Available online at www.sciencedirect.com

ScienceDirect

Journal of Computational and Applied Mathematics 215 (2008) 28–37

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

Limiting accuracy of segregated solution methods for nonsymmetric saddle point problems

Pavel Jiránek[a,*,1], Miroslav Rozložník[a,b,2]

[a] *Department of Modelling of Processes, Technical University of Liberec, Hálkova 6, CZ-461 17 Liberec, Czech Republic*
[b] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic*

Received 25 October 2006

Abstract

Nonsymmetric saddle point problems arise in a wide variety of applications in computational science and engineering. The aim of this paper is to discuss the numerical behavior of several nonsymmetric iterative methods applied for solving the saddle point systems via the Schur complement reduction or the null-space projection approach. Krylov subspace methods often produce the iterates which fluctuate rather strongly. Here we address the question whether large intermediate approximate solutions reduce the final accuracy of these two-level (inner–outer) iteration algorithms. We extend our previous analysis obtained for symmetric saddle point problems and distinguish between three mathematically equivalent back-substitution schemes which lead to a different numerical behavior when applied in finite precision arithmetic. Theoretical results are then illustrated on a simple model example.

MSC: 65N22; 65F10; 65G50; 15A06

Keywords: Saddle point problems; Schur complement reduction method; Null-space projection method; Rounding error analysis

1. Introduction

We consider a solution of the generalized saddle point problem with 2×2 block structure

$$\begin{pmatrix} A & B \\ B^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{1}$$

where $A \in \mathbb{R}^{n,n}$ is generally nonsymmetric ($A^{\mathrm{T}} \neq A$) nonsingular matrix and $B \in \mathbb{R}^{n,m}$ has a full column rank $m \leqslant n$. These systems arise in many application areas including the discretizations of partial differential equations in computational fluid dynamics and solid state physics, constrained optimization, optimal control, etc. For a wide overview of applications leading to (nonsymmetric) saddle point problems and solution approaches, we refer to [2].

* Corresponding author.
E-mail address: pavel.jiranek@tul.cz (P. Jiránek).

[1] The work of this author was supported by the MSMT CR under the project 1M0554 “Advanced Remedial Technologies”.

[2] The work of this author was supported by the project 1ET400300415 within the National Program of Research “Information Society” and by the Institutional Research Plan AV0Z10300504 “Computer Science for the Information Society: Models, Algorithms, Applications”.

0377-0427/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2007.03.021

In this paper we look at the numerical behavior of two main representatives of a segregated solution approach called the Schur complement reduction and the null-space projection method. They are both based on the transformation of the original system (1) to the reduced (and still a nonsymmetric) system for one from the two solution components x and y which is then solved with some iterative method. Such a process produces a sequence of iterates y_{k+1} for the Schur complement method and x_{k+1} for the null-space method ($k = 0, 1, \ldots$). At each step of a segregated method we need to multiply a vector by a reduced system matrix, which involves a solution of some inner system (a nonsingular system with the matrix A or the full-rank least squares problem with the matrix B). The approximation to the corresponding second component of the solution vector is then obtained by the back-substitution into (1) which leads to the solution of another linear or the least squares system either with the block A or with the block B, respectively. Since we cannot solve such problems exactly, in practice we have only the approximations to their solutions. These vectors, however, can be interpreted as the (exact) solutions to (slightly) perturbed systems, where the perturbation matrix (or the error matrix) measuring the inexactness of the computation changes every time when the application of inner solver is required.

Our motivation here is to analyze what is the best accuracy we can get from such (inexact) schemes when implemented in finite precision arithmetic (see also the connection to [20,17]). Due to the associated rounding errors we must expect that there is a limitation to the accuracy of computed approximate solutions $\bar{x}_{k+1}$ and $\bar{y}_{k+1}$. In fact, typically the norms of so-called true residuals $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ and $-B^{\mathrm{T}}\bar{x}_{k+1}$ stagnate from certain point on. We say then that the level of maximum attainable (or limiting) accuracy has been reached. This quantity was analyzed [11] for inexact saddle point solvers, where in addition to our current assumptions we assume the symmetric positive-definite block A. The reduced Schur complement and null-space projection systems are then symmetric positive (semi)definite and therefore the energy-norm minimizing conjugate gradient (CG) method [10] or the related conjugate residual (CR) method [19] (minimizing the residual norm) can be efficiently applied. Indeed it was shown in [11] that the bounds on the maximum attainable accuracy depend very much on the actual implementation of back-substitution formulas for computing the corresponding approximate solutions (see also the discussion in next two sections). In addition, all developed bounds depend on the largest norm of the iterates (either $\bar{x}_i$ or $\bar{y}_i$) during the full iteration $i = 0, 1, \ldots, k + 1$. For CG or CR methods (also for the steepest descent method or any other error/residual norm minimizing method—with respect to any fixed norm), where the error norm or the residual norm, respectively, are known to converge monotonously, these bounds depend actually on the initial approximate solutions x_0 and y_0. In such cases this term does not play an important role. The situation is significantly different when considering a nonsymmetric block A. It turns out that for general nonsingular systems the GMRES [16] method with iterates satisfying the residual minimizing property cannot be implemented with short recurrences so that work and storage requirements per iteration would be low and roughly constant. On the other hand, there are nonsymmetric iterative methods (such as Bi-CG [6] or CGS [18]) that are known to produce very large intermediate approximate solutions (and residuals). The oscillation in their norms may then affect the maximum attainable accuracy of the scheme. Our aim here is to analyze this effect for various implementations of back-substitution in the Schur complement reduction and null-space projection method and for various iterative methods for solving the reduced systems. In particular, we consider the residual minimizing GMRES method, the Bi-CG, CGS and CGNE [4] methods and we illustrate our theoretical results on a simple numerical example.

Throughout the text we denote the unit roundoff by u, $\|x\|$ denotes the 2-norm of a real vector x. If A is a real (rectangular) matrix, $\|A\|$ stands for the 2-norm of A and $\kappa(A)$ is its condition number. For the notation of a least squares problem we use the symbol $\approx$, i.e., $Bu \approx v$ means that we are looking for the least squares solution $u = \arg\min_w \|v - Bw\| = B^{\dagger}v$, where $B^{\dagger}$ is the Moore–Penrose pseudoinverse of B.

2. Segregated methods for the solution of saddle point problems

The Schur complement reduction method is based on the elimination of the unknown vector x from system (1). This leads to the system for the unknown vector y

$$B^{\mathrm{T}}A^{-1}By = B^{\mathrm{T}}A^{-1}f, \tag{2}$$

where $B^{\mathrm{T}}A^{-1}B$ is negative Schur complement of the block A in the whole system matrix of (1). Assume that the iterative method applied to the Schur complement system (2) produces approximations to y in the form $y_{k+1} = y_k + \alpha_k p_k^{(y)}$ ($k = 0, 1, 2, \ldots$). The corresponding approximate solution component x_{k+1} is computed using the first

equation of (1) via

$$x_{k+1} = A^{-1}(f - By_{k+1}) = x_k - A^{-1}Bp_k^{(y)} = x_k + A^{-1}(f - Ax_k - By_{k+1}) \tag{3}$$

with $x_0 = A^{-1}(f - By_0)$. The algorithm of the Schur complement reduction method is as follows.

Algorithm 1. *The Schur complement reduction method*

1. *choose* y_0 (*e.g.*, $y_0 = 0$)
2. *solve* $Ax_0 = f - By_0$
3. *compute* $r_0^{(y)} = -B^{\mathrm{T}}x_0$
4. *for* $k = 0, 1, 2, \ldots$ *until convergence*
5. *compute* α_k *and* $p_k^{(y)}$
6. *solve* $Ap_k^{(x)} = -Bp_k^{(y)}$
7. *compute* $v_k^{(y)} = B^{\mathrm{T}}p_k^{(x)}$
8. *update* $y_{k+1} = y_k + \alpha_k p_k^{(y)}$
9. *update* $r_{k+1}^{(y)} = r_k^{(y)} - \alpha_k v_k^{(y)}$
10. *compute* x_{k+1}:
    (a) *update* $x_{k+1} = x_k + \alpha_k p_k^{(x)}$ *(generic update)*
    (b) *solve* $Ax_{k+1} = f - By_{k+1}$ *(direct substitution)*
    (c) *solve* $Au_{k+1} = f - Ax_k - By_{k+1}$,
    update $x_{k+1} = x_k + u_k$ *(corrected direct substitution)*
11. *end*

In lines 6 and 7 as well as in lines 2 and 10 implementing the back-substitution (3) the multiplication of a direction vector $p_k^{(y)}$ by the Schur complement matrix $-B^{\mathrm{T}}A^{-1}B$ is performed which requires a solution of a system with the matrix A. Since in practice these systems cannot be solved exactly, we assume that they are solved with a backward error τ, i.e., that the computed solution $\bar{v}$ of the system $Av = b$ is the exact solution of a perturbed system $(A + \Delta A)\bar{v} = b + \Delta b$ with $\|\Delta A\| \leqslant \tau\|A\|$ and $\|\Delta b\| \leqslant \tau\|b\|$. To preserve a nonsingularity of $A + \Delta A$ we also assume that $\tau\kappa(A) \ll 1$. Based on (3) here we consider three different but mathematically equivalent schemes for computation of the approximate solution component x_{k+1}. They are listed in line 10 of Algorithm 1. The cheapest scheme, called generic update [14,1], computes the approximation x_{k+1} via only one saxpy operation (since we need the computation of the vector $p_k^{(x)}$ anyway). The second scheme, called direct substitution [5], requires the additional solution of a system with the matrix A. Note that since this scheme depends only on the actually computed iterate y_{k+1}, we can perform the computation of x_{k+1} only when we need it (e.g., at the end of the iteration process). The third scheme is called corrected direct substitution [3]. As in direct substitution, its formula needs to solve an additional system with A at each step (together with two extra saxpy operations).

The null-space projection method is based on the projection of the first equation in (1) onto the null-space of the matrix B^{T} (denoted as $N(B^{\mathrm{T}})$). Since the second equation of (1) ensures that the unknown x can be found in this subspace, we look for the approximate solution $x_k \in N(B^{\mathrm{T}})$ and we obtain the system

$$(I - \Pi)A(I - \Pi)x = (I - \Pi)f, \tag{4}$$

where $I - \Pi = I - BB^{\dagger}$ is the orthogonal projector onto $N(B^{\mathrm{T}})$. This system is solved by an iterative method producing the sequence of iterates $x_{k+1} = x_k + \alpha_k p_k^{(x)}$ ($k = 0, 1, \ldots$). Considering (1) the approximate solution y_{k+1} is then obtained from x_{k+1} using the formulas

$$y_{k+1} = B^{\dagger}(f - Ax_{k+1}) = y_k + B^{\dagger}\left(r_k^{(x)} - \alpha_k Ap_k^{(x)}\right) = y_k + B^{\dagger}(f - Ax_{k+1} - By_k). \tag{5}$$

Here we summarize the algorithm of the null-space projection method.

Algorithm 2. *The null-space projection method*

1. *choose an initial guess* $x_0 \in N(B^{\mathrm{T}})$ (*e.g.*, $x_0 = 0$)
2. *solve* $By_0 \approx f - Ax_0$
3. *compute* $r_0^{(x)} = f - Ax_0 - By_0$
4. *for* $k = 0, 1, 2, \ldots$ *until convergence*
5. *compute* α_k *and* $p_k^{(x)} \in N(B^{\mathrm{T}})$
6. *set* $v_k^{(x)} = r_k^{(x)} - \alpha_k A p_k^{(x)}$
7. *solve* $Bp_k^{(y)} \approx v_k^{(x)}$
8. *update* $x_{k+1} = x_k + \alpha_k p_k^{(x)}$
9. *update* $r_{k+1}^{(x)} = v_k^{(x)} - Bp_k^{(y)}$
10. *compute* y_{k+1}:
    (a) *update* $y_{k+1} = y_k + p_k^{(y)}$ *(generic update)*
    (b) *solve* $By_{k+1} \approx f - Ax_{k+1}$ *(direct substitution)*
    (c) *solve* $Bv_k \approx f - Ax_{k+1} - By_k$,
    update $y_{k+1} = y_k + v_k$ *(corrected direct substitution)*
11. *end*

Note that in lines 6, 7 and 9 we in fact compute the update $r_{k+1}^{(x)} = r_k^{(x)} - \alpha_k(I - \Pi)A(I - \Pi)p_k^{(x)}$, where $r_k^{(x)}$ and $p_k^{(x)}$ are supposed to belong to $N(B^{\mathrm{T}})$. In lines 7 and 10 we need to solve the least squares problem with B. Since in practice such a problem cannot be solved exactly, we assume that the computed solution $\bar{w}$ of the least squares problem $Bw \approx c$ is the exact solution of a perturbed problem $(B + \Delta B)\bar{w} \approx c + \Delta c$, where the relative perturbations are bounded by τ with $\|\Delta B\| \leqslant \tau\|B\|$ and $\|\Delta c\| \leqslant \tau\|c\|$. To leave the rank of $B + \Delta B$ unchanged the parameter τ is supposed to satisfy the inequality $\tau\kappa(B) \ll 1$. Similarly as in the case of the Schur complement reduction we analyze three different but mathematically equivalent schemes for the computation of the iterate y_{k+1} which is performed in line 10. For corresponding schemes we use the same names as in the case of the Schur complement reduction method. They have similar properties also when considering their computational costs. However, as we will see in the next section, their numerical behavior can be completely different. The generic update was used in [7] (under the name "residual update"). The corrected direct substitution is closely connected to the constraint preconditioning of saddle point systems, see e.g., [13,12,15].

3. Behavior in finite precision arithmetic

In this section we recall the results on the limiting accuracy of the computed iterates $\bar{x}_{k+1}$ and $\bar{y}_{k+1}$. As we have noted similar results were developed in [11] for the case of symmetric saddle point problems and they can be extended also for the nonsymmetric case. Here we give several (not exactly straightforward) reformulations and due to the simplicity of presentation we do not include their proofs and refer to the paper [11]. By bars we denote here the quantities computed in finite precision arithmetic.

First we look at the Schur complement reduction method. Independently of the back-substitution scheme for the approximation x_{k+1} we can state a bound on the gap between the true residual $-B^{\mathrm{T}}A^{-1}f + B^{\mathrm{T}}A^{-1}B\bar{y}_{k+1}$ and the updated residual $\bar{r}_{k+1}^{(y)}$ associated with the Schur complement system (2) as follows:

$$\| - B^{\mathrm{T}}A^{-1}f + B^{\mathrm{T}}A^{-1}B\bar{y}_{k+1} - \bar{r}_{k+1}^{(y)}\| \leqslant \frac{\mathrm{O}(\tau)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_{k+1}), \tag{6}$$

where $\bar{Y}_{k+1} \equiv \max\{\|\bar{y}_i\| | i = 0, 1, \ldots, k+1\}$. It is a well-known fact [8,9] that the norm of the updated residual decreases far below the level of unit roundoff. Hence, from the estimate for the gap (6) we obtain an estimate for the maximum accuracy level for the true Schur complement residual which ultimately stagnates on the level proportional to the parameter τ. A similar argumentation is used throughout the whole paper, where in the corresponding bound by $\lesssim$ we denote that $\bar{r}_{k+1}^{(y)}$ (or $\bar{r}_{k+1}^{(x)}$ in the case of the null-space projection method) has already converged below a level

of unit roundoff and thus the estimate for the gap between the true and updated residuals leads to the bound for the maximum attainable accuracy of the true one.

In the following we give bounds on the limiting accuracy level for the norms of the residuals $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ and $-B^{\mathrm{T}}\bar{x}_{k+1}$, i.e., for two components of the true residual associated with the original saddle point system (1).

Generic update $x_{k+1} = x_k - \alpha_k A^{-1} B p_k^{(y)}$: Provided that the updated residual $\bar{r}_{k+1}^{(y)}$ drops beyond the unit roundoff level, the true residuals in (1) satisfy

$$\|f - A\bar{x}_{k+1} - B\bar{y}_{k+1}\| \leqslant \frac{\mathrm{O}(\tau)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|\bar{Y}_{k+1}), \tag{7a}$$

$$\| - B^{\mathrm{T}}\bar{x}_{k+1}\| \lesssim \frac{\mathrm{O}(u)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_{k+1}). \tag{7b}$$

The use of the simple generic update thus leads ultimately to the second component of the residual in (1) on the level proportional to the unit roundoff. However, the first component of the residual is proportional to the parameter τ since it accumulates the errors from solving the inner systems in line 6. Note that the maximum attainable accuracy level of this residual depends on the whole history of the norms of the iterations $\bar{y}_i$ $(i = 0, 1, \ldots, k+1)$.

Direct substitution $x_{k+1} = A^{-1}(f - By_{k+1})$: Provided that the updated residual $\bar{r}_{k+1}^{(y)}$ drops beyond the unit roundoff level, the true residuals in (1) satisfy

$$\|f - A\bar{x}_{k+1} - B\bar{y}_{k+1}\| \leqslant \frac{\mathrm{O}(\tau)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|\|\bar{y}_{k+1}\|), \tag{8a}$$

$$\| - B^{\mathrm{T}}\bar{x}_{k+1}\| \lesssim \frac{\mathrm{O}(\tau)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_{k+1}). \tag{8b}$$

Now both residuals ultimately stagnate on the level proportional to the parameter τ. However, the first component of the residual depends only on the norm of the actual iterate $\bar{y}_{k+1}$.

Corrected direct substitution $x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1})$: Provided that the updated residual $\bar{r}_{k+1}^{(y)}$ drops beyond the unit roundoff level and for sufficiently large $k \geqslant k_0$ the true residuals in (1) satisfy

$$\|f - A\bar{x}_{k+1} - B\bar{y}_{k+1}\| \leqslant \frac{\mathrm{O}(u)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|\bar{Y}_{k+1}^{(k_0)}), \tag{9a}$$

$$\| - B^{\mathrm{T}}\bar{x}_{k+1}\| \lesssim \frac{\mathrm{O}(\tau)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_{k+1}), \tag{9b}$$

where $\bar{Y}_{k+1}^{(k_0)} \equiv \max\{\|\bar{y}_i\| | i = k_0, k_0 + 1, \ldots, k+1\}$. This scheme gives a similar accuracy in the second component of the residual as the direct substitution (8b) but it is significantly more accurate in the first component. Again, for k large enough the attainable accuracy level of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ is proportional to unit roundoff u and it depends only on the norm of the iterates in the last few steps $\bar{y}_i$ $(i = k_0, k_0 + 1, \ldots, k+1)$. Choosing k_0 large enough makes possible to eliminate large initial oscillations of the iterates from the quantity $\bar{Y}_{k+1}$ leading to the much optimistic bound with $\bar{Y}_{k+1}^{(k_0)}$.

Now we continue with the results on the null-space projection method. When the norm of the updated residual $\bar{r}_{k+1}^{(x)}$ falls below the unit roundoff level, we can bound (independently of the computation scheme for the iterate y_{k+1}) the gap between the projected residual $(I - \Pi)(f - A\bar{x}_{k+1})$ and the updated residual $(I - \Pi)\bar{r}_{k+1}^{(x)}$ as follows:

$$\|(I - \Pi)(f - A\bar{x}_{k+1} - \bar{r}_{k+1}^{(x)})\| \leqslant \frac{\mathrm{O}(u)\kappa(B)}{1 - \tau\kappa(B)}(\|f\| + \|A\|\bar{X}_{k+1}), \tag{10}$$

where $\bar{X}_{k+1} \equiv \max\{\|\bar{x}_i\| | i = 0, 1, \ldots, k+1\}$. Hence, the residual of the projected system ultimately stagnates on the level proportional to unit roundoff. It is clear that the size of the residual $-B^{\mathrm{T}}\bar{x}_{k+1}$ depends on the "quality" of the computed direction vectors $\bar{p}_k^{(x)}$ which are the results of the projections of some vector onto $N(B^{\mathrm{T}})$. This is strongly dependent on the method used to compute these vectors and therefore we do not consider such analysis here. We expect that $\| - B^{\mathrm{T}}\bar{x}_{k+1}\| \leqslant \mathrm{O}(\tau)\|B\|\bar{X}_{k+1}$ but more precise analysis should be done in every specific situation. The accuracy measured by the residual of the projected system (4) depends also on the departure of the iterates $\bar{x}_{k+1}$ (or the direction

vectors $\bar{p}_{k+1}^{(x)}$) from $N(B^{\mathrm{T}})$ as it follows from

$$\|(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_{k+1}\|\leqslant\|(I-\Pi)(f-A\bar{x}_{k+1})\|+\|(I-\Pi)A\Pi\bar{x}_{k+1}\|.$$

In the following we give bounds on the limiting accuracy of computed approximations $\bar{x}_{k+1}$ and $\bar{y}_{k+1}$ measured by the norm of the true residual $f-A\bar{x}_{k+1}-B\bar{y}_{k+1}$.

Generic update $y_{k+1}=y_k+B^{\dagger}(r_k^{(x)}-\alpha_k Ap_k^{(x)})$: Provided that the updated residual $\bar{r}_{k+1}^{(x)}$ drops beyond the unit roundoff level, the norm of the first component of the true residual in (1) satisfies

$$\|f-A\bar{x}_{k+1}-B\bar{y}_{k+1}\|\lesssim\frac{\mathrm{O}(u)\kappa(B)}{1-\tau\kappa(B)}(\|f\|+\|A\|\bar{X}_{k+1}).\tag{11}$$

Thus, the simple generic update formula makes the first component of the true residual in (1) stagnating ultimately on the level proportional to unit roundoff. Note that the maximum attainable level of this residual depends on the whole history of the norms of the iterations $\bar{x}_i$ $(i=0,1,\ldots,k+1)$.

Direct substitution $y_{k+1}=B^{\dagger}(f-Ax_{k+1})$: Provided that the updated residual $\bar{r}_{k+1}^{(x)}$ drops beyond the roundoff unit level, the norm of the first component of the true residual in (1) satisfies

$$\|f-A\bar{x}_{k+1}-B\bar{y}_{k+1}\|\lesssim\frac{\mathrm{O}(\tau)\kappa(B)}{1-\tau\kappa(B)}(\|f\|+\|A\|\|\bar{x}_{k+1}\|)+\frac{\mathrm{O}(u)\kappa(B)}{1-\tau\kappa(B)}(\|f\|+\|A\|\bar{X}_{k+1}).\tag{12}$$

In contrast to the bound for the generic update formula (11) the first component of the residual in (1) ultimately stagnates on the level proportional to the parameter τ. Note that the term proportional to τ depends only on the norm of the actual iterate $\bar{x}_{k+1}$ (not on the maximum over all $i=0,1,\ldots,k+1$).

Corrected direct substitution $y_{k+1}=y_k+B^{\dagger}(f-Ax_{k+1}-By_k)$: Provided that the updated residual $\bar{r}_{k+1}^{(x)}$ drops beyond the unit roundoff level and for sufficiently large k, the norm of the first component of the true residual in (1) satisfies

$$\|f-A\bar{x}_{k+1}-B\bar{y}_{k+1}\|\lesssim\frac{\mathrm{O}(u)\kappa(B)}{1-\tau\kappa(B)}(\|f\|+\|A\|\bar{X}_{k+1}).\tag{13}$$

This scheme gives a similar accuracy as the generic update (11) but it costs one additional solution of the least squares problem with B. For k large enough the attainable accuracy level of the norm of $f-A\bar{x}_{k+1}-B\bar{y}_{k+1}$ is proportional to unit roundoff but in contrast to the corresponding scheme (9a) in the Schur complement reduction method it depends on the whole history of the norms of the iterates $\bar{x}_i$, $i=0,1,\ldots,k+1$.

4. Numerical experiments

We have applied several nonsymmetric iterative methods for the solution of the Schur complement system (2) and the projected system (4). Here we demonstrate our theoretical results on a simple numerical example of the nonsymmetric system (1) with

$$A=\mathrm{tridiag}\ (1,10^{-5},-1)\in\mathbb{R}^{100,100},\quad B=\mathrm{rand}\ (100,50),\quad f=(1,\ldots,1)^{\mathrm{T}}.$$

These matrices are quite well conditioned since $\kappa(A)=\|A\|\|A^{-1}\|=2.00\cdot 32.15=64.27$, $\kappa(B)=\|B\|\|B^{\dagger}\|=7.39\cdot 0.75=5.55$. The conditioning of matrices A and B thus has no significant effect on the behavior of considered schemes. For each test we have chosen the zero initial guess $y_0=0$ and $x_0=0$ for the Schur complement reduction method and for the null-space projection method, respectively.

As we have noted in previous section, the norms of the updated residual vectors converge usually to zero or at least become orders of magnitude smaller than unit roundoff. It follows from our theory that in such cases the true residuals associated with the approximate solutions $\bar{x}_{k+1}$ and $\bar{y}_{k+1}$ stagnate on the level proportional to the maximum norms (measured either by $\bar{X}_{k+1}$ or $\bar{Y}_{k+1}$) of iterates computed during the whole iteration process. It is also a well-known fact that for methods in which some (fixed) norm of the error or the residual decreases monotonically the maximum attainable accuracy level depends then on the norm of the initial residual.

One of the most straightforward methods to solve a general nonsymmetric system is the CGNE method which transforms the solution of a general square system to the symmetric positive (semi)definite system of normal equations. Since the CGNE method is nothing but the CG method applied to the system of normal equations, its approximate solution minimizes the 2-norm of the error over the associated Krylov subspace. Because the condition number of the system matrix is squared, we can expect rather slow convergence of CGNE in general. Therefore, the use of the GMRES method is preferred where the residual norm is minimized over the entire Krylov subspace generated with the original system matrix and corresponding right-hand side. Indeed, due to the optimality of iterates the quantities $\bar{X}_{k+1}$ and $\bar{Y}_{k+1}$ in CGNE and GMRES applied either to (2) or (4) cannot be significantly larger than the size of the initial approximations x_0, y_0 and unknowns x and y. Depending on the actual implementation of (3) or (5) the maximum attainable accuracy level is then proportional either to roundoff unit u or to the parameter τ and the quantities $\bar{Y}_{k+1}$ and $\bar{X}_{k+1}$ do not play an important role in the bounds from previous section.

Unfortunately, for general nonsymmetric systems the GMRES method cannot be implemented without full recurrences. In order to reduce the storage and computational work several classes of nonsymmetric iterative methods have been proposed including very popular methods based on the nonsymmetric Lanczos process such as Bi-CG, QMR or CGS. These methods compute the iterates and residual vectors using short recurrences keeping the computational cost constant at each iteration step (in contrast to the linear growth for the case of GMRES). The approximate solutions of such methods are, however, no longer optimal and their convergence behavior can be quite irregular (they even may occasionally fail to converge). In practice the norms of iterates can become (very) large during the initial phase of the computation until the iterates begin to converge and finally to stagnate (hopefully) near the true solution. For this reason one cannot give an a priori bound on $\bar{X}_{k+1}$ and $\bar{Y}_{k+1}$, and indeed the algorithms for solving (2) or (4) such as the Bi-CG or CGS method may fail to obtain small ultimate residuals even if the updated residuals converged beyond the unit roundoff. So the possibility of large iterates may correspondingly affect the maximum attainable accuracy level for such nonsymmetric iterative methods.

An example of these effects is shown in Fig. 1 where we consider GMRES, CGNE, Bi-CG and CGS in the Schur complement reduction method with the inner systems solved by the direct method based on the LU factorization of the matrix A (on the left). Similarly in Fig. 1 we report the results for the null-space projection method (on the right), where the inner systems were solved using the Householder QR factorization of the matrix B. We have plotted the true residual $B^{\mathrm{T}}A^{-1}f - B^{\mathrm{T}}A^{-1}B\bar{y}_{k+1}$ and $(I-\Pi)(f - A\bar{x}_{k+1})$ and the updated residuals $\bar{r}^{(y)}_{k+1}$ and $\bar{r}^{(x)}_{k+1}$, respectively, for GMRES (solid lines), CGNE (dash-dotted lines), Bi-CG (dotted lines) and CGS (dashed lines). As the computed residuals converge to zero for all methods (or to the unit roundoff level in the case of the GMRES method), the true residuals in (2) and (4) behave as indicated by the estimates (6) and (10). It is clear from Fig. 1 that for the error norm minimizing CGNE and the residual minimizing GMRES is the maximum attainable accuracy level proportional

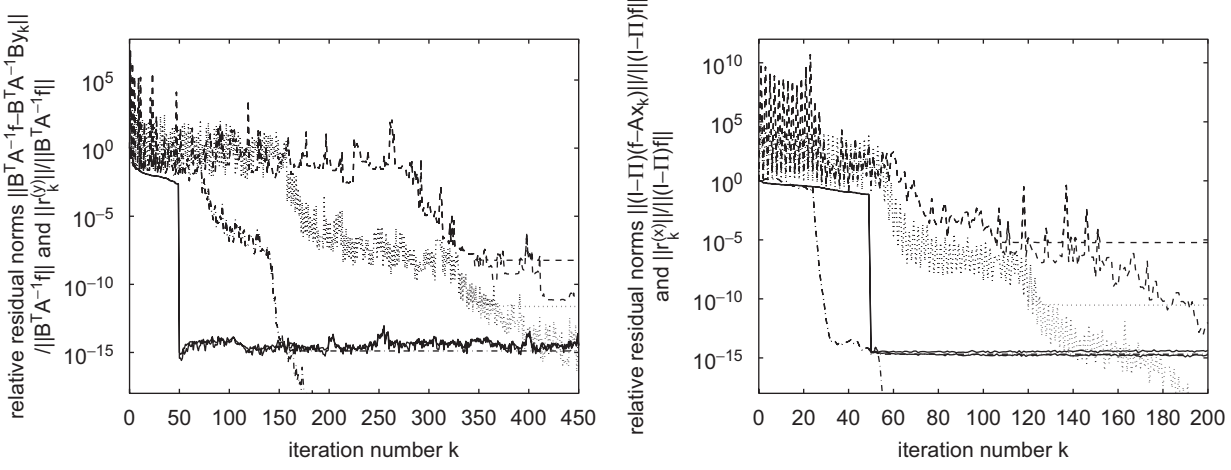

Fig. 1. Relative norms of the residual $B^{\mathrm{T}}A^{-1}f - B^{\mathrm{T}}A^{-1}B\bar{y}_{k+1}$ in the Schur complement reduction method (on the left) and of the projected residual $(I-\Pi)(f - A\bar{x}_{k+1})$ in the null-space projection method (on the right) with respect to the iteration number for GMRES (solid lines), CGNE (dash-dotted lines), Bi-CG (dotted lines) and CGS (dashed lines) with a direct solver used for the solution of inner systems.

Table 1
Quantities $\bar{Y}_{k+1}$ and $\bar{X}_{k+1}$ in the Schur complement method and in the null-space projection method, respectively, for GMRES, CGNE, Bi-CG and CGS

	Schur complement reduction		Null-space projection	
	$\bar{Y}_{k+1}$ (dir. sol.)	$\bar{Y}_{k+1}$ ($\tau = 10^{-12}$)	$\bar{X}_{k+1}$ (dir. sol.)	$\bar{X}_{k+1}$ ($\tau = 10^{-9}$)
GMRES	$1.6155 \cdot 10^{1}$	$1.6155 \cdot 10^{1}$	$3.9445 \cdot 10^{1}$	$3.9445 \cdot 10^{1}$
CGNE	$1.6157 \cdot 10^{1}$	$1.6156 \cdot 10^{1}$	$3.9445 \cdot 10^{1}$	$3.9445 \cdot 10^{1}$
Bi-CG	$9.8556 \cdot 10^{4}$	$1.5190 \cdot 10^{6}$	$6.5733 \cdot 10^{5}$	$6.5733 \cdot 10^{5}$
CGS	$3.3247 \cdot 10^{7}$	$7.7455 \cdot 10^{9}$	$5.2896 \cdot 10^{10}$	$5.2896 \cdot 10^{10}$

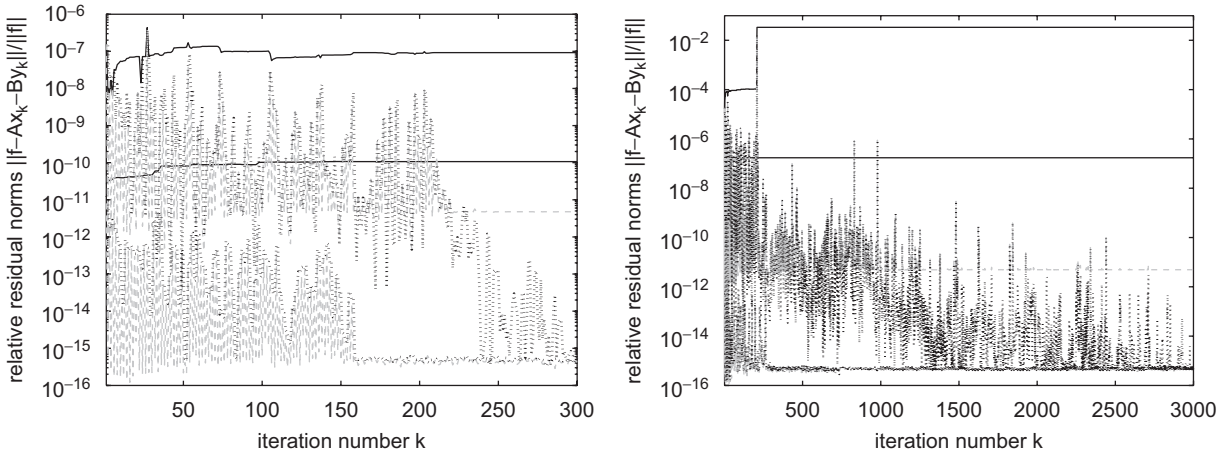

Fig. 2. The Schur complement reduction method: relative norms of the residual $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ for the Bi-CG (on the left) and for the CGS (on the right) methods using the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines) with the inner systems solved either by a direct solver or by an iterative method where $\tau = 10^{-12}$.

to the unit roundoff. The quantities $\bar{Y}_{k+1}$ and $\bar{X}_{k+1}$ are comparable to the size of unknowns y and x and they do not affect the limiting accuracy of computed approximate solutions. The situation is completely different for the Bi-CG and CGS methods where the size of iterates grows approximately to 10^5 (for Bi-CG) and to 10^7 (for CGS) in the Schur complement reduction method, or to 10^6 (for Bi-CG) and to 10^{11} (for CGS) in the null-space projection method (see the corresponding Table 1). Indeed, the results confirm that the final residuals reach the levels which are roughly equal to $\mathrm{O}(u)\bar{Y}_{k+1}$ or $\mathrm{O}(u)\bar{X}_{k+1}$ instead of $\mathrm{O}(u)|x|$ or $O(u)|y|$. Note that the matrices A and B are well conditioned and thus the norms of the Schur complement matrix and the projected matrix do not affect the final accuracy level for this example.

In Fig. 2 we report the norms of the residual $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ in the Schur complement reduction method where system (2) is solved by the Bi-CG method (on the left) or by the CGS method (on the right). In each plot we show the norms of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ for the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines). The inner systems are solved either by the direct solver (LU factorization) or by the Bi-CG method with $\tau = 10^{-12}$. The presented results confirm our estimates from the previous section. From Fig. 2 we can see the difference between the final accuracy levels of the norm of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ for the generic update (7a) and for the direct substitution (8a). In the first case, where the ultimate accuracy level depends on the maximum norm of the iterates $\bar{Y}_{k+1}$, the residual is essentially growing due to the accumulation of the residuals in inner systems. On the other hand, for the direct substitution (8a) the maximum attainable accuracy of the first equation in (1) is bounded by the norm of the actual iterate $\bar{y}_{k+1}$. The norms of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ are somewhat oscillating which reflects the jumps of $\|\bar{y}_{k+1}\|$ in the initial phase of the iteration process. When the norms of $\bar{y}_{k+1}$ begin to stagnate, the norms of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ do so but on much smaller level than for the generic update (7a). This difference between the accuracy levels is even more significant for the CGS method which exhibits much larger oscillations of iterates. Note

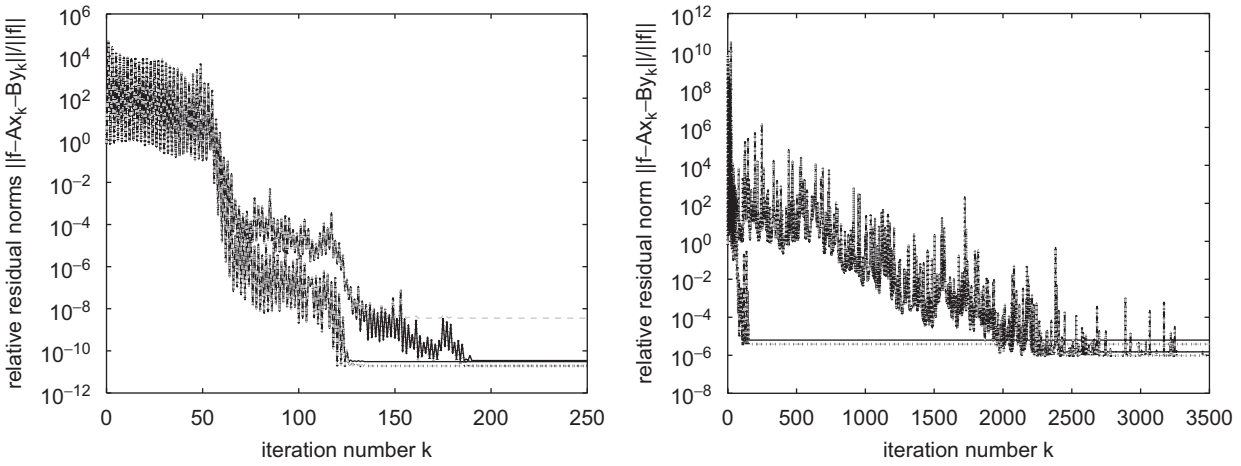

Fig. 3. Null-space projection method: relative norms of the residual $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ for the Bi-CG (on the left) and for the CGS (on the right) methods using the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines) with the inner systems solved either by a direct solver or by an iterative method where $\tau = 10^{-9}$.

that both for Bi-CG and CGS the residual norms for the corrected direct substitution converge to the unit roundoff level and they are not affected by the oscillations in the initial phase.

In Fig. 3 we report the norms of the residual $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ for the null-space projection method where the system (4) is solved either by the Bi-CG method (on the left) or by the CGS method (on the right). In each plot we show the norms of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ for the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines). The inner systems are solved either by the direct solver (Householder QR factorization) or by the CGLS method with $\tau = 10^{-9}$. The results confirm our estimates discussed in the previous section. For the direct substitution (12) the bound for the attainable accuracy level of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ depends on two terms. The first term is proportional to the unit roundoff u and to the quantity $\bar{X}_{k+1}$, while the second term is proportional to τ and to the norm of the actual iterate $\bar{x}_{k+1}$.Therefore, if the convergence behavior is very dramatic, the maximum attainable accuracy can be significantly affected by the rounding errors proportional to u dominating the bound over the terms dependent on the parameter τ. However, when the convergence behavior is quite regular the ultimate level of the norm of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ does depend also on τ. This can be seen in Fig. 3. The final level of the residual $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ in Bi-CG (with the direct substitution scheme and $\tau = 10^{-9}$) is still dependent on τ (on the left), while the same quantity for CGS (with more irregular convergence behavior) is actually dominated only by the rounding errors (on the right). For other two back-substitution formulas the norms of $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ ultimately stagnates on the level proportional to u. In contrast to the Schur complement reduction method for both Bi-CG and CGS the residuals in the corrected direct substitution (13) converge to the level of unit roundoff which is affected, however, by the oscillations of the iterates.

5. Conclusions

In this paper we have extended the previous results of [11] developed for saddle point systems where the leading diagonal block is symmetric positive definite. It appears that in the context of nonsymmetric saddle point systems the situation is significantly more complicated. Indeed, the limiting (maximum attainable) accuracy of these algorithms depends very much on intermediate iterates which may oscillate rather strongly for some methods (such as Bi-CG or CGS) where the error or residual norms decrease far from monotonously. On the other hand, for the GMRES or CGNE methods which are known to converge monotonously, these bounds depend actually on the initial error or residual. In addition, we have analyzed three mathematically equivalent back-substitution formulas and studied the influence of inexact solution of inner systems in the Schur complement reduction or in the null-space projection method onto the limiting accuracy level of residuals associated with the computed approximate solutions. Our results confirm the observed fact that the schemes with the direct substitution formula can be significantly less accurate than their generic

or corrected counterparts frequently used in such computational schemes as inexact Uzawa algorithms or stationary iterative methods with the constraint preconditioner used as smoothers.

References

[1] J. Atanga, D. Silvester, Iterative methods for stabilized mixed velocity–pressure finite elements, Internat. J. Numer. Methods Fluids 14 (1992) 71–81.

[2] M. Benzi, G.H. Golub, J. Liesen, Numerical solution of saddle point problems, Acta Numerica 14 (2005) 1–137.

[3] J.H. Bramble, J.E. Pasciak, A.T. Vassilev, Analysis of the inexact Uzawa algorithm for saddle point problems, SIAM J. Numer. Anal. 34 (3) (1997) 1072–1092.

[4] E.J. Craig, The N-step iteration procedures, J. Math. Phys. 34 (1955) 64–73.

[5] H.C. Elman, G.H. Golub, Inexact and preconditioned Uzawa algorithms for saddle point problems, SIAM J. Numer. Anal. 31 (6) (1661) 1645–1661.

[6] R. Fletcher, Conjugate gradient methods for indefinite systems, in: G.A. Watson (Ed.), Proceedings of the Dundee Biennial Conference on Numerical Analysis, Springer, New York, 1975, pp. 73–89.

[7] N.I.M. Gould, M.E. Hribar, J. Nocedal, On the solution of equality constrained quadratic programming problems arising in optimization, SIAM J. Sci. Comput. 23 (4) (2001) 1376–1395.

[8] A. Greenbaum, Accuracy of computed solutions from conjugate-gradient-like methods, in: M. Natori, T. Nodera (Eds.), Advances in Numerical Methods for Large Sparse Sets of Linear Systems, vol. 10, Keio University, Yokahama, Japan, 1994, pp. 126–138.

[9] A. Greenbaum, Estimating the attainable accuracy of recursively computed residual methods, SIAM J. Matrix Anal. Appl. 18 (3) (1997) 535–551.

[10] M.R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, J. Res. Nat. Bur. Standards 49 (1952) 409–436.

[11] P. Jiránek, M. Rozložník, Maximum attainable accuracy of inexact saddle point solvers, Technical Report, Institute of Computer Science, Academy of Sciences of the Czech Republic, 2006. www: ⟨http://www.cs.cas.cz/mweb/download/publi/JiRo2006.pdf⟩.

[12] C. Keller, N.I.M. Gould, A.J. Wathen, Constraint preconditioning for indefinite linear systems, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1300–1317.

[13] L. Lukšan, J. Vlček, Conjugate gradient methods for saddle point systems, Technical Report 778, ICS CAS CR, 1999.

[14] A. Ramage, A.J. Wathen, Iterative solution techniques for the stokes and Navier–Stokes equations, Internat. J. Numer. Methods Fluids 19 (1994) 67–83.

[15] M. Rozložník, V. Simoncini, Krylov subspace methods for saddle point problems with indefinite preconditioning, SIAM J. Matrix Anal. Appl. 24 (2) (2002) 368–391.

[16] Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 7 (1986) 856–869.

[17] V. Simoncini, D. Szyld, Theory of inexact Krylov subspace methods and applications to scientific computing, SIAM J. Sci. Comput. 25 (2) (2003) 454–477.

[18] P. Sonneveld, CGS, a fast Lanczos-type solver for nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 10 (1989) 36–52.

[19] E. Stiefel, Relaxationsmethoden bester Strategie zur Lösung linearer Gleichungssysteme, Comment Math. Helv. 29 (1955) 157–179.

[20] J. van den Eshof, G. Sleijpen, Inexact Krylov subspace methods for linear systems, SIAM J. Matrix Anal. Appl. 26 (1) (2004) 125–153.