

## Using approximate inverses in algebraic multilevel methods<sup>\*</sup>

Y. Notay

Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165), 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium; e-mail: ynotay@ulb.ac.be

Received March 3, 1997 / Revised version received July 16, 1997

**Summary.** This paper deals with the iterative solution of large sparse symmetric positive definite systems. We investigate preconditioning techniques of the two-level type that are based on a block factorization of the system matrix. Whereas the basic scheme assumes an exact inversion of the submatrix related to the first block of unknowns, we analyze the effect of using an approximate inverse instead. We derive condition number estimates that are valid for any type of approximation of the Schur complement and that do *not* assume the use of the hierarchical basis. They show that the two-level methods are stable when using approximate inverses based on modified ILU techniques, or explicit inverses that meet some row-sum criterion. On the other hand, we bring to the light that the use of standard approximate inverses based on convergent splittings can have a dramatic effect on the convergence rate. These conclusions are numerically illustrated on some examples.

*Mathematics Subject Classification (1991):* 65F10, 65B99, 65N20

### 1. Introduction

To solve large sparse symmetric positive definite systems

$$(1.1) \quad \mathbf{A}\mathbf{u} = \mathbf{b}$$

arising from discrete elliptic PDEs, many recent works focus on the design of efficient algebraic multilevel preconditioners. Like standard multigrid [18,30], these are based on a two-level method recursively used in a  $V$  or

---

<sup>\*</sup> Supported by the “Fonds National de la Recherche Scientifique”, Chercheur qualifié

$W$  cycle algorithm. However, the basic scheme originates here from a block factorization of the system matrix

$$(1.2) \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & \\ & S_A \end{pmatrix} \begin{pmatrix} I & A_{11}^{-1} A_{12} \\ & I \end{pmatrix},$$

where the first block of unknowns corresponds to the fine grid nodes, the second block of unknowns to the coarse grid nodes, and where  $S_A = A_{22} - A_{21} A_{11}^{-1} A_{12}$  is the Schur complement of  $A$ ; since the latter is a dense matrix, some relevant sparse approximation  $S$  has to be supplied and the preconditioner writes

$$(1.3) \quad \tilde{B} = \begin{pmatrix} A_{11} & \\ & S \end{pmatrix} \begin{pmatrix} I & A_{11}^{-1} A_{12} \\ & I \end{pmatrix}.$$

Whether considered as a “stand alone” method or as an intermediate step in a full multilevel cycle, the potentialities of this two-level scheme essentially depend on the spectral condition number

$$(1.4) \quad \kappa(\tilde{B}^{-1}A) = \frac{\lambda_{\max}(\tilde{B}^{-1}A)}{\lambda_{\min}(\tilde{B}^{-1}A)}$$

which should be close to 1 and bounded independently of the grid size <sup>1</sup>.

A straightforward computation shows that

$$(1.5) \quad \tilde{B}^{-1}A = \begin{pmatrix} I & * \\ S^{-1} & S_A \end{pmatrix}.$$

and an essential step in the analysis consists therefore in proving  $O(1)$  upper and lower bounds on the spectrum of  $S^{-1}S_A$ . Such results exist for the methods that originate from the hierarchical basis multigrid method [5, 7–11, 26–28, 31], but also for some other approaches where  $S$  is computed according an incomplete Gaussian elimination process [23, 29].

Now, the preconditioner (1.3) involves solving two systems with  $A_{11}$  each time it is needed to solve a system with  $\tilde{B}$ . Some works pay little attention to that point. The key argument is that  $A_{11}$  has condition number  $O(1)$  [1, 5, 10], so that any reasonable iterative scheme converges quickly in a number of steps independent of the grid size. In practice however, the cost of such inner iterations becomes rapidly prohibitive. Most works take this into account and consider preconditioners of the form

$$(1.6) \quad B = \begin{pmatrix} P & \\ & S \end{pmatrix} \begin{pmatrix} I & P^{-1} A_{12} \\ & I \end{pmatrix},$$

<sup>1</sup> Throughout this paper,  $\lambda_{\max}(C)$  and  $\lambda_{\min}(C)$  denote respectively the largest and the smallest eigenvalue of  $C$

where  $P^{-1}$  stands for the used approximate inverse of  $A_{11}$ , possibly implicitly defined by a very few steps of some iterative procedure. The argument about the conditioning of  $A_{11}$  serves then only to show that it is easy to define cost effective approximate inverses such that the spectrum of  $P^{-1} A_{11}$  is very well clustered around 1.

Preconditioners of the form (1.6) also arise in another family of algebraic multilevel methods [1,3,4,6,24], whose primary step consists in approximating  $A_{11}^{-1}$  by some sparse matrix  $B_{11}$ . The basic scheme assumes then that the Schur complement  $A_{22} - A_{21} B_{11} A_{12}$  of the so approximated matrix is computed exactly, but some recent developments [6,13] introduce versions for which  $B_{11}$  is less sparse, leading to a block incomplete factorization (1.6) in which both  $A_{11}$  and the Schur complement are approximated.

Now, although numerous works analyze two-level methods of the type (1.6), so far available theoretical results involve the so-called strengthened CBS constant [9,10,17] and can therefore be useful only when  $A$  is a finite element matrix computed using the *hierarchical* finite element basis functions.

In fact, as is well known [26], the Schur complement  $S_A$  is the same for both nodal and hierarchical representations. Hence, as easily seen from (1.5), these analyzes apply to the version (1.3) with exact inversion of  $A_{11}$  whatever the used basis.

This is however not true anymore in the general case. In [8,26,27], it is shown that the equivalence of the preconditioner (1.6) in the usual nodal basis with its counterpart in the hierarchical basis can be restored by adding to  $A_{12}$  and  $A_{21}$  a term equal to  $(A_{11} - P)$  times some interpolation matrix. As  $(A_{11} - P)$  is small when a sufficiently accurate approximate inverse is used, the question of the necessity of this additional term is raised [8]. The purpose of the present paper is to develop a finer analysis of this issue.

In this view, we investigate the eigenvalue distribution associated to preconditioners (1.6) in function of the spectral properties of  $P^{-1} A_{11}$  and  $S^{-1} S_A$ , using only assumptions compatible with  $A$  defined from a standard finite difference or finite element scheme.

If  $A$  is a (non strictly) diagonally dominant M-matrix, our analysis proves in particular the stability of the two-level method when  $P$  is a modified (possibly blockwise) ILU factorization of  $A_{11}$  (e.g. [2,14–16,19]), or  $P^{-1}$  an explicit approximate inverse of  $A_{11}$  satisfying some row-sum criterion [2, Chapter 8].

On the other hand, we show that the use of standard approximate inverses based on convergent splittings can have a much more dramatic effect than expected, leading to a condition number that may grow up to  $O(h^{-4})$  when the mesh size parameter  $h$  decreases.

The paper is organized as follows: in Sect. 2, we develop a simple analysis of the behavior of the condition number when  $P^{-1}$  is a standard approximate inverse of  $A_{11}$ ; our main theoretical results are proved in Sect. 3, and their application to approximate inverses satisfying a row-sum criterion is discussed in Sect. 4. Section 5 is devoted to numerical results.

### *Terminology and notation*

Throughout this paper, inequalities between matrices or vectors of the same dimensions are to be understood elementwise, whereas *positive (nonnegative)*, when applied to a matrix or a vector, means *elementwise positive (nonnegative)*.

## 2. Standard approximate inverses

In this section, we derive a *lower* bound on the condition number by evaluating the Rayleigh quotients

$$\bar{r} = \frac{\bar{\mathbf{v}}^T A \bar{\mathbf{v}}}{\bar{\mathbf{v}}^T B \bar{\mathbf{v}}} \quad , \quad \underline{r} = \frac{\underline{\mathbf{v}}^T A \underline{\mathbf{v}}}{\underline{\mathbf{v}}^T B \underline{\mathbf{v}}}$$

associated to vectors of the form

$$\bar{\mathbf{v}} = \begin{pmatrix} -P^{-1} A_{12} \mathbf{v}_2 \\ \mathbf{v}_2 \end{pmatrix} \quad , \quad \underline{\mathbf{v}} = \begin{pmatrix} -A_{11}^{-1} A_{12} \mathbf{v}_2 \\ \mathbf{v}_2 \end{pmatrix} .$$

A straightforward computation shows that

$$\bar{\mathbf{v}}^T A \bar{\mathbf{v}} = \mathbf{v}_2^T S_A \mathbf{v}_2 + \mathbf{v}_2^T A_{21} \left( A_{11}^{-1} - 2P^{-1} + P^{-1} A_{11} P^{-1} \right) A_{12} \mathbf{v}_2$$

$$\bar{\mathbf{v}}^T B \bar{\mathbf{v}} = \mathbf{v}_2^T S \mathbf{v}_2 ,$$

$$\underline{\mathbf{v}}^T A \underline{\mathbf{v}} = \mathbf{v}_2^T S_A \mathbf{v}_2 ,$$

$$\underline{\mathbf{v}}^T B \underline{\mathbf{v}} = \mathbf{v}_2^T S \mathbf{v}_2 + \mathbf{v}_2^T A_{21} \left( P^{-1} - 2A_{11}^{-1} + A_{11}^{-1} P A_{11}^{-1} \right) A_{12} \mathbf{v}_2 .$$

Further,  $A_{11}^{-1} - 2P^{-1} + P^{-1} A_{11} P^{-1} = (I - P^{-1} A_{11})^2 A_{11}^{-1}$ , whereas, letting  $\lambda_M = \lambda_{\max}(P^{-1} A_{11})$ , one has for all  $\mathbf{w}_1$

$$\begin{aligned} & \mathbf{w}_1^T \left( P^{-1} - 2A_{11}^{-1} + A_{11}^{-1} P A_{11}^{-1} \right) \mathbf{w}_1 \\ &= \mathbf{w}_1^T \left( A_{11}^{-1} P \right) \left( P^{-1} A_{11} P^{-1} - 2P^{-1} + A_{11}^{-1} \right) \mathbf{w}_1 \\ &\geq \lambda_M^{-1} \mathbf{w}_1^T \left( I - P^{-1} A_{11} \right)^2 A_{11}^{-1} \mathbf{w}_1 . \end{aligned}$$

Hence, with

$$q_{\mathbf{v}_2} = \frac{\mathbf{v}_2^T S_A \mathbf{v}_2}{\mathbf{v}_2^T S \mathbf{v}_2},$$

$$g_{\mathbf{v}_2} = \frac{\mathbf{v}_2^T A_{21} (I - P^{-1} A_{11})^2 A_{11}^{-1} A_{12} \mathbf{v}_2}{\mathbf{v}_2^T S_A \mathbf{v}_2},$$

one obtains

$$\bar{r} = q_{\mathbf{v}_2} (1 + g_{\mathbf{v}_2}), \quad r \leq \frac{1}{q_{\mathbf{v}_2}^{-1} + \lambda_M^{-1} g_{\mathbf{v}_2}}.$$

Since  $\lambda_{\max}(B^{-1} A) \geq \bar{r}$ ,  $r \geq \lambda_{\min}(B^{-1} A)$ , this implies

$$(2.1) \quad \kappa(B^{-1} A) \geq (1 + g_{\mathbf{v}_2}) \left(1 + \frac{q_{\mathbf{v}_2}}{\lambda_M} g_{\mathbf{v}_2}\right),$$

showing that the two-level method can be stable with respect to the use of the approximate inverse only if  $g_{\mathbf{v}_2}$  is bounded above independently of the mesh size.

Now, as is well known, in discrete PDE applications,  $S_A$  is spectrally equivalent to the coarse grid discretization matrix. Therefore,  $\mathbf{v}_2^T S_A \mathbf{v}_2 = O(h^2)$  for smooth vectors, which further implies  $\mathbf{v}_2^T A_{21} A_{11}^{-1} A_{12} \mathbf{v}_2 \simeq \mathbf{v}_2^T A_{22} \mathbf{v}_2 = O(1)$  when one does *not* make use of the hierarchical basis. Hence,  $g_{\mathbf{v}_2}$  can be bounded independently of the mesh size only if  $P^{-1}$  acts nearly as an exact inverse for these *smooth* vectors, whereas usual convergent splittings are on the contrary known to converge fast for *rough* vectors [18, 30].

To illustrate this, let  $A$  be more specifically the matrix associated with the standard five point finite difference approximation of the Laplacian on the unit square with Dirichlet boundary conditions and a uniform mesh size  $h$  in both directions, where  $h$  is such that  $h^{-1}$  is an even integer.

Assume further for simplicity that  $P^{-1}$  is the approximate inverse resulting from  $m \geq 1$  stationary iterations with the Jacobi preconditioning, i.e.

$$(I - P^{-1} A_{11}) = (M^{-1} N)^m$$

where  $M = \text{diag}(A_{11}) = 4I$  and  $N = M - A_{11} = -\text{offdiag}(A_{11})$ .

The vector  $\mathbf{v}_2^{(\nu\mu)}$  which interpolates on the coarse grid nodes the function  $\sin \nu\pi x \sin \mu\pi y$  is shown in [22] to be an eigenvector of  $S_A$  for  $\nu = 1, \dots, \frac{h^{-1}}{2} - 1, \mu = 1, \dots, \frac{h^{-1}}{2} - 1$ . The corresponding eigenvalue writes

$$\lambda_{\nu\mu} = 4 \left( \frac{1}{4 - 2(c_\nu + c_\mu)} + \frac{1}{4 - 2(c_\nu - c_\mu)} + \frac{1}{4 + 2(c_\nu - c_\mu)} + \frac{1}{4 + 2(c_\nu + c_\mu)} \right)^{-1}$$

where  $c_\nu = \cos \nu\pi h$ ,  $c_\mu = \cos \mu\pi h$ . For  $\nu = \mu$ , it takes the simpler expression

$$\lambda_{\nu\nu} = \frac{8(1 - c_\nu^2)}{2 - c_\nu^2},$$

and it is also easy to check that

$$(M^{-1}N)^2 (A_{12} \mathbf{v}_2^{(\nu\nu)}) = \frac{c_\nu^2}{2} (A_{12} \mathbf{v}_2^{(\nu\nu)}).$$

Hence, since  $A_{22} = 4I$  and  $(M^{-1}N)^T = M^{-1}N$ ,

$$\begin{aligned} g_{\mathbf{v}_2^{(\nu\nu)}} &= \left(\frac{c_\nu^2}{2}\right)^m \frac{\mathbf{v}_2^{(\nu\nu)t} (A_{22} - S_A) \mathbf{v}_2^{(\nu\nu)}}{\mathbf{v}_2^{(\nu\nu)t} S_A \mathbf{v}_2^{(\nu\nu)}} \\ &= \left(\frac{c_\nu^2}{2}\right)^m \frac{c_\nu^2}{2(1 - c_\nu^2)}, \end{aligned}$$

that is, for the smoothest mode  $\nu = 1$

$$g_{\mathbf{v}_2^{(11)}} \simeq \left(\frac{1}{2}\right)^{m+1} \frac{h^{-2}}{\pi^2}$$

showing with (2.1) that  $\kappa(B^{-1}A) \geq O(h^{-4})$  except if one increases  $m$  as the mesh is refined.

*Remark.*  $g_{\mathbf{v}_2} = O(h^{-2})$  implies  $\kappa(B^{-1}A) \geq O(h^{-4})$  because we implicitly assume that  $q_{\mathbf{v}_2} = O(1)$ . On the other hand,  $q_{\mathbf{v}_2} = O(h^2)$ , which can be easily obtained (e.g.  $S = A_{22}$ ), leads then to a lower bound  $\kappa(B^{-1}A) \geq O(h^{-2})$ . Thus, a worse coarse grid approximation potentially yields a better method. Note however that only a *lower* bound on the condition number is improved, which gives no guarantee on the behavior of the actual conditioning. As  $\kappa(B^{-1}A) \geq O(h^{-2})$  remains anyway by far too large, we leave this point as an open question.

### 3. Eigenvalue bounds

In this section, we analyze the spectrum of  $B^{-1}A$ . The only restrictive assumptions are

$$(3.1) \quad \mathbf{v}_1^T P \mathbf{v}_1 \leq \mathbf{v}_1^T A_{11} \mathbf{v}_1 \quad \forall \mathbf{v}_1$$

and

$$(3.2) \quad \begin{aligned} \mathbf{v}_2^T A_{21} P^{-1} A_{12} \mathbf{v}_2 &\leq (1 - \xi) \mathbf{v}_2^T A_{22} \mathbf{v}_2 \\ &+ \xi \mathbf{v}_2^T A_{21} A_{11}^{-1} A_{12} \mathbf{v}_2 \quad \forall \mathbf{v}_2 \end{aligned}$$

for some  $\xi < 1$ . This is relatively weak since  $\xi$  may be negative, but nevertheless sufficient to entail an acceptable behavior, as it can be seen on an intuitive basis: whenever  $\mathbf{v}_2^T A_{22} \mathbf{v}_2 = (1 + O(h^2)) \mathbf{v}_2^T A_{21} A_{11}^{-1} A_{12} \mathbf{v}_2$ , (3.1), (3.2) imply

$$\begin{aligned} \mathbf{v}_2^T A_{21} A_{11}^{-1} A_{12} \mathbf{v}_2 &\leq \mathbf{v}_2^T A_{21} P^{-1} A_{12} \mathbf{v}_2 \\ &\leq \left(1 + (1 - \xi) O(h^2)\right) \mathbf{v}_2^T A_{21} A_{11}^{-1} A_{12} \mathbf{v}_2 , \end{aligned}$$

i.e.  $P^{-1}$  has to act nearly as an exact inverse for the corresponding modes as long as  $-\xi \leq O(1)$ .

How to compute a lower bound on  $\xi$  for some classes of preconditioners that meet (3.1) is addressed in the next section.

Letting

$$\begin{aligned} \zeta &= \lambda_{\min}^{-1} \left( S^{-1} S_A \right) , \\ \eta &= \lambda_{\max}^{-1} \left( S^{-1} S_A \right) , \\ \beta &= \lambda_{\max}^{-1} \left( P^{-1} A_{11} \right) = \kappa^{-1} \left( P^{-1} A_{11} \right) , \end{aligned}$$

Theorem 3.1 below proves upper and lower eigenvalue bounds for  $B^{-1}A$  that are only function of  $\zeta, \eta, \beta$  and  $\xi$ . Their exact expression looks complicated, but a further analysis is provided, leading to the following simplified but nevertheless rigorous bounds (3.9), (3.10):

$$\begin{aligned} \lambda_{\max} \left( B^{-1}A \right) &\leq \frac{\eta + 1 - \xi(1 - \beta)}{\eta \beta} , \\ \lambda_{\min} \left( B^{-1}A \right) &\geq \zeta^{-1} \left( 1 + \frac{(1 - \beta)(1 - \xi)}{\zeta - \beta} \right)^{-1} . \end{aligned}$$

Hence, since we assume  $\eta \leq 1 \leq \zeta$ ,

$$\begin{aligned} \kappa \left( B^{-1}A \right) &\leq \frac{\zeta}{\eta \beta} (2 - \xi(1 - \beta)) (2 - \xi) \\ &= \kappa \left( S^{-1} S_A \right) \kappa \left( P^{-1} A_{11} \right) (2 - \xi(1 - \beta)) (2 - \xi) . \end{aligned}$$

Thus, when our basic assumptions (3.1), (3.2) are satisfied with  $-\xi \leq O(1)$ , the problems illustrated in Sect. 2 are prevented and the condition number of  $B^{-1}A$  cannot be much larger than the product of the condition numbers related to the two involved approximation processes.

In Sect. 5, we show on an example that our “exact” bounds (3.7), (3.8) allow to guarantee a nice conditioning for the two-level method when  $\kappa(P^{-1}A_{11})$  is sufficiently close to 1.

**Theorem 3.1** *Let*

$$(3.3) \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} P & \\ & S \end{pmatrix} \begin{pmatrix} I & P^{-1}A_{12} \\ & I \end{pmatrix}$$

*be symmetric nonnegative definite matrices such that  $A_{11}$  and  $P$  are invertible.*

*Assume that*

$$(3.4) \quad \beta \mathbf{v}_1^T A_{11} \mathbf{v}_1 \leq \mathbf{v}_1^T P \mathbf{v}_1 \leq \mathbf{v}_1^T A_{11} \mathbf{v}_1 \quad \forall \mathbf{v}_1$$

*for some  $\beta$  such that  $0 < \beta < 1$  and that*

$$(3.5) \quad \begin{aligned} \eta \mathbf{v}_2^T (A_{22} - A_{21} A_{11}^{-1} A_{12}) \mathbf{v}_2 &\leq \mathbf{v}_2^T S \mathbf{v}_2 \\ &\leq \zeta \mathbf{v}_2^T (A_{22} - A_{21} A_{11}^{-1} A_{12}) \mathbf{v}_2 \quad \forall \mathbf{v}_2 \end{aligned}$$

*for some  $\eta, \zeta$  such that  $0 \leq \eta \leq 1 \leq \zeta$ .*

*If*

$$(3.6) \quad \begin{aligned} \mathbf{v}_2^T A_{21} P^{-1} A_{12} \mathbf{v}_2 &\leq (1 - \xi) \mathbf{v}_2^T A_{22} \mathbf{v}_2 \\ &+ \xi \mathbf{v}_2^T A_{21} A_{11}^{-1} A_{12} \mathbf{v}_2 \quad \forall \mathbf{v}_2 \end{aligned}$$

*for some  $\xi < 1$ , then*

$$(3.7) \quad \mathbf{v}^T B \mathbf{v} \geq \gamma \mathbf{v}^T A \mathbf{v} \quad \forall \mathbf{v}$$

*where  $\gamma$  is the smallest root of*

$$\gamma^2 - \gamma(\eta + 1 - \xi + \beta\xi) + \beta\eta$$

*and*

$$(3.8) \quad \alpha \mathbf{v}^T A \mathbf{v} \geq \mathbf{v}^T B \mathbf{v} \quad \forall \mathbf{v}$$

*where  $\alpha$  is the largest root of*

$$\alpha^2 - \alpha(\zeta + 1 - \xi + \beta\xi) + \beta\zeta.$$

*Moreover,*

$$(3.9) \quad \gamma \geq \frac{\eta\beta}{\eta + 1 - \xi(1 - \beta)}$$

*and*

$$(3.10) \quad \alpha \leq \zeta \left( 1 + \frac{(1-\beta)(1-\xi)}{\zeta-\beta} \right)$$

*with, in addition,*

$$(3.11) \quad \gamma \geq \begin{cases} \eta\beta & \text{if } \xi \geq \eta \\ \eta \left( \beta - \sqrt{(1-\beta)(1-\xi)} \right) & \text{if } 0 \leq \xi \leq \eta. \end{cases}$$

*Proof.* Define the family of second degree polynomials

$$P_a(t) = t^2 - t(a + 1 - \xi + \xi\beta) + a\beta,$$

where  $a > 0$ . Since  $P_a(a) = -a(1 - \beta)(1 - \xi) < 0$  and  $P_a(\beta) = -\beta(1 - \xi)(1 - \beta) < 0$ ,  $P_a(t)$  admits two positive roots  $t_1 < \min(\beta, a) \leq \max(\beta, a) < t_2$ , which shows that  $\gamma, \alpha$  exist, are positive and satisfy

$$(3.12) \quad \gamma < \min(\eta, \beta) < 1 \leq \zeta < \alpha.$$

Further, it is easily checked that

$$(3.13) \quad \gamma = \eta - \frac{\gamma(1 - \beta)(1 - \xi)}{\beta - \gamma}, \quad \alpha = \zeta + \frac{\alpha(1 - \beta)(1 - \xi)}{\alpha - \beta}.$$

We now prove (3.7).

$$B - \gamma A = \begin{pmatrix} P - \gamma A_{11} & (1 - \gamma) A_{12} \\ (1 - \gamma) A_{21} S + A_{21} P^{-1} A_{12} - \gamma A_{22} \end{pmatrix}$$

and, since  $\gamma < \beta$ , (3.4) implies that  $P - \gamma A_{11}$  is positive definite. Hence  $B - \gamma A$  is nonnegative definite if so is its Schur complement

$$S_{B-\gamma A} = S - \gamma A_{22} + A_{21} \left( P^{-1} - (1 - \gamma)^2 (P - \gamma A_{11})^{-1} \right) A_{12}.$$

Now,  $\mathbf{v}_2^T S \mathbf{v}_2 \geq \eta \mathbf{v}_2^T (A_{22} - A_{21} A_{11}^{-1} A_{12}) \mathbf{v}_2$  while, since  $\eta - \gamma > 0$ , (3.6), (3.13) imply

$$\begin{aligned} (\eta - \gamma) \mathbf{v}_2^T A_{22} \mathbf{v}_2 &= \frac{\gamma(1 - \beta)(1 - \xi)}{\beta - \gamma} \mathbf{v}_2^T A_{22} \mathbf{v}_2 \\ &\geq \frac{\gamma(1 - \beta)}{\beta - \gamma} \mathbf{v}_2^T A_{21} \left( P^{-1} - \xi A_{11}^{-1} \right) A_{12} \mathbf{v}_2. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbf{v}_2^T S_{B-\gamma A} \mathbf{v}_2 &\geq \mathbf{v}_2^T A_{21} \left\{ \left( 1 + \frac{\gamma(1 - \beta)}{\beta - \gamma} \right) P^{-1} \right. \\ &\quad \left. - \left( \eta + \frac{\xi\gamma(1 - \beta)}{\beta - \gamma} \right) A_{11}^{-1} - (1 - \gamma)^2 (P - \gamma A_{11})^{-1} \right\} A_{12} \mathbf{v}_2, \end{aligned}$$

and, using (3.13) to eliminate  $\eta$ ,

$$\begin{aligned} \mathbf{v}_2^T S_{B-\gamma A} \mathbf{v}_2 &\geq \mathbf{v}_2^T A_{21} \left\{ \frac{\beta(1 - \gamma)}{\beta - \gamma} P^{-1} - \frac{\gamma(1 - \gamma)}{\beta - \gamma} A_{11}^{-1} \right. \\ &\quad \left. - (1 - \gamma)^2 (P - \gamma A_{11})^{-1} \right\} A_{12} \mathbf{v}_2 \\ &= \frac{1 - \gamma}{\beta - \gamma} \mathbf{v}_2^T A_{21} P^{-\frac{1}{2}} \left\{ \beta I - \gamma X^{-1} \right. \\ &\quad \left. - (1 - \gamma)(\beta - \gamma)(I - \gamma X)^{-1} \right\} P^{-\frac{1}{2}} A_{12} \mathbf{v}_2, \end{aligned}$$

where  $X = P^{-\frac{1}{2}} A_{11} P^{-\frac{1}{2}}$ . Since  $X$  is symmetric positive definite,

$$f(X) = \beta I - \gamma X^{-1} - (1 - \gamma)(\beta - \gamma)(I - \gamma X)^{-1}$$

will be non negative definite if  $f(\lambda) \geq 0$  for all  $\lambda \in \sigma(X) \subset [1, \beta^{-1}]$ ; (3.7) follows then because

$$\begin{aligned} f(\lambda) &= (1 - \gamma \lambda)^{-1} \left( (1 - \gamma \lambda) (\beta - \gamma \lambda^{-1}) - (1 - \gamma)(\beta - \gamma) \right) \\ &= (1 - \gamma \lambda)^{-1} (1 - \lambda^{-1}) (1 - \beta \lambda). \end{aligned}$$

The proof of (3.8) is similar.

$$\alpha A - B = \begin{pmatrix} \alpha A_{11} - P & (\alpha - 1) A_{12} \\ (\alpha - 1) A_{21} & \alpha A_{22} - S - A_{21} P^{-1} A_{12} \end{pmatrix}$$

and, since  $\alpha > 1$ , (3.4) implies that  $\alpha A_{11} - P$  is positive definite. Hence,  $\alpha A - B$  is nonnegative definite if so is its Schur complement

$$S_{\alpha A - B} = \alpha A_{22} - S - A_{21} \left( P^{-1} + (\alpha - 1)^2 (\alpha A_{11} - P)^{-1} \right) A_{12}.$$

We may use  $\mathbf{v}_2^T S \mathbf{v}_2 \leq \zeta \mathbf{v}_2^T \left( A_{22} - A_{21} A_{11}^{-1} A_{12} \right) \mathbf{v}_2$ , while, since  $(\alpha - \zeta) > 0$ , (3.6), (3.13) imply

$$\begin{aligned} (\alpha - \zeta) \mathbf{v}_2^T A_{22} \mathbf{v}_2 &= \frac{\alpha(1-\beta)(1-\xi)}{\alpha-\beta} \mathbf{v}_2^T A_{22} \mathbf{v}_2 \\ &\geq \frac{\alpha(1-\beta)}{\alpha-\beta} \mathbf{v}_2^T A_{21} \left( P^{-1} - \xi A_{11}^{-1} \right) A_{12} \mathbf{v}_2, \end{aligned}$$

whence

$$\begin{aligned} \mathbf{v}_2^T S_{\alpha A - B} \mathbf{v}_2 &\geq \mathbf{v}_2^T A_{21} \left\{ \left( \frac{\alpha(1-\beta)}{\alpha-\beta} - 1 \right) P^{-1} + \left( \zeta - \frac{\xi \alpha(1-\beta)}{\alpha-\beta} \right) A_{11}^{-1} \right. \\ &\quad \left. - (\alpha - 1)^2 (\alpha A_{11} - P)^{-1} \right\} A_{12} \mathbf{v}_2 \\ &= \frac{\alpha-1}{\alpha-\beta} \mathbf{v}_2^T A_{21} P^{-\frac{1}{2}} \left\{ -\beta I + \alpha X^{-1} \right. \\ &\quad \left. - (\alpha-1)(\alpha-\beta)(\alpha X - I)^{-1} \right\} P^{-\frac{1}{2}} A_{12} \mathbf{v}_2. \end{aligned}$$

Letting

$$g(X) = -\beta I + \alpha X^{-1} - (\alpha - 1)(\alpha - \beta)(\alpha X - I)^{-1},$$

one has

$$\begin{aligned} g(\lambda) &= (\alpha \lambda - 1)^{-1} \left( (\alpha \lambda - 1) (-\beta + \alpha \lambda^{-1}) - (\alpha - 1)(\alpha - \beta) \right) \\ &= \alpha (\alpha \lambda - 1)^{-1} (1 - \lambda^{-1}) (1 - \beta \lambda) \end{aligned}$$

which is nonnegative for all  $\lambda \in [1, \beta^{-1}]$ , whence (3.8).

Finally, (3.10) holds by virtue of (3.13), together with  $\alpha > \zeta$  and the fact that  $\frac{x}{x-\beta}$  is a decreasing function of  $x$  for  $x > \beta$ ; (3.9) is deduced from  $\gamma(\eta + 1 - \xi(1 - \beta)) - \beta\eta = \gamma^2 \geq 0$ ; (3.11) is proved by checking that

$$\begin{aligned} P_\eta(\eta\beta) &= \eta\beta(\xi - \eta)(1 - \beta), \\ P_\eta\left(\eta(\beta - \sqrt{(1-\beta)(1-\xi)})\right) &= \eta(1 - \beta)(\xi(1 - \eta) + (1 - \beta)(\eta - \xi)) \\ &\quad + \eta\sqrt{(1 - \beta)(1 - \xi)}(1 - \eta\beta + (\eta - \xi)(1 - \beta)) \end{aligned}$$

are both nonnegative when respectively  $\xi \geq \eta$  and  $\eta \geq \xi \geq 0$ .  $\square$

#### 4. Approximate inverses satisfying a row-sum criterion

In this section, we examine how to check our main assumptions (3.1), (3.2). Of course (3.1) may hold for any preconditioner with an appropriate scaling. However, since this requires in general an eigenvalue estimation, we find more practical to focus on the classes of approximate inverses for which this condition holds by construction.

About (3.2), note that  $(1 - \xi)A_{22} - A_{21}(P^{-1} - \xi A_{11}^{-1})A_{12}$  is the Schur complement of

$$\begin{pmatrix} P & A_{12} \\ A_{21} & A_{22} - \xi S_A \end{pmatrix},$$

i.e. (3.2) holds if and only if the latter matrix is nonnegative definite. Further, to prove this at once for a wide class of preconditioners it suffice to check that

$$(4.1) \quad \mathbf{v}^T \begin{pmatrix} \Delta & A_{12} \\ A_{21} & A_{22} - \xi S_A \end{pmatrix} \mathbf{v} \geq 0 \quad \forall \mathbf{v}$$

for some matrix  $\Delta$  satisfying

$$(4.2) \quad \mathbf{v}_1^T P \mathbf{v}_1 \geq \mathbf{v}_1^T \Delta \mathbf{v}_1 \quad \forall \mathbf{v}_1$$

for any preconditioner of the class.

If  $A$  is an M-matrix, there always exists a positive vector  $\mathbf{x} = (\mathbf{x}_1 \ \mathbf{x}_2)^T$  such that  $A\mathbf{x}$  is nonnegative (e.g.  $\mathbf{x} = (1 \ 1 \ \dots \ 1)^T$  when  $A$  has nonnegative row-sum) [12]. It is then interesting to consider the diagonal matrix  $\Delta$  such that  $\Delta\mathbf{x}_1 = A_{11}\mathbf{x}_1$ . Indeed,

$$\tilde{A} = \begin{pmatrix} \Delta & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

has then nonpositive offdiagonal entries and nonnegative generalized row-sum since  $\tilde{A}\mathbf{x} = A\mathbf{x}$ . It is thus a (possibly singular<sup>2</sup>) symmetric M-matrix,

<sup>2</sup> According to [12], an M-matrix can be singular, and a symmetric matrix  $C$  that has nonpositive offdiagonal entries is an M-matrix if and only if it is nonnegative definite or, equivalently, if and only if  $C\mathbf{x} \geq 0$  for some  $\mathbf{x} > 0$

therefore nonnegative definite [12], i.e.  $\xi = 0$  at the very least. In Sect. 4.4 below, we show that in fact (4.1) holds with  $\xi \geq \frac{1}{2}$  when  $\mathbf{x} = (1 \ 1 \ \dots \ 1)^T$  and  $A$  arises from the linear finite element discretization of a 2D second order elliptic PDE.

Before, we show that (3.1), (4.2) are met when  $P$  is a (possibly blockwise) MILU factorization of  $A_{11}$  computed with respect to  $\mathbf{x}_1$  as modification vector (Sect. 4.1), or when  $P^{-1}$  is an explicit inverse of  $A_{11}$  satisfying a generalized row-sum criterion based on  $\mathbf{x}_1$  (Sect. 4.2).

Thus, combining the results in Sect. 4.1 and 4.4,  $\xi \geq \frac{1}{2}$  when  $A$  is a linear finite element matrix and  $P$  a MILU factorization of  $A_{11}$  satisfying the usual row-sum criterion.

Further, one is not restricted to a single application of the so defined preconditioner. We indeed consider in Sect. 4.3 approximating  $A_{11}^{-1}$  by a few steps of some inner iterative procedure with a preconditioner satisfying (3.1), (3.2). It turns out that the so defined approximate inverse still satisfies (3.1), (3.2) (for the same value of  $\xi$ ) when a proper polynomial acceleration is used.

#### 4.1. MILU factorizations

Assume  $A_{11}$  is an M-matrix and let  $\mathbf{x}_1$  be a positive vector such that  $A_{11} \mathbf{x}_1$  is nonnegative.

Then, if  $P$  results from a possibly blockwise MILU factorization of  $A_{11}$  (e.g. [2, 14–16, 19]) computed to satisfy the generalized row-sum criterion

$$P \mathbf{x}_1 = A_{11} \mathbf{x}_1 ,$$

it is well known that  $A_{11} - P$  is a symmetric M-matrix, whence (3.1). On the other hand,  $P$  writes

$$P = (Q - F^T) Q^{-1} (Q - F)$$

where  $F$  is strictly upper triangular and nonnegative and  $Q$  a nonsingular symmetric M-matrix (which reduces to a diagonal matrix in case of a point-wise factorization). For such matrices, the following theorem shows that  $P - \Delta$  is nonnegative definite.

**Theorem 4.1** *Let  $P = (Q - F^T) Q^{-1} (Q - F)$  be such that  $Q$  is a nonsingular symmetric M-matrix and  $F$  is nonnegative and strictly upper triangular.*

*Assume that  $P \mathbf{x}_1$  is nonnegative for some positive vector  $\mathbf{x}_1$ .*

*Then,*

$$\mathbf{v}_1^T P \mathbf{v}_1 \geq \mathbf{v}_1^T \Delta \mathbf{v}_1 \quad \forall \mathbf{v}_1$$

*where  $\Delta$  is the diagonal matrix such that  $P \mathbf{x}_1 = \Delta \mathbf{x}_1$ .*

*Proof.* First, since  $Q$  is an M-matrix,  $Q^{-1}$  is nonnegative [12]. From  $(I - Q^{-1}F) \mathbf{x}_1 = Q^{-1} (\Delta \mathbf{x}_1 + F^T (I - Q^{-1}F) \mathbf{x}_1)$ , it is then easily seen by induction that  $(I - Q^{-1}F) \mathbf{x}_1$  is nonnegative. Let  $T$  be the diagonal matrix such that  $T \mathbf{x}_1 = Q^{-1}F \mathbf{x}_1$ . One has

$$P - \Delta = (T - F^T Q^{-1}) Q (T - Q^{-1} F) + \left\{ Q - T Q T - (I - T) F - F^T (I - T) - \Delta \right\}.$$

The first term of the r.h.s. is nonnegative definite and has zero generalized row-sum by construction. On the other hand, the term under brackets, which has therefore also a zero generalized row-sum, has nonpositive offdiagonal entries since  $T$  is diagonal with diagonal entries less or equal to 1. Therefore, the term under brackets is a symmetric M-matrix, hence nonnegative definite.  $\square$

#### 4.2. “Compensated” explicit inverses

Assume  $A_{11}$  is an M-matrix, and let  $\mathbf{x}_1$  be a positive vector such that  $A_{11} \mathbf{x}_1$  is positive.

Since  $A_{11}^{-1}$  is nonnegative [12], it is not restrictive to assume from an explicit approximate inverses  $P^{-1}$  that it has nonnegative offdiagonal entries not larger than the corresponding ones in  $A_{11}^{-1}$ . Assume further that the neglected entries have been “compensated” on the diagonal in such a way that  $P^{-1}(A_{11} \mathbf{x}_1) = A_{11}^{-1}(A_{11} \mathbf{x}_1) = \mathbf{x}_1$ ; (3.1) follows then from [2, Theorem 8.4]. On the other hand, since  $(\Delta - \Delta P^{-1} \Delta) \mathbf{x}_1 = 0$ , it is easily seen that  $\Delta - \Delta P^{-1} \Delta$  is a symmetric M-matrix, which implies (4.2) since both  $\Delta$  and  $P$  are positive definite.

#### 4.3. Inner iterations

Here we consider approximate inverses of  $A_{11}$  implicitly defined by a very few steps of some preconditioned iterative method.

This means that

$$(4.3) \quad P^{-1} = \mathcal{P}_m \left( \tilde{P}^{-1} A_{11} \right) A_{11}^{-1},$$

where  $\tilde{P}$  stands for the used preconditioner and where  $\mathcal{P}_m$  is a polynomial such that  $\mathcal{P}_m(0) = 0$ .

Assume that  $\tilde{P}$  satisfies (3.1), (3.2) for some  $\xi$ . Then, letting  $b = \lambda_{\max}^{-1}(\tilde{P}^{-1} A_{11})$ , (3.1) will hold for  $P$  defined by (4.3) if

$$1 \leq \mathcal{P}_m(t) \quad \forall t \in [1, b^{-1}] ,$$

whereas (3.2) will hold for the same value of  $\xi$  if

$$\mathcal{P}_m(t) \leq t \quad \forall t \in [1, b^{-1}] .$$

As is well known [25], the best polynomial acceleration is obtained with shifted Chebyshev polynomials of the first kind, i.e.,

$$\mathcal{P}_m(t) = c \left( 1 - \frac{T_m \left( \frac{b^{-1}+1-2t}{b^{-1}-1} \right)}{T_m \left( \frac{b^{-1}+1}{b^{-1}-1} \right)} \right)$$

where  $c$  is some scaling factor and where  $T_m(x)$ ,  $m = 1, 2, \dots$  obeys the recurrence relation

$$T_0 = 1, \quad T_1 = x, \quad T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x), \quad m = 1, 2, \dots$$

Since  $|T_m(x)| \leq 1$  for  $|x| \leq 1$ , we set

$$c = \left( 1 - T_m^{-1} \left( \frac{b^{-1}+1}{b^{-1}-1} \right) \right)^{-1} ,$$

so that

$$\mathcal{P}_m(t) = \frac{T_m \left( \frac{b^{-1}+1}{b^{-1}-1} \right) - T_m \left( \frac{b^{-1}+1-2t}{b^{-1}-1} \right)}{T_m \left( \frac{b^{-1}+1}{b^{-1}-1} \right) - 1}$$

satisfies  $\mathcal{P}(t) \geq 1$  for all  $t \in [1, b^{-1}]$ .

We have no general proof that  $\mathcal{P}_m(t) \leq t \forall t \in [1, b^{-1}]$ . Nevertheless, for  $m = 2$ , one finds

$$\mathcal{P}_2(t) = (b+1)t - bt^2 ,$$

whence  $t - \mathcal{P}_2(t) = bt(t-1) \geq 0$  for  $t \geq 1$ . For  $m = 3$ , an explicit check is harder, but one may use the following reasoning: since  $T_3(-1) = -1$  and  $T_3 \left( \frac{b^{-1}+1}{b^{-1}-1} \right) > \left( \frac{b^{-1}+1}{b^{-1}-1} \right)$ , one has  $\mathcal{P}_3(b^{-1}) < b^{-1}$ ; therefore  $t - \mathcal{P}_3(t)$  is a third degree polynomial that is zero for  $t = 0$  and  $t = 1$ , positive for  $t = b^{-1}$ , and negative for  $t \rightarrow \infty$  (the leading coefficient is  $< 0$ ); hence,  $t - \mathcal{P}_3(t)$  cannot have an additional root in  $]1, b^{-1}[$ , i.e. is uniformly positive in that interval.

On the other hand, one will seldom consider in practice polynomials with degree larger than 3. Indeed, letting

$$\beta^{-1} = \max_{t \in ]1, b^{-1}[} \mathcal{P}(t) ,$$

one readily obtains, since  $|T_k(x)| \leq 1$  for  $|x| \leq 1$ ,

$$1 - \beta = \frac{2}{T_m\left(\frac{1+b}{1-b}\right) + 1} ,$$

and the resulting value is generally sufficiently small for  $m = 2$  or 3. For instance, with  $b = \frac{1}{2}$ , (which is already pessimistic since  $A_{11}$  is well conditioned), one gets  $1 - \beta = \frac{1}{9}$  for  $m = 2$  and  $1 - \beta = \frac{1}{50}$  for  $m = 3$ .

#### 4.4. Linear finite elements

Here we prove a better lower bound on  $\xi$  in the case  $A$  arises from the linear finite element discretization of a two dimensional second order elliptic PDE

$$(4.4) \quad -\partial_x a_x \partial_x u - \partial_y a_y \partial_y u = f$$

on an arbitrary “coarse” mesh to which one has added one level of uniform refinement.

Let  $A_c$  be the finite element matrix associated to the discretization of (4.4) on this coarse mesh. As is well known [26],  $\mathbf{v}_2^T A_c \mathbf{v}_2 \geq \mathbf{v}_2^T S_A \mathbf{v}_2$  so that (4.1) holds for some  $\xi > 0$  if

$$(4.5) \quad \mathbf{v}^T \begin{pmatrix} \Delta & A_{12} \\ A_{21} & A_{22} - \xi A_c \end{pmatrix} \mathbf{v} \geq 0 \quad \forall \mathbf{v} .$$

Now, the latter matrix corresponds to the assembly of elementary matrices associated to each coarse triangle. Therefore, (4.5) may be checked considering only these elementary matrices. If we assume further that  $a_x(x, y)$  and  $a_y(x, y)$  are piecewise constant on the coarse triangulation, one readily obtains that each of these elementary matrices writes (the vertex opposite to the mid edge node  $i$  is the node  $i + 3, i = 1, \dots, 3$ )

$$\begin{pmatrix} 2c_1 & 0 & 0 & 0 & -c_1 & -c_1 \\ 0 & 2c_2 & 0 & -c_2 & 0 & -c_2 \\ 0 & 0 & 2c_3 & -c_3 & -c_3 & 0 \\ \hline 0 & -c_2 & -c_3 & (c_2 + c_3)(1 - \xi) & \xi c_3 & \xi c_2 \\ -c_1 & 0 & -c_3 & \xi c_3 & (c_1 + c_3)(1 - \xi) & \xi c_1 \\ -c_1 & -c_2 & 0 & \xi c_2 & \xi c_1 & (c_1 + c_2)(1 - \xi) \end{pmatrix} \\ = c_1 T_1 + c_2 T_2 + c_3 T_3 ,$$

where  $c_1, c_2, c_3$  are positive provided there are no interior angles larger than  $\frac{\pi}{2}$ . On the other hand, deleting the zero lines and columns in  $T_1, T_2, T_3$ , one gets the same  $3 \times 3$  matrix

$$\tilde{T} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 - \xi & \xi \\ -1 & \xi & 1 - \xi \end{pmatrix},$$

which is readily found nonnegative definite for  $\xi = \frac{1}{2}$ .

## 5. Numerical results and conclusions

We first report the results of a numerical experiment that illustrates our theoretical investigations. In this experiment,  $A$  is the matrix resulting from the five point finite difference discretization of the Laplacian on the unit square with Dirichlet boundary conditions and a uniform mesh size  $h$  in both directions, where  $h$  is such that  $h^{-1}$  is an even integer.

We consider two-level preconditioners of the form (1.6), where the second block is formed with the nodes present in the grid of mesh size  $2h$ , and where  $S$  is taken equal to the matrix associated with the discretization of the Laplacian on this coarse grid.

For  $P$ , we consider preconditioners based on a simple pointwise incomplete factorization of  $A_{11}$  without fill-in and with a natural ordering of the concerned nodes. Two choices are included: a MILU factorization for which  $P$  and  $A_{11}$  have the same row-sum, and an ILU factorization for which the error matrix  $A_{11} - P$  has a zero diagonal.

The results are given in Table 1. Recalling that  $\kappa(S^{-1}S_A) = 2$  [20], they illustrate the stability of the two-grid method when using MILU preconditioning for  $A_{11}$ . On the other hand, the results obtained with ILU preconditioning are quite disastrous, although  $P$  is an excellent preconditioner of  $A_{11}$  alone.

For the sake of completeness, we have also performed some tests with scaled ILU preconditioners, that is  $P = cP_{\text{ILU}}$  where: (1)  $c = \lambda_{\max}(P_{\text{ILU}}^{-1}A_{11})$ , as one would have use on the basis of some theoretical results for hierarchical finite elements [8, 26–28], which require  $\mathbf{v}_1^T P \mathbf{v}_1 \geq \mathbf{v}_1^T A_{11} \mathbf{v}_1 \quad \forall \mathbf{v}$ ; (2)  $c = \lambda_{\min}(P_{\text{ILU}}^{-1}A_{11})$ , to satisfy on the contrary  $\mathbf{v}_1^T P \mathbf{v}_1 \leq \mathbf{v}_1^T A_{11} \mathbf{v}_1 \quad \forall \mathbf{v}$ ; (3)  $c = \lambda_{\min}(P_{\text{ILU}}^{-1}A_{11}) + 0.03$  and  $c = \lambda_{\min}(P_{\text{ILU}}^{-1}A_{11}) - 0.03$ , which may be seen as attempts to realize the above choice, but with a rough eigenvalue estimation. The results are given in Table 2, and clearly show that  $c = \lambda_{\min}(P_{\text{ILU}}^{-1}A_{11})$  is the only viable choice.

**Table 1.** Results for the model Poisson problem

$h^{-1}$	$P^{-1}A_{11}$			$B^{-1}A$		
	$\lambda_{\min}$	$\lambda_{\max}$	$\kappa$	$\lambda_{\min}$	$\lambda_{\max}$	$\kappa$
MILU preconditioning of $A_{11}$						
16	1.00	1.20	1.20	0.51	1.25	2.45
32	1.00	1.21	1.21	0.50	1.27	2.54
64	1.00	1.21	1.21	0.50	1.29	2.58
128	1.00	1.21	1.21	0.50	1.29	2.58
ILU preconditioning of $A_{11}$						
16	0.88	1.09	1.24	0.510	1.42	2.78
32	0.88	1.09	1.24	0.380	2.28	6.00
64	0.87	1.09	1.25	0.176	4.97	28.30
128	0.87	1.09	1.25	0.058	15.00	258.00

### *Towards efficient multilevel preconditioning*

Returning to the case where  $P$  is a MILU factorization of  $A_{11}$ , it is interesting to discuss to what extent one may trust its use in multilevel methods that were primarily designed assuming an exact inversion of  $A_{11}$ . Of course, a complete analysis lies beyond the scope of the present paper, as it would require a careful examination of each concerned preconditioner. However, on the whole, we expect that most methods could be used with approximate inversion of  $A_{11}$  as long as the results for the basic two-level scheme remain similar to those obtained above for the model Poisson problem.

Now, multilevel methods were developed for much harder problems such as problems with jumping coefficients or anisotropy. Usual analyzes of  $\kappa(S^{-1}S_A)$  (e.g. [20]), as well as our analysis of  $\xi$  in Sect. 4.4, only resort to local estimations and are independent of possible jumps or anisotropy in the PDE coefficients as long as the latter are piecewise constant on the coarse mesh. At the light of Theorem 3.1, one sees then that the nice behavior observed for the model problem will be essentially preserved if  $P^{-1}A_{11}$  remains nicely conditioned. For instance, in the case of a linear finite element discretization of (4.4), one has  $\xi = \frac{1}{2}$ , and taking  $S$  equal to the coarse grid discretization matrix leads to  $\eta = 1, \zeta = 2$  [20]. Then, with  $\beta^{-1} = 1.21$ , our theoretical bounds (3.7), (3.8) imply

$$\begin{aligned}\lambda_{\max}(B^{-1}A) &\leq 1.52, \\ \lambda_{\min}(B^{-1}A) &\geq 0.47,\end{aligned}$$

which is not too far from the values observed in Table 1.

For harder problems, one may fear an increase of  $\kappa(P^{-1}A_{11})$ , so it worths mentioning that, as shown in [21],  $\kappa(P^{-1}A_{11})$  can in fact never be

**Table 2.** Results for the model Poisson problem with scaled ILU preconditioning of  $A_{11}$ 

$\kappa(B^{-1}A)$				
$h^{-1}$	$c = 1.09$	$c = 0.87$	$c = 0.84$	$c = 0.90$
16	4.07	2.56	2.69	2.095
32	15.8	2.63	2.90	2.55
64	120.	2.65	3.66	3.02
128	1475	2.66	9.03	5.20

larger than  $\frac{3}{2}$  in the case of equation (4.4) with linear finite elements on right triangles; again, this proof is based on local estimations and assumes only that the PDE coefficients are piecewise constant on the coarse mesh.

Now, with  $\beta = \frac{2}{3}$ , (3.7) and (3.8) give

$$(5.1) \quad \lambda_{\max}(B^{-1}A) \leq 2.00,$$

$$(5.2) \quad \lambda_{\min}(B^{-1}A) \geq 0.45.$$

Hence, in the worst cases, the condition number will be at most slightly larger than twice that of the “ideal” method with exact inversion of  $A_{11}$ .

Although this is reasonable for very difficult problems, this might be too much in the context of some multilevel algorithms. One may then possibly rely on the fact that these bounds are likely to be pessimistic, but a robust alternative is easily obtained thanks to the results of Sect. 4.3. Indeed, one may use

$$P^{-1} = (1 + b)\tilde{P}^{-1} - b\tilde{P}^{-1}A_{11}\tilde{P}^{-1}$$

where  $\tilde{P}$  stands for the simple MILU factorization of  $A_{11}$  considered so far, and where  $b^{-1}$  is a known upper bound on  $\lambda_{\max}(\tilde{P}^{-1}A_{11})$ . Since

$$\beta = \lambda_{\max}^{-1}(P^{-1}A_{11}) = \frac{4b}{(b+1)^2},$$

this means that  $\beta = \frac{24}{25}$  when  $b = \frac{2}{3}$ . Our bounds (3.7), (3.8) imply then

$$\begin{aligned} \lambda_{\max}(B^{-1}A) &\leq 1.179, \\ \lambda_{\min}(B^{-1}A) &\geq 0.497. \end{aligned}$$

Thus, just by adding a single inner iteration step, we *guarantee* better conditioning properties than those observed in Table 1 for the model problem.

To illustrate this, we consider the linear finite element discretization of the PDE (4.4) on the unit square with the boundary conditions

$$\begin{cases} u = 0 & \text{on } \Gamma_0 = \{(x, y) \mid 0 \leq x \leq 1, y = 0\} \\ \frac{\partial u}{\partial n} = 0 & \text{on } \Gamma_1 = \partial\Omega \setminus \Gamma_0 \end{cases}.$$

and

Problem 1:

$$a_x = a_y = \begin{cases} 1000 & \text{in } \left(\frac{1}{4}, \frac{3}{4}\right) \times \left(\frac{1}{4}, \frac{3}{4}\right) \\ 1 & \text{elsewhere ;} \end{cases}$$

Problem 2:

$$a_x = \begin{cases} 1000 & \text{in } \left(\frac{1}{4}, \frac{1}{2}\right) \times \left(\frac{1}{4}, \frac{1}{2}\right) \\ 1 & \text{elsewhere ,} \end{cases}$$

$$a_y = \begin{cases} 1000 & \text{in } \left(\frac{1}{2}, \frac{3}{4}\right) \times \left(\frac{1}{2}, \frac{3}{4}\right) \\ 1 & \text{elsewhere .} \end{cases}$$

We use right triangles and a uniform mesh size  $h$  in both directions, where  $h$  is such that  $h^{-1}$  is an even integer;  $S$  is taken equal to the coarse grid discretization matrix.

The results are reported in Tables 3 and 4. It turns out that the increase of  $\kappa(B^{-1}A)$  is indeed more moderate than indicated by the bounds (5.1), (5.2). On the other hand, with a simple and still cheap inner iteration, the deviation from the “ideal” condition number is less than 6%, which is likely to be acceptable for most multilevel schemes.

**Table 3.** Results for Problem 1

$h^{-1}$	$P^{-1}A_{11}$			$B^{-1}A$		
	$\lambda_{\min}$	$\lambda_{\max}$	$\kappa$	$\lambda_{\min}$	$\lambda_{\max}$	$\kappa$
$P^{-1} = \tilde{P}^{-1}$						
32	1.00	1.34	1.34	0.50	1.44	2.85
64	1.00	1.34	1.34	0.50	1.44	2.87
128	1.00	1.34	1.34	0.50	1.44	2.87
$P^{-1} = \frac{5}{3}\tilde{P}^{-1} - \frac{2}{3}\tilde{P}^{-1}A_{11}\tilde{P}^{-1}$						
32	1.00	1.04	1.04	0.50	1.06	2.10
64	1.00	1.04	1.04	0.50	1.06	2.11
128	1.00	1.04	1.04	0.50	1.06	2.11

We refer to [21] for an example of multilevel preconditioner that even avoids this inner iteration and turns out to be particularly cost effective thanks to this dramatic reduction of the number of operations associated to the fine grid nodes.

**Table 4.** Results for Problem 2

$h^{-1}$	$P^{-1}A_{11}$			$B^{-1}A$		
	$\lambda_{\min}$	$\lambda_{\max}$	$\kappa$	$\lambda_{\min}$	$\lambda_{\max}$	$\kappa$
$P^{-1} = \tilde{P}^{-1}$						
32	1.00	1.34	1.34	0.50	1.47	2.93
64	1.00	1.31	1.31	0.50	1.53	3.04
128	1.00	1.31	1.31	0.50	1.56	3.10
$P^{-1} = \frac{5}{3}\tilde{P}^{-1} - \frac{2}{3}\tilde{P}^{-1}A_{11}\tilde{P}^{-1}$						
32	1.00	1.04	1.04	0.50	1.06	2.11
64	1.00	1.04	1.04	0.50	1.06	2.12
128	1.00	1.04	1.04	0.50	1.06	2.12

*Acknowledgements.* We are indebted to some anonymous referee for the remark at the end of Sect. 2 and some other useful comments.

## References

1. Axelsson, O. (1981): On multigrid methods of the two-level type. In: Hackbusch, W., Trottenberg, U. (eds.) Multigrid Methods, Lectures Notes in Mathematics No. 960, Berlin Heidelberg New York, Springer-Verlag, pp. 352–367
2. Axelsson, O. (1994): Iterative Solution Methods. University Press, Cambridge
3. Axelsson, O., Eijkhout, V. (1990): Analysis of recursive 5-point/9-point factorization method. In: Axelsson, O., Kolotilina, L. (eds.) Preconditioned Conjugate Gradient Methods, Lectures Notes in Mathematics No. 1457, Springer-Verlag, pp. 154–173
4. Axelsson, O., Eijkhout, V. (1991): The nested recursive two level factorization for nine-point difference matrices. SIAM J. Sci. Stat. Comput. **12**, 1373–1400
5. Axelsson, O., Gustafsson, I. (1983): Preconditioning and two-level multigrid methods of arbitrary degree of approximation. Math. Comp. **40**, 214–242
6. Axelsson, O., Neytcheva, M. (1994): Algebraic multilevel iterations for Stieltjes matrices. Num. Lin. Alg. Appl. **1**, 213–236
7. Axelsson, O., Vassilevski, P.S. (1989): Algebraic multilevel preconditioning methods, I. Numer. Math. **56**, 157–177
8. Axelsson, O., Vassilevski, P.S. (1990): Algebraic multilevel preconditioning methods, II. SIAM J. Numer. Anal. **27**, 1569–1590
9. Bank, R.E. (1996): Hierarchical bases and the finite element method. Acta Numerica **5**, 1–43
10. Bank, R.E., Dupont, T.F. (1980): Analysis of a two-level scheme for solving finite element equations. Tech. Rep. CNA-159, Center for Numerical Analysis, The University of Texas at Austin, Texas, USA
11. Bank, R.E., Dupont, T.F., Yserentant, H. (1988): The hierarchical basis multigrid method. Numer. Math. **52**, 427–458
12. Berman, A., Plemmons, R.J. (1979): Nonnegative Matrices in the Mathematical Sciences. Academic Press, New York

13. Botta, E., van der Ploeg, A. (1995): Preconditioning techniques for matrices with arbitrary sparsity patterns. Preprint, Department of Mathematics, University of Groningen, The Netherlands
14. Chan, T.C., van der Vorst, H.A. (1994): Approximate and incomplete factorizations. Preprint 871, Department of Mathematics, University of Utrecht, The Netherlands
15. Concus, P., Golub, G.H., Meurant, G. (1985): Block preconditioning for the conjugate gradient method. *SIAM J. Sci. Statist. Comput.* **6**, 220–252
16. Dupont, T., Kendall, R.P., Rachford, H.H. (1968): An approximate factorization procedure for solving self-adjoint elliptic difference equations. *SIAM J. Numer. Anal.* **5**, 559–573
17. Eijkhout, V., Vassilevski, P. (1991): The role of the strengthened c.b.s. inequality in multilevel methods. *SIAM Review* **33**, 405–419
18. Hackbusch, W. (1985): *Multi-grid Methods and Applications*. Springer, Berlin
19. Magolu, M.M. (1995): Ordering strategies for modified block incomplete factorization. *SIAM J. Sci. Comput.* **16**, 378–399
20. Maitre, J., Musy, F. (1981): The contraction number of a class of two-level methods; an exact evaluation for some finite element subspaces and model problems. In: Hackbusch, W., Trottenberg, U. (eds.) *Multigrid Methods, Lectures Notes in Mathematics No. 960*, Berlin Heidelberg New York, Springer-Verlag, pp. 535–544
21. Notay, Y. (1997): Optimal order preconditioning of finite difference matrices. Tech. Rep. GANMN 97-02, Université Libre de Bruxelles, Brussels, Belgium. <http://homepages.ulb.ac.be/~ynotay>
22. Reusken, A. (1995): Fourier analysis of a robust multigrid method for convection-diffusion equations. *Numer. Math.* **71**, 365–398
23. Reusken, A. (1996): A multigrid method based on incomplete Gaussian elimination. *J. Num. Lin. Alg. with Appl.* **3**, 369–390
24. van der Ploeg, A., Botta, E., Wubs, F. (1996): Nested grids ILU-decomposition (NGILU). *J. Comput. Appl. Math.* **66**, 515–526
25. Varga, R.S. (1962): *Matrix iterative analysis*. Prentice Hall, Englewood Cliffs
26. Vassilevski, P. (1989): Nearly optimal iterative methods for solving finite element elliptic equations based on the multilevel splitting of the matrix. Tech. Rep. # 1989-09, Institute for Scientific Computation, University of Wyoming, Laramie, USA
27. Vassilevski, P. (1992): Hybrid V-cycle algebraic multilevel preconditioners. *Math. Comp.* **58**, 489–512
28. Vassilevski, P. (1997): On two ways of stabilizing the hierarchical basis multilevel methods. *SIAM Review* **39**, 18–53
29. Wagner, C., Kinzelbach, W., Wittum, G. (1997): Schur-complement multigrid. *Numer. Math.* **75**, 523–545
30. Wesseling, P. (1992): *An Introduction to Multigrid Methods*. J. Wiley and Sons, Chichester
31. Yserentant, H. (1986): On the multi-level splitting of finite element spaces. *Numer. Math.* **49**, 379–412