,
\]

which is essentially a Dirichlet problem on each element \( K \) with the data on \( \partial K \) already obtained.

The reduced system (9) is also called the Schur complement system to element interiors, and the above procedure to reduce (7) into (9) is called elimination of internal modes. In practice, elimination of internal modes is performed on all elements locally, and the reduced linear system (9) is obtained by assembling all the reduced local stiffness matrices and load vectors.

3. Preconditioner with inexact element face solvers

The reduced system (9) may still be too large to be solved by direct methods. In this case, iterative methods, e.g., conjugate gradient method, is the only viable option. A key to the success of iterative methods is the use of suitable preconditioning techniques, which reduces the condition number of (9) while not causing substantial extra work. During the past two decades, various efficient and effective preconditioners have been developed for the finite element approximation. We here mainly follow the idea of Pavarino and Widlund [15,16] based on the iterative substructures for the \( p \)-version and spectral element approximations.

3.1. Preconditioner of Pavarino and Widlund

The basic idea used in [15,16] to construct the preconditioner is first to decrease the degree of coupling between the basis functions associated with the wirebasket and those with the element faces. Then the preconditioner is defined as a block diagonal matrix with each diagonal block associated with either the wirebasket or one element face. In order to make the condition number of the preconditioned system be independent of the element sizes, it is crucial to choose the subspace associated with the wirebasket to contain the constant function on each element. Unfortunately, this is not true for the subspace spanned by the basis functions \( \{ \psi_{w_j} \} \). To remedy, it is noted that the constant function 1 can be decomposed as

\[
1 = \sum \psi_{w_j} + \sum_{k=1}^{6} \mathcal{F}_{f_k}
\]

with

\[
\mathcal{F}_{f_k} = \sum_{i,j=1}^{p-1} \psi_{f_k}^{(i,j)}
\]

and \( \psi_{f_k}^{(i,j)}, 1 \leq i, j \leq p - 1 \), are all the basis functions associated with face \( f_k \).

Now we define the new basis function associated with the wirebasket as

\[
\tilde{\psi}_{w_j} = \psi_{w_j} + \sum_{k=1}^{6} m_{jk} \mathcal{F}_{f_k},
\]
where \( m_{jk} \) is the average of \( \psi_{wj} \) on the edges of face \( f_k \), i.e.,
\[
m_{jk} = \frac{\int_{\partial f_k} \psi_{wj} \, ds}{\int_{\partial f_k} ds}.
\]

Let \([I, R]\) be the transformation matrix for the change of bases from \( \{\psi_{wj}\} \) to \( \{\tilde{\psi}_{wj}\} \), i.e.,
\[
\tilde{\psi}_w = \psi_w + R\psi_f.
\]

Then under the new bases \( \{\tilde{\psi}_{wj}\} \) and \( \{\psi_{f_k}\} \), the coefficient matrix in (9) will be changed into
\[
\begin{bmatrix}
I & R \\
0 & I
\end{bmatrix}
\begin{bmatrix}
\tilde{S}_{ww} & \tilde{S}_{wf} \\
\tilde{S}_{fw}^T & \tilde{S}_{ff}
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
R^T & I
\end{bmatrix}
= \begin{bmatrix}
\tilde{S}_{ww} & \tilde{S}_{wf} \\
\tilde{S}_{fw}^T & \tilde{S}_{ff}
\end{bmatrix}.
\]

(10)

Now, the preconditioner for (10) can be constructed by: (1) deleting from (10) all the coupling between the wirebasket and faces, and all the coupling between different element faces; (2) replacing the block \( \tilde{S}_{ww} \) with a simpler but spectrally equivalent matrix. More specifically, the preconditioner for matrix (10) is defined as
\[
\begin{bmatrix}
C_{ww} & 0 \\
0 & C_{ff}
\end{bmatrix},
\]

(11)

where
\[
C_{ff} = \begin{bmatrix}
\tilde{S}_{fi,fi} & 0 & \cdots \\
0 & \tilde{S}_{fi,fi} & \cdots \\
\cdots & \cdots & \cdots
\end{bmatrix}
\]

(12)

with each block \( \tilde{S}_{fi,fi} \) corresponding to one element face. The block \( C_{ww} \) is the matrix whose quadratic form is equivalent to the following scaled \( L^2 \)-norm on the wirebasket \( w \): if \( u = \sum u_w \hat{\psi}_w \), then
\[
\sum u_w^T C_{ww} u_w = (1 + \ln p) \inf_c \|u - c\|^2_{L^2(w)}.
\]

Note that for polynomials of degree \( \leq p \), their \( L^2 \)-norms are equivalent to the discrete \( L^2 \)-norms calculated with GLL quadrature formula, thus it is easy to derive an explicit form for \( C_{ww} \) as follows
\[
C_{ww} = (1 + \ln p) \left( D - \frac{(Dz)(Dz)^T}{z^T Dz} \right),
\]

(13)

where \( z = (1, 1, \ldots, 1)^T \) and \( D \) is the diagonal matrix of dimension equal to the total number of GLL points on all edges of the element. The diagonal entries of \( D \) are the sum of the GLL quadrature weights, i.e.,
\[
d_{mm} = \begin{cases}
\frac{6}{p(p+1)} & \text{if } m \text{ corresponds to a vertex of } K, \\
\frac{2}{p(p+1)(L_p(t_j))^2} & \text{if } m \text{ corresponds to } j^{th} \text{ GLL point on an edge of } K, 1 \leq j \leq p - 1.
\end{cases}
\]
Finally, we return to the bases $\psi_{w_j}$ and $\phi_{f_k}$. The preconditioner for system (9) is defined as

$$C(0) = \begin{bmatrix} I & 0 \\ -R & I \end{bmatrix} \begin{bmatrix} C_{ww} & 0 \\ 0 & C_{ff} \end{bmatrix} \begin{bmatrix} I & -R^T \\ 0 & I \end{bmatrix}$$

(14)

and the global matrix for the preconditioner (still denoted as $C(0)$) is obtained by assembling all the above local preconditioners on elements.

**Theorem 3.1** (Pavarino and Widlund [16, Theorem 2]). *Let $A$ be the stiffness matrix in (9) defined with Lagrange basis (3). For the preconditioner $C(0)$ defined by (12)–(14), there holds

$$\kappa((C(0))^{-1}A) \leq c(1 + \ln p)^2.$$*

### 3.2. Preconditioner within exact element face solvers

Consider the action of the preconditioner $C(0)$ to a residual vector $r = [r_w, r_f]^T$, i.e., to compute the matrix–vector product $(C(0))^{-1}r$. Clearly, we need to calculate for each element face $S^{-1}_{f_k,f_k}r_{f_k}$, or equivalently, to solve a subproblem $S^{-1}_{f_k,f_k}u_{f_k} = r_{f_k}$. Note that $S_{f_k,f_k}$ is dense and of dimension $(p - 1)^2 \times (p - 1)^2$. A direct calculation by Gaussian elimination would involve $O((p - 1)^6)$ operations. Even if $S_{f_k,f_k}^{-1}$ has been obtained prior to the iteration and available at each iteration step, the multiplication of $S_{f_k,f_k}^{-1}$ with $r_{f_k}$ still involves $2(p - 1)^4$ operations for each face. To save the work required in this part of preconditioning, we will not solve these subproblems exactly. Instead we may replace $S_{f_k,f_k}$ with a simpler but spectrally equivalent matrix. This is the idea of *inexact element face solvers*. Perhaps the simplest choice for the replacement is a diagonal matrix. Several types of diagonal matrices have already been investigated in [5,10]. Unfortunately, the equivalence between these diagonal matrices and $S_{f_k,f_k}$ involves a constant at least of order $p$. This equivalence implies that the condition number for the preconditioned system will have an extra factor of order at least $p$. This can possibly undermine the overall performances of preconditioning. On the other hand, a nondiagonal inexact solver on element faces was also introduced in [5] based on the representation of $H^{1/2}_{00}$-norms. Though it is conjectured that the matrix for the inexact face solver is spectrally equivalent (with equivalence constants independent of $p$) to $S_{f_k,f_k}$, we were only able to show that the equivalence constants are bounded by a constant of order $p$, see Remark 4.2 in [5]. Here, we integrate the inexact face solver into preconditioner (14). We also present a refined analysis to prove the spectral equivalence, with constants independent of $p$, between the matrix for the inexact face solver and $S_{f_k,f_k}$, which implies readily that the modified preconditioner using the inexact face solver will keep the same logarithmic estimate as for $C(0)$ in Theorem 3.1.

We note that $S_{f_k,f_k}$ is the stiffness matrix for the discrete harmonic basis functions associated with face $f_k$. By the so-called *discrete harmonic extension* theorem [15] for the $p$-version approximation, if $u = \sum u_{f_k} \psi_{f_k}$, then

$$u_{f_k}^T S_{f_k,f_k} u_{f_k} = \|u\|_{H^1(K)}^2 \asymp \|u\|_{H^{1/2}_{00}(f_k)}^2.$$ 

Here and in the below we use symbol “$\asymp$” to denote equivalence with equivalent constants independent of $p$. For any function $u$ defined on a face $f_k$, let $\hat{u}$ be the image of $u$ under the inverse
of the affine mapping from the reference face \( \hat{f} = (-1, 1)^2 \) onto \( f_k \). By a scaling argument, it is easy to see that \( \|u\|_{H^{1/2}_{00}(\hat{f}_k)} \) is equivalent to \( \|\hat{u}\|_{H^{1/2}_{00}(\hat{f})} \) with equivalence constants depending only on the shape of the face. Therefore, if we replace \( \hat{S}_{\hat{f}_k} \) with the matrix \( \hat{S}_{\hat{f}} \) associated with \( H^{1/2}_{00}(\hat{f}) \)-norm on \( \hat{f} \), the condition number of the new preconditioned system will be kept in the same logarithmic order as in Theorem 3.1.

Bear this argument in mind, we start to look for the matrix representation for the \( H^{1/2}_{00}(\hat{f}) \)-norm. By Theorem 13 in [11], an equivalent norm for \( H^{1/2}_{00}(\hat{f}) \) is

\[
\|v\|_{H^{1/2}_{00}(\hat{f})} \asymp \left( \|v\|_{L^2(I_s;H^{1/2}_{00}(I_s))}^2 + \|v\|_{L^2(I_s;H^{1/2}_{00}(I_s))}^2 \right)^{1/2},
\]

where

\[
\|v\|_{L^2(I_s;H^{1/2}_{00}(I_s))} = \int_{I_s} \|v(x, \cdot)\|_{H^{1/2}_{00}(I_s)}^2 \, dx
\]

and \( \|v\|_{L^2(I_s;H^{1/2}_{00}(I_s))} \) is defined similarly.

Let \( \{\hat{f}\}_j \) be the set of Lagrange basis \( (3) \). Define the one-dimensional mass and stiffness matrices as \( M = (m_{ij}) \) and \( Q = (q_{ij}) \) with

\[
m_{ij} = \int_{-1}^1 \hat{f}_i(t)\hat{f}_j(t) \, dt, \quad q_{ij} = \int_{-1}^1 \hat{f}'_i(t)\hat{f}'_j(t) \, dt.
\]

Then for \( v = \sum_{j=1}^{p-1} v_j \hat{f}_j \) and \( v = (v_1, v_2, \ldots, v_{p-1}) \), we have

\[
v^T M v = \|v\|_{L^2(I)}^2, \quad v^T Q v = \|v\|_{H^1(I)}^2.
\]

Define

\[
J = M^{1/2} (M^{-1/2} Q M^{-1/2})^{1/2} M^{1/2}.
\]

It can be proved that (see Section 4)

\[
v^T J v \asymp \|v\|_{H^{1/2}_{00}(I)}^2.
\]

If we define

\[
C_{\hat{f}} = M \otimes J + J \otimes M,
\]

then for any \( v = \sum v_j \psi_j \), we have

\[
v_j^T J v_j = \|v_j\|_{H^{1/2}_{00}(\hat{f}_j)}^2.
\]

Consequently, replacing \( \hat{S}_{\hat{f}_k} \) in the preconditioner \( C \) with \( C_{\hat{f}} \) will not cause an increase in the order of condition number of the preconditioned system.

We define the new preconditioner with inexact element face solvers as follows:

\[
C^{(1)} = \begin{bmatrix} I & 0 \\ -R & I \end{bmatrix} \begin{bmatrix} C_{ww} & 0 \\ 0 & \hat{C}_{ff} \end{bmatrix} \begin{bmatrix} I & -R^T \\ 0 & I \end{bmatrix},
\]
where
\[
\tilde{C}_{ff} = \begin{bmatrix}
C_{\hat{f}\hat{f}} & 0 \\
\vdots & \ddots & \ddots \\
0 & \ddots & C_{\hat{f}\hat{f}} \\
0 & \ddots & \ddots & \ddots 
\end{bmatrix}
\]
and the global matrix for the new preconditioner (still denoted by \( C^{(1)} \)) is also formed by assembling these local ones.

4. Analysis of condition number

Theorem 4.1. Let \( A \) be the stiffness matrix in (9) defined with Lagrange basis (3). For the new preconditioner defined by (17), we have
\[
\kappa((C^{(1)})^{-1}A) \leq c(1 + \ln p)^2.
\]
Proof. Since \( C^{(1)} \) is obtained by modifying Pavarino and Widlund’s preconditioner \( C^{(0)} \) by replacing \( S_{\hat{f}\hat{f}} \) with \( \hat{S}_{\hat{f}\hat{f}} \) given in (16), we need only to show that these two matrices are spectrally equivalent with equivalence constants independent of \( p \). Because \( S_{\hat{f}\hat{f}} \) corresponds to the \( H_{00}^{1/2} \)-norm on \( f_k \), by a scaling argument it is spectrally equivalent to the matrix corresponding to the \( H_{00}^{1/2}(\hat{f}) \)-norm. From Lemma 4.2 below, we see that matrix \( \hat{C}_{\hat{f}\hat{f}} \) gives a quadratic form equivalent to the \( H_{00}^{1/2}(\hat{f}) \)-norm. Therefore \( \hat{C}_{\hat{f}\hat{f}} \) is spectrally equivalent to \( S_{\hat{f}\hat{f}} \).

Lemma 4.2. Let \( \{\phi_j\}_{j=1}^{p-1} \) be any set of basis functions of \( P_0^p(I) \). \( M = (m_{ij}) \) and \( Q = (q_{ij}) \) with \( m_{ij} = (\phi_i, \phi_j) \) and \( q_{ij} = (\phi'_i, \phi'_j) \). Then for any \( v = \sum_{j=1}^{p-1} v_j \phi_j \) and \( v = (v_1, v_2, \ldots, v_{p-1})^T \),
\[
z_1 v^T J v \leq \|v\|_{H_{00}^{1/2}(I)}^2 \leq z_2 v^T J v,
\]
where \( z_1 \) and \( z_2 \) are positive constants independent of \( p \), and
\[
J = M^{1/2}(M^{-1/2}QM^{-1/2})^{1/2}M^{1/2} = Q^{1/2}(Q^{-1/2}MQ^{-1/2})^{1/2}Q^{1/2}. \tag{19}
\]
Proof. The second equality in (19) for the definition of \( J \) can be verified by direct matrix computation. Now we consider (18). First, we prove for the case \( Q = I \) and \( J = M^{1/2} \). Consider the following eigenvalue problem:
\[
(\Phi_j', \Phi_j') = \lambda_j(\Phi_j, \Phi_j), \quad \forall \Phi \in P_0^p(I).
\]
We normalize the eigenfunctions \( \{\Phi_j\}_{j=1}^{p-1} \) such that \( \|\Phi_j\|_{L^2(I)} = 1 \). Note that both the eigenvalues \( \lambda_j \) and eigenfunctions \( \Phi_j \) depend on \( p \). Nevertheless, it has been proved [13,1] that for any...
\[ v = \sum_{j=1}^{p-1} \tilde{v}_j \Phi_j, \] there holds that
\[ \|v\|_{H_{00}^1(I)}^2 \geq \sum_{j=1}^{p-1} \frac{1}{\sqrt{\lambda_j}} (\tilde{v}_j)^2 = \tilde{v}^T A^{-1/2} \tilde{v} \]

with the equivalence constants independent of \( p \). Here
\[ A = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p-1}). \]

Now we make a change of the bases from \( \{ y_{RSj} \} \) to \( \{ y_{BSj} \} \). Suppose
\[ v = \sum_{j=1}^{p-1} \tilde{v}_j \Phi_j = \sum_{j=1}^{p-1} \tilde{v}_j y_{BSj} \]
and
\[ \tilde{v} = Tv. \]

Then
\[ \|v\|_{L^2(I)}^2 = v^T M v = \tilde{v}^T A^{-1} \tilde{v}, \]
\[ \|v\|_{H^1(I)}^2 = v^T Q v = \tilde{v}^T \tilde{v}. \]

Hence we have
\[ M = T^T A^{-1} T, \quad T^T T = I, \]
which implies \( T^T = T^{-1} \) and \( M^{1/2} = T^T A^{-1/2} T \), Consequently
\[ \|v\|_{H_{00}^1(I)}^2 \approx \tilde{v}^T A^{-1/2} \tilde{v} = v^T (T^T A^{-1/2} T) v \]
\[ = v^T M^{-1/2} v = v^T J v. \]

Next, we prove (18) for the general basis functions \( \{ \phi_j \} \) where \( Q \neq I \). Define a set of new basis functions as follows
\[ \{ \psi_1, \psi_2, \ldots, \psi_{p-1} \} = Q^{-1/2} \{ \phi_1, \phi_2, \ldots, \phi_{p-1} \}. \]

Let \( \tilde{M} = (\tilde{m}_{ij}) \) and \( \tilde{Q} = (\tilde{q}_{ij}) \) with \( \tilde{m}_{ij} = (\psi_i, \psi_j) \) and \( \tilde{q}_{ij} = (\psi'_i, \psi'_j) \). Then
\[ \tilde{Q} = Q^{-1/2} Q Q^{-1/2} = I, \quad \tilde{M} = Q^{-1/2} M Q^{-1/2}. \]

For any \( v = \sum_{j=1}^{p-1} u_j \phi_j = \sum_{j=1}^{p-1} c_j \psi_j \) with \( v = Q^{-1/2} c \), we have
\[ \|v\|_{H_{00}^1(I)}^2 \approx c^T \tilde{M}^{-1/2} c = (Q^{1/2} v)^T (Q^{-1/2} M Q^{-1/2} T)^{1/2} (Q^{1/2} v) \]
\[ = v^T J v. \]

which completes the proof of this lemma. \( \square \)

**Lemma 4.3.** Let \( \{ \phi_j \}_{j=1}^{p-1} \) be a set of basis functions of \( P^0_p(I) \), and let \( P^0_p(f) = \text{span} \{ \phi_i(x) \phi_j(y), 1 \leq i, j \leq p-1 \} \). For any
\[ v = \sum_{i,j=1}^{p-1} u_{ij} \phi_i(x) \phi_j(y), \]
Table 1
The ratios $\max(x_2)/\min(x_1)$ and $\max(x_4)/\min(x_3)$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\max(x_2)/\min(x_1)$</th>
<th>$\max(x_4)/\min(x_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>4</td>
<td>1.0187</td>
<td>1.1268</td>
</tr>
<tr>
<td>6</td>
<td>1.0097</td>
<td>1.2071</td>
</tr>
<tr>
<td>8</td>
<td>1.0065</td>
<td>1.2709</td>
</tr>
<tr>
<td>10</td>
<td>1.0053</td>
<td>1.3133</td>
</tr>
<tr>
<td>12</td>
<td>1.0048</td>
<td>1.3409</td>
</tr>
<tr>
<td>16</td>
<td>1.0043</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.0041</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1.0040</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1.0039</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1.0039</td>
<td></td>
</tr>
</tbody>
</table>

let vector $v$ be defined as $(v)_{j+i(p-1)} = v_{ij}$. Then there are two positive constants $\alpha_3$ and $\alpha_4$ independent of $p$ such that

$$\alpha_3 v^T C_{jj} v \leq \|v\|_{H_0^{1/2}(\hat{j})}^2 \leq \alpha_4 v^T C_{jj} v,$$

where $C_{jj}$ is the matrix defined in (16).

**Proof.** The proof of this lemma follows readily from the above lemma and the equivalent form (15) for the $H_0^{1/2}(\hat{j})$-norm. □

We checked numerically the equivalence relations (18) and (20). The ratios $\max(x_2)/\min(x_1)$ and $\max(x_4)/\min(x_3)$ are listed in Table 1 for various $p$, which suggests that these two ratios are bounded from the above by constants independent of $p$.

5. Efficient algorithm for implementation

Now we consider how to apply $C^{(1)}$ efficiently in the preconditioned iterative method. Clearly for each element face $f_k$, we have to compute the multiplication of $(C_{jj})^{-1}$ with the residual vector $r_{f_k}$ associated with face $f_k$. We describe here a fast algorithm developed in [5] to compute $(C_{jj})^{-1} r_{f_k}$. Let $u_{f_k} = (C_{jj})^{-1} r_{f_k}$. We rewrite the vectors $u_{f_k}$ and $r_{f_k}$ (which are of length $(p-1)^2$) into the matrix form. Define $U = (U_{ij})$ and $R = (R_{ij})$ with $U_{ij} = (u_{f_k})_{j+i(p-1)}$ and $R_{ij} = (r_{f_k})_{j+i(p-1)}$. Then the equation $C_{jj} u_{f_k} = r_{f_k}$ is equivalent to

$$MUJ + JUM = R.$$  (21)

Substitute into the above equation $J = M^{1/2}(M^{-1/2}QM^{-1/2})^{1/2}M^{1/2}$, and define

$$\hat{U} = M^{1/2}UM^{1/2}, \quad \hat{R} = M^{-1/2}RM^{-1/2}.$$

Then (21) is changed into

$$\hat{U}(M^{-1/2}QM^{-1/2})^{1/2} + (M^{-1/2}QM^{-1/2})^{1/2}\hat{U} = \hat{R}.$$  (22)
Let $G$ be the unitary matrix composed of all the eigenvectors of $M^{-1/2}QM^{-1/2}$, then

$$M^{-1/2}QM^{-1/2} = GAG', \quad (23)$$

where $A = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p-1})$ with $\lambda_i$'s being the eigenvalues of $M^{-1/2}QM^{-1/2}$. Substitute (23) into (22) and define

$$\tilde{U} = G'\tilde{U}G,$$
$$\tilde{R} = G'\tilde{RG} = E'RE,$$

where $E = M^{-1/2}G$. Then (22) can be changed further into

$$\tilde{A}^{1/2} + A^{1/2}\tilde{U} = \tilde{R},$$

whose solution is readily obtained by

$$(\tilde{U})_{ij} = \frac{1}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}(\tilde{R})_{ij}, \quad 1 \leq i, j \leq p - 1.$$ 

Finally, $U$ can be recovered from $\tilde{U}$ as follows

$$U = M^{-1/2}G\tilde{U}G'M^{-1/2} = E\tilde{U}E'.$$

By using the above algorithm, the total number of arithmetic operations to evaluate $(C_{ij})^{-1}r_{fk}$ is only $8(p - 1)^3$ (four $(p - 1) \times (p - 1)$ matrix–matrix multiplications). This is an order of $p$ reduction compared to the work required to calculate the action of $S_{f_k}$ in Pavarino and Widlund's preconditioner. In addition, note that in the $p$-version finite element method or spectral element method in three dimensions, the number of unknowns in an (open) element is $(p - 1)^3$. Therefore by the above algorithm the cost to evaluate the action of $C_{ij}$ is very small. The computation of the eigenpairs of the matrices $M$ and $B$ in the preprocessing step is small, since they are matrices only of order $p - 1$. Besides, this step is needed only once for all elements and all iterations. We believe that the inexact element face solver with $C_{ij}$, together with the fast algorithm, is a desirable one for the $p$-version finite element method and spectral element method in three dimensions.

References


