Reduced order solution of structured linear systems arising in certain PDE-constrained optimization problems

V. Simoncini

Abstract The solution of PDE-constrained optimal control problems is a computationally challenging task, and it involves the solution of structured algebraic linear systems whose blocks stem from the discretized first-order optimality conditions. In this paper we analyze the numerical solution of this large-scale system: we first perform a natural order reduction, and then we solve the reduced system iteratively by exploiting specifically designed preconditioning techniques. The analysis is accompanied by numerical experiments on two application problems.

Keywords Structured linear systems · Iterative methods · PDE-constraints · Optimization · Preconditioning

1 Introduction

We consider solving the following structured linear system

$$
\begin{bmatrix} A_1 & 0 & B_1^\top \\ 0 & A_2 & B_2^\top \\ B_1 & B_2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}, \tag{1.1}
$$

where A_1 is symmetric and positive definite matrix, A_2 is symmetric positive semidefinite and possibly highly singular, and B_2 is square and nonsingular, so that

V. Simoncini (✉)
Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato 5, 40127 Bologna, Italy
e-mail: valeria.simoncini@unibo.it

V. Simoncini
CIRSA, Ravenna, Italy

the whole matrix is nonsingular; $B_1^\top$ denotes the transpose of B_1. This type of structure is very common in a large number of application problems; here we are mainly interested in linear systems stemming from the discretization of PDE-constrained optimal control problems. In this context, the x_1, x_2 and x_3 usually refer to the control, state and adjoint state variables. Several authors have addressed the problem of efficiently solving (1.1), and various preconditioners have very recently been explored that try to fully take into account the special coefficient matrix structure; we refer to, e.g., [3, 26, 29, 33, 39, 40] and references therein.

Within a discretize-then-optimize framework, systems similar to (1.1) need to be solved many times, usually either because they are used to tune a numerical model, or because they represent a step in an iterative (outer) nonlinear solution procedure. The computational cost associated with the solution of (1.1) is therefore a major indicator of the overall suitability of the numerical model for practical engineering purposes. We stress that the possible high singularity of A_2, with no other hypotheses, poses an additional challenge for the linear system solution.

A key fact in the solution of (1.1) is that the matrix blocks can be extremely large already for rather coarse mesh discretizations, because of a three or higher dimensional setting. This fact is particularly crucial in so-called "all-at-once" strategies, where the time frame is also discretized, and the matrix blocks in (1.1) include subblocks corresponding to different time instants [17, 26, 39]. In this framework, an initial reduction of the matrix size may lead to significant savings, as long as this reduction does not entail extra computational burden, such as additional inner solves or approximation of variables. In this paper we report on our numerical experience with a particularly simple and effective reduction. The reduction process fully relies on the assumption that A_1 is easy to solve with (e.g., A_1 is diagonal), which is usually the case in the problems we are addressing. We discuss two realistic and popular specific problems, that require solving (1.1). The first one is a PDE-constrained optimal control problem with distributed controls, with linear quadratic functional [41, Sect. 2]. This simplified problem has become a popular benchmark for analyzing several numerical devices. We will see that in this case, the reduction process does not even need to solve with A_1. The second application we consider is a simplified Monge-Kantorovich mass transfer problem, used in the context of image registration [3]. Although we focus on two very specific applications, our discussion can be adapted to other settings, and we hope it may serve as a general platform for related and more general frameworks.

The solution of the reduced system exploits the obtained structure to build effective preconditioning techniques, which allow us to solve medium size 3D problems in a small amount of time on a commodity workstation. We also perform an ad-hoc spectral analysis that in some cases ensures optimality of the adopted preconditioning strategy. As it will be clear in the discussion of the specific problems, the use of the reduced formulation allows us to eliminate some redundant information of the original formulation, so that the preconditioning step need only be concerned with the relevant blocks in the matrix.

A synopsis of the paper is as follows. Section 2 shows the reduction step, based on the Schur complement formulation, and reviews and analyzes some general preconditioning strategies that are employed in later sections. Section 3 introduces the matrix

setting of the PDE-constrained optimal control problem, and provides a detailed theoretical analysis of the proposed preconditioning approaches; Sect. 3.1 reports on our numerical experience with these solution strategies. Section 4 is devoted to the matrix description of the simplified Monge-Kantorovich mass transfer problem, and of its numerical solution, while Sect. 4.1 reports on our thorough numerical experience with these data. Finally, Sect. 5 collects our concluding remarks.

Throughout the paper we shall use the Euclidean norm for vectors, and the induced norm for matrices. We shall denote by $\mathrm{spec}(A)$ the set of eigenvalues (spectrum) of a given matrix A.

2 The reduced order problem and preconditioning strategies

The size of each block of the coefficient matrix in (1.1) depends on the fineness of the given domain (2D or 3D in space possibly plus time), so that the problem dimension becomes very high already for rather coarse discretizations. The prospect of reducing the dimension of (1.1) is therefore of great interest, especially if this does not entail higher computational complexity. Indeed, we can decrease the number of blocks from three to two, by using the Schur complement formulation with respect to A_1 (so-called "static condensation"): the first equation block gives $x_1 = A_1^{-1}(f_1 - B_1^\top x_3)$, which substituted into the original system yields the reduced system

$$\begin{bmatrix} A_2 & B_2^\top \\ B_2 & -B_1 A_1^{-1} B_1^\top \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f_2 \\ f_3 - B_1 A_1^{-1} f_1 \end{bmatrix} \quad \Leftrightarrow \quad Sx = b, \qquad (2.1)$$

which, in the optimal control language, amounts to eliminating the *control* variables. We recall here that we assume that A_2 is positive semidefinite and possibly singular, while $B_1 A_1^{-1} B_1^\top$ is positive semidefinite, with the possibility of being nonsingular if B_1 is full row-rank.[1] It is interesting that this is a slightly different setting than the one often encountered in the literature: A_2 is often required to be positive definite on the kernel of B_2, but since B_2 is nonsingular here, such condition is empty. Thanks to this nonsingularity, both diagonal blocks can be highly singular. A sufficient condition for the solvability of (2.1) is that $B_2^\top + A_2 B_2^{-1} B_1 A_1^{-1} B_1^\top$ be nonsingular. Under the hypothesis of nonsingularity of the original system (1.1), such condition is satisfied.

This reduction procedure is classical, and it is often discarded because in general it requires explicitly dealing with the inverse in $B_1 A_1^{-1} B_1^\top$. However, it turns out that for various PDE-constrained optimization problems the matrix $B_1 A_1^{-1} B_1^\top$ is very cheap to deal with, so that the reduction becomes extremely attractive.

The reduced problem (2.1) may also be derived by using the following decomposition

$$\left[\begin{array}{c|cc} A_1 & 0 & B_1^\top \\ \hline 0 & A_2 & B_2^\top \\ B_1 & B_2 & 0 \end{array}\right] = \left[\begin{array}{c|cc} I & 0 & 0 \\ \hline 0 & I & 0 \\ B_1 A_1^{-1} & 0 & I \end{array}\right] \left[\begin{array}{c|cc} A_1 & 0 & B_1^\top \\ \hline 0 & A_2 & B_2^\top \\ 0 & B_2 & -B_1 A_1^{-1} B_1^\top \end{array}\right] \equiv \mathcal{L}\mathcal{U}.$$

---

[1] In optimal control problems, the full rank assumption is actually an exceptional case, since it corresponds to having as many control variables as states, which is seldom encountered in practice.

The system in (1.1) can be rewritten as

$$\mathcal{L}\mathcal{U}x = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \quad \Leftrightarrow \quad \mathcal{U}x = \mathcal{L}^{-1} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \quad \Leftrightarrow \quad \mathcal{U}x = \begin{bmatrix} f_1 \\ f_2 \\ f_3 - B_1 A_1^{-1} f_1 \end{bmatrix},$$

from which the reduced system is obtained.

The system in (2.1) is once again in symmetric saddle point form, where the first diagonal block is only positive semidefinite, while the nonsingularity of the second diagonal block depends on the rank of $B_1^\top$. Note that the system may be written in a more familiar form after row and column permutation of the blocks, so that the two diagonal blocks are switched; a change of sign could also be applied, so as to have a positive definite leading block whenever B_1 is full rank. We do not actually apply these changes, but we keep them in mind when choosing the preconditioners, as the choice of a suitable preconditioning strategy heavily depends on whether the second diagonal block is nonsingular. It is also important to realize that effective preconditioners can exploit the nonsingularity of the non-diagonal block.

A lot of effort is usually devoted to the determination of "ideal" preconditioners, with which the preconditioned problem can be solved in approximately the same number of iterations, irrespective of the size, leading to a mesh independent performance of the solver;[2] cf., e.g., [12, 21, 43]. On the other hand, we also stress that these ideal preconditioners may be difficult to either explicitly compute, or to cheaply approximate in a way so as to maintain their optimality properties. Therefore, great care should be put in choosing the ideal preconditioner, so that it can be then transformed into a feasible one.

Nonsingular (2, 2) block   If the matrix $B_1 A_1^{-1} B_1^\top$ is nonsingular, then a natural preconditioner is given by the following block diagonal matrix (see, e.g., [43, Sect. 3.2]):

$$P_d = \begin{bmatrix} \widetilde{B}_2 C^{-1} \widetilde{B}_2^\top & 0 \\ 0 & C \end{bmatrix}, \qquad C \approx B_1 A_1^{-1} B_1^\top, \qquad \widetilde{B}_2 \approx B_2. \qquad (2.2)$$

Note that the block $\widetilde{B}_2 C^{-1} \widetilde{B}_2^\top$ is an approximation to the ideal matrix $A_2 + \widetilde{B}_2 C^{-1} \widetilde{B}_2^\top$. We refer to [36] for a detailed analysis of the algebraic spectral properties of the preconditioned matrix $S P_d^{-1}$. From a computational standpoint, we observe that systems of the form $\widetilde{B}_2 C^{-1} \widetilde{B}_2^\top v = w$ arising when employing P_d can be solved in sequence as $v = \widetilde{B}_2^{-1}(C(\widetilde{B}_2^{-\top} w))$. Such computational advantage drove our choice of the first diagonal block in P_d, as opposed to the ideal one mentioned above. The actual matrices C, $\widetilde{B}_2$ used in computation depend on the properties of the original blocks, and this will be discussed for the specific problems.

Singular (2, 2) block   If the matrix $B_1 A_1^{-1} B_1^\top$ is singular, and since the first diagonal block is also singular, a common choice consists of augmenting one of the diagonal

---

[2]The computational cost will still grow with the problem size, as basic operations are performed with longer vectors when finer discretizations are used.

blocks in the preconditioner. For instance, in the block diagonal case,

$$P_{ad} := \begin{bmatrix} D & 0 \\ 0 & \widetilde{C}(D) \end{bmatrix}, \qquad \widetilde{C}(D) \approx C(D) := B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top, \quad (2.3)$$

where the symmetric and positive definite matrix D is chosen appropriately and $\widetilde{C}(D)$ is positive definite; see [13, 16, 28] and references therein. We also refer to [9] for the augmentation case when the $(1, 1)$ block of S is zero. For this augmented block diagonal preconditioner, we can apply the results in [13] to derive the following general spectral bounds for the preconditioned matrix $P_{ad}^{-\frac{1}{2}} S P_{ad}^{-\frac{1}{2}}$. Refined bounds will be obtained in Sect. 4 for specific choices of the blocks.

Proposition 2.1 [13, Prop. 3.3] *Let*

$$\widehat{S} := \begin{bmatrix} \widehat{A}_2 & \widehat{B}_2^\top \\ \widehat{B}_2 & -\widehat{T} \end{bmatrix},$$

and let $\lambda_{\max}, \gamma_{\max}$ be the largest eigenvalues of $\widehat{A}_2$ and $\widehat{T}$, respectively. Let $\sigma_{\min}, \sigma_{\max}$ be the largest and smallest (nonzero) singular values of the square matrix $\widehat{B}_2$. Then $\text{spec}(\widehat{S}) \subseteq \mathcal{I}^- \cup \mathcal{I}^+$, *where*

$$\mathcal{I}^- = \left[\frac{1}{2}\left(-\gamma_{\max} - \sqrt{\gamma_{\max}^2 + 4\sigma_{\max}^2}\right), \frac{1}{2}\left(\lambda_{\max} - \sqrt{\lambda_{\max}^2 + 4\sigma_{\min}^2}\right) \right]$$

and

$$\mathcal{I}^+ = \left[\frac{1}{2}\left(-\gamma_{\max} + \sqrt{\gamma_{\max}^2 + 4\sigma_{\min}^2}\right), \frac{1}{2}\left(\lambda_{\max} + \sqrt{\lambda_{\max}^2 + 4\sigma_{\max}^2}\right) \right].$$

This quite general result already indicates that the length of both positive and negative intervals mainly depends on the magnitude of the singular values of $\widehat{B}_2$, which in turn strongly depends on the choice of D, as the following result shows.

Proposition 2.2 *Let $\widetilde{C}(D) = C(D)$, with D chosen so that $C(D)$ in (2.3) is symmetric and positive definite, and $\widehat{S} = P_{ad}^{-\frac{1}{2}} S P_{ad}^{-\frac{1}{2}}$, with S defined in (2.1). With the notation of Proposition 2.1, let σ be a singular value of $\widehat{B}_2$, which is explicitly given as $\widehat{B}_2 = C(D)^{-\frac{1}{2}} B_2 D^{-\frac{1}{2}}$. Then either $\sigma = 1$ or $\sigma^2 = \mu/(\mu + 1) < 1$, where μ are the finite eigenvalues of the problem $B_2 D^{-1} B_2^\top x = \mu B_1 A_1^{-1} B_1^\top x$, which thus satisfy*

$$\frac{\sigma_{\min}(B_2)^2}{\|D\| \, \|B_1 A_1^{-1} B_1^\top\|} \leq \mu \leq \frac{\|D^{-1}\| \, \|B_2\|^2}{\lambda_{\min}(B_1 A_1^{-1} B_1^\top)},$$

where $\lambda_{\min}(B_1 A_1^{-1} B_1^\top)$ is the smallest nonzero eigenvalue of $B_1 A_1^\top B_1^{-1}$.

Proof The singular values of $\widehat{B}_2$ are the square roots of the eigenvalues in the eigenvalue problem $B_2 D^{-1} B_2^\top x = \lambda(B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top)x$. For x in the null space

of $B_1^\top$, we obtain $\lambda = 1$. Otherwise, assuming that $\lambda \neq 1$, we can rewrite this eigenvalue problem as $B_2 D^{-1} B_2^\top x = \mu B_1 A_1^{-1} B_1^\top x$ with $\lambda = \mu/(1 + \mu)$, and the result follows.   □

The proposition above shows that for $\widetilde{C}(D) = C(D)$, $\sigma_{\max} = 1$ in Proposition 2.1, so that the external extremes of the two intervals $\mathcal{I}^-, \mathcal{I}^+$ only depend on the norms of the diagonal blocks. The lower bound for μ, and thus for $\sigma_{\min}$ in Proposition 2.2, provides a good indicator towards the selection of D so as to improve the clustering of the two spectral intervals $\mathcal{I}^-, \mathcal{I}^+$. We will explore this further in Sect. 4.

Regardless of the singularity of $B_1 A_1^{-1} B_1$, the following indefinite (constraint-type) block preconditioner may be considered:

$$P_{indef} = \begin{bmatrix} D & \widetilde{B}_2^\top \\ \widetilde{B}_2 & -C \end{bmatrix}, \qquad \widetilde{B}_2 \approx B_2, \tag{2.4}$$

with $\widetilde{B}_2$ nonsingular, where the symmetric matrix C is appropriately chosen, as in the case of P_{ad}. We restrict the discussion to the choice $D = 0$. Since $\widetilde{B}_2$ is nonsingular, systems with P_{indef} can be solved by solving with the non-diagonal blocks, in sequence. In particular, C is not required to be nonsingular. The indefiniteness of the symmetric preconditioner enforces the use either of a nonsymmetric solver such as GMRES or Simplified QMR (see [38] and references therein), or of a variant of the projected PCG [14]; see also [19] for additional alternatives. Experiments in literature show that the preconditioner P_{indef} is particularly well suited for various classes of optimization problems, such as constrained quadratic optimization, cf., e.g., [5, 10, 11, 18, 20, 25], and this is fully confirmed by our numerical experiments in Sect. 3.1. On the other hand, it should be mentioned that the positive definite block diagonal preconditioner P_d allows one to use a short-term symmetric solver, such as MINRES, whose convergence behavior is better understood than that of nonsymmetric solvers.

3 A PDE-constrained optimal control problem

In this section we discuss the algebraic setting of a benchmark PDE-constrained optimal control problem described in [40]. We briefly introduce the problem here, while we refer to [40] for a more complete presentation.

The problem can be stated as follows.[3] Given the bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, and $\hat{u}$ (the desired state) defined in $\widehat{\Omega} \subseteq \Omega$, find u such that

$$\min_{u, f} \frac{1}{2} \|u - \hat{u}\|_{L_2(\widehat{\Omega})}^2 + \beta \|f\|_{L_2(\Omega)}^2 \tag{3.1}$$

$$\text{s.t.} \quad -\nabla^2 u = f \quad \text{in } \Omega \tag{3.2}$$

$$u = \hat{u} \quad \text{on } \partial\Omega, \tag{3.3}$$

---

[3]With a little abuse of notation, the same letters will be used here to denote scalar functions and vectors.

where β is the regularization parameter. This linear quadratic optimal control problem is a simplified model, as more complex formulations may include, e.g., box constraints for f, more complex PDE constraints and/or other PDE boundary conditions; see, e.g., [18, 27, 33] for some sample treatments of the resulting algebraic problems, and [41] for a general discussion of the PDE and optimal control framework. After appropriate discretization, which among other things should ensure that the same matrix M in the first and third columns of the first block row in (3.4) arises, the discrete cost functional associated with (3.1) is given by

$$\min_{u_h, f_h} \frac{1}{2}\|u_h - \hat{u}\|_2^2 + \beta\|f_h\|_2^2 = \min_{u, f} \frac{1}{2}u^\top \bar{M}u - u^\top b + \alpha_u + \beta f^\top Mf$$

with $\alpha_u = \frac{1}{2}\|\hat{u}\|_2^2$, $b = M\hat{u}$, M the mass matrix, and $\bar{M}$ a portion of the mass matrix corresponding to the part of the domain where $\hat{u}$ is defined. The constraint $-\nabla^2 u = f$ in Ω gives the algebraic equation $Ku = Mf + d_b$, where K is the stiffness matrix, in this case the discretization of the Laplacian, d_b accounts for the boundary conditions, and b contains the integrals expressing $\bar{u}$ in the discretization basis [40]. The solution with the Lagrange multipliers method deals with the following Lagrange function

$$\mathcal{L}(f, u, \ell) = \frac{1}{2}u^\top \bar{M}u - u^\top b + \alpha_u + \beta f^\top Mf + \ell^\top(Ku - Mf - d_b),$$

and the associated optimality first order conditions give the following algebraic saddle point linear system

$$\begin{bmatrix} 2\beta M & 0 & -M \\ 0 & \bar{M} & K^\top \\ -M & K & 0 \end{bmatrix} \begin{bmatrix} f \\ u \\ \ell \end{bmatrix} = \begin{bmatrix} 0 \\ b \\ d_b \end{bmatrix}. \tag{3.4}$$

A lot of attention has been recently devoted to structured matrices of this form. Known block preconditioners, such as block diagonal, constraint-style and Bramble-Pasciak preconditioners have been successfully adapted to this setting [30, 33, 40] and to somewhat more general ones, see. e.g., [18, 21]. However, in the recent related linear system literature it has been somehow overlooked that the structure in (3.4) perfectly matches that in (1.1), so that the order reduction in (2.1) can be performed; but see [37], and also [2], [43, Sect. 4.1], [6, Sect. 8.2] for similar considerations. It should be mentioned, however, that a similar elimination technique is the background of classical "indirect solution methods" for (nonlinear) optimal control problems; see, e.g., the classical software [23].[4]

We first note that the presence of the nonsingular matrix M in two of the three places in the first block row of (3.4) allows one to readily eliminate the first unknown, as $2\beta f = \ell$. Therefore, the reduction process described in (2.1) is in this case particularly convenient. The reduction gives the system

$$\begin{bmatrix} \bar{M} & K^\top \\ K & -\frac{1}{2\beta}M \end{bmatrix} \begin{bmatrix} u \\ \ell \end{bmatrix} = \begin{bmatrix} b \\ d_b \end{bmatrix},$$

---

[4]We thank one of the referees for pointing this out.

where $\bar{M}$ may be singular, in case $\hat{u}$ is not defined in the whole Ω [40]; note also the particularly simple structure of the (2, 2) block compared with the general case in (2.1). In addition, the matrix M is nonsingular and its inverse can either be cheaply computed (if M is diagonal) or easily approximated; see, e.g., [29, 42]. Although the matrix structure for more complex PDE-constrained optimization problems may vary significantly, the reduction performed above is made possible simply by the presence of the regularizing term $\beta\|u\|^2$, which is often employed; see, e.g., [41]. We thus stress that the reduction step is extremely cheap to perform, and does not lead to a more difficult problem. Indeed, the main ingredients of the preconditioning step are the same as those used to solve the original problem (1.1). The simplified strategy however, eliminates some of the "redundant" information in the system. We will show that the resulting reduced system can be solved efficiently at low cost.

An ideal block diagonal preconditioner in this setting is given by P_d in (2.2) with the following blocks:

$$P_d = \begin{bmatrix} KC^{-1}K^\top & 0 \\ 0 & C \end{bmatrix}, \qquad C = \frac{1}{2\beta}M. \tag{3.5}$$

In practice, a more cost efficient preconditioner is obtained for $\widetilde{K} \approx K$ a, say, (algebraic) multigrid operator, and $\widetilde{C} \approx C$, with for instance $\widetilde{C} = \frac{1}{2\beta}\text{diag}(M)$; Note that with the discretization used in [40], the diagonal of M is constant. Other more sophisticated approaches have been devised, if more accurate approximations are desired, e.g., [30]. Under natural hypotheses on the given blocks (cf. [40]), the following corollary of Proposition 2.1 can be derived, ensuring mesh independence of the spectrum of the preconditioned matrix.

Corollary 3.1 *Under the hypotheses of Proposition* 2.1, *let h be the mesh parameter and* $d \in \{2, 3\}$. *Assume that* $C = \frac{\alpha h^d}{2\beta}I$ *for some* $\alpha > 0$ *independent of h, and that there exist* $\gamma_1, \gamma_2, \delta_1, \delta_2, \bar{\delta}_2$ *all positive and independent of h such that*

$$\gamma_1 h^d \le \lambda(K) \le \gamma_2 h^{d-2}, \qquad \delta_1 h^d \le \lambda(M) \le \delta_2 h^d, \qquad 0 \le \lambda(\bar{M}) \le \bar{\delta}_2 h^d.$$

Finally, assume that $\widetilde{K}$ *is chosen so that there exist* $\tau_1, \tau_2 > 0$ *such that* $\tau_1 x^*\widetilde{K}x \le x^*Kx \le \tau_2 x^*\widetilde{K}x$ *for all nonzero vectors x. Then the spectrum of* SP_d^{-1} *with* $\widetilde{K}$ *in place of K in* P_d *is bounded independently of h.*

Proof Following Proposition 2.1, the matrix $\widehat{S}$ takes the form

$$\widehat{S} = \begin{bmatrix} \widetilde{K}^{-1}C^{\frac{1}{2}}\bar{M}C^{\frac{1}{2}}\widetilde{K}^{-1} & \widetilde{K}^{-1}C^{\frac{1}{2}}KC^{-\frac{1}{2}} \\ C^{-\frac{1}{2}}KC^{\frac{1}{2}}\widetilde{K}^{-1} & -\frac{1}{2\beta}C^{-\frac{1}{2}}MC^{-\frac{1}{2}} \end{bmatrix}.$$

We will show that the spectral intervals of Proposition 2.1 applied to $\widehat{S}$ do not depend on h. To this end, we only need to show that the relevant eigenvalues of the (1, 1) and (2, 2) blocks are bounded independently of h, similarly for the singular values of the nondiagonal block. Since C is a multiple of the identity, these singular values satisfy $\sigma_i(\widetilde{K}^{-1}C^{\frac{1}{2}}KC^{-\frac{1}{2}}) = \lambda_i(\widetilde{K}^{-1}K)$, and these are bounded from below and from above by τ_1 and τ_2, respectively, showing mesh independence.

For the (2, 2) block we have $\gamma_{\max} = \lambda_{\max}(\frac{1}{2\beta}C^{-\frac{1}{2}}MC^{-\frac{1}{2}})$ so that

$$\gamma_{\max} = \frac{2\beta}{\alpha h^d}\frac{1}{2\beta}\lambda_{\max}(M) \le \frac{\delta_2 h^d}{\alpha h^d} = \frac{\delta_2}{\alpha}.$$

Finally, for the first diagonal block we have $\lambda_{\max} = \lambda_{\max}(\widetilde{K}^{-1}C^{\frac{1}{2}}\bar{M}C^{\frac{1}{2}}\widetilde{K}^{-1})$ so that

$$\lambda_{\max} = \frac{\alpha h^2}{2\beta}\lambda_{\max}(\widetilde{K}^{-1}\bar{M}\widetilde{K}^{-1}) \le \frac{\alpha h^d}{2\beta}\bar{\delta}_2 h^d \frac{\tau_2^2}{\gamma_1^2 h^{2d}} = \frac{\alpha \bar{\delta}_2 \tau_2^2}{2\beta \gamma_1^2}.$$

The result thus follows. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

The existence of τ_1, τ_2 is called *spectral equivalence condition* between K and $\widetilde{K}$, and it is often met, at least numerically, when $\widetilde{K}$ is chosen as a multigrid type operator [7, 8, 31].

Finally, we notice that our choice of P_d in (3.5) is closely related to the *robust* theoretical preconditioner in [43, Sect. 3.2] for the same problem, where the matrix $\bar{M} + KC^{-1}K^{\top}$ is suggested for the (1, 1) block instead of $KC^{-1}K^{\top}$. Indeed, under the hypotheses of Corollary 3.1, the two matrices are spectrally equivalent.

As mentioned in Sect. 2, an alternative to P_d is an indefinite preconditioner. The one in (2.4) takes a convenient form in this case since $\widetilde{K}$ is square and nonsingular, making the analytic computation of its inverse particularly simple:

$$P_{indef} = \begin{bmatrix} 0 & \widetilde{K} \\ \widetilde{K} & -C \end{bmatrix}, \qquad P_{indef}^{-1} = \begin{bmatrix} \widetilde{K}^{-1}C\widetilde{K}^{-1} & \widetilde{K}^{-1} \\ \widetilde{K}^{-1} & 0 \end{bmatrix}. \qquad (3.6)$$

The explicit form for P_{indef}^{-1} shows that the application of P_{indef}^{-1} to a vector requires two solves with $\widetilde{K}$, one multiplication with C and one sum of vectors. Since there are no solves with C, we can select $C = \frac{1}{2\beta}M$. The following proposition gives an explicit form for the preconditioned matrix, from which spectral information can be deduced. The proof is a simple matrix-matrix multiplication and is therefore omitted.

Proposition 3.2 *Let S be defined in* (2.1). *Consider the preconditioner P_{indef} with $C = \frac{1}{2\beta}M$. If $\widetilde{K} = K$, the preconditioned matrix has the form*

$$SP_{indef}^{-1} = \begin{bmatrix} G+I & \bar{M}K^{-1} \\ 0 & I \end{bmatrix}, \qquad G = \bar{M}K^{-1}CK^{-1},$$

and G singular. Therefore, $spec(SP_{indef}^{-1}) = \{1\} \cup spec(G+I)$.

Since G is singular, unit eigenvalues also arise in the (1, 1) block of SP_{indef}^{-1}, so that some of the unit eigenvalues may have geometric multiplicity larger than one; we refer to [4, Sect. 10.2] for a detailed description of the class of indefinite preconditioners and for other relevant references in the context of spectral analysis.

We observe that again for $\widetilde{K} = K$ and since C is symmetric and positive definite, G is similar to a symmetric and positive semidefinite matrix, from which it follows

that the eigenvalues of SP_{indef}^{-1} are real and not smaller than one. The following results ensures that the spectrum of SP_{indef}^{-1} (for $\widetilde{K} = K$) is bounded independently of the mesh parameter, under natural conditions on the involved matrices; cf., e.g., [40, Th. 2.1], [12, Sect. 5.5].

Corollary 3.3 *Under the hypotheses of Proposition 3.2, let h be the mesh parameter and $d \in \{2, 3\}$. Assume there exist $\gamma_1, \gamma_2, \delta_1, \delta_2, \bar{\delta}_2$ all positive and independent of h such that*

$$\gamma_1 h^d \leq \lambda(K) \leq \gamma_2 h^{d-2}, \qquad \delta_1 h^d \leq \lambda(M) \leq \delta_2 h^d, \qquad 0 \leq \lambda(\bar{M}) \leq \bar{\delta}_2 h^d,$$

then $\lambda(SP_{indef}^{-1}) = 1 + \eta$ with $0 \leq \eta \leq \delta_2 \bar{\delta}_2 / (2\beta\gamma_1^2)$.

Proof From Proposition 3.2 we know that either $\eta = 0$ or η is a positive eigenvalue of $\bar{M} K^{-1} M K^{-1}$. Since $\eta \leq \|\bar{M} K^{-1} M K^{-1}\| \leq \frac{1}{2\beta} \|\bar{M}\| \|K^{-1}\| \|M\| \|K^{-1}\|$, the result follows. $\qquad\square$

For $\beta \ll 1$, the eigenvalues of SP_{indef}^{-1} are bounded by a possibly quite large quantity; cf. also [2] for similar remarks, but on a problem with nonsingular first diagonal block. Our experiments showed that indeed the eigenvalues do increase for small β, in agreement with this bound. This justifies the need for a few extra iterations in the experiments reported in Tables 1–2. On the other hand, since for these data $\bar{M}$ has very low rank, then only few non-unit eigenvalues arise in SP_{indef}^{-1}, at least for $\widetilde{K} = K$, explaining the overall low number of iterations.

As already mentioned, indefinite preconditioners try to mimic as much as possible the original matrix, S here. Practical computationally feasible preconditioners however, replace the diagonal blocks of the original matrix with convenient approximations. In addition to taking zero as (1, 1) block, our choice of P_{indef} has the particular feature that also the nondiagonal blocks are approximated by means of $\widetilde{K}$;

Table 1 2D problem. Number of iterations and CPU time for solvers with P_{indef} and P_d and AMG, for various dimensions and values of the parameter β

β	n	GMRES w/P_{indef}		MINRES w/P_d	
		# it	time	# it	time
10^{-2}	961	3	0.15	31	0.18
	3969	3	0.17	28	0.37
	16129	3	0.29	25	1.22
	65025	3	0.88	25	4.28
	261121	4	4.54	27	21.31
10^{-5}	961	10	0.19	35	0.20
	3969	10	0.25	32	0.41
	16129	10	0.65	29	1.19
	65025	10	2.21	27	4.52
	261121	10	9.80	27	19.74

Table 2 3D problem. Number of iterations and CPU time for solvers with P_{indef} and P_d, for various dimensions and values of the parameter β

β	n	GMRES w/P_{indef}		MINRES w/P_d	
		# it	time	# it	time
10^{-2}	343	3	0.16	59	0.20
	3375	3	0.20	63	1.09
	29791	3	0.94	63	11.20
	250047	3	9.96	63	129.48
10^{-5}	343	8	0.13	57	0.20
	3375	9	0.30	61	1.03
	29791	9	2.04	64	11.28
	250047	9	23.02	65	132.36

this is similar in spirit to the approach taken in [5, 35]. For general symmetric and positive definite $\widetilde{K}$, we can write the preconditioned problem as

$$SP_{indef}^{-1} = \begin{bmatrix} I + \frac{1}{2\beta}\bar{M}\widetilde{K}^{-1}M\widetilde{K}^{-1} & \bar{M}\widetilde{K}^{-1} \\ 0 & I \end{bmatrix} + \begin{bmatrix} E & 0 \\ E & E \end{bmatrix}, \qquad E = K\widetilde{K}^{-1} - I.$$

This explicit form shows how the spectrum of SP_{indef}^{-1} changes from $\mathrm{spec}(I + \frac{1}{2\beta}\bar{M}\widetilde{K}^{-1}M\widetilde{K}^{-1}) \cup \{1\}$ as a (possibly nonlinear) function of $\|E\|$. Note that if for instance $\widetilde{K}^{-1}$ is spectrally equivalent to K, then $\mathrm{spec}(I + \frac{1}{2\beta}\bar{M}\widetilde{K}^{-1}M\widetilde{K}^{-1})$ satisfies a result similar to that of Corollary 3.3, since

$$\|\bar{M}\widetilde{K}^{-1}M\widetilde{K}^{-1}\| \leq \|\bar{M}\| \|\widetilde{K}^{-1}K\|^2 \|K^{-1}\|^2 \|M\|.$$

Moreover, $\|E\|$ does not depend on the mesh parameter either, ensuring mesh independence of the spectrum of SP_{indef}^{-1}. As is well known, however, eigenvalues alone may not completely explain the behavior of a nonsymmetric solver, and additional information on the eigenvectors should also be included. Nonetheless, our numerical experiments did not seem to indicate that the eigenbasis played a role. We refer to [34] for a more complete invariant subspace analysis of similarly perturbed structured matrices.

3.1 Numerical experiments

We consider the algebraic problem stemming from the finite element discretization with Q_1 elements on a uniform mesh of a distributed control problem analyzed in [40]. In particular, we focus on the data described in [40, Sect. 4], the 2D and 3D problems labelled Target 2. The domain is $\Omega = [0, 1]^d$, $d = 2, 3$. In 2D the subset $\widehat{\Omega}$ is given by a circle centered at $(\frac{5}{8}, \frac{3}{4})$ and radius $\frac{1}{5}$, union the boundary of Ω. In 3D, $\widehat{\Omega}$ is a sphere with center at $(\frac{5}{8}, \frac{3}{4}, \frac{7}{10})$ and radius $\frac{1}{4}$, union the boundary of Ω. The target $\hat{u}$, defined in $\widehat{\Omega}$ is set equal to a constant value within the circle or sphere, and zero on $\partial\Omega$.

We analyze the performance of MINRES preconditioned by P_d (cf. (2.2) and (3.5), with $\widetilde{C} = \frac{1}{2\beta}\text{diag}(M)$), and of GMRES right preconditioned by P_{indef} defined in (3.6) with $C = \frac{1}{2\beta}M$; in both cases the action of K^{-1} is substituted with that of $\widetilde{K}^{-1}$ computed as the Algebraic Multigrid method HSL_MI20 [7] with all default parameters except for `control.one_pass_coarsen=1`, `control.v_iterations=5`, and `control.st_parameter=.85`. The stopping criterion is based on the relative residual norm. However, it should be noticed that while during the GMRES iterations it is possible to monitor the Euclidean norm of the residual, in MINRES the P_d^{-1}-norm of the residual can be monitored. This discrepancy implies that the final solutions may have different accuracies. For the smaller dimension problems we observed that the GMRES solution was at least one order of magnitude more accurate than the MINRES solution.

The number of iterations and CPU time to meet the stopping criteria with a tolerance of 10^{-8} are reported in Tables 1 and 2, for the 2D and the 3D examples, respectively. For various mesh discretizations we considered two typical values of β, namely $\beta = 10^{-2}, 10^{-5}$, according to [40]. Note that the reported n implies that the actual problem (1.1) has size $3n$, whereas the reduced one in (2.1) has size $2n$.

The digits show the mesh independence of the solver preconditioned with P_{indef}, in agreement with the discussion towards the end of Sect. 3. Also MINRES preconditioned by P_d appears to have a mesh independent performance; thanks to Corollary 3.1 this was to be expected since, for the used discretization, all hypotheses of Corollary 3.1 are satisfied, at least for $\widetilde{K} = K$. For $\widetilde{K}$ being the HSL_MI20 approximation, we refer to [7] for a more detailed analysis of the (approximate) spectral equivalence to the stiffness matrix, for $d = 2, 3$.

The cost of applying P_{indef} and P_d is very similar, therefore the much lower number of iterations when using P_{indef} makes this approach far more appealing, with CPU times that are one order of magnitude lower for $\beta = 10^{-2}$, and less than 20% than those of the problem solved with MINRES and P_d for $\beta = 10^{-5}$. It is also interesting to note that the number of iterations with P_{indef} seems to be rather insensitive to the physical dimension of the problem (2D or 3D), whereas the performance of P_d degrades significantly when passing from the 2D to the 3D case.

We remark that in [40], the indefinite preconditioning strategy implemented within the Projected Preconditioned Conjugate Gradient (PPCG) method was used on the $3n$ size problem [14]. If one is not willing to use a memory consuming nonsymmetric solver with P_{indef}, then PPCG is a viable effective strategy. Since the number of iterations of GMRES preconditioned by P_{indef} is extremely low, we did not find it necessary to employ PPCG; see also the discussion in [10].

For the sake of completeness, we report in Table 3 the results obtained when solving the original $3n \times 3n$ system with the following preconditioner

$$P_{full} = \begin{bmatrix} 0 & 0 & -\widetilde{M} \\ 0 & 2\beta\widetilde{K}^\top M^{-1}\widetilde{K} & K^\top \\ -\widetilde{M} & K & 0 \end{bmatrix},$$

which was used in [40] for the same 3D problem. For these experiments, we used GMRES and we chose $\widetilde{K}$ as the HSL_MI20 algebraic multigrid approximation to K, as

Table 3 3D problem. Number of iterations and CPU time for solvers with P_{full} on the original unreduced problem, for various dimensions and values of the parameter β

β	n	GMRES w/P_{full}	
		# it	time
10^{-2}	343	5	0.14
	3375	5	0.31
	29791	5	2.41
	250047	5	47.29
10^{-5}	343	9	0.15
	3375	10	0.40
	29791	10	3.55
	250047	10	60.81

above, and $\widetilde{M}$ as the Matlab incomplete Cholesky decomposition of M with threshold 10^{-2} (cholinc); note that the choice $\widetilde{M} = \mathrm{diag}(M)$ lead to a rather slow convergence of the iterative process and it was therefore discarded. In [40] 20 iterations of the Chebyshev semi-iterative method were used as $\widetilde{M}$. In terms of number of iterations, the method with our choices of $\widetilde{K}$ and $\widetilde{M}$ behaved basically the same as with the exact matrices, but it was largerly cheaper. The digits in the table show that number of iterations for each value of β remains constant irrespective of the problem size, confirming the good quality of this preconditioner already observed in [40]. On the other hand, the timings show the price to be payed when dealing with the unreduced problem, when compared with the timings of the indefinite preconditioner P_{indef} in Table 2 for the smaller (reduced) problem.

4 A simplified Monge-Kantorovich mass transfer problem

In [3] the authors have considered the numerical solution of a simplified Monge-Kantorovich mass transfer problem. The proposed approach applies a Newton-type iteration to approximate the solution to a nonlinear system of equations, which considers an "all-at-once" discretization strategy of the original problem, stated in two space dimensions and in time. Here we provide a very brief review of the setting that leads to a 3×3 block structured system similar to that in (1.1), while we refer to [3] for a full description of the application and of the employed discretization strategy. The linear system stems from a parameter identification problem, which in the most general form may be stated as a constrained optimization problem as $\min_{u,m} \frac{1}{2}\|Qu - b\|^2 + \alpha_R R(m - m_r)$, with the constraint $\mathcal{A}(m)u - q = 0$. The latter driving equation is selected as the hyperbolic equation $u_t + \nabla \cdot (um) = 0$, $t \in [0, T]$, and m, the time-dependent velocity field, represents the set of parameters to be identified, which are also space dependent. As of the other quantities involved, b contains the observed data affected by noise; Q is the projector onto the portion of the space(-time) domain associated with b; R is the regularization functional, together with its weight α_R; and $m_r = 0$. According to [3], $\alpha_R = 10$ in our experiments. After discretization and for appropriate choices of the operators, the discretized optimization

problem discussed in [3] is posed as

$$\min_{u,m} \frac{1}{2}\|Qu - b\|^2 + \frac{1}{2}\xi u^\top L \operatorname{diag}(m)m \tag{4.1}$$

$$\text{s.t.} \quad u_t + \nabla \cdot (um) = 0, \qquad u|_{t=0} = u_0, \tag{4.2}$$

with $\xi = \alpha_R T h_t h_x^2$, where h_t, h_x are the mesh parameters in time and space, respectively (in our experiments we shall use $h_t = T/n_t$ and $h_x = 1/(n_x - 1)$ for a finite difference discretization on the unit square with n_t and n_x points, and the same number n_x of points in the x and y space directions). Here $A(m)$ represents the matrix discretization in space and time of the constraint, using an implicit Lax-Friedrichs scheme. Moreover, L is a constant rectangular matrix with identity matrices as diagonal and upper diagonal blocks. The associated Lagrangian function is given by $\mathcal{L}(u, m, p) = \frac{1}{2}\|Qu - b\|^2 + \frac{1}{2}\xi u^\top L \operatorname{diag}(m)m + p^\top V(A(m)u - q)$, where V is a diagonal matrix such that q is a grid discretization of the continuous (in space) Lagrange multiplier. By applying a Newton-like method to the nonlinear system given by the first order optimality conditions, one obtains a sequence of linear systems in the form:

$$\begin{bmatrix} Q^\top Q & 0 & A^\top V \\ 0 & \xi \operatorname{diag}(L^\top u) & G^\top V \\ V A & V G & 0 \end{bmatrix} \begin{bmatrix} \delta u \\ \delta m \\ \delta p \end{bmatrix} = -\begin{bmatrix} \mathcal{L}_u \\ \mathcal{L}_m \\ \mathcal{L}_p \end{bmatrix},$$

where G is the discretization of the Jacobian of $A(m)u$ with respect to m, and $A = A(m)$. Note that the coefficient matrix is an approximation of the Jacobian of the nonlinear function defining the first order optimality conditions. Clearly, both these matrices depend on the current approximation of m and u, and thus the whole system changes at each Newton iteration. By reordering rows and columns we obtain a linear system in the form (1.1):

$$\begin{bmatrix} \xi \operatorname{diag}(L^\top u) & 0 & G^\top V \\ 0 & Q^\top Q & A^\top V \\ V G & V A & 0 \end{bmatrix} \begin{bmatrix} \delta m \\ \delta u \\ \delta p \end{bmatrix} = -\begin{bmatrix} \mathcal{L}_m \\ \mathcal{L}_u \\ \mathcal{L}_p \end{bmatrix},$$

in which the first diagonal block is diagonal and nonsingular, while the second block $Q^\top Q$ is diagonal and highly singular, since we assume that b is only available in a small portion of the domain. Here $A^\top V$ is square and nonsingular, whereas $G^\top V$ is tall rectangular and possibly rank deficient, depending, e.g., on the boundary conditions imposed to define the discretized constraint, here periodic conditions. The reduction described in Sect. 2 yields

$$\begin{bmatrix} Q^\top Q & A^\top V \\ V A & -V G(\xi \operatorname{diag}(L^\top u))^{-1}G^\top V \end{bmatrix} \begin{bmatrix} \delta u \\ \delta p \end{bmatrix} = -\begin{bmatrix} \mathcal{L}_u \\ \mathcal{L}_p + V G(\xi \operatorname{diag}(L^\top u))^{-1}\mathcal{L}_m \end{bmatrix}.$$

Therefore, the linear system above can be restated as that in (2.1) with $A_2 = Q^\top Q$, $B_2 = V A$, $A_1 = \xi \operatorname{diag}(L^\top u)$ and $B_1 = V G$; note that here B_2 is nonsymmetric. The reduced system is once again obtained at a negligible cost, while the dimension reduction is very significant, since in this application problem $\xi \operatorname{diag}(L^\top u)$ is twice as large as $Q^\top Q$.

Since both diagonal blocks are only semi-definite, we consider using the augmented block diagonal preconditioner P_{ad} in (2.3). We are thus left to appropriately select the block D in (2.3). The matrix $Q^\top Q$ is diagonal, with all zero diagonal entries in the top part, namely $Q^\top Q = \mathrm{blkdiag}(0, \Omega)$ with Ω diagonal and nonsingular; if not already in this form, such ordering of the zero diagonal elements may be achieved by row and column permutation of the data in the reduced system. Let $\mathcal{I}_Q$ be the projector onto the null space of $Q^\top Q$ (this is nothing but a diagonal matrix with ones where the diagonal elements of $Q^\top Q$ are zero). Following the approach in [3], we define

$$D = Q^\top Q + \gamma \mathcal{I}_Q = \begin{bmatrix} \gamma I & 0 \\ 0 & \Omega \end{bmatrix}, \tag{4.3}$$

with $\gamma > 0$ to be chosen. Proposition 2.2 suggests that if $\gamma^{-1} \approx \frac{\|B_1 A_1^{-1} B_1^\top\|}{\sigma_{\min}(B_2)^2}$, then all singular values of the (1, 2) block of the preconditioned matrix will be close to one, and thus the spectrum of SP_{ad}^{-1} will be nicely clustered. Since estimating $\sigma_{\min}(B_2)$ is expensive, we content ourselves with the consideration that the finite eigenvalues μ and corresponding eigenvectors x of the pencil $(B_2 D^{-1} B_2^\top, B_1 A_1^{-1} B_1^\top)$ satisfy

$$\mu = \frac{x^\top B_2 D^{-1} B_2^\top x}{x^\top B_1 A_1^{-1} B_1^\top x},$$

(cf. Proposition 2.2) where B_2 depends on $O(h_t / h_x)$ and B_1 is independent of the mesh [3]. Therefore, for $h_x \approx h_t$, if D mimics A_1 then all μ are not too far from one, and the matrix $\widehat{B}_2$ is not ill-conditioned. This reasoning leads us to set

$$\gamma := \|A_1\|, \tag{4.4}$$

and our numerical experiments seem to support this choice. This choice is in agreement with the argument in [3, Sect. 6] where it is stated that γ should scale like $h_x^2 h_t$, the way A_1 does indeed. General qualitative arguments for other choices of this augmentation parameter can also be found in [15].

To continue our analysis of P_{ad} as introduced in (2.3), we rewrite the eigenvalue problem $P_{ad}^{-\frac{1}{2}} S P_{ad}^{-\frac{1}{2}} u = \lambda u$ as

$$\begin{bmatrix} 0 & & \widehat{B}_{12}^\top \\ & I & \widehat{B}_{22}^\top \\ \widehat{B}_{12} & \widehat{B}_{22} & -\widehat{T} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \text{with} \quad \begin{aligned} \widehat{B}_{12} &= \tfrac{1}{\sqrt{\gamma}} C^{-\frac{1}{2}} B_{12}, \\ \widehat{B}_{22} &= C^{-\frac{1}{2}} B_{22} \Omega^{-\frac{1}{2}}, \end{aligned} \tag{4.5}$$

and $\widehat{T} = C^{-\frac{1}{2}} B_1 A_1^{-1} B_1^\top C^{-\frac{1}{2}}$; here we have split the original B_2 as $[B_{12}, B_{22}]$ conformingly with A_2 and also D in (4.3). This explicit form shows that a necessary condition for the nonsingularity of the reduced system is that $\widehat{B}_{12}$, and thus also B_{12}, be full column rank. For this type of structure it is possible to derive sharp estimates for its positive and negative spectral intervals.

Proposition 4.1 *Consider the preconditioner P_{ad} in (2.3) with*

$$\widetilde{C}(D) = C(D) = B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top \quad \text{and} \quad D = Q^\top Q + \gamma \mathcal{I}_Q, \quad \gamma > 0.$$

Then the eigenvalues of $S P_{ad}^{-1}$ are contained in $\mathcal{I}^- \cup \mathcal{I}^+$, with

$$\mathcal{I}^- = \left[-1, \frac{1}{2}(1 - \sqrt{5}) \right], \qquad \mathcal{I}^+ = \left[\frac{-1 + \sqrt{1 + 4\sigma_{\min}^2}}{2}, \frac{1}{2}(1 + \sqrt{5}) \right],$$

where $\sigma_{\min} \leq 1$ is the smallest nonzero singular value of $\widehat{B}_{12}$ in (4.5).

Proof We readily notice that if $z = 0$ in (4.5), then from the first block equation it also follows that $x = 0$ ($\lambda \neq 0$, because of the nonsingularity of the matrix). Moreover, from the second and third equations we obtain $y = \lambda y$ and $\widehat{B}_{22} y = 0$. Since $\widehat{B}_{22}$ has full column rank, then the second equation is only satisfied for $y = 0$, which cannot hold as it would imply a zero eigenvector. From now on we can thus assume that $z \neq 0$.

With the notation introduced in Proposition 2.1, we first notice that the $(1, 1)$ block in (4.5), namely $\widehat{A}_2 = \text{blkdiag}(0, I)$, satisfies $\lambda_{\max} = 1$. Moreover, since $\widehat{T} + \widehat{B}_2 \widehat{B}_2^\top = I$ and both matrices on the left-hand side are positive semidefinite, we have $\gamma_{\max} \leq 1$. A similar reasoning shows that $\sigma_{\max}(\widehat{B}_2) \leq 1$, where $\widehat{B}_2 = [\widehat{B}_{12}, \widehat{B}_{22}]$. Therefore, Proposition 2.1 provides the right extreme of $\mathcal{I}^+$. Finding a sharp left positive extreme is more cumbersome, as a mere application of Proposition 2.1 would give a loose bound.

Let $\lambda > 0$ be an eigenvalue of (4.5). Substituting the first and second matrix equations in the third one for $\lambda \neq 0$, $\lambda \neq 1$, we obtain

$$\frac{1}{\lambda} \widehat{B}_{12} \widehat{B}_{12}^\top z + \frac{1}{\lambda - 1} \widehat{B}_{22} \widehat{B}_{22}^\top z - \widehat{T} z - \lambda z = 0.$$

We thus eliminate the denominator and noticing once again that $\widehat{T} + \widehat{B}_{12} \widehat{B}_{12}^\top + \widehat{B}_{22} \widehat{B}_{22}^\top = I$, we obtain

$$-\lambda^3 z + \lambda^2 \widehat{B}_2 \widehat{B}_2^\top z + \lambda z - \widehat{B}_{12} \widehat{B}_{12}^\top z = 0. \tag{4.6}$$

Let us write $z = z_0 + z_B$ with $\widehat{B}_{12}^\top z_0 = 0$ and $z_B \perp z_0$. We multiply (4.6) from the left by $z_0^\top$ and by $z_B^\top$ in sequence, that is

$$\left(-\lambda^3 + \lambda \right) \|z_0\|^2 + \lambda^2 z_0^\top \widehat{B}_{22} \widehat{B}_{22}^\top z_0 + \lambda^2 z_0^\top \widehat{B}_{22} \widehat{B}_{22}^\top z_B = 0 \tag{4.7}$$

and

$$\left(-\lambda^3 + \lambda \right) \|z_B\|^2 + \lambda^2 z_B^\top \widehat{B}_{22} \widehat{B}_{22}^\top z_0 \tag{4.8}$$

$$+ \lambda^2 z_B^\top \widehat{B}_{12} \widehat{B}_{12}^\top z_B + \lambda^2 z_B^\top \widehat{B}_{22} \widehat{B}_{22}^\top z_B - z_B^\top \widehat{B}_{12} \widehat{B}_{12}^\top z_B = 0. \tag{4.9}$$

We note that for $z_B = 0$ (and thus $z_0 \neq 0$), and using $\lambda > 0$, equation (4.7) gives $(-\lambda^2 + 1)\|z_0\|^2 + \lambda \|\widehat{B}_{22}^\top z_0\|^2 = 0$. Using $\lambda \|\widehat{B}_{22}^\top z_0\|^2 \geq 0$ we get $(-\lambda^2 + 1)\|z_0\|^2 \leq 0$,

so that $\lambda \geq 1$. In the following we can assume that $z_B \neq 0$. Moreover, we can assume that $\lambda < 1$, otherwise 1 would be the sought after extreme. From (4.7) we get

$$\lambda^2 z_0^\top \widehat{B}_{22} \widehat{B}_{22}^\top z_B = -\lambda \|z_0\|^2 + \lambda^3 \|z_0\|^2 - \lambda^2 \|z_0^\top \widehat{B}_{22}\|^2$$
$$= -\lambda(1 - \lambda^2) \|z_0\|^2 - \lambda^2 \|z_0^\top B_{22}\|^2 \leq 0,$$

so that from (4.8)–(4.9) we obtain

$$0 \leq (-\lambda^3 + \lambda)\|z_B\|^2 + \lambda^2 z_B^\top \widehat{B}_{12} \widehat{B}_{12}^\top z_B + \lambda^2 z_B^\top \widehat{B}_{22} \widehat{B}_{22}^\top z_B - z_B^\top \widehat{B}_{12} \widehat{B}_{12}^\top z_B,$$

that is, $\lambda^3 \|z_B\|^2 - \lambda^2 z_B^\top \widehat{B}_2 \widehat{B}_2^\top z_B - \lambda \|z_B\|^2 + \|\widehat{B}_{12}^\top z_B\|^2 \leq 0$. Therefore we obtain the inequality (for $z_B \neq 0$)

$$\lambda^3 - \lambda^2 \|\widehat{B}_2\|^2 - \lambda + \sigma_{\min}^2 \leq 0,$$

where $\sigma_{\min}$ is the smallest nonzero singular value of $\widehat{B}_{12}$. Finally, using $\lambda^3 \geq 0$ and $\|\widehat{B}_2\| \leq 1$ (cf. Proposition 2.2) we obtain $\lambda^2 + \lambda - \sigma_{\min}^2 \geq 0$ from which the lower extreme of $\mathcal{I}_+$ follows.

Let $\lambda < 0$. We multiply (4.6) from the left by $z^\top$ and notice that $z^\top \widehat{B}_2 \widehat{B}_2^\top z \leq \|z\|^2$, and $-z^\top \widehat{B}_{12} \widehat{B}_{12}^\top z \leq 0$. Therefore,

$$0 \leq \lambda(-\lambda^2 + \lambda + 1)\|z\|^2.$$

Since $\lambda < 0$, it must hold that $-\lambda^2 + \lambda + 1 \leq 0$, from which the bound $\lambda \leq \frac{1}{2}(1 - \sqrt{5})$ follows. To obtain the left-hand bound, we observe that in (4.6) it holds that $z^\top \widehat{B}_2 \widehat{B}_2^\top z \geq z^\top \widehat{B}_{12} \widehat{B}_{12}^\top z$. Therefore, multiplying (4.6) from the left by $z^\top$ we obtain

$$0 \geq \lambda(1 - \lambda^2)\|z\|^2 + (\lambda^2 - 1)z^\top \widehat{B}_{12} \widehat{B}_{12}^\top z,$$

that is, $(1 - \lambda^2)(\lambda \|z\|^2 - z^\top \widehat{B}_{12} \widehat{B}_{12}^\top z) \leq 0$. Since $\lambda \|z\|^2 - z^\top \widehat{B}_{12} \widehat{B}_{12}^\top z \leq 0$ for all $z \neq 0$, it must be $1 - \lambda^2 \geq 0$, from which the bound $-1 \leq \lambda$ follows. $\square$

Remark 4.2 Interval bounds of the type as in Proposition 4.1 could be obtained either directly, or upon matrix row and column permutations, from Proposition 2.1. However, they are either loose, or they require further estimates for some of the involved quantities.

We note that all bounds of Proposition 4.1 except for the left extreme of $\mathcal{I}^+$ are independent of the spectral properties of the blocks and are thus mesh independent. Therefore, the performance of the *exact P_{ad}* (i.e., with $\widetilde{C}(D) = C(D)$) will solely depend on $\sigma_{\min}$, and to a large extent on the choice of γ.

The following example shows that the bounds of Proposition 4.1 are sharp.

Example 4.3 We consider the following data:

$$S = \left[\begin{array}{cc|cc} 0 & 0 & -1 & \delta \\ 0 & 2 & 10 & 1 \\ \hline -1 & 10 & -20 & 0 \\ \delta & 1 & 0 & 0 \end{array} \right]$$

and δ is freely chosen. The preconditioner P_{ad} is determined with $D = \text{diag}([10, 2])$ ($\gamma = 10$) and $C = B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top$. The table below shows the true eigenvalues of SP_{ad}^{-1} and the estimates of Proposition 4.1 for various selections of δ.

δ		Spectral values			
10^{-2}	$\text{spec}(SP_{ad}^{-1})$	-1.0000	-0.61805	0.00603	1.6180
	$\mathcal{I}^-, \mathcal{I}^+$	-1.0000	-0.61803	0.00599	1.6180
10^0	$\text{spec}(SP_{ad}^{-1})$	-1.0000	-0.66646	0.42423	1.5774
	$\mathcal{I}^-, \mathcal{I}^+$	-1.0000	-0.61803	0.33426	1.6180
10^2	$\text{spec}(SP_{ad}^{-1})$	-1.0000	-0.70458	0.99980	1.4194
	$\mathcal{I}^-, \mathcal{I}^+$	-1.0000	-0.61803	0.61797	1.6180

We note that the accuracy of all bounds can vary somewhat with δ, but that for $\delta = 10^{-2}$ all bounds are accurate.

If the exact diagonal block C of P_{ad} is replaced by an accurate approximation $\widetilde{C}$, then the spectrum of SP_{ad}^{-1} does not change significantly, and the four extremes of Proposition 4.1 are each multiplied by a modest constant; cf., e.g., [25]. On the other hand, if $\widetilde{C}$ does not fully capture all spectral features of an ill-conditioned matrix C, then the matrix $C\widetilde{C}^{-1}$ may have an isolated, small cluster of eigenvalues close to zero. This is indeed the situation we encounter with the data in the application problem we discuss in the next section. Theorem 2.4 in [24] ensures that in this case, all eigenvalues of SP_{ad}^{-1} with $\widetilde{C}$ remain quite close to those obtained when using the exact C (cf. Proposition 4.1), except for a small cluster of eigenvalues that scale like the cluster near zero in $C\widetilde{C}^{-1}$. The number of these eigenvalues is strictly related to the number of small eigenvalues of $C\widetilde{C}^{-1}$. We will explore in detail this phenomenon in the next section.

4.1 Numerical experiments

In this section we report on our experience with the reduced dimension linear system stemming from the Monge-Kantorovich problem. The data stem from an image registration problem, where u represents the image density, which is known at time zero (initial image) and at time $T = 1/8$ (final image). The data and discretization procedure by means of finite differences are as described in [3]. We stress here that since discretization in time is performed as for the other variables, once the number of nodes n_x, n_t is selected, each block of the reduced linear systems will have dimension $n_x^2 n_t$. Since our main interest is in the solution of the system, we will not report results for the whole simulation, corresponding to the numerical solution of the nonlinear problem. Instead, in most experiments we select a Newton cycle, typically the first one, and analyze the algebraic linear system to be solved at that cycle. It is important to realize that the numerical results do change at different cycles, although the qualitative performance is the same when comparing different strategies.

We first consider solving the reduced system (4.1) by means of MINRES, preconditioned by P_{ad}, where the matrix D is obtained by replacing zero diagonal entries of

Fig. 1 Value of $\sigma_{\min}(\widehat{B}_{12})$ as γ varies (cf. Proposition 4.1)
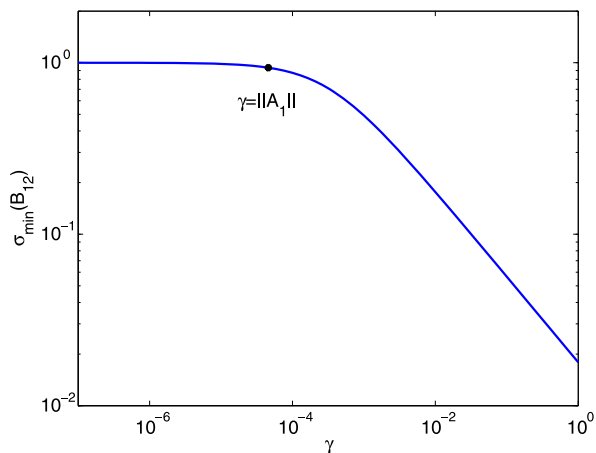
$Q^\top Q$ with γ (cf. (4.3) and (4.4)), and the second diagonal block of P_{ad} is given by $C = B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top$. For the proposed discretization A_1 is diagonal, therefore its Euclidean norm can be computed very cheaply to get γ. To fully appreciate our choice of γ, in Fig. 1 we show the value of $\sigma_{\min}(\widehat{B}_{12})$ in Proposition 4.1, for $\gamma \in [10^{-7}, 1]$ and $n = 1000$ (see below for more explanation on the data). The thick dot refers to the value $\gamma = \|A_1\|$ in (4.4). Recalling that in general $\sigma_{\min} \leq 1$, the estimates of Proposition 4.1 suggest that $\sigma_{\min}$ should be as close as possible to one to have a good clustering, and this is clearly achieved by tiny values of γ. On the other hand, since γ affects the condition number of C and D, it should not be chosen to be too small, otherwise dealing with an inexact version of P_{ad} may become numerically difficult. The plot shows that our choice of $\gamma = \|A_1\|$ appears to be very close to the trade-off value of γ: the largest possible value yielding $\sigma_{\min} \approx 1$. The choice of γ proposed in [3], $\gamma = \text{mean}(\text{diag}(B_2 B_2^\top)/\text{diag}(B_1 A_1^{-1} B_1))) = 0.00249$ clearly gives a smaller value of $\sigma_{\min}(\widehat{B}_{12})$.

The exact application of the preconditioner yields the results shown in Table 4, for different discretization meshes: $n_x \times n_x$ in 2D space, and n_t in time, yielding diagonal blocks both of size $n_x^2 n_t \times n_x^2 n_t$. In addition, VG is of size $2n_x^2 n_t$. Various combinations of n_x, n_t are reported.

The convergence tolerance, based on the preconditioned residual Euclidean norm, was set to 10^{-10}. Note that this is a quite small tolerance, as significantly larger values are used within the Newton iteration; For instance, 10^{-4} was used in [3]. We decided to set a strict tolerance because it allowed us to uncover some phenomena that were not visible otherwise, at least for small size problems and during the first few Newton iterations. In particular, the stagnation phases we will discuss later in this section are visible at the first Newton cycle only for a low residual norm (cf. left plot of Fig. 2), as they also depend on the spectral distribution of the right-hand side. However, for finer discretizations or for different data, these phenomena were visible also at an earlier convergence stage, as shown in the right plot of Fig. 2), where data from the last Newton step were used.

According to Table 4, both discretizations show that this preconditioner is "optimal", in the sense that the number of iterations does not grow with the problem

Table 4 Monge-Kantorovich problem. Performance (CPU time in seconds and # iterations) of MINRES preconditioned by the exact version of P_{ad}. n is the size of each of the (1, 1) and (2, 2) blocks. A dash stands for excessive memory requirements

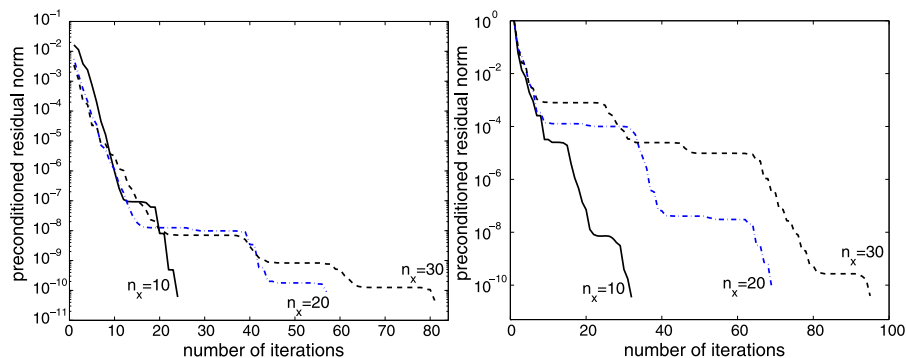| n_x	n_t	n	# it	time	n_x	n_t	n	# it	time
10	10	1000	12	0.19	10	10	1000	12	0.19
15	15	3375	13	1.28	20	10	4000	16	3.44
20	20	8000	14	8.17	30	10	9000	21	34.16
25	25	15625	15	33.31	40	20	32000	21	376.54
30	30	27000	16	109.25	50	20	50000	–	–
35	35	42875	16	309.97					
40	40	64000	–	–					

Fig. 2 Convergence history of MINRES with AMG-based P_{ad}, for three mesh refinements (here $n_x = n_t$). *Left*: System data from first Newton iteration. *Right*: System data from last Newton iteration (at convergence)

size. On the other hand, the computational costs of this "exact" preconditioner become quickly prohibitive, mostly due to the factorization and solution with the matrix $C = C(D) = B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top$. As in the previous application problem, we thus substitute the use of the exact C and its inverse with that of an Algebraic Multigrid operator; in the following experiments we selected once again the code HSL_MI20 [7], with the parameters `control.one_pass_coarsen=1`, `control.v_iterations=5`, `control.st_parameter=.85`. In the following, the cost for generating the Multigrid operator will always be included in the reported timing. We note that C does not stem from the discretization of a standard second order elliptic operator, therefore we do not expect optimality of HSL_MI20. We also remark that the matrix C also arises in the numerical solution of the original problem in [3]: in there, the authors opted for an inner-outer procedure, in which a system with C is solved by means of a Krylov subspace method at each (outer) iteration. We do not explore this possibility here, as this involves the introduction of an extra parameter, the inner stopping tolerance; moreover, it was shown in [3] that this approach still displays some mesh dependence.

Table 6 shows that the computational costs decrease significantly when using the multigrid operator, although mesh independence seems to be lost, as the number of iterations grows with the grid refinement. This phenomenon will be analyzed more closely in the following. But first we would like to notice that our numerical experiments seem to support our choice of γ. Indeed, for γ as in [3] and $n = 1000$, MINRES with P_{ad} and AMG converged in 69 iterations and 0.47 seconds; for $n = 27,000$ the method converged in 139 iterations and 38 seconds. These figures should be compared with the corresponding ones in Table 6. We notice, however, that the value of γ proposed by the authors of [3] was taylored towards their solution method, which differs from ours, therefore different performance may be expected in the two cases.

A close look at the convergence history of MINRES reveals that the convergence behavior is a little more subtle than the numbers in Table 6 suggest, and that phases of (almost) stagnation alternate with much steeper convergence rates. This pattern can be appreciated in Fig. 2, where the MINRES convergence history for three mesh refinements is reported. The left plot uses data from the first Newton iteration, while the right plot uses those from the last Newton iteration, at convergence. We recall that both the coefficient matrix and the right-hand side change at each Newton iteration. Especially for the data in the first Newton iteration, we see that for a large portion of the convergence history, all curves look alike, suggesting an almost mesh independence. Moreover, only at some stage stagnation occurs in each of the histories. Such a pattern seems to indicate that the spectral distribution of $S P_{ad}^{-1}$ is good, except perhaps for a few badly behaved eigenvalues, which delay convergence. Note also that stagnation occurs at different convergence stages in the two plots, and this mainly depends on the right-hand side eigendecomposition. Therefore, to completely avoid the stagnation phase, it would not be sufficient to stop the MINRES iteration earlier by setting a looser stopping tolerance, as stagnation may still occur.

As already mentioned, this setting was recently analyzed in [24], where it is shown that the occurrence of such stagnation pattern may be associated with the non-full optimality of the preconditioning matrix. More precisely, assume that $\widetilde{C}$ is an approximation to C with both matrices symmetric and positive definite. It is shown in [24] that if the pencil $(C, \widetilde{C})$ has a small group of *positive* eigenvalues close to zero, well separated from the other eigenvalues, then the preconditioned matrix $S P_{ad}^{-1}$ will inherit such a cluster: an isolated group of *negative* eigenvalues close to zero arises. Such a cluster of "badly" behaved eigenvalues causes temporary stagnation of MINRES, while the method tries to resolve the cluster.

Table 5 reports the approximation by means of 50 iterations of the Lanczos process (see, e.g., [1]) of the smallest 10 eigenvalues of $(C, \widetilde{C})$ for the three mesh refinements considered in Fig. 2. In all cases the small group of well separated eigenvalues is readily visible. It is also interesting to notice that the cluster moves towards zero with the mesh refinements, causing a more severe stagnation phase for finer meshes. This observation is in agreement with the curves shown in Fig. 2.

We also notice that the occurrence of the small cluster does not seem to be due to some anomalous behavior of the AMG preconditioner. Indeed, the original matrix C does have a few small eigenvalues, which are apparently not captured by the approximation $\widetilde{C}$, and thus they are still visible in the pencil $(C, \widetilde{C})$. We observed a similar behavior when using an incomplete LU decomposition as $\widetilde{C}$. The occurrence of this

Table 5 First approximate eigenvalues of the pencil $(C, \widetilde{C})$

approx λ_i	$n_x = 10$	$n_x = 20$	$n_x = 30$
$i = 1$	9.4632e-04	1.7676e-05	1.3247e-05
$i = 2$	9.4999e-04	1.8274e-05	1.3937e-05
$i = 3$	8.5302e-01	1.5449e-04	4.6116e-05
$i = 4$	8.5400e-01	1.5484e-04	4.6200e-05
$i = 5$	8.5722e-01	6.2973e-01	3.7707e-01
$i = 6$	8.5752e-01	6.3079e-01	3.7791e-01
$i = 7$	8.6410e-01	6.6089e-01	3.8408e-01
$i = 8$	8.6825e-01	6.6091e-01	3.8409e-01
$i = 9$	8.7875e-01	6.6557e-01	4.0425e-01
$i = 10$	8.8097e-01	6.6674e-01	4.0514e-01

Table 6 Performance results of exact, inexact and deflated preconditioning variants. CPU time is in seconds. Data are taken from the first Newton iteration. The last column reports the time used to obtain the approximate eigenvectors of $(C, \widetilde{C})$ and the (automatically chosen) number of approximate eigenvectors retained

n_x	n_t	n	P_{ad}		P_{ad} w/AMG		P_{ad} w/AMG+defl		
			# it	time	# it	time	# it	time	
10	10	1000	12	0.19	24	0.29	17	0.26	+0.52, 2 eigs
15	15	3375	13	1.28	52	1.43	25	0.80	+1.53, 5 eigs
20	20	8000	14	8.17	57	4.64	21	1.95	+4.51, 4 eigs
25	25	15625	15	33.31	85	14.64	36	6.68	+9.57, 5 eigs
30	30	27000	16	109.25	81	28.37	28	10.61	+19.03, 4 eigs
35	35	42875	16	309.97	122	65.82	47	27.33	+29.82, 5 eigs
40	40	64000	–		90	88.63	37	37.82	+52.23, 4 eigs
20	10	4000	16	3.44	46	1.49	18	0.70	+1.74, 4 eigs
30	10	9000	21	34.16	68	5.26	25	2.12	+4.36, 4 eigs
40	20	32000	21	376.54	76	31.63	32	12.55	+21.05, 4 eigs
50	20	50000	–		94	59.96	39	24.28	+33.21, 4 eigs

small cluster in C is problem and discretization dependent, therefore such a difficulty might not arise when analyzing slightly different application problems.

To cure the delay due to small tiny clusters, a deflation strategy is proposed in [24], which can be easily incorporated in the usual preconditioned MINRES iteration. The approach consists of employing *approximate* eigenvectors of the negative cluster of SP_{ad}^{-1} to enrich the method recurrence, so that no iterations are used to approximate the corresponding quantities. It is important to realize that the approximate eigenvector does not have to be accurate, so that cheap devices can be used: if y is an eigen-approximation to an exact eigenvector of the pencil $(C, \widetilde{C})$, then the vector $[-D^{-1}B_1^\top y; y]$ may be a good approximation to an eigenvector of the pencil (S, P_{ad}) [24]. We adopted this strategy in the following experiments, where the

Fig. 3 Convergence history of MINRES with AMG-based P_{ad} without and with approximate deflation strategy; $n_x = n_t = 20$
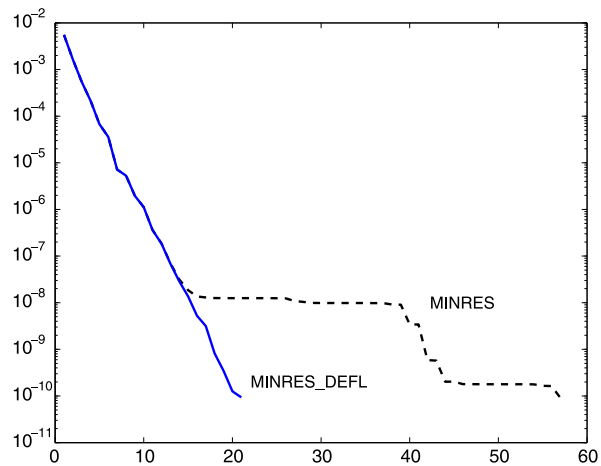

Table 7 j_*th Newton iteration. Solution via MINRES with the augmentation procedure when the approximate eigenvectors are computed either at iteration j_* or at the first iteration. Reported are number of iterations and CPU time in seconds

n_x	n_t	j_*	defl. vec. at j_*th it.		defl. vec. at first it.	
			# it	time	# it	time
20	10	4	21	0.80	21	0.79
30	10	5	30	2.29	30	2.26
40	20	5	38	15.81	38	16.55

vectors y were obtained with the Lanczos iteration [1], with a single cycle of dimension 50 on $(C, \widetilde{C})$. The retained approximate eigenvectors were selected as those corresponding to eigenvalues of $(C, \widetilde{C})$ less than 10^{-1}. A typical convergence history is depicted in Fig. 3 where the convergence improvement can be fully appreciated. A more complete account of the results obtained with this approach is summarized in the right portion of Table 6. The rightmost values report the CPU time used to generate these approximations, and the number of eigenvectors retained. Although these timings are high with respect to the solution time, it should be kept in mind that this information can be recycled in subsequent Newton steps and thus can be fully amortized. As an example, Table 7 reports the results from solving the system at the j_*th Newton iteration (last iteration before convergence) when the augmenting approximate eigenvectors are computed: (i) at the same j_* iteration, or (ii) at the first iteration. No significant difference is observed between the two choices, suggesting that the approximate eigenvectors computed at the first Newton iteration can be effectively used throughout the whole process. If additional spectral information on the problem is available beforehand, then further savings can be obtained by avoiding the use of an eigenvalue routine.

For the sake of comparison, we also tried different parameter values in `hsl_mi20`, namely we selected `control.v_iterations=10` while all other values were selected as default (and not the ones chosen in previous experiments). The results

Table 8 Performance results of P_{ad} w/AMG, where the only non-default parameter was set to `control.v_iterations` =10. The digits report number of iterations and CPU time with data at the first Newton iteration

n_x	n_t	n	P_{ad} w/AMG(V_IT=10)	
			# it	time
10	10	1000	25	0.52
15	15	3375	29	1.82
20	20	8000	35	6.20
25	25	15625	43	16.81
30	30	27000	31	25.78
35	35	42875	56	70.61
40	40	64000	48	93.38
20	10	4000	33	2.54
30	10	9000	44	9.79
40	20	32000	40	45.22
50	20	50000	35	67.06

are reported in Table 8, and show that a lower number of iterations can be achieved by allowing more V-cycles within the AMG operator; note however that the number of iterations fluctuates somewhat, although it does not seem to have an increasing pattern. We also notice that no clear stagnation phases could be detected during the convergence history in all cases. However, the costs of the AMG-based preconditioner are significantly higher, compared with those obtained with the setting used in previous experiments. These results show the expected fact that by allowing a more accurate AMG preconditioner, performance in terms of number of iterations improves, at the price of higher costs. Nonetheless, other preconditioning techniques specifically adapted to this type of matrices may be better suited than the generic AMG operator we have used. With the current setting, the use of the deflation strategy appears to be more effective.

Remark 4.4 We also implemented the indefinite preconditioner P_{indef} in (2.4) for this problem. However, the eigenvalue distribution of the preconditioned matrix, even in the exact case (that is, with $\widetilde{B}_2 = B_2$) was not favorable, as the (nonsymmetric) solver showed some stagnation phase. We could adopt a deflation-type strategy similar to what we proposed for the inexact P_{ad}, however the occurrence of a nonsymmetric matrix requires much more care in the implementation and accuracy of the deflation strategy, which is beyond the aim of this paper.

Remark 4.5 Due to the more complex nature of the problem, we also experimented with nonsymmetric preconditioners, which require the use of a nonsymmetric solver. In particular, we explored a block (augmented) triangular preconditioner in the style of [3], namely

$$P_{ab} = \begin{bmatrix} D & B_2^\top \\ 0 & -C \end{bmatrix}, \quad C \approx B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top, \tag{4.10}$$

and D is a symmetric positive definite matrix completing the singular matrix A_2. In our tests, D and C were selected as in P_{ad}. However, for the realistic case where C

is an (AMG) approximation to $B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top$, we observed convergence delays very similar to those described in Remark 4.4. For these reasons, we did not explore this preconditioner further. Note that in [3], where (4.10) was used to precondition (1.1), approximate solves with $B_1 A_1^{-1} B_1^\top + B_2 D^{-1} B_2^\top$ were performed by an AMG-preconditioned Conjugate Gradient (PCG) iteration, giving rise to an inner-outer method, in which the inner PCG stopping tolerance was an extra parameter to be tuned; The outer method was necessarily a flexible solver [32].

5 Conclusions

We have shown that certain PDE-constrained optimization problems lead to structured algebraic linear systems, whose dimension can be significantly reduced without major computational overhead, thanks to the properties of the matrices defining the blocks. We have explored the numerical solution of these reduced problems by using different structured preconditioning techniques, reporting the effectiveness of the proposed strategies on two typical problems with data stemming from realistic applications.

References

1. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.): Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM, Philadelphia (2000)
2. Bai, Z.-Z., Benzi, M., Chen, F., Wang, Z.-Q.: Preconditioned MHSS iteration methods for a class of block two-by-two linear systems with applications to distributed control problems. Technical Report 2011-001, Math/CS Department, Emory University (January 2011). To appear in IMA J. Numer. Anal.
3. Benzi, M., Haber, E., Taralli, L.: A preconditioning technique for a class of PDE-constrained optimization problems. Adv. Comput. Math. **35**, 149–173 (2011)
4. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
5. Bergamaschi, L., Gondzio, J., Venturin, M., Zilli, G.: Inexact constraint preconditioners for linear systems arising in interior point methods. Comput. Optim. Appl. **36**(2–3), 137–147 (2007)
6. Borzì, A., Schulz, V.: Multigrid methods for PDE optimization. SIAM Rev. **51**(2), 361–395 (2009)
7. Boyle, J., Mihajlović, M.D., Scott, J.A.: HSL_MI20: an efficient AMG preconditioner. Int. J. Numer. Methods Eng., **82**(1), 64–98 (2010)
8. Brandt, A.: Algebraic multigrid theory: the symmetric case. Appl. Math. Comput. **19**, 23–56 (1986)
9. Cao, Z.-H.: Augmentation block preconditioners for saddle point-type matrices with singular (1, 1) blocks. Numer. Linear Algebra Appl. **15**, 515–533 (2008)
10. Dollar, H.S., Gould, N.I.M., Schilders, W.H.A., Wathen, A.J.: Implicit-factorization preconditioning and iterative solvers for regularized saddle-point systems. SIAM J. Matrix Anal. Appl. **28**, 170–189 (2006)

11. Durazzi, C., Ruggiero, V.: Indefinitely preconditioned conjugate gradient method for large sparse equality and inequality constrained quadratic problems. Numer. Linear Algebra Appl. **10**, 673–688 (2003)
12. Elman, H.C., Silvester, D.J., Wathen, A.J.: Finite Elements and Fast Iterative Solvers, with Applications in Incompressible Fluid Dynamics. Numerical Mathematics and Scientific Computation, vol. 21. Oxford University Press, London (2005)
13. Gould, N., Simoncini, V.: Spectral analysis of saddle point matrices with indefinite leading blocks. SIAM J. Matrix Anal. Appl. **31**, 1152–1171 (2009)
14. Gould, N.I.M., Hribar, M.E., Nocedal, J.: On the solution of equality constrained quadratic programming problems arising in optimization. SIAM J. Sci. Comput. **23**, 1376–1395 (2001)
15. Greif, C., Schötzau, D.: Preconditioners for saddle point linear systems with highly singular (1, 1) blocks. Electron. Trans. Numer. Anal. **22**, 114–121 (2006)
16. Greif, C., Overton, M.: An analysis of low-rank modifications of preconditioners for saddle point systems. Electron. Trans. Numer. Anal. **37**, 307–320 (2010)
17. Haber, E., Ascher, U.: Preconditioned all-at-once methods for large, sparse parameter estimation problems. Inverse Probl. **17**, 1847–1864 (2001)
18. Herzog, R., Sachs, E.: Preconditioned conjugate gradient method for optimal control problems with control and state constraints. SIAM J. Matrix Anal. Appl. **31**(5), 2291–2317 (2010)
19. Krzyzanowski, P.: On block preconditioners for saddle point problems with singular or indefinite (1, 1) block. Numer. Linear Algebra Appl. **18**(1), 123–140 (2011)
20. Lukšan, L., Vlček, J.: Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems. Numer. Linear Algebra Appl. **5**, 219–247 (1998)
21. Mardal, K.-A., Winther, R.: Preconditioning discretizations of systems of partial differential equations. Numer. Linear Algebra Appl. **18**(1), 1–40 (2011)
22. The MathWorks, Inc. *MATLAB 7*, September 2004
23. Oberle, H.J., Grimm, W.: BNDSCO-A Program for the numerical solution of optimal control problems. Technical report, Institute for Flight Systems Dynamics, DLR, Oberpfaffenhofen (1989)
24. Olshanskii, M., Simoncini, V.: Acquired clustering properties and solution of certain saddle point systems. SIAM J. Matrix Anal. Appl. **31**, 2754–2768 (2010)
25. Perugia, I., Simoncini, V.: Block–diagonal and indefinite symmetric preconditioners for mixed finite element formulations. Numer. Linear Algebra Appl. **7**, 585–616 (2000)
26. Rees, T., Stoll, M., Wathen, A.: All-at-once preconditioning in PDE-constrained optimization. Kybernetika **46**(2), 341–360 (2010)
27. Rees, T., Wathen, A.J.: Preconditioning iterative methods for the optimal control of the Stokes equations. SIAM J. Sci. Comput. **33**, 2903–2926 (2011)
28. Rees, T., Greif, C.: A preconditioner for linear systems arising from interior point optimization methods. SIAM J. Sci. Comput. **29**(5), 1992–2007 (2007)
29. Rees, T.: Preconditioning iterative methods for PDE constrained optimization. Ph.D. thesis, University of Oxford (2010)
30. Rees, T., Stoll, M.: Block-triangular preconditioners for PDE-constrained optimization. Numer. Linear Algebra Appl. **17**(6), 977–996 (2010)
31. Ruge, J.W., Stüben, K.: Algebraic multigrid. In: Multigrid Methods, Frontiers Appl. Math. SIAM, Philadelphia (1987)
32. Saad, Y.: A flexible inner-outer preconditioned GMRES. SIAM J. Sci. Comput. **14**, 461–469 (1993)
33. Schöberl, J., Zulehner, W.: Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. SIAM J. Matrix Anal. Appl. **29**(3), 752–773 (2007)
34. Sesana, D., Simoncini, V.: Spectral analysis of inexact constraint preconditioning for symmetric saddle point matrices. Technical report, Dipartimento di Matematica, Università di Bologna (2010)
35. Silvester, D., Mihajlovic, M.: A black-box multigrid preconditioner for the biharmonic equation. BIT Numer. Math. **44**, 151–163 (2004)
36. Silvester, D., Wathen, A.: Fast iterative solution of stabilized Stokes systems part II: using general block preconditioners. SIAM J. Numer. Anal. **31**, 1352–1367 (1994)
37. Simoncini, V.: Solution of structured algebraic linear systems in PDE-constrained optimization problems. Available at http://www.dm.unibo.it/~simoncin/, July 2010. Slides of the talk given at the "Erice 2010 Workshop on Nonlinear Optimization, Variational Inequalities and Equilibrium Problems, 2–10 July 2010"
38. Simoncini, V., Szyld, D.B.: Recent computational developments in Krylov subspace methods for linear systems. Numer. Linear Algebra Appl. **14**(1), 1–59 (2007)

39. Stoll, M., Wathen, A.: All-at-once solution of time-dependent PDE-constrained optimization problems. Technical Report 1017, The Mathematical Institute, University of Oxford (2010)
40. Thorne, H.S.: Distributed control and constraint preconditioners. Technical Report RAL-TR-2010-016, Rutherford Appleton Laboratory (2010)
41. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods and Applications. Am. Math. Soc., Providence (2010). Translated by J. Sprekels
42. Wathen, A., Rees, T.: Chebyshev semi-iteration in preconditioning for problems including the mass matrix. Electron. Trans. Numer. Anal. **34**, 125–135 (2008–2009)
43. Zulehner, W.: Nonstandard norms and robust estimates for saddle point problems. SIAM. J. Matrix Anal. Appl. **32**, 536 (2011)