Rapid error reduction for block Gauss-Seidel based on p-hierarchical bases

(revised version: June 2011)

by

Sabine Le Borne, and Jeffrey Ovall

Preprint no.: 33 2011
RAPID ERROR REDUCTION FOR BLOCK GAUSS-SEIDEL BASED ON P-HIERARCHICAL BASIS

SABINE LE BORNE * AND JEFFREY S. OVALL †

Abstract. We consider a two-level block Gauss-Seidel iteration for solving systems arising from finite element discretizations employing higher-order elements. A p-hierarchical basis is used to induce this block structure. Using superconvergence results normally employed in the analysis of gradient recovery schemes, we argue that a massive reduction of $H^1$-error occurs in the first iterate, so that the discrete solution is adequately resolved in very few iterates—sometimes a single iteration is sufficient. Numerical experiments support these claims.

Key words. higher-order finite elements, hierarchical bases, block Gauss-Seidel, hierarchical matrices

AMS subject classifications. 65N22, 65N55, 65N30, 65F08, 65F05

1. Introduction. We consider linear systems arising from the discretization of second-order, linear, elliptic problems via Lagrange finite elements of order $> 1$. We use a $p$-hierarchical basis of the finite element space $V$ to split it as $V = V_1 \oplus V_2$, with $V_1$ containing the lower order elements (i.e., at least the affine elements). This splitting of $V$ induces a natural $2 \times 2$ block structure and is designed so that the ill-conditioning of the system matrix is concentrated in its $(1,1)$-block and the $(2,2)$-block is well-conditioned (in the ideal case, it may even be spectrally equivalent to its own diagonal). We propose a block Gauss-Seidel iteration, either on its own as a classical linear iteration, or as a preconditioner for a Krylov iteration. The standard convergence analysis for such two-level schemes, at least in case of symmetric problems, is presented for example in [3].

Our analysis differs from others in our arguments concerning the significant error reduction in the first step of the iterative method, which in some cases already resolves the discrete solution up to the discretization error with respect to the $H^1$-norm. These arguments use superconvergence results commonly appearing in the analysis of gradient-recovery schemes. Our analysis requires that the block lower-triangular system (i.e., the system representing the preconditioner of the forward block Gauss-Seidel method) is solved accurately. In practice, this means that more effort must be spent on the $(1,1)$-block, in particular on that portion corresponding to the affine elements, and something cheap can be applied to the $(2,2)$-block. This approach shares a similar philosophical principle with cascadic multigrid [8] and other cascadic methods [12, 11, 28], in which more effort is taken to solve “coarser” problems accurately in order that less effort might be spent on “finer” problems, for a net reduction in solve-time. We also mention the so-called $p$-multigrid and $hp$-multigrid methods, which have the same basic structure as the classical $h$-multigrid methods, but with a $p$-hierarchical basis used to generate the hierarchy of levels. We refer interested readers to [26] and references therein for discussions of these multigrid variants.

The rest of this paper is organized as follows: In Section 2, we introduce our basic notation related to the elliptic problems under consideration and their finite element
discretization. Section 3 introduces the two-level block Gauss-Seidel iteration and contains our main results concerning the massive error reduction in the first iteration. Hierarchical matrices, and how we employ them in this context, are described in Section 4. Finally, in Section 5, we present numerical results to demonstrate our claims on both symmetric and non-symmetric problems for quadratic and quartic elements.

2. Preliminaries. We are interested in efficiently and reliably solving linear systems associated with finite element discretizations of problems of the form

\[
\text{Find } u \in \mathcal{H} \text{ such that } B(u, v) = F(v) \text{ for all } v \in \mathcal{H},
\]

where

\[
B(u, v) = \int_{\Omega} K \nabla u \cdot \nabla v + (b \cdot \nabla u + cu)v \, dx,
\]

\[
F(v) = \int_{\Omega} fv \, dx + \int_{\Gamma_N} gv \, ds,
\]

\[
\mathcal{H} = H^1_{0,D}(\Omega) = \{ v \in H^1(\Omega) : u = 0 \text{ on } \Gamma_D \text{ in the sense of trace} \}.
\]

Here, \( \Omega \subset \mathbb{R}^2 \) is open, bounded and (for simplicity) polygonal, having boundary \( \Gamma = \Gamma_D \cup \Gamma_N \). The Dirichlet and Neumann portions of the boundary, \( \Gamma_D \) and \( \Gamma_N \) respectively, are disjoint, with \( \Gamma_D \) closed in the relative topology. In the present work, for the sake of analysis, we make the assumption that the data functions, \( K, b, c, f, g \), are smooth on \( \Omega \). The matrix \( K \) is also assumed to be symmetric and uniformly positive definite throughout the domain. We assume that both \( F \) and \( B \) are bounded, and we assume, for convenience of exposition, that \( B \) is coercive. The key results, Lemma 3.1 and Theorem 3.3, carry over naturally to the more general inf-sup setting, assuming that similar inf-sup conditions are also uniformly satisfied by the discrete problems described below.

We discretize (2.1) by choosing a finite dimensional subspace \( V \subset \mathcal{H} \) and restricting the problem:

\[
\text{Find } \hat{u} \in V \text{ such that } B(\hat{u}, v) = F(v) \text{ for all } v \in V.
\]

Having chosen a basis \( \{ \psi_k : 1 \leq k \leq N \} \), we make the obvious identification between coefficient vectors \( v \in \mathbb{R}^N \) and functions \( v \in V \), and obtain the following linear system corresponding to (2.5):

\[
Au = f \text{ where } A_{ij} = B(\phi_j, \phi_i) \text{ and } f_i = F(\phi_i).
\]

It is clear that \( B(v, w) = w^t A v \), and when \( B(\cdot, \cdot) \) is an inner-product \( (b = 0) \), \( A \) is symmetric, positive-definite and we have

\[
\| v \|^2 = B(v, v) = v^t A v = \| v \|^2_A.
\]

For finite element discretizations, the stiffness matrix \( A \) is large and sparse, and it is well-known that \( \kappa(A) \) grows without bound as \( N \) increases.

Standard finite element discretizations begin with a conforming triangulation (mesh) \( T \) consisting of:
- Triangles \( T \in T \).
- Edges \( e \in \mathcal{E} = \mathcal{E}_D \cup \mathcal{E}, \) with \( \mathcal{E}_D \) denoting those edges contained in \( \Gamma_D \), and \( \mathcal{E} \) denoting those edges which are not.
• Vertices \( z \in \mathcal{V} = \mathcal{V}_D \cup \mathcal{V} \) with \( \mathcal{V}_D \) denoting those vertices contained in \( \Gamma_D \), and \( \mathcal{V} \) denoting those vertices which are not.

Here, we have implicitly assumed that (the interior of) each boundary edge is contained entirely within either \( \Gamma_D \) or \( \Gamma_N \). The conformity condition on \( T \) means that for any two triangles \( T, T' \in T \), \( T \cap T' \) is either: empty, consists of a (complete) shared edge \( e \in \mathcal{E} \), or consists of a single shared vertex \( z \in \mathcal{V} \). We consider families of quasi-uniform, shape-regular meshes.

The \( p \)'th order, continuous Lagrange finite element spaces associated with the triangulation \( T \) are defined as

\[
S_p = \{ v \in C(\overline{\Omega}) \cap \mathcal{H} : v|_T \in \mathbb{P}_p \text{ for each } T \in T \},
\]

where \( \mathbb{P}_p \) is the collection of all polynomials of (total) degree \( \leq p \). For simple reference, we give the dimension of \( S_p \), \( |S_p| \), for \( p = 1, 2, 3, 4 \), in terms of the cardinalities of \( \mathcal{V}, \mathcal{E} \) and \( T \):  

\[
|S_1| = |\mathcal{V}|, \quad |S_2| = |\mathcal{V}| + |\mathcal{E}|, \quad |S_3| = |\mathcal{V}| + 2|\mathcal{E}| + |T|, \quad |S_4| = |\mathcal{V}| + 3|\mathcal{E}| + 3|T|.
\]

For quasi-uniform, shape-regular meshes \( \{ T_h \} \), and \( \hat{u} \in S_p \) satisfying (2.5) the standard \textit{a priori} error estimate is

\[
\| u - \hat{u} \| \lesssim h^p \| u \|_{p+1},
\]

where \( \| \cdot \|_j \) denotes the usual Sobolev norm on \( H^j(\Omega) \). For quasi-uniform meshes in \( \mathbb{R}^2 \) it holds that \( h \sim N^{-1/2} \), where \( N \) is the number of vertices, edges, or triangles in the mesh. In the case of adapted meshes, \( N^{-p/2} \) is considered the optimal convergence rate.

3. **Lagrange vs. Hierarchical Bases, Two-Level Block Gauss-Seidel.**

The most common basis for \( S_p \) is the so-called \textit{Lagrange (nodal) basis}. Given an appropriate set of (vertex, edge and interior) nodes, all basis functions are globally continuous, piecewise of degree \( p \), and the basis function associated with a given node has value 1 at that node and vanishes at all other nodes. These basis functions are most readily described on a given triangle in terms of its barycentric coordinates \( (\ell_1, \ell_2, \ell_3) \). Although in principle the \( \binom{p+2}{2} \) nodes on triangle \( T \) may be distributed in a number of ways—for example, at Fekete points [30]—we consider the usual case in which they are evenly distributed at the following barycentric coordinates

\[
\left\{ \left( \frac{i}{p}, \frac{j}{p}, \frac{k}{p} \right) : i, j \geq 0, \quad i + j + k = p \right\}.
\]

The corresponding basis functions, when restricted to \( T \), may be expressed as

\[
\psi^{(i,j,k)}(\cdot, \cdot, \cdot) = \left( \prod_{\ell=0}^{i-1} \frac{p\ell_1 - i'}{i - i'} \right) \left( \prod_{j'=0}^{j-1} \frac{p\ell_2 - j'}{j - j'} \right) \left( \prod_{k'=0}^{k-1} \frac{p\ell_3 - k'}{k - k'} \right).
\]

Such a function takes the value 1 at its associated node, and 0 at all other nodes. In contrast, a \textit{p-hierarchical basis} for \( S_p \) is built up from functions of various degrees. In this work we consider hierarchical bases of \( S_2 \) and \( S_4 \) suggested by the hierarchical splittings

\[
S_2 = S_1 \oplus (S_2 \setminus S_1), \quad S_4 = S_1 \oplus (S_2 \setminus S_1) \oplus (S_4 \setminus S_2).
\]
These matrices are singular, with respective nullspaces chosen many systematic ways of generating hierarchical bases (cf. [1, 2, 29, 13]), and we have and (sub-)spaces in these cases.

More precisely, on a triangle, the p-hierarchical basis of $S_2$ consists of linear functions $\psi_{\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)}$ associated with the nodes numbered 1, 2, 3 in Figure 3.1, and the quadratic functions $\psi_{\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)}$ associated with the nodes numbered 4, 5, 6. For $S_4$, we add to these the quartic functions $\psi_{\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)}$ associated with nodes 7 through 15. There are many systematic ways of generating hierarchical bases (cf. [1, 2, 29, 13]), and we have chosen $p = 2^n$ and the given splittings because of the convenient nesting of nodes and (sub-)spaces in these cases.

It is clear that the sparsity pattern of the stiffness matrix $A$ and the error estimate (2.9) are independent of the choice of basis in the computation of $\hat{u} \in S_p$, but we argue that a hierarchical basis can provide significant gains in terms of solving the associated linear systems. In order to provide an intuitive sense of why hierarchical bases may be advantageous, we first consider the $6 \times 6$ element stiffness matrices for $V = S_2$ for both the standard Lagrange as well as the p-hierarchical basis. As before, the vertex unknowns are ordered first. Let $(\theta_1, \theta_2, \theta_3)$ denote the interior angles of the triangle $T$ and define $s_j = \cot \theta_j$, $r_k = s_{k-1} + s_{k+1}$ and $s = s_1 + s_2 + s_3$. Then the element stiffness matrices are given by

$$A_T^{LB} = \frac{1}{6} \begin{pmatrix} 3r_1 & s_3 & s_2 & 0 & -4s_2 & -4s_3 \\ s_3 & 3r_2 & s_1 & -4s_1 & 0 & -4s_3 \\ s_2 & s_1 & 3r_3 & -4s_1 & -4s_2 & 0 \\ 0 & -4s_1 & -4s_1 & 8s & -8s_3 & -8s_2 \\ -4s_2 & 0 & -4s_2 & -8s_3 & 8s & -8s_1 \\ -4s_3 & -4s_3 & 0 & -8s_2 & -8s_1 & 8s \end{pmatrix}, \quad (3.1)$$

$$A_T^{HB} = \frac{1}{6} \begin{pmatrix} 3r_1 & -3s_3 & -3s_2 & -4r_1 & 4s_3 & 4s_2 \\ -3s_3 & 3r_2 & -3s_1 & 4s_3 & -4r_2 & 4s_1 \\ -3s_2 & -3s_1 & 3r_3 & 4s_2 & 4s_1 & -4r_3 \\ -4r_1 & 4s_3 & 4s_2 & 8s & -8s_3 & -8s_2 \\ 4s_3 & -4r_2 & 4s_1 & -8s_3 & 8s & -8s_1 \\ 4s_2 & 4s_1 & -4r_3 & -8s_2 & -8s_1 & 8s \end{pmatrix}. \quad (3.2)$$

These matrices are singular, with respective nullspaces $N(A_T^{LB}) = \text{span}\{(1, 1, 1, 1, 1, 1)\}$ and $N(A_T^{HB}) = \text{span}\{(1, 1, 1, 0, 0, 0)\}$, which leads to the global stiffness matrices being poorly conditioned. To put it another way, if the element stiffness matrices are

![Node numbering for bases of degrees 1, 2 and 4.](image)
non-singular, it can be proved that the corresponding global stiffness matrix is spec-
trally equivalent to its diagonal (cf. [3, 19]), requiring no sophisticated machinery to
solve the associated linear systems. In both cases, vectors in the nullspace represent
(not surprisingly) functions which are locally constant. The key difference is how
these constant functions are represented. In the first case, they are distributed evenly
among the basis functions, and in the second they are concentrated in the linear basis
functions. Suppose that the basei functions of $B^{LB}$ and $B^{HB}$ are ordered with those
associated with vertices before those associated with edges. This induces the natural
block structures

$$A^{LB} = \begin{pmatrix} A_{11}^{LB} & A_{12}^{LB} \\ A_{21}^{LB} & A_{22}^{LB} \end{pmatrix}, \quad A^{HB} = \begin{pmatrix} A_{11}^{HB} & A_{12}^{HB} \\ A_{21}^{HB} & A_{22}^{HB} \end{pmatrix}$$

(3.3)
on the global stiffness matrices. It can be shown that both $A_{11}^{LB}$ and $A_{22}^{LB}$ are well-
conditioned, so the ill-conditioning of $A^{LB}$ is due to strong coupling by the off-diagonal
blocks. In contrast, the ill-conditioning of $A^{HB}$ is “concentrated” in $A_{11}^{HB}$ while $A_{22}^{HB}$
is well-conditioned and the off-diagonal coupling can be shown to be mild because of a
strong Cauchy inequality in the $H^1$-inner-product between the spaces $S_1$ and $S_2 \setminus S_1$.
We explain this further below.

Motivated by the discussion above, we consider a generic splitting of $V = S_p$
$V = V_1 \oplus V_2$, $V_1 \cap V_2 = \{0\}$. Choosing bases for $V_1$ and $V_2$, this splitting induces a
natural $2 \times 2$ block structure for (2.6):

$$\begin{pmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$  

(3.4)

Throughout, we make the obvious identifications between coefficient vectors and
functions:

$$v \in \mathbb{R}^{|V|} \leftrightarrow v \in V, \quad v_1 \in \mathbb{R}^{|V_1|} \leftrightarrow v_1 \in V_1, \quad v_2 \in \mathbb{R}^{|V_2|} \leftrightarrow v_2 \in V_2,$$

often without additional clarification. The block Gauss-Seidel iteration for solving $A\mathbf{u} = \mathbf{f}$ is described in linear-algebraic terms by the iteration

$$M\mathbf{u}^{k+1} = \mathbf{f} + N\mathbf{u}^k, \quad M = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \quad N = -\begin{pmatrix} 0 & A_{12} \\ 0 & 0 \end{pmatrix},$$

(3.5)

and in terms of the bilinear form, by the iteration

$$B(u_1^{k+1}, v) = F(v) - B(u_2^k, v) \text{ for all } v \in V_1,$$

(3.6)

$$B(u_2^{k+1}, v) = F(v) - B(u_1^{k+1}, v) \text{ for all } v \in V_2.$$

(3.7)

When $B(\cdot, \cdot)$ is an inner-product, it can be shown (e.g. [14, 25, 3]) that there is
a constant $\gamma = \gamma(B, V_1, V_2) \in [0, 1)$ for which

$$\max_{v_1 \in V_1, v_2 \in V_2} \frac{|B(v_1, v_2)|}{\|v_1\|\|v_2\|} \leq \gamma,$$

(3.8)

and the error in each block Gauss-Seidel iteration is guaranteed to contract by (at
least) a factor of $\gamma^2$. This inequality is often called a strong Cauchy inequality, and
it has played an important role in the analysis of multi-level linear solvers, as well
as hierarchical basis error estimation. The optimal constant \( \gamma \) is often estimated in terms of local quantities \( \gamma_T \),

\[
\gamma \leq \max_{T \in \mathcal{T}} \gamma_T, \quad \max_{v_1 \in V_1, v_2 \in V_2} \frac{|B_T(v_1, v_2)|}{\|v_1\|_T \|v_2\|_T} \leq \gamma_T, \tag{3.9}
\]

where \( B_T(\cdot, \cdot) \) and \( \| \cdot \|_T \) are defined by restricting to the element \( T \) the integrals defining \( B(\cdot, \cdot) \) and \( \| \cdot \| \). It is clear that, if \( v_1 + v_2 \) is constant on \( T \), with \( v_j \in V_j \) and \( v_j \neq 0 \), then

\[
\frac{\| \int_T \nabla v_1 \cdot \nabla v_2 \, dx \|}{\| v_1 \|_T \| v_2 \|_T} = 1, \tag{3.10}
\]

which indicates that one should require that locally constant functions should lie entirely in one these spaces, which we take to be \( V_1 \). In terms of the Laplacian example above, the local constant \( \gamma_T \) for \( A_{LB}^T \) is 1, and it can be shown that the global constant \( \gamma \) for \( A_{LB}^T \) approaches 1 as the mesh is refined. In contrast, it is well-known [25] that \( \gamma_T \) for \( A_{LB}^T \) is given by

\[
\gamma_T^2 = \frac{1}{2} + \frac{1}{3} \sqrt{\frac{d-3}{4}}, \quad d = \sum_{k=1}^3 \cos^2 \theta_k
\]

where \( \theta_k \) denote the interior angles of the triangle element. If the angles in \( T \) are non-obtuse, then \( 3/4 \leq d \leq 1 \), so \( 1/2 \leq \gamma_T^2 \leq 2/3 \). The optimal value of \( 1/2 \) is achieved for equilateral triangles. For isosceles right triangles, \( \gamma_T^2 = 2/3 \).

At this stage we see that, at least in the case of symmetric problems, block Gauss-Seidel, either in the form presented or in its symmetric form, can provide an effective preconditioner for Krylov methods. A key benefit of hierarchical decompositions having \( S_1 \subset V_1 \) is that only the \( A_{11} \) block requires a “sophisticated” solver—such as a direct method, multi-grid method, or an \( H \)-matrix method as used in this work—for each Gauss-Seidel iteration. The remaining block is spectrally-equivalent to its own diagonal, and therefore simple to handle.

We now argue that the initial step of the block Gauss-Seidel iteration provides error reduction which is significantly greater than what can be explained by these by now standard contraction arguments. These arguments follow the pattern given in [27], which were given in the context of hierarchical error estimation.

**Lemma 3.1.** Suppose that \( u_0 = 0 \), and \( \phi = \phi_1 + \phi_2 \in V_2 \). Then

\[
\| u - u^1 \|_1 \lesssim \inf_{w \in V_2} \| u - u^1 - w \|_1 \leq \inf_{\phi \in V} (\| u - \phi \|_1 + \| \phi_1 - u^1_1 \|_1). 
\]

The constant hidden in \( \lesssim \) is determined by the constants of boundedness (continuity) and either coercivity or the inf-sup condition.

**Proof.** For the sake of clarity, let us assume that \( B \) is coercive, with constant \( m \), and that \( M \) is its continuity constant. Using Galerkin orthogonality, we see that

\[
m\| u - u^1 \|_1 \leq B(u - u^1, u - u^1) = B(u - u^1, u - u^1 - v) \leq M\| u - u^1 \|_1 \| u - u^1 - v \|_1,
\]

for all \( v \in V_2 \). Setting \( w = u^1_2 + v \) establishes the first inequality. Now, for any \( \phi = \phi_1 + \phi_2 \), we see that

\[
\inf_{w \in V_2} \| u - u^1 - w \|_1 = \inf_{w \in V_2} \| u - \phi + \phi_1 - u^1 + \phi_2 - w \|_1 \leq \| u - \phi \|_1 + \| \phi_1 - u^1_1 \|_1,
\]
which completes the proof. □

To obtain a practical estimate from Lemma 3.1, one generally chooses \( \phi \) so that it and \( \phi_1 \) are both suitable (quasi-)interpolants of \( u \). Our arguments below employ super-convergence results of the sort found in [4, 31, 22], which in turn assume a mild approximate mesh symmetry condition. Although the results we present below allow for Neumann boundary conditions, for simplicity the mesh condition below is stated for homogeneous Dirichlet boundary conditions—the more technical condition for boundary edges/vertices can be found in [4].

**Definition 3.2** (\( O(h^{2\sigma}) \)-irregular Triangulation). Let \( \mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2 \) denote the set of interior edges in \( \mathcal{T} = \mathcal{T}_h \). For each \( e \in \mathcal{E}_1 \), the adjacent triangles \( T \) and \( T' \) form an approximate parallelogram, in which the opposite sides differ in length by \( O(h^2) \). For \( e \in \mathcal{E}_2 \), we have the overall bound

\[
\sum_{e \in \mathcal{E}_2} (|T| + |T'|) = O(h^{2\sigma}).
\]

**Theorem 3.3.** Suppose that \( \mathcal{T} = \mathcal{T}_h \) satisfies the approximate mesh symmetry condition above, and \( u \) is sufficiently regular. In the case \( \mathcal{V}_1 = \mathcal{S}_1 \) and \( \mathcal{V} = \mathcal{S}_p \) for some \( p > 1 \), the standard nodal quadratic interpolant \( u_q = u_\ell + u_b \), \( u_\ell \in \mathcal{S}_1 \), \( u_b \in \mathcal{S}_2 \setminus \mathcal{S}_1 \), satisfies

\[
\| u - u^1 \|_1 \lesssim \| u - u_q \|_1 + \| u_\ell - u_b \|_1 \lesssim h^{1+\min(1, \sigma)} \log h^{1/2} \| u \|_{W^{3, \infty}(\Omega)}. \tag{3.11}
\]

In the case \( \mathcal{V}_1 = \mathcal{S}_2 \) and \( \mathcal{V} = \mathcal{S}_p \) for some \( p > 2 \), there is a cubic quasi-interpolant \( u_c = u_q + u_w \), \( u_q \in \mathcal{S}_2 \) (not the nodal interpolant), \( u_w \in \mathcal{S}_3 \setminus \mathcal{S}_2 \) such that

\[
\| u - u^1 \|_1 \lesssim \| u - u_q \|_1 + \| u_\ell - u_q \|_1 \lesssim h^{2+\min(1/2, \sigma)} (\| u \|_4 + \| u \|_{W^{3, \infty}(\Omega)}). \tag{3.12}
\]

**Proof.** See [4] for (3.11) and [22] for (3.12). □

The results of primary interest from [4, 22] concern what we have denoted as \( \| u^1 - \phi_1 \|_1 \) in Lemma 3.1. In both cases \( u^1 = u_h \) is the finite element solution. Those comparing [4] and [22] should note that what is called \( 2\sigma \) here and in [4] is called \( \sigma \) in [22]. Careful reading of these papers reveals when one can (slightly) relax the regularity assumptions in their proofs, and under what conditions optimal convergence rates can be expected. In brief, it is not unreasonable in practice to expect to see \( \| u - u^1 \|_1 = O(h^2) \) or \( \| u - u^1 \|_1 = O(h^2) \log h^{1/2} \) when \( \mathcal{V}_1 = \mathcal{S}_1 \), and \( \| u - u^1 \|_1 = O(h^2) \) or \( \| u - u^1 \|_1 = O(h^2) \log h^{1/2} \) when \( \mathcal{V}_1 = \mathcal{S}_2 \). Even in more realistic situations in which \( u \) possesses a few isolated singularities, and the triangulation is appropriately refined near those singularities, similar convergence is observed when one replaces \( h \) with \( N^{-1/2} \), where \( N \) is the number of vertices, edges, or triangles. A partial explanation of this phenomenon, in the case of the Dirichlet Laplacian, is provided in [10]. The proofs of such superconvergence properties are generally quite technical and so have been limited to lower-order finite elements. One difficulty in extending these arguments to higher-order elements is finding a proper choice of quasi-interpolant—it is known that the standard Lagrange interpolant does not yield the desired superconvergence properties for higher-order elements [24]. Finally, it is clear that Lemma 3.1 is dimension-independent, so any superconvergence result of the sort above which holds in \( \mathbb{R}^n \), would imply a similarly large first step in the corresponding block Gauss-Seidel iteration. One such superconvergence result for mildly structured tetrahedral meshes in \( \mathbb{R}^3 \) is provided in [9].
4. Concerning Hierarchical Matrices. The block Gauss-Seidel iteration (3.5) requires the solution of systems of equations involving the diagonal blocks $A_{11}$ and $A_{22}$. In the numerical tests in the following section 5, both systems will be solved only approximately, i.e., they will be replaced by approximations $\tilde{A}_{11}$ and $\tilde{A}_{22}$. The $2 \times 2$ splitting of the matrix results in an ill-conditioned block $A_{11}$ and a well-conditioned block $A_{22}$. We therefore propose to construct $A_{11}$ as a (highly accurate) approximate Cholesky factorization using the technique of hierarchical ($H$-)matrices to be described in this section. Since $A_{22}$ is well-conditioned, it can be replaced by an approximation $\tilde{A}_{22}$ representing a few steps of an inner iteration.

In the remainder of this section, we will specify the particular variant of $H$-matrix construction which is proposed for the $H$-Cholesky (or $H$-LU factorization) of the $A_{11}$-block, and point to the literature for further details on $H$-matrices.

$H$-matrices were introduced in [20] and since then entered into a wide range of applications. The basic $H$-matrix construction and corresponding arithmetic have reached a relatively mature state and are documented in the comprehensive lecture notes [7] and books [6, 21].

$H$-matrices are based on a hierarchical subdivision of the matrix into subblocks and low-rank approximations of matrix data within (most of) these subblocks. Originally, $H$-matrices were introduced in the context of fully populated matrices arising from solution operators of elliptic differential equations and in boundary element methods. In the finite element context, the stiffness matrix itself does not require an efficient approximation by an $H$-matrix since it is sparse. Its Cholesky- or LU-factors, however, suffer from unacceptable fill-in if computed exactly and can be efficiently computed or approximated by $H$-Cholesky or $H$-LU factors, resp. [18, 5]. The $H$-matrix construction and arithmetic as originally developed for fully populated matrices have a straightforward generalization to sparse matrices. However, there are two modifications for $H$-matrices which have been designed for sparse matrices in particular and have been used here.

The first modification concerns the construction of the block structure of the $H$-matrix. In the classical $H$-matrix, the block structure is generated through a repeated bisection of the respective index sets, i.e., row and column index sets are divided into two subsets, respectively, which leads to four matrix subblocks. In the case of sparse matrices, the bissection has been replaced by a nested dissection approach in which row and column index sets are divided into three subsets each; two subsets $S_1$, $S_2$ of indices that are pairwise disconnected in the sense that $a_{ij} = 0 = a_{ji}$ if $i \in S_1$, $j \in S_2$ where $A = (a_{ij})$ denotes the stiffness matrix, and a third subset $S_3$ containing the remaining indices of the interior boundary. Such a subdivision results in a $3 \times 3$ matrix structure with zero blocks in the $1 \times 2$ and $2 \times 1$ positions. These zero blocks remain zero in a subsequent Cholesky or LU factorization which results in considerably faster ($H$-)Cholesky factorizations compared to bisection-based $H$-matrices [23, 17]. The blocks are subdivided recursively until a minimum blocksize $n_{\text{min}}$ is reached, i.e., a block is further subdivided if the minimum of its number of rows and columns is greater than $n_{\text{min}}$. In the subsequent numerical results in section 5, we set $n_{\text{min}} = 8$.

The second modification concerns the development of a “blackbox” clustering algorithm. The classical construction of $H$-matrices requires geometric information associated with the underlying indices in order to determine a suitable block structure. For sparse matrices, the information contained in the associated matrix graph can replace the need for geometric information [16].

Whereas the classical $H$-matrix uses a fixed rank for the low rank approximations
within matrix subblocks, it is possible to replace it by adaptive ranks in order to enforce a desired accuracy within the individual blocks. In particular, given a matrix block $C$ and a desired $H$-accuracy $0 < \delta_H < 1$, we set the rank $k_C$ of the approximation to $C$ as

$$k_C := \min\{k' \mid \sigma_{k'} \leq \delta_H \sigma_1\}$$

where $\sigma_i$ denotes the $i$'th largest singular value of $C$. These modifications from the classical $H$-matrix setting have led to highly efficient $H$-Cholesky/$H$-LU preconditioners for a wide range of sparse matrices [15].

5. Numerical Experiments. As a model problem, we consider the Laplace problem

$$-\Delta u = f \text{ in } \Omega = (-1,1) \times (-1,1) \text{ with } u = 0 \text{ on } \partial \Omega. \quad (5.1)$$

We choose the right hand side function $f$ so that the exact solution is known to be

$$u = \cos\left(\frac{\pi}{2} x\right) \cos\left(\frac{\pi}{2} y\right) \quad (5.2)$$

with semi-norms

$$|u|_k^2 = \int_\Omega \sum_{j=0}^{k} \binom{k}{j} \left(\frac{\partial^k u}{\partial x^{k-j} \partial y^j}\right)^2 = \pi^{2k} \cdot \frac{2^k}{2^k}.$$
such $H$-Cholesky factorizations for various adaptive $H$-accuracies $\delta_H$. Both the setup times and storage requirements increase almost linearly in the problem size. The set-up times will be put into perspective and be compared with the iteration times in later experiments.

Table 5.1

<table>
<thead>
<tr>
<th>$\delta_H$</th>
<th>setup time (seconds)</th>
<th>storage (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1e-2$</td>
<td>$1e-4$</td>
</tr>
<tr>
<td>9,801</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>19,321</td>
<td>1.5</td>
<td>1.7</td>
</tr>
<tr>
<td>39,601</td>
<td>3.5</td>
<td>4.0</td>
</tr>
<tr>
<td>77,841</td>
<td>7.1</td>
<td>8.2</td>
</tr>
<tr>
<td>159,201</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>312,481</td>
<td>35</td>
<td>38</td>
</tr>
<tr>
<td>638,401</td>
<td>71</td>
<td>83</td>
</tr>
</tbody>
</table>

In Figures 5.1 and 5.3, we show the energy norm of the iteration error for quadratic and quartic finite elements, resp., on a log-log scale, i.e., given the exact solution $u$ (5.2) and the iterative solution $\hat{u} \in V$ as determined by its coefficient vector $u^i \in \mathbb{R}^N$, we approximate the energy norm $\|e^i\|$ of the iteration error $e^i = u - u^i$ using a 25-point quadrature on each triangle in $\Omega_h$. We show results for various accuracies for the solvers for the diagonal blocks $A_{11}$ and $A_{22}$. The block $A_{11}$ is replaced by its $H$-Cholesky factors using $H$-accuracies $1e-2$, $1e-4$ and $1e-8$, resp.. The block $A_{22}$ is either replaced by its upper triangular part (backward Gauss-Seidel preconditioner) - which is denoted by “inner=1”, or by (4 steps of) an inner (Jacobi preconditioned) GMRES iteration - which is denoted by “inner=4”. The results in Fig. 5.1 show that an insufficient accuracy for solving the first diagonal block (i.e., $H$-acc=$1e-2$), does not lead to a successful method. When we solve the first diagonal block with an accuracy of $H$-acc=$1e-4$ (or better), we clearly see the big decrease in the first iteration step. If we solve both diagonal blocks highly accurately, i.e., $H$-acc=$1e-8$ for $A_{11}$ and four steps of an inner GMRES method to solve for $A_{22}$, then the iteration error is
essentially of the same order as the discretization error within the first iteration step which is also supported by the following Figure 5.2. It illustrates the energy norm of the iteration error after the first step for various problem sizes. Here, both diagonal blocks have been solved very accurately (H-acc=1e-8, inner=4 and inner=8, resp.).

![Fig. 5.2. H1-error after the first iteration step for quadratic elements](image1)

In Fig. 5.3, we repeat this set of tests using quartic finite elements in the discretization (and up to 8 steps in the inner GMRES iteration for \( A_{22} \)). Initially, \( A_{11} \) contains the coupling of linear basis functions whereas \( A_{22} \) contains both quadratic and quartic basis functions. We see a big decrease of the error in the first iteration step as long as we choose an H-accuracy H-acc of 1e-4 or smaller for \( \tilde{A}_{11} \). If the \( A_{22} \) block is not solved accurately (i.e., inner=1), then the better accuracy for \( A_{11} \) does not help to improve the convergence. In fact, the results for H-acc=1e-4, inner=1 and H-acc=1e-8, inner=1, are identical (the filled circle within the empty square almost appears to be a filled square).

![Fig. 5.3. H1-error per iteration step for quartic elements](image2)

While the first step is quite large, Fig. 5.4 shows that this step is only of order \( \mathcal{O}(h^2) \), i.e., the same as for quadratic elements (here, the problem sizes are N=150201,
Le Borne, Ovall

312481 and 638401, resp.). This is in agreement with Lemma 3.1. In order to achieve a higher order of convergence in the first step, we simply regroup our basis functions to include the quadratic functions with the linear functions in block $A_{11}$. In this case, we see an error of $O(h^3)$ after the first step which is also shown in Fig. 5.4 (curve with filled squares).

We have also performed experiments for a convection-diffusion problem

$$-\epsilon \Delta u + b \cdot \nabla u = f \quad \text{in } \Omega = (-1,1) \times (-1,1) \quad \text{with } u = 0 \text{ on } \partial\Omega \tag{5.3}$$

with a recirculating convection

$$b(x, y) = \begin{pmatrix}
\frac{0.1 e^{0.1 y}}{2 \pi (e^y - 1)} \sin \left( \frac{2 \pi (e^{0.1 y} - 1)}{e^y - 1} \right) \\
-\frac{4 e^x}{2 \pi (e^x - 1)} \sin \left( \frac{2 \pi (e^{4 x} - 1)}{e^x - 1} \right) \\
\end{pmatrix}
\begin{pmatrix}
1 - \cos \left( \frac{2 \pi (e^{0.1 y} - 1)}{e^y - 1} \right) \\
1 - \cos \left( \frac{2 \pi (e^{4 x} - 1)}{e^x - 1} \right)
\end{pmatrix}.
$$

Moderate convection $\epsilon \geq 10^{-3}$ in (5.3) hardly causes any difference in the iteration errors at all, and results very similar to those in Fig. 5.1 are shown in Figure 5.5 (strong convection $\epsilon \leq 10^{-4}$ led to meaningless results since we did not implement any upwinding in our discretization).

Finally, we performed tests using the block Gauss-Seidel preconditioned GMRES method. In the following tests, $A_{11}$ is again approximated by its $H$-Cholesky factors (of varying accuracy), and $A_{22}$ is simply replaced by its lower triangular part. We did not perform inner iteration for $A_{22}$ within the solver since the above approximation leads to quite good results already so that we do not expect any (timewise) improvements through inner iterations. In addition to the setup times for the $H$-Cholesky factorization of $A_{11}$, we show the time necessary to reduce the initial residual by a factor of $1e^{-6}$. Figure 5.6 shows the results for quadratic elements while Fig. 5.7 shows results for quartic elements. In either case we observe a significant acceleration when we improve the $H$-accuracy for the $H$-Cholesky factorization of $A_{11}$ from $1e^{-2}$ down to $1e^{-4}$. A further improvement down to $1e^{-8}$ yields only a slight improvement for quadratic elements and none for quartic elements. Whereas the setup times dominate the actual iteration times for quadratic elements, the setup and iteration times are quite similar for quartic elements.
REFERENCES


Fig. 5.7. for quartic elements


